

---

**sagemaker**

*Release 1.9.3.1*

**Dec 13, 2018**



---

## Contents

---

|          |                                         |           |
|----------|-----------------------------------------|-----------|
| <b>1</b> | <b>Overview</b>                         | <b>3</b>  |
| 1.1      | Estimators . . . . .                    | 3         |
| 1.2      | HyperparameterTuner . . . . .           | 8         |
| 1.3      | Predictors . . . . .                    | 12        |
| 1.4      | Transformer . . . . .                   | 13        |
| 1.5      | Session . . . . .                       | 15        |
| 1.6      | Model . . . . .                         | 24        |
| 1.7      | Analytics . . . . .                     | 26        |
| <b>2</b> | <b>MXNet</b>                            | <b>29</b> |
| 2.1      | MXNet . . . . .                         | 29        |
| <b>3</b> | <b>TensorFlow</b>                       | <b>33</b> |
| 3.1      | TensorFlow . . . . .                    | 33        |
| <b>4</b> | <b>SageMaker First-Party Algorithms</b> | <b>37</b> |
| 4.1      | K-means . . . . .                       | 37        |
| 4.2      | PCA . . . . .                           | 42        |
| 4.3      | LinearLearner . . . . .                 | 46        |
| 4.4      | Amazon Estimators . . . . .             | 54        |
| 4.5      | FactorizationMachines . . . . .         | 56        |
| 4.6      | LDA . . . . .                           | 61        |
| 4.7      | NTM . . . . .                           | 66        |
|          | <b>Python Module Index</b>              | <b>71</b> |



Amazon SageMaker Python SDK is an open source library for training and deploying machine-learned models on Amazon SageMaker.

With the SDK, you can train and deploy models using popular deep learning frameworks: **Apache MXNet** and **TensorFlow**. You can also train and deploy models with **algorithms provided by Amazon**, these are scalable implementations of core machine learning algorithms that are optimized for SageMaker and GPU training. If you have **your own algorithms** built into SageMaker-compatible Docker containers, you can train and host models using these as well.

Here you'll find API docs for SageMaker Python SDK. The project home-page is in Github: <https://github.com/aws/sagemaker-python-sdk>, there you can find the SDK source, installation instructions and a general overview of the library there.



The SageMaker Python SDK consists of a few primary interfaces:

## 1.1 Estimators

A high level interface for SageMaker training

```
class sagemaker.estimator.Estimator(image_name, role, train_instance_count,  
                                     train_instance_type, train_volume_size=30,  
                                     train_volume_kms_key=None, train_max_run=86400,  
                                     input_mode='File', output_path=None, output_kms_key=None,  
                                     base_job_name=None, sagemaker_session=None, hyperparameters=None,  
                                     tags=None, subnets=None, security_group_ids=None)
```

Bases: `sagemaker.estimator.EstimatorBase`

A generic Estimator to train using any supplied algorithm. This class is designed for use with algorithms that don't have their own, custom class.

Initialize an `Estimator` instance.

### Parameters

- **image\_name** (*str*) – The container image to use for training.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if it needs to access an AWS resource.
- **train\_instance\_count** (*int*) – Number of Amazon EC2 instances to use for training.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.

- **train\_volume\_size** (*int*) – Size in GB of the EBS volume to use for storing input data during training (default: 30). Must be large enough to store training data if File Mode is used (which is the default).
- **train\_volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting EBS volume attached to the training instance (default: None).
- **train\_max\_run** (*int*) – Timeout in seconds for training (default: 24 \* 60 \* 60). After this amount of time Amazon SageMaker terminates the job regardless of its current status.
- **input\_mode** (*str*) – The input mode that the algorithm supports (default: 'File'). Valid modes:
  - 'File' - Amazon SageMaker copies the training dataset from the S3 location to a local directory.
  - 'Pipe' - Amazon SageMaker streams data directly from S3 to the container via a Unix-named pipe.
- **output\_path** (*str*) – S3 location for saving the training result (model artifacts and output files). If not specified, results are stored to a default bucket. If the bucket with the specific name does not exist, the estimator creates the bucket during the `fit()` method execution.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the training output (default: None).
- **base\_job\_name** (*str*) – Prefix for training job name when the `fit()` method launches. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.
- **sagemaker\_session** (*sagemaker.session.Session*) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.
- **hyperparameters** (*dict*) – Dictionary containing the hyperparameters to initialize this estimator with.
- **tags** (*list[dict]*) – List of tags for labeling a training job. For more, see [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_Tag.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Tag.html).
- **subnets** (*list[str]*) – List of subnet ids. If not specified training job will be created without VPC config.
- **security\_group\_ids** (*list[str]*) – List of security group ids. If not specified training job will be created without VPC config.

**train\_image** ()

Returns the docker image to use for training.

The `fit()` method, that does the model training, calls this method to find the image to use for model training.

**set\_hyperparameters** (\*\*kwargs)

**hyperparameters** ()

Returns the hyperparameters as a dictionary to use for training.

The `fit()` method, that does the model training, calls this method to find the hyperparameters you specified.

**create\_model** (*role=None, image=None, predictor\_cls=None, serializer=None, deserializer=None, content\_type=None, accept=None, \*\*kwargs*)

Create a model to deploy.

### Parameters



- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the `Model`, which is also used during transform jobs. If not specified, the role from the `Estimator` will be used.
- **image** (*str*) – An container image to use for deploying the model. Defaults to the image used for training.
- **predictor\_cls** (`RealTimePredictor`) – The predictor class to use when deploying the model.
- **serializer** (*callable*) – Should accept a single argument, the input data, and return a sequence of bytes. May provide a `content_type` attribute that defines the endpoint request content type
- **deserializer** (*callable*) – Should accept two arguments, the result data and the response content type, and return a sequence of bytes. May provide a `content_type` attribute that defines th endpoint response Accept content type.
- **content\_type** (*str*) – The invocation `ContentType`, overriding any `content_type` from the `serializer`
- **accept** (*str*) – The invocation `Accept`, overriding any `accept` from the `deserializer`.
- **serializer, deserializer, content\_type, and accept arguments are only used to define a default (The) –**
- **They are ignored if an explicit predictor class is passed in. Other arguments (RealTimePredictor.) –**
- **passed through to the Model class. (are) –**

Returns: a `Model` ready for deployment.

**classmethod** `attach(training_job_name, sagemaker_session=None)`

Attach to an existing training job.

Create an `Estimator` bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching, if the training job has a `Complete` status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, `attach` will block and display log messages from the training job, until the training job completes.

#### Parameters

- **training\_job\_name** (*str*) – The name of the training job to attach to.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

#### Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling `Estimator` Class with the attached training job.

**delete\_endpoint ()**

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

**deploy (initial\_instance\_count, instance\_type, endpoint\_name=None, \*\*kwargs)**

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

**Parameters**

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, 'ml.c4.xlarge'.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model ()`. Implementations may customize `create_model ()` to accept `**kwargs` to customize model creation during deploy. For more, see the implementation docs.

**Returns**

A predictor that provides a **predict ()** method, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** `sagemaker.predictor.RealTimePredictor`

**fit (inputs, wait=True, logs=True, job\_name=None)**

Train a model using the input training dataset.

The API calls the Amazon SageMaker `CreateTrainingJob` API to start model training. The API uses configuration you provided to create the estimator and the specified input training data to send the `CreateTrainingJob` request to Amazon SageMaker.

This is a synchronous operation. After the model training successfully completes, you can call the `deploy ()` method to host the model using the Amazon SageMaker hosting services.

**Parameters**

- **inputs** (*str or dict or sagemaker.session.s3\_input*) – Information about the training data. This can be one of three types:
  - (*str*) the S3 location where training data is saved.
  - (**dict[str, str]** or **dict[str, sagemaker.session.s3\_input]**) **If using multiple channels for training data**, you can specify a dict mapping channel names to strings or `s3_input ()` objects.
  - (**sagemaker.session.s3\_input**) - **channel configuration for S3 data sources that can provide additional information** as well as the path to the training dataset. See `sagemaker.session.s3_input ()` for full details.
- **wait** (*bool*) – Whether the call should wait until the job completes (default: True).
- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when `wait` is True (default: True).
- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

**model\_data**

The model location in S3. Only set if Estimator has been `fit()`.

**Type** `str`

**training\_job\_analytics**

Return a `TrainingJobAnalytics` object for the current training job.

**transformer** (*instance\_count*, *instance\_type*, *strategy=None*, *assemble\_with=None*, *output\_path=None*, *output\_kms\_key=None*, *accept=None*, *env=None*, *max\_concurrent\_transforms=None*, *max\_payload=None*, *tags=None*, *role=None*, *volume\_kms\_key=None*)

Return a `Transformer` that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

**Parameters**

- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the `Model`, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

## 1.2 HyperparameterTuner

**class** `sagemaker.tuner.HyperparameterTuner` (*estimator, objective\_metric\_name, hyperparameter\_ranges, metric\_definitions=None, strategy='Bayesian', objective\_type='Maximize', max\_jobs=1, max\_parallel\_jobs=1, tags=None, base\_tuning\_job\_name=None*)

Bases: `object`

A class for creating and interacting with Amazon SageMaker hyperparameter tuning jobs, as well as deploying the resulting model(s).

Initialize a `HyperparameterTuner`. It takes an estimator to obtain configuration information for training jobs that are created as the result of a hyperparameter tuning job.

### Parameters

- **estimator** (*sagemaker.estimator.EstimatorBase*) – An estimator object that has been initialized with the desired configuration. There does not need to be a training job associated with this instance.
- **objective\_metric\_name** (*str*) – Name of the metric for evaluating training jobs.
- **hyperparameter\_ranges** (*dict[str, sagemaker.tuner.\_ParameterRange]*) – Dictionary of parameter ranges. These parameter ranges can be one of three types: Continuous, Integer, or Categorical. The keys of the dictionary are the names of the hyperparameter, and the values are the appropriate parameter range class to represent the range.
- **metric\_definitions** (*list[dict]*) – A list of dictionaries that defines the metric(s) used to evaluate the training jobs (default: None). Each dictionary contains two keys: 'Name' for the name of the metric, and 'Regex' for the regular expression used to extract the metric from the logs. This should be defined only for hyperparameter tuning jobs that don't use an Amazon algorithm.
- **strategy** (*str*) – Strategy to be used for hyperparameter estimations (default: 'Bayesian').
- **objective\_type** (*str*) – The type of the objective metric for evaluating training jobs. This value can be either 'Minimize' or 'Maximize' (default: 'Maximize').
- **max\_jobs** (*int*) – Maximum total number of training jobs to start for the hyperparameter tuning job (default: 1).
- **max\_parallel\_jobs** (*int*) – Maximum number of parallel training jobs to start (default: 1).
- **tags** (*list[dict]*) – List of tags for labeling the tuning job (default: None). For more, see [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_Tag.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Tag.html).
- **base\_tuning\_job\_name** (*str*) – Prefix for the hyperparameter tuning job name when the `fit()` method launches. If not specified, a default job name is generated, based on the training image name and current timestamp.

**TUNING\_JOB\_NAME\_MAX\_LENGTH** = 32

**SAGEMAKER\_ESTIMATOR\_MODULE** = 'sagemaker\_estimator\_module'

**SAGEMAKER\_ESTIMATOR\_CLASS\_NAME** = 'sagemaker\_estimator\_class\_name'

**DEFAULT\_ESTIMATOR\_MODULE** = 'sagemaker.estimator'

```
DEFAULT_ESTIMATOR_CLS_NAME = 'Estimator'
```

```
fit (inputs, job_name=None, include_cls_metadata=True, **kwargs)
    Start a hyperparameter tuning job.
```

#### Parameters

- **inputs** – Information about the training data. Please refer to the `fit()` method of the associated estimator, as this can take any of the following forms:
  - (str) - The S3 location where training data is saved.
  - (dict[str, str] or dict[str, sagemaker.session.s3\_input]) - If using multiple channels for training data, you can specify a dict mapping channel names to strings or `s3_input()` objects.
  - (sagemaker.session.s3\_input) - Channel configuration for S3 data sources that can provide additional information about the training dataset. See `sagemaker.session.s3_input()` for full details.
  - (sagemaker.amazon.amazon\_estimator.RecordSet) - A collection of Amazon :class:~'Record' objects serialized and stored in S3. For use with an estimator for an Amazon algorithm.
  - (list[sagemaker.amazon.amazon\_estimator.RecordSet]) - A list of :class:~'sagemaker.amazon.amazon\_estimator.RecordSet' objects, where each instance is a different channel of training data.
- **job\_name** (str) – Tuning job name. If not specified, the tuner generates a default job name, based on the training image name and current timestamp.
- **include\_cls\_metadata** (bool) – Whether or not the hyperparameter tuning job should include information about the estimator class (default: True). This information is passed as a hyperparameter, so if the algorithm you are using cannot handle unknown hyperparameters (e.g. an Amazon ML algorithm that does not have a custom estimator in the Python SDK), then set `include_cls_metadata` to False.
- **\*\*kwargs** – Other arguments needed for training. Please refer to the `fit()` method of the associated estimator to see what other arguments are needed.

```
classmethod attach (tuning_job_name, sagemaker_session=None, job_details=None, estimator_cls=None)
```

Attach to an existing hyperparameter tuning job.

Create a HyperparameterTuner bound to an existing hyperparameter tuning job. After attaching, if there exists a best training job (or any other completed training job), that can be `deploy()`ed to create an Amazon SageMaker Endpoint and return a `Predictor`.

#### Parameters

- **tuning\_job\_name** (str) – The name of the hyperparameter tuning job to attach to.
- **sagemaker\_session** (sagemaker.session.Session) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, one is created using the default AWS configuration chain.
- **job\_details** (dict) – The response to a `DescribeHyperParameterTuningJob` call. If not specified, the `HyperparameterTuner` will perform one such call with the provided hyperparameter tuning job name.

- **estimator\_cls** (*str*) – The estimator class name associated with the training jobs, e.g. ‘sagemaker.estimator.Estimator’. If not specified, the `HyperparameterTuner` will try to derive the correct estimator class from training job metadata, defaulting to `:class::~sagemaker.estimator.Estimator` if it is unable to determine a more specific class.

## Examples

```
>>> my_tuner.fit()
>>> job_name = my_tuner.latest_tuning_job.name
Later on:
>>> attached_tuner = HyperparameterTuner.attach(job_name)
>>> attached_tuner.deploy()
```

### Returns

A `HyperparameterTuner` instance with the attached hyperparameter tuning job.

**Return type** *sagemaker.tuner.HyperparameterTuner*

**deploy** (*initial\_instance\_count*, *instance\_type*, *endpoint\_name=None*, *\*\*kwargs*)

Deploy the best trained or user specified model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

For more information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, ‘ml.c4.xlarge’.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Other arguments needed for deployment. Please refer to the `create_model()` method of the associated estimator to see what other arguments are needed.

### Returns

A predictor that provides a `predict()` method, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** *sagemaker.predictor.RealTimePredictor*

**stop\_tuning\_job()**

Stop latest running hyperparameter tuning job.

**wait()**

Wait for latest hyperparameter tuning job to finish.

**best\_training\_job()**

Return name of the best training job for the latest hyperparameter tuning job.

**Raises** `Exception` – If there is no best training job available for the hyperparameter tuning job.

**delete\_endpoint** (*endpoint\_name=None*)

Delete an Amazon SageMaker endpoint.

If an endpoint name is not specified, this defaults to looking for an endpoint that shares a name with the best training job for deletion.

**Parameters** **endpoint\_name** (*str*) – Name of the endpoint to delete

**hyperparameter\_ranges** ()

Return the hyperparameter ranges in a dictionary to be used as part of a request for creating a hyperparameter tuning job.

**sagemaker\_session**

~'sagemaker.session.Session' object associated with the estimator for the `HyperparameterTuner`.

**Type** Convenience method for accessing the

**Type** class

**analytics** ()

An instance of `HyperparameterTuningJobAnalytics` for this latest tuning job of this tuner. `Analytics` object gives you access to tuning results summarized into a pandas dataframe.

**class** `sagemaker.tuner.ContinuousParameter` (*min\_value, max\_value*)

Bases: `sagemaker.tuner._ParameterRange`

A class for representing hyperparameters that have a continuous range of possible values.

Initialize a `ContinuousParameter`.

**Parameters**

- **min\_value** (*float*) – The minimum value for the range.
- **max\_value** (*float*) – The maximum value for the range.

**class** `sagemaker.tuner.IntegerParameter` (*min\_value, max\_value*)

Bases: `sagemaker.tuner._ParameterRange`

A class for representing hyperparameters that have an integer range of possible values.

Initialize a `IntegerParameter`.

**Parameters**

- **min\_value** (*int*) – The minimum value for the range.
- **max\_value** (*int*) – The maximum value for the range.

**class** `sagemaker.tuner.CategoricalParameter` (*values*)

Bases: `sagemaker.tuner._ParameterRange`

A class for representing hyperparameters that have a discrete list of possible values.

Initialize a `CategoricalParameter`.

**Parameters** **values** (*list or object*) – The possible values for the hyperparameter. This input will be converted into a list of strings.

**as\_tuning\_range** (*name*)

Represent the parameter range as a dictionary suitable for a request to create an Amazon SageMaker hyperparameter tuning job.

**Parameters** **name** (*str*) – The name of the hyperparameter.

**Returns** A dictionary that contains the name and values of the hyperparameter.

**Return type** `dict[str, list[str]]`

**as\_json\_range** (*name*)

Represent the parameter range as a dictionary suitable for a request to create an Amazon SageMaker hyperparameter tuning job using one of the deep learning frameworks.

The deep learning framework images require that hyperparameters be serialized as JSON.

**Parameters** **name** (*str*) – The name of the hyperparameter.

**Returns**

A dictionary that contains the name and values of the hyperparameter, where the values are serialized as JSON.

**Return type** `dict[str, list[str]]`

## 1.3 Predictors

Make real-time predictions against SageMaker endpoints with Python objects

```
class sagemaker.predictor.RealTimePredictor(endpoint, sagemaker_session=None,
                                             serializer=None, deserializer=None,
                                             content_type=None, accept=None)
```

Bases: `object`

Make prediction requests to an Amazon SageMaker endpoint.

Initialize a `RealTimePredictor`.

Behavior for serialization of input data and deserialization of result data can be configured through initializer arguments. If not specified, a sequence of bytes is expected and the API sends it in the request body without modifications. In response, the API returns the sequence of bytes from the prediction result without any modifications.

**Parameters**

- **endpoint** (*str*) – Name of the Amazon SageMaker endpoint to which requests are sent.
- **sagemaker\_session** (`sagemaker.session.Session`) – A SageMaker Session object, used for SageMaker interactions (default: `None`). If not specified, one is created using the default AWS configuration chain.
- **serializer** (*callable*) – Accepts a single argument, the input data, and returns a sequence of bytes. It may provide a `content_type` attribute that defines the endpoint request content type. If not specified, a sequence of bytes is expected for the data.
- **deserializer** (*callable*) – Accepts two arguments, the result data and the response content type, and returns a sequence of bytes. It may provide a `content_type` attribute that defines the endpoint response’s “Accept” content type. If not specified, a sequence of bytes is expected for the data.
- **content\_type** (*str*) – The invocation’s “ContentType”, overriding any `content_type` from the serializer (default: `None`).
- **accept** (*str*) – The invocation’s “Accept”, overriding any `accept` from the deserializer (default: `None`).

**predict** (*data*)

Return the inference from the specified endpoint.



**Parameters** `data` (*object*) – Input data for which you want the model to provide inference. If a serializer was specified when creating the RealTimePredictor, the result of the serializer is sent as input data. Otherwise the data must be sequence of bytes, and the predict method then sends the bytes in the request body as is.

#### Returns

**Inference for the given input. If a deserializer was specified when creating** the RealTimePredictor, the result of the deserializer is returned. Otherwise the response returns the sequence of bytes as is.

**Return type** `object`

`delete_endpoint()`

Delete the Amazon SageMaker endpoint backing this predictor.

## 1.4 Transformer

```
class sagemaker.transformer.Transformer(model_name, instance_count, instance_type,
                                         strategy=None, assemble_with=None, out-
                                         put_path=None, output_kms_key=None, ac-
                                         cept=None, max_concurrent_transforms=None,
                                         max_payload=None, tags=None, env=None,
                                         base_transform_job_name=None, sage-
                                         maker_session=None, volume_kms_key=None)
```

Bases: `object`

A class for handling creating and interacting with Amazon SageMaker transform jobs.

Initialize a Transformer.

#### Parameters

- **model\_name** (*str*) – Name of the SageMaker model being used for the transform job.
- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.

- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **tags** (*list[dict]*) – List of tags for labeling a transform job (default: None). For more, see [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_Tag.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Tag.html).
- **base\_transform\_job\_name** (*str*) – Prefix for the transform job when the `transform()` method launches. If not specified, a default prefix will be generated based on the training image name that was used to train the model associated with the transform job.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

**transform** (*data*, *data\_type='S3Prefix'*, *content\_type=None*, *compression\_type=None*, *split\_type=None*, *job\_name=None*)  
Start a new transform job.

#### Parameters

- **data** (*str*) – Input data location in S3.
- **data\_type** (*str*) – What the S3 location defines (default: 'S3Prefix'). Valid values:
  - 'S3Prefix' - the S3 URI defines a key name prefix. All objects with this prefix will be used as inputs for the transform job.
  - 'ManifestFile' - the S3 URI points to a single manifest file listing each S3 object to use as an input for the transform job.
- **content\_type** (*str*) – MIME type of the input data (default: None).
- **compression** (*str*) – Compression type of the input data, if compressed (default: None). Valid values: 'Gzip', None.
- **split\_type** (*str*) – The record delimiter for the input object (default: 'None'). Valid values: 'None', 'Line', and 'RecordIO'.
- **job\_name** (*str*) – job name (default: None). If not specified, one will be generated.

**wait** ()

**classmethod attach** (*transform\_job\_name*, *sagemaker\_session=None*)  
Attach an existing transform job to a new Transformer instance

#### Parameters

- **transform\_job\_name** (*str*) – Name for the transform job to be attached.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, one will be created using the default AWS configuration chain.

**Returns** The Transformer instance with the specified transform job attached.

**Return type** `sagemaker.transformer.Transformer`

## 1.5 Session

```
class sagemaker.session.LogState
```

```
Bases: object
```

```
STARTING = 1
```

```
WAIT_IN_PROGRESS = 2
```

```
TAILING = 3
```

```
JOB_COMPLETE = 4
```

```
COMPLETE = 5
```

```
class sagemaker.session.Session (boto_session=None, sagemaker_client=None, sage-
maker_runtime_client=None)
```

```
Bases: object
```

Manage interactions with the Amazon SageMaker APIs and any other AWS services needed.

This class provides convenient methods for manipulating entities and resources that Amazon SageMaker uses, such as training jobs, endpoints, and input datasets in S3.

AWS service calls are delegated to an underlying Boto3 session, which by default is initialized using the AWS configuration chain. When you make an Amazon SageMaker API call that accesses an S3 bucket location and one is not specified, the `Session` creates a default bucket based on a naming convention which includes the current AWS account ID.

Initialize a SageMaker `Session`.

### Parameters

- **boto\_session** (*boto3.session.Session*) – The underlying Boto3 session which AWS service calls are delegated to (default: `None`). If not provided, one is created with default AWS configuration chain.
- **sagemaker\_client** (*boto3.SageMaker.Client*) – Client which makes Amazon SageMaker service calls other than `InvokeEndpoint` (default: `None`). Estimators created using this `Session` use this client. If not provided, one will be created using this instance's `boto_session`.
- **sagemaker\_runtime\_client** (*boto3.SageMakerRuntime.Client*) – Client which makes `InvokeEndpoint` calls to Amazon SageMaker (default: `None`). Predictors created using this `Session` use this client. If not provided, one will be created using this instance's `boto_session`.

**boto\_region\_name**

**upload\_data** (*path, bucket=None, key\_prefix='data'*)

Upload local file or directory to S3.

If a single file is specified for upload, the resulting S3 object key is `{key_prefix}/{filename}` (filename does not include the local path, if any specified).

If a directory is specified for upload, the API uploads all content, recursively, preserving relative structure of subdirectories. The resulting object key names are: `{key_prefix}/{relative_subdirectory_path}/filename`.

### Parameters

- **path** (*str*) – Path (absolute or relative) of local file or directory to upload.

- **bucket** (*str*) – Name of the S3 Bucket to upload to (default: None). If not specified, the default bucket of the `Session` is used (if default bucket does not exist, the `Session` creates it).
- **key\_prefix** (*str*) – Optional S3 object key name prefix (default: 'data'). S3 uses the prefix to create a directory structure for the bucket content that it display in the S3 console.

### Returns

**The S3 URI of the uploaded file(s). If a file is specified in the path argument, the URI format is:**  
`s3://{bucket name}/{key_prefix}/{original_file_name}`. If a directory is specified in the path argument, the URI format is `s3://{bucket name}/{key_prefix}`.

### Return type *str*

#### **default\_bucket** ()

Return the name of the default bucket to use in relevant Amazon SageMaker interactions.

**Returns** The name of the default bucket, which is of the form:  
`sagemaker-{region}-{AWS account ID}`.

### Return type *str*

**train** (*image, input\_mode, input\_config, role, job\_name, output\_config, resource\_config, vpc\_config, hyperparameters, stop\_condition, tags*)  
Create an Amazon SageMaker training job.

### Parameters

- **image** (*str*) – Docker image containing training code.
- **input\_mode** (*str*) – The input mode that the algorithm supports. Valid modes:
  - 'File' - Amazon SageMaker copies the training dataset from the S3 location to a directory in the Docker container.
  - 'Pipe' - Amazon SageMaker streams data directly from S3 to the container via a Unix-named pipe.
- **input\_config** (*list*) – A list of Channel objects. Each channel is a named input source. Please refer to the format details described: [https://botocore.readthedocs.io/en/latest/reference/services/sagemaker.html#SageMaker.Client.create\\_training\\_job](https://botocore.readthedocs.io/en/latest/reference/services/sagemaker.html#SageMaker.Client.create_training_job)
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. You must grant sufficient permissions to this role.
- **job\_name** (*str*) – Name of the training job being created.
- **output\_config** (*dict*) – The S3 URI where you want to store the training results and optional KMS key ID.
- **resource\_config** (*dict*) – Contains values for ResourceConfig:
  - **instance\_count** (*int*): **Number of EC2 instances to use for training.** The key in `resource_config` is 'InstanceCount'.
  - **instance\_type** (*str*): **Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.** The key in `resource_config` is 'InstanceType'.
- **vpc\_config** (*dict*) – Contains values for VpcConfig:

- **subnets (list[str]): List of subnet ids.** The key in `vpc_config` is ‘Subnets’.
- **security\_group\_ids (list[str]): List of security group ids.** The key in `vpc_config` is ‘SecurityGroupIds’.
- **hyperparameters (dict)** – Hyperparameters for model training. The hyperparameters are made accessible as a `dict[str, str]` to the training code on SageMaker. For convenience, this accepts other types for keys and values, but `str()` will be called to convert them before training.
- **stop\_condition (dict)** – Defines when training shall finish. Contains entries that can be understood by the service like `MaxRuntimeInSeconds`.
- **tags (list[dict])** – List of tags for labeling a training job. For more, see [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_Tag.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Tag.html).

**Returns** ARN of the training job, if it is created.

**Return type** `str`

**tune** (*job\_name, strategy, objective\_type, objective\_metric\_name, max\_jobs, max\_parallel\_jobs, parameter\_ranges, static\_hyperparameters, image, input\_mode, metric\_definitions, role, input\_config, output\_config, resource\_config, stop\_condition, tags*)  
Create an Amazon SageMaker hyperparameter tuning job

#### Parameters

- **job\_name (str)** – Name of the tuning job being created.
- **strategy (str)** – Strategy to be used for hyperparameter estimations.
- **objective\_type (str)** – The type of the objective metric for evaluating training jobs. This value can be either ‘Minimize’ or ‘Maximize’.
- **objective\_metric\_name (str)** – Name of the metric for evaluating training jobs.
- **max\_jobs (int)** – Maximum total number of training jobs to start for the hyperparameter tuning job.
- **max\_parallel\_jobs (int)** – Maximum number of parallel training jobs to start.
- **parameter\_ranges (dict)** – Dictionary of parameter ranges. These parameter ranges can be one of three types: Continuous, Integer, or Categorical.
- **static\_hyperparameters (dict)** – Hyperparameters for model training. These hyperparameters remain unchanged across all of the training jobs for the hyperparameter tuning job. The hyperparameters are made accessible as a dictionary for the training code on SageMaker.
- **image (str)** – Docker image containing training code.
- **input\_mode (str)** – The input mode that the algorithm supports. Valid modes:
  - ‘File’ - Amazon SageMaker copies the training dataset from the S3 location to a directory in the Docker container.
  - ‘Pipe’ - Amazon SageMaker streams data directly from S3 to the container via a Unix-named pipe.
- **metric\_definitions (list[dict])** – A list of dictionaries that defines the metric(s) used to evaluate the training jobs. Each dictionary contains two keys: ‘Name’ for the name of the metric, and ‘Regex’ for the regular expression used to extract the metric from the logs. This should be defined only for hyperparameter tuning jobs that don’t use an Amazon algorithm.

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. You must grant sufficient permissions to this role.
- **input\_config** (*list*) – A list of Channel objects. Each channel is a named input source. Please refer to the format details described: [https://botocore.readthedocs.io/en/latest/reference/services/sagemaker.html#SageMaker.Client.create\\_training\\_job](https://botocore.readthedocs.io/en/latest/reference/services/sagemaker.html#SageMaker.Client.create_training_job)
- **output\_config** (*dict*) – The S3 URI where you want to store the training results and optional KMS key ID.
- **resource\_config** (*dict*) – Contains values for ResourceConfig:
  - **instance\_count** (*int*): **Number of EC2 instances to use for training.** The key in resource\_config is 'InstanceCount'.
  - **instance\_type** (*str*): **Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.** The key in resource\_config is 'InstanceType'.
- **stop\_condition** (*dict*) – When training should finish, e.g. MaxRuntimeInSeconds.
- **tags** (*list[dict]*) – List of tags for labeling the tuning job. For more, see [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_Tag.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Tag.html).

**stop\_tuning\_job** (*name*)

Stop the Amazon SageMaker hyperparameter tuning job with the specified name.

**Parameters** **name** (*str*) – Name of the Amazon SageMaker hyperparameter tuning job.

**Raises** `ClientError` – If an error occurs while trying to stop the hyperparameter tuning job.

**transform** (*job\_name, model\_name, strategy, max\_concurrent\_transforms, max\_payload, env, input\_config, output\_config, resource\_config, tags*)

Create an Amazon SageMaker transform job.

**Parameters**

- **job\_name** (*str*) – Name of the transform job being created.
- **model\_name** (*str*) – Name of the SageMaker model being used for the transform job.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request. Possible values are 'MULTI\_RECORD' and 'SINGLE\_RECORD'.
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **env** (*dict*) – Environment variables to be set for use during the transform job.
- **input\_config** (*dict*) – A dictionary describing the input data (and its location) for the job.
- **output\_config** (*dict*) – A dictionary describing the output location for the job.
- **resource\_config** (*dict*) – A dictionary describing the resources to complete the job.

- **tags** (*list[dict]*) – List of tags for labeling a training job. For more, see [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_Tag.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_Tag.html).

**create\_model** (*name, role, primary\_container*)

Create an Amazon SageMaker Model.

Specify the S3 location of the model artifacts and Docker image containing the inference code. Amazon SageMaker uses this information to deploy the model in Amazon SageMaker.

#### Parameters

- **name** (*str*) – Name of the Amazon SageMaker Model to create.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. You must grant sufficient permissions to this role.
- **primary\_container** (*str or dict[str, str]*) – Docker image which defines the inference code. You can also specify the return value of `sagemaker.container_def()`, which is used to create more advanced container configurations, including model containers which need artifacts from S3.

**Returns** Name of the Amazon SageMaker Model created.

**Return type** *str*

**create\_model\_from\_job** (*training\_job\_name, name=None, role=None, primary\_container\_image=None, model\_data\_url=None, env={}*)

Create an Amazon SageMaker Model from a SageMaker Training Job.

#### Parameters

- **training\_job\_name** (*str*) – The Amazon SageMaker Training Job name.
- **name** (*str*) – The name of the SageMaker Model to create (default: None). If not specified, the training job name is used.
- **role** (*str*) – The ExecutionRoleArn IAM Role ARN for the Model, specified either by an IAM role name or role ARN. If None, the RoleArn from the SageMaker Training Job will be used.
- **primary\_container\_image** (*str*) – The Docker image reference (default: None). If None, it defaults to the Training Image in `training_job_name`.
- **model\_data\_url** (*str*) – S3 location of the model data (default: None). If None, defaults to the ModelS3Artifacts of `training_job_name`.
- **env** (*dict[string, string]*) – Model environment variables (default: {}).

**Returns** The name of the created Model.

**Return type** *str*

**create\_endpoint\_config** (*name, model\_name, initial\_instance\_count, instance\_type*)

Create an Amazon SageMaker endpoint configuration.

The endpoint configuration identifies the Amazon SageMaker model (created using the `CreateModel` API) and the hardware configuration on which to deploy the model. Provide this endpoint configuration to the `CreateEndpoint` API, which then launches the hardware and deploys the model.

#### Parameters

- **name** (*str*) – Name of the Amazon SageMaker endpoint configuration to create.
- **model\_name** (*str*) – Name of the Amazon SageMaker Model.

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to launch. The actual number of active instances for an endpoint at any given time varies due to autoscaling.
- **instance\_type** (*str*) – Type of EC2 instance to launch, for example, 'ml.c4.xlarge'.

**Returns** Name of the endpoint point configuration created.

**Return type** *str*

**create\_endpoint** (*endpoint\_name, config\_name, wait=True*)

Create an Amazon SageMaker Endpoint according to the endpoint configuration specified in the request.

Once the Endpoint is created, client applications can send requests to obtain inferences. The endpoint configuration is created using the CreateEndpointConfig API.

**Parameters**

- **endpoint\_name** (*str*) – Name of the Amazon SageMaker Endpoint being created.
- **config\_name** (*str*) – Name of the Amazon SageMaker endpoint configuration to deploy.
- **wait** (*bool*) – Whether to wait for the endpoint deployment to complete before returning (default: True).

**Returns** Name of the Amazon SageMaker Endpoint created.

**Return type** *str*

**delete\_endpoint** (*endpoint\_name*)

Delete an Amazon SageMaker Endpoint.

**Parameters** **endpoint\_name** (*str*) – Name of the Amazon SageMaker Endpoint to delete.

**wait\_for\_job** (*job, poll=5*)

Wait for an Amazon SageMaker training job to complete.

**Parameters**

- **job** (*str*) – Name of the training job to wait for.
- **poll** (*int*) – Polling interval in seconds (default: 5).

**Returns** Return value from the DescribeTrainingJob API.

**Return type** (*dict*)

**Raises** *ValueError* – If the training job fails.

**wait\_for\_tuning\_job** (*job, poll=5*)

Wait for an Amazon SageMaker hyperparameter tuning job to complete.

**Parameters**

- **job** (*str*) – Name of the tuning job to wait for.
- **poll** (*int*) – Polling interval in seconds (default: 5).

**Returns** Return value from the DescribeHyperParameterTuningJob API.

**Return type** (*dict*)



**Raises** `ValueError` – If the hyperparameter tuning job fails.

**wait\_for\_transform\_job** (*job*, *poll=5*)

Wait for an Amazon SageMaker transform job to complete.

**Parameters**

- **job** (*str*) – Name of the transform job to wait for.
- **poll** (*int*) – Polling interval in seconds (default: 5).

**Returns** Return value from the `DescribeTransformJob` API.

**Return type** (`dict`)

**Raises** `ValueError` – If the transform job fails.

**wait\_for\_endpoint** (*endpoint*, *poll=5*)

Wait for an Amazon SageMaker endpoint deployment to complete.

**Parameters**

- **endpoint** (*str*) – Name of the Endpoint to wait for.
- **poll** (*int*) – Polling interval in seconds (default: 5).

**Returns** Return value from the `DescribeEndpoint` API.

**Return type** `dict`

**endpoint\_from\_job** (*job\_name*, *initial\_instance\_count*, *instance\_type*, *deployment\_image=None*,  
*name=None*, *role=None*, *wait=True*, *model\_environment\_vars=None*)

Create an `Endpoint` using the results of a successful training job.

Specify the job name, Docker image containing the inference code, and hardware configuration to deploy the model. Internally the API, creates an Amazon SageMaker model (that describes the model artifacts and the Docker image containing inference code), endpoint configuration (describing the hardware to deploy for hosting the model), and creates an `Endpoint` (launches the EC2 instances and deploys the model on them). In response, the API returns the endpoint name to which you can send requests for inferences.

**Parameters**

- **job\_name** (*str*) – Name of the training job to deploy the results of.
- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to launch. The actual number of active instances for an endpoint at any given time varies due to autoscaling.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, 'ml.c4.xlarge'.
- **deployment\_image** (*str*) – The Docker image which defines the inference code to be used as the entry point for accepting prediction requests. If not specified, uses the image used for the training job.
- **name** (*str*) – Name of the `Endpoint` to create. If not specified, uses the training job name.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. You must grant sufficient permissions to this role.
- **wait** (*bool*) – Whether to wait for the endpoint deployment to complete before returning (default: True).

- **model\_environment\_vars** (*dict[str, str]*) – Environment variables to set on the model container (default: None).

**Returns** Name of the `Endpoint` that is created.

**Return type** `str`

**endpoint\_from\_model\_data** (*model\_s3\_location, deployment\_image, initial\_instance\_count, instance\_type, name=None, role=None, wait=True, model\_environment\_vars=None*)

Create and deploy to an `Endpoint` using existing model data stored in S3.

#### Parameters

- **model\_s3\_location** (*str*) – S3 URI of the model artifacts to use for the endpoint.
- **deployment\_image** (*str*) – The Docker image which defines the runtime code to be used as the entry point for accepting prediction requests.
- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to launch. The actual number of active instances for an endpoint at any given time varies due to autoscaling.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, e.g. ‘ml.c4.xlarge’.
- **name** (*str*) – Name of the `Endpoint` to create. If not specified, uses a name generated by combining the image name with a timestamp.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. You must grant sufficient permissions to this role.
- **wait** (*bool*) – Whether to wait for the endpoint deployment to complete before returning (default: True).
- **model\_environment\_vars** (*dict[str, str]*) – Environment variables to set on the model container (default: None).

**Returns** Name of the `Endpoint` that is created.

**Return type** `str`

**endpoint\_from\_production\_variants** (*name, production\_variants, tags=None, wait=True*)

Create an SageMaker `Endpoint` from a list of production variants.

#### Parameters

- **name** (*str*) – The name of the `Endpoint` to create.
- **production\_variants** (*list[dict[str, str]]*) – The list of production variants to deploy.
- **tags** (*list[dict[str, str]]*) – A list of key-value pairs for tagging the endpoint (default: None).
- **wait** (*bool*) – Whether to wait for the endpoint deployment to complete before returning (default: True).

**Returns** The name of the created `Endpoint`.

**Return type** `str`

**expand\_role** (*role*)

Expand an IAM role name into an ARN.

If the role is already in the form of an ARN, then the role is simply returned. Otherwise we retrieve the full ARN and return it.

**Parameters** **role** (*str*) – An AWS IAM role (either name or full ARN).

**Returns** The corresponding AWS IAM role ARN.

**Return type** *str*

**get\_caller\_identity\_arn** ()

Returns the ARN user or role whose credentials are used to call the API. :returns: The ARN user or role :rtype: (*str*)

**logs\_for\_job** (*job\_name*, *wait=False*, *poll=10*)

Display the logs for a given training job, optionally tailing them until the job is complete. If the output is a tty or a Jupyter cell, it will be color-coded based on which instance the log entry is from.

**Parameters**

- **job\_name** (*str*) – Name of the training job to display the logs for.
- **wait** (*bool*) – Whether to keep looking for new log entries until the job completes (default: *False*).
- **poll** (*int*) – The interval in seconds between polling for new log entries and job completion (default: *5*).

**Raises** *ValueError* – If waiting and the training job fails.

`sagemaker.session.container_def` (*image*, *model\_data\_url=None*, *env=None*)

Create a definition for executing a container as part of a SageMaker model.

**Parameters**

- **image** (*str*) – Docker image to run for this container.
- **model\_data\_url** (*str*) – S3 URI of data required by this container, e.g. SageMaker training job model artifacts (default: *None*).
- **env** (*dict[str, str]*) – Environment variables to set inside the container (default: *None*).

**Returns** A complete container definition object usable with the CreateModel API.

**Return type** *dict[str, str]*

`sagemaker.session.production_variant` (*model\_name*, *instance\_type*, *initial\_instance\_count=1*, *variant\_name='AllTraffic'*, *initial\_weight=1*)

Create a production variant description suitable for use in a `ProductionVariant` list as part of a `CreateEndpointConfig` request.

**Parameters**

- **model\_name** (*str*) – The name of the SageMaker model this production variant references.
- **instance\_type** (*str*) – The EC2 instance type for this production variant. For example, 'ml.c4.8xlarge'.
- **initial\_instance\_count** (*int*) – The initial instance count for this production variant (default: *1*).

- **variant\_name** (*string*) – The VariantName of this production variant (default: 'AllTraffic').
- **initial\_weight** (*int*) – The relative InitialVariantWeight of this production variant (default: 1).

**Returns** An SageMaker ProductionVariant description

**Return type** dict[str, str]

`sagemaker.session.get_execution_role(sagemaker_session=None)`

Return the role ARN whose credentials are used to call the API. Throws an exception if :param sagemaker\_session: Current sagemaker session :type sagemaker\_session: Session

**Returns** The role ARN

**Return type** (str)

```
class sagemaker.session.s3_input(s3_data, distribution='FullyReplicated', compression=None, content_type=None, record_wrapping=None, s3_data_type='S3Prefix')
```

Bases: object

Amazon SageMaker channel configurations for S3 data sources.

**config**

A SageMaker DataSource referencing a SageMaker S3DataSource.

**Type** dict[str, dict]

Create a definition for input data used by an SageMaker training job.

See AWS documentation on the CreateTrainingJob API for more details on the parameters.

**Parameters**

- **s3\_data** (*str*) – Defines the location of s3 data to train on.
- **distribution** (*str*) – Valid values: 'FullyReplicated', 'ShardedByS3Key' (default: 'FullyReplicated').
- **compression** (*str*) – Valid values: 'Gzip', None (default: None). This is used only in Pipe input mode.
- **content\_type** (*str*) – MIME type of the input data (default: None).
- **record\_wrapping** (*str*) – Valid values: 'RecordIO' (default: None).
- **s3\_data\_type** (*str*) – Valid values: 'S3Prefix', 'ManifestFile'. If 'S3Prefix', s3\_data defines a prefix of s3 objects to train on. All objects with s3 keys beginning with s3\_data will be used to train. If 'ManifestFile', then s3\_data defines a single s3 manifest file, listing each s3 object to train on. The Manifest file format is described in the SageMaker API documentation: [https://docs.aws.amazon.com/sagemaker/latest/dg/API\\_S3DataSource.html](https://docs.aws.amazon.com/sagemaker/latest/dg/API_S3DataSource.html)

## 1.6 Model

```
class sagemaker.model.Model(model_data, image, role, predictor_cls=None, env=None, name=None, sagemaker_session=None)
```

Bases: object

A SageMaker Model that can be deployed to an Endpoint.

Initialize an SageMaker Model.

### Parameters

- **model\_data** (*str*) – The S3 location of a SageMaker model data `.tar.gz` file.
- **image** (*str*) – A Docker image URI.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if it needs to access an AWS resource.
- **predictor\_cls** (*callable[string, sagemaker.session.Session]*) – A function to call to create a predictor (default: None). If not None, `deploy` will return the result of invoking this function on the created endpoint name.
- **env** (*dict[str, str]*) – Environment variables to run with `image` when hosted in SageMaker (default: None).
- **name** (*str*) – The model name. If None, a default model name will be selected on each `deploy`.
- **sagemaker\_session** (*sagemaker.session.Session*) – A SageMaker Session object, used for SageMaker interactions (default: None). If not specified, one is created using the default AWS configuration chain.

**prepare\_container\_def** (*instance\_type*)

Return a dict created by `sagemaker.container_def()` for deploying this model to a specified instance type.

Subclasses can override this to provide custom container definitions for deployment to a specific instance type. Called by `deploy()`.

**Parameters** **instance\_type** (*str*) – The EC2 instance type to deploy this Model to. For example, `'ml.p2.xlarge'`.

**Returns** A container definition object usable with the CreateModel API.

**Return type** `dict`

**deploy** (*initial\_instance\_count, instance\_type, endpoint\_name=None, tags=None*)

Deploy this Model to an Endpoint and optionally return a Predictor.

Create a SageMaker Model and EndpointConfig, and deploy an Endpoint from this Model. If `self.predictor_cls` is not None, this method returns a the result of invoking `self.predictor_cls` on the created endpoint name.

The name of the created endpoint is accessible in the `endpoint_name` field of this Model after `deploy` returns.

### Parameters

- **instance\_type** (*str*) – The EC2 instance type to deploy this Model to. For example, `'ml.p2.xlarge'`.
- **initial\_instance\_count** (*int*) – The initial number of instances to run in the Endpoint created from this Model.
- **endpoint\_name** (*str*) – The name of the endpoint to create (default: None). If not specified, a unique endpoint name will be created.
- **tags** (*list[dict[str, str]]*) – A list of key-value pairs for tagging the endpoint (default: None).

**Returns**

Invocation of `self.predictor_cls` on the created endpoint name, if `self.predictor_cls` is not `None`. Otherwise, return `None`.

**Return type** callable[string, *sagemaker.session.Session*] or `None`

## 1.7 Analytics

**class** `sagemaker.analytics.AnalyticsMetricsBase`

Bases: `object`

Base class for tuning job or training job analytics classes. Understands common functionality like persistence and caching.

**export\_csv** (*filename*)

Persists the analytics dataframe to a file.

**Parameters** `filename` (*str*) – The name of the file to save to.

**dataframe** (*force\_refresh=False*)

A pandas dataframe with lots of interesting results about this object. Created by calling SageMaker List and Describe APIs and converting them into a convenient tabular summary.

**Parameters** `force_refresh` (*bool*) – Set to `True` to fetch the latest data from SageMaker API.

**clear\_cache** ()

Clear the object of all local caches of API methods, so that the next time any properties are accessed they will be refreshed from the service.

**class** `sagemaker.analytics.HyperparameterTuningJobAnalytics` (*hyperparameter\_tuning\_job\_name*,  
*sagemaker\_session=None*)

Bases: `sagemaker.analytics.AnalyticsMetricsBase`

Fetch results about a hyperparameter tuning job and make them accessible for analytics.

Initialize a `HyperparameterTuningJobAnalytics` instance.

**Parameters**

- **hyperparameter\_tuning\_job\_name** (*str*) – name of the `HyperparameterTuningJob` to analyze.
- **sagemaker\_session** (*sagemaker.session.Session*) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, one is created using the default AWS configuration chain.

**name**

Name of the `HyperparameterTuningJob` being analyzed

**clear\_cache** ()

Clear the object of all local caches of API methods.

**tuning\_ranges**

A dictionary describing the ranges of all tuned hyperparameters. The keys are the names of the hyperparameter, and the values are the ranges.

**description** (*force\_refresh=False*)

Call `DescribeHyperParameterTuningJob` for the hyperparameter tuning job.

**Parameters** `force_refresh` (*bool*) – Set to True to fetch the latest data from SageMaker API.

**Returns** The Amazon SageMaker response for `DescribeHyperParameterTuningJob`.

**Return type** `dict`

**training\_job\_summaries** (*force\_refresh=False*)

A (paginated) list of everything from `ListTrainingJobsForTuningJob`.

**Parameters** `force_refresh` (*bool*) – Set to True to fetch the latest data from SageMaker API.

**Returns** The Amazon SageMaker response for `ListTrainingJobsForTuningJob`.

**Return type** `dict`

```
class sagemaker.analytics.TrainingJobAnalytics (training_job_name,           met-
                                               ric_names=None,           sage-
                                               maker_session=None)
```

Bases: `sagemaker.analytics.AnalyticsMetricsBase`

Fetch training curve data from CloudWatch Metrics for a specific training job.

Initialize a `TrainingJobAnalytics` instance.

#### Parameters

- **training\_job\_name** (*str*) – name of the `TrainingJob` to analyze.
- **metric\_names** (*list, optional*) – string names of all the metrics to collect for this training job. If not specified, then it will use all metric names configured for this job.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, one is specified using the default AWS configuration chain.

```
CLOUDWATCH_NAMESPACE = '/aws/sagemaker/HyperParameterTuningJobs'
```

#### **name**

Name of the `TrainingJob` being analyzed

#### **clear\_cache** ()

Clear the object of all local caches of API methods, so that the next time any properties are accessed they will be refreshed from the service.





A managed environment for MXNet training and hosting on Amazon SageMaker

## 2.1 MXNet

### 2.1.1 MXNet Estimator

```
class sagemaker.mxnet.estimator.MXNet (entry_point,          source_dir=None,          hyperpa-  
                                       rameters=None,        py_version='py2',        frame-  
                                       work_version='1.2', image_name=None, **kwargs)
```

Bases: `sagemaker.estimator.Framework`

Handle end-to-end training and deployment of custom MXNet code.

This `Estimator` executes an MXNet script in a managed MXNet execution environment, within a SageMaker Training Job. The managed MXNet environment is an Amazon-built Docker container that executes functions defined in the supplied `entry_point` Python script.

Training is started by calling `fit()` on this Estimator. After training is complete, calling `deploy()` creates a hosted SageMaker endpoint and returns an `MXNetPredictor` instance that can be used to perform inference against the hosted model.

Technical documentation on preparing MXNet scripts for SageMaker training and using the MXNet Estimator is available on the project home-page: <https://github.com/aws/sagemaker-python-sdk>

#### Parameters

- **entry\_point** (*str*) – Path (absolute or relative) to the Python source file which should be executed as the entry point to training. This should be compatible with either Python 2.7 or Python 3.5.
- **source\_dir** (*str*) – Path (absolute or relative) to a directory with any other training source code dependencies aside from the entry point file (default: None). Structure within this directory are preserved when training on Amazon SageMaker.

- **hyperparameters** (*dict*) – Hyperparameters that will be used for training (default: None). The hyperparameters are made accessible as a dict[str, str] to the training code on SageMaker. For convenience, this accepts other types for keys and values, but `str()` will be called to convert them before training.
- **py\_version** (*str*) – Python version you want to use for executing your model training code (default: 'py2'). One of 'py2' or 'py3'.
- **framework\_version** (*str*) – MXNet version you want to use for executing your model training code. List of supported versions <https://github.com/aws/sagemaker-python-sdk#mxnet-sagemaker-estimators>
- **image\_name** (*str*) – If specified, the estimator will use this image for training and hosting, instead of selecting the appropriate SageMaker official image based on framework\_version and py\_version. It can be an ECR url or dockerhub image and tag.  
  
**Examples:** 123.dkr.ecr.us-west-2.amazonaws.com/my-custom-image:1.0  
custom-image:latest.
- **\*\*kwargs** – Additional kwargs passed to the `Framework` constructor.

**create\_model** (*model\_server\_workers=None, role=None*)

Create a SageMaker `MXNetModel` object that can be deployed to an Endpoint.

#### Parameters

- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the Model, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **model\_server\_workers** (*int*) – Optional. The number of worker processes used by the inference server. If None, server will use one worker per vCPU.

#### Returns

A SageMaker `MXNetModel` object. See `MXNetModel()` for full details.

**Return type** `sagemaker.mxnet.model.MXNetModel`

## 2.1.2 MXNet Model

```
class sagemaker.mxnet.model.MXNetModel (model_data, role, entry_point, image=None, py_version='py2', framework_version='1.2', predictor_cls=<class 'sagemaker.mxnet.model.MXNetPredictor'>, model_server_workers=None, **kwargs)
```

Bases: `sagemaker.model.FrameworkModel`

An MXNet SageMaker Model that can be deployed to a SageMaker Endpoint.

Initialize an MXNetModel.

#### Parameters

- **model\_data** (*str*) – The S3 location of a SageMaker model data `.tar.gz` file.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if it needs to access an AWS resource.

- **entry\_point** (*str*) – Path (absolute or relative) to the Python source file which should be executed as the entry point to model hosting. This should be compatible with either Python 2.7 or Python 3.5.
- **image** (*str*) – A Docker image URI (default: None). If not specified, a default image for MXNet will be used.
- **py\_version** (*str*) – Python version you want to use for executing your model training code (default: 'py2').
- **framework\_version** (*str*) – MXNet version you want to use for executing your model training code.
- **predictor\_cls** (*callable[str, sagemaker.session.Session]*) – A function to call to create a predictor with an endpoint name and SageMaker Session. If specified, `deploy()` returns the result of invoking this function on the created endpoint name.
- **model\_server\_workers** (*int*) – Optional. The number of worker processes used by the inference server. If None, server will use one worker per vCPU.
- **\*\*kwargs** – Keyword arguments passed to the `FrameworkModel` initializer.

**prepare\_container\_def** (*instance\_type*)

Return a container definition with framework configuration set in model environment variables.

**Parameters** **instance\_type** (*str*) – The EC2 instance type to deploy this Model to. For example, 'ml.p2.xlarge'.

**Returns** A container definition object usable with the CreateModel API.

**Return type** `dict[str, str]`

### 2.1.3 MXNet Predictor

**class** `sagemaker.mxnet.model.MXNetPredictor` (*endpoint\_name, sagemaker\_session=None*)

Bases: `sagemaker.predictor.RealTimePredictor`

A RealTimePredictor for inference against MXNet Endpoints.

This is able to serialize Python lists, dictionaries, and numpy arrays to multidimensional tensors for MXNet inference.

Initialize an MXNetPredictor.

#### Parameters

- **endpoint\_name** (*str*) – The name of the endpoint to perform inference on.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.



A managed environment for TensorFlow training and hosting on Amazon SageMaker

## 3.1 TensorFlow

### 3.1.1 TensorFlow Estimator

```
class sagemaker.tensorflow.estimator.TensorFlow(training_steps=None,          eval-
                                                uation_steps=None,          check-
                                                point_path=None, py_version='py2',
                                                framework_version='1.10', require-
                                                ments_file="", image_name=None,
                                                **kwargs)
```

Bases: `sagemaker.estimator.Framework`

Handle end-to-end training and deployment of user-provided TensorFlow code.

Initialize an `TensorFlow` estimator. :param training\_steps: Perform this many steps of training. *None*, the default means train forever. :type training\_steps: int :param evaluation\_steps: Perform this many steps of evaluation. *None*, the default means that evaluation

runs until input from `eval_input_fn` is exhausted (or another exception is raised).

#### Parameters

- **checkpoint\_path** (*str*) – Identifies S3 location where checkpoint data during model training can be saved (default: *None*). For distributed model training, this parameter is required.
- **py\_version** (*str*) – Python version you want to use for executing your model training code (default: 'py2').

- **framework\_version** (*str*) – TensorFlow version you want to use for executing your model training code. List of supported versions <https://github.com/aws/sagemaker-python-sdk#tensorflow-sagemaker-estimators>
- **requirements\_file** (*str*) – Path to a `requirements.txt` file (default: `''`). The path should be within and relative to `source_dir`. Details on the format can be found in the [Pip User Guide](#).
- **image\_name** (*str*) – If specified, the estimator will use this image for training and hosting, instead of selecting the appropriate SageMaker official image based on `framework_version` and `py_version`. It can be an ECR url or dockerhub image and tag.

**Examples:** `123.dkr.ecr.us-west-2.amazonaws.com/my-custom-image:1.0`  
`custom-image:latest`.

- **\*\*kwargs** – Additional kwargs passed to the Framework constructor.

**fit** (*inputs*, *wait=True*, *logs=True*, *job\_name=None*, *run\_tensorboard\_locally=False*)  
 Train a model using the input training dataset.

See `fit()` for more details.

#### Parameters

- **inputs** (*str* or *dict* or `sagemaker.session.s3_input`) – Information about the training data. This can be one of three types: (*str*) - the S3 location where training data is saved. (`dict[str, str]` or `dict[str, sagemaker.session.s3_input]`) - If using multiple channels for

training data, you can specify a dict mapping channel names to strings or `s3_input()` objects.

(`sagemaker.session.s3_input`) - channel configuration for S3 data sources that can provide additional information as well as the path to the training dataset. See `sagemaker.session.s3_input()` for full details.

- **wait** (*bool*) – Whether the call should wait until the job completes (default: `True`).
- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).
- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.
- **run\_tensorboard\_locally** (*bool*) – Whether to execute TensorBoard in a different process with downloaded checkpoint information (default: `False`). This is an experimental feature, and requires TensorBoard and AWS CLI to be installed. It terminates TensorBoard when execution ends.

**create\_model** (*model\_server\_workers=None*, *role=None*)

Create a SageMaker TensorFlowModel object that can be deployed to an Endpoint.

#### Parameters

- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the Model, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **model\_server\_workers** (*int*) – Optional. The number of worker processes used by the inference server. If `None`, server will use one worker per vCPU.

#### Returns

A SageMaker `TensorFlowModel` object. See `TensorFlowModel()` for full details.

**Return type** `sagemaker.tensorflow.model.TensorFlowModel`

**hyperparameters()**

Return hyperparameters used by your custom TensorFlow code during model training.

### 3.1.2 TensorFlow Model

```
class sagemaker.tensorflow.model.TensorFlowModel(model_data, role, entry_point,
                                                image=None, py_version='py2',
                                                framework_version='1.10',
                                                predictor_cls=<class 'sage-
                                                maker.tensorflow.model.TensorFlowPredictor'>,
                                                model_server_workers=None,
                                                **kwargs)
```

Bases: `sagemaker.model.FrameworkModel`

Initialize an `TensorFlowModel`.

#### Parameters

- **model\_data** (*str*) – The S3 location of a SageMaker model data `.tar.gz` file.
- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if it needs to access an AWS resource.
- **entry\_point** (*str*) – Path (absolute or relative) to the Python source file which should be executed as the entry point to model hosting. This should be compatible with either Python 2.7 or Python 3.5.
- **image** (*str*) – A Docker image URI (default: None). If not specified, a default image for TensorFlow will be used.
- **py\_version** (*str*) – Python version you want to use for executing your model training code (default: 'py2').
- **framework\_version** (*str*) – TensorFlow version you want to use for executing your model training code.
- **predictor\_cls** (*callable*[*str*, `sagemaker.session.Session`]) – A function to call to create a predictor with an endpoint name and SageMaker `Session`. If specified, `deploy()` returns the result of invoking this function on the created endpoint name.
- **model\_server\_workers** (*int*) – Optional. The number of worker processes used by the inference server. If None, server will use one worker per vCPU.
- **\*\*kwargs** – Keyword arguments passed to the `FrameworkModel` initializer.

**prepare\_container\_def** (*instance\_type*)

Return a container definition with framework configuration set in model environment variables.

This also uploads user-supplied code to S3.

**Parameters** **instance\_type** (*str*) – The EC2 instance type to deploy this Model to. For example, 'ml.p2.xlarge'.

**Returns** A container definition object usable with the `CreateModel` API.

**Return type** `dict[str, str]`

### 3.1.3 TensorFlow Predictor

**class** `sagemaker.tensorflow.model.TensorFlowPredictor` (*endpoint\_name*, *sagemaker\_session=None*)

Bases: `sagemaker.predictor.RealTimePredictor`

A `RealTimePredictor` for inference against TensorFlow “Endpoint”s.

This is able to serialize Python lists, dictionaries, and numpy arrays to multidimensional tensors for MXNet inference

Initialize an `TensorFlowPredictor`.

#### Parameters

- **endpoint\_name** (*str*) – The name of the endpoint to perform inference on.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.



---

## SageMaker First-Party Algorithms

---

Amazon provides implementations of some common machine learning algorithms optimized for GPU architecture and massive datasets.

### 4.1 K-means

The Amazon SageMaker K-means algorithm.

```
class sagemaker.KMeans(role, train_instance_count, train_instance_type, k, init_method=None,  
                       max_iterations=None, tol=None, num_trials=None, local_init_method=None,  
                       half_life_time_size=None, epochs=None, center_factor=None, eval_metrics=None, **kwargs)
```

Bases: `sagemaker.amazon.amazon_estimator.AmazonAlgorithmEstimatorBase`

A k-means clustering `AmazonAlgorithmEstimatorBase`. Finds k clusters of data in an unlabeled dataset.

This Estimator may be fit via calls to `fit_ndarray()` or `fit()`. The former allows a `KMeans` model to be fit on a 2-dimensional numpy array. The latter requires `AmazonRecord` protobuf serialized data to be stored in S3.

To learn more about the Amazon protobuf `Record` class and how to prepare bulk data in this format, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>.

After this Estimator is fit, model data is stored in S3. The model may be deployed to an Amazon SageMaker Endpoint by invoking `deploy()`. As well as deploying an Endpoint, `deploy` returns a `KMeansPredictor` object that can be used to k-means cluster assignments, using the trained k-means model hosted in the SageMaker Endpoint.

`KMeans` Estimators can be configured by setting hyperparameters. The available hyperparameters for `KMeans` are documented below. For further information on the AWS `KMeans` algorithm, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/k-means.html>.

#### Parameters

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access

training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if accessing AWS resource.

- **train\_instance\_count** (*int*) – Number of Amazon EC2 instances to use for training.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, ‘ml.c4.xlarge’.
- **k** (*int*) – The number of clusters to produce.
- **init\_method** (*str*) – How to initialize cluster locations. One of ‘random’ or ‘kmeans++’.
- **max\_iterations** (*int*) – Maximum iterations for Lloyds EM procedure in the local kmeans used in finalize stage.
- **tol** (*float*) – Tolerance for change in *ssd* for early stopping in local kmeans.
- **num\_trials** (*int*) – Local version is run multiple times and the one with the best loss is chosen. This determines how many times.
- **local\_init\_method** (*str*) – Initialization method for local version. One of ‘random’, ‘kmeans++’
- **half\_life\_time\_size** (*int*) – The points can have a decayed weight. When a point is observed its weight, with regard to the computation of the cluster mean is 1. This weight will decay exponentially as we observe more points. The exponent coefficient is chosen such that after observing *half\_life\_time\_size* points after the mentioned point, its weight will become 1/2. If set to 0, there will be no decay.
- **epochs** (*int*) – Number of passes done over the training data.
- **center\_factor** (*int*) – The algorithm will create *num\_clusters \* extra\_center\_factor* as it runs and reduce the number of centers to *k* when finalizing
- **eval\_metrics** (*list*) – JSON list of metrics types to be used for reporting the score for the model. Allowed values are “msd” Means Square Error, “ssd”: Sum of square distance. If test data is provided, the score shall be reported in terms of all requested metrics.
- **\*\*kwargs** – base class keyword argument values.

```
repo_name = 'kmeans'
```

```
repo_version = 1
```

```
classmethod attach(training_job_name, sagemaker_session=None)
```

Attach to an existing training job.

Create an Estimator bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching, if the training job has a Complete status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, attach will block and display log messages from the training job, until the training job completes.

#### Parameters

- **training\_job\_name** (*str*) – The name of the training job to attach to.

- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

## Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling `Estimator` Class with the attached training job.

**data\_location**

**delete\_endpoint()**

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

**deploy** (*initial\_instance\_count*, *instance\_type*, *endpoint\_name=None*, *\*\*kwargs*)

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, ‘ml.c4.xlarge’.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model()`. Implementations may customize `create_model()` to accept `**kwargs` to customize model creation during deploy. For more, see the implementation docs.

### Returns

**A predictor that provides a `predict()` method**, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** `sagemaker.predictor.RealTimePredictor`

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

`records` should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set()` to construct a `RecordSet` object from `ndarray` arrays.

#### Parameters

- **records** (`RecordSet`) – The records to train this `Estimator` on
- **mini\_batch\_size** (`int` or `None`) – The size of each mini-batch to use when training. If `None`, a default value will be used.
- **wait** (`bool`) – Whether the call should wait until the job completes (default: `True`).
- **logs** (`bool`) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).
- **job\_name** (`str`) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

#### `model_data`

The model location in S3. Only set if `Estimator` has been `fit()`.

**Type** `str`

#### `record_set(train, labels=None, channel='train')`

Build a `RecordSet` from a `numpy ndarray` matrix and label vector.

For the 2D `ndarray train`, each row is converted to a `Record` object. The vector is stored in the “values” entry of the `features` property of each `Record`. If `labels` is not `None`, each corresponding label is assigned to the “values” entry of the `labels` property of each `Record`.

The collection of `Record` objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the `train_instance_count` property on this `Estimator`. One S3 object is created per training instance.

#### Parameters

- **train** (`numpy.ndarray`) – A 2D `numpy` array of training data.
- **labels** (`numpy.ndarray`) – A 1D `numpy` array of labels. Its length must be equal to the number of rows in `train`.
- **channel** (`str`) – The SageMaker TrainingJob channel this `RecordSet` should be assigned to.

**Returns** A `RecordSet` referencing the encoded, uploading training and label data.

**Return type** `RecordSet`

#### `train_image()`

Return the Docker image to use for training.

The `fit()` method, which does the model training, calls this method to find the image to use for model training.

**Returns** The URI of the Docker image.

**Return type** `str`

#### `training_job_analytics`

Return a `TrainingJobAnalytics` object for the current training job.

**transformer** (*instance\_count*, *instance\_type*, *strategy=None*, *assemble\_with=None*, *output\_path=None*, *output\_kms\_key=None*, *accept=None*, *env=None*, *max\_concurrent\_transforms=None*, *max\_payload=None*, *tags=None*, *role=None*, *volume\_kms\_key=None*)

Return a `Transformer` that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

#### Parameters

- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the Model, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

#### **eval\_metrics**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

#### **create\_model()**

Return a `KMeansModel` referencing the latest s3 model data produced by this Estimator.

#### **hyperparameters()**

Return the SageMaker hyperparameters for training this KMeans Estimator

**class** `sagemaker.KMeansModel` (*model\_data*, *role*, *sagemaker\_session=None*)

Bases: `sagemaker.model.Model`

Reference KMeans s3 model data. Calling `deploy()` creates an Endpoint and return a Predictor to performs k-means cluster assignment.

**class** `sagemaker.KMeansPredictor` (*endpoint*, *sagemaker\_session=None*)

Bases: `sagemaker.predictor.RealTimePredictor`

Assigns input vectors to their closest cluster in a KMeans model.

The implementation of `predict()` in this `RealTimePredictor` requires a numpy `ndarray` as input. The array should contain the same number of columns as the feature-dimension of the data used to fit the model this Predictor performs inference on.

`predict()` returns a list of `Record` objects, one for each row in the input `ndarray`. The nearest cluster is stored in the `closest_cluster` key of the `Record.label` field.

## 4.2 PCA

The Amazon SageMaker PCA algorithm.

```
class sagemaker.PCA(role, train_instance_count, train_instance_type, num_components, algo-
                    rithm_mode=None, subtract_mean=None, extra_components=None, **kwargs)
Bases: sagemaker.amazon.amazon_estimator.AmazonAlgorithmEstimatorBase
```

A Principal Components Analysis (PCA) `AmazonAlgorithmEstimatorBase`.

This Estimator may be fit via calls to `fit_ndarray()` or `fit()`. The former allows a PCA model to be fit on a 2-dimensional numpy array. The latter requires Amazon `Record` protobuf serialized data to be stored in S3.

To learn more about the Amazon protobuf `Record` class and how to prepare bulk data in this format, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

After this Estimator is fit, model data is stored in S3. The model may be deployed to an Amazon SageMaker Endpoint by invoking `deploy()`. As well as deploying an Endpoint, `deploy` returns a `PCAPredictor` object that can be used to project input vectors to the learned lower-dimensional representation, using the trained PCA model hosted in the SageMaker Endpoint.

PCA Estimators can be configured by setting hyperparameters. The available hyperparameters for PCA are documented below. For further information on the AWS PCA algorithm, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/pca.html>

This Estimator uses Amazon SageMaker PCA to perform training and host deployed models. To learn more about Amazon SageMaker PCA, please read: <https://docs.aws.amazon.com/sagemaker/latest/dg/how-pca-works.html>

### Parameters

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if accessing AWS resource.
- **train\_instance\_count** (*int*) – Number of Amazon EC2 instances to use for training.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.
- **num\_components** (*int*) – The number of principal components. Must be greater than zero.
- **algorithm\_mode** (*str*) – Mode for computing the principal components. One of 'regular' or 'randomized'.
- **subtract\_mean** (*bool*) – Whether the data should be unbiased both during train and at inference.

- **extra\_components** (*int*) – As the value grows larger, the solution becomes more accurate but the runtime and memory consumption increase linearly. If this value is unset or set to -1, then a default value equal to the maximum of 10 and num\_components will be used. Valid for randomized mode only.
- **\*\*kwargs** – base class keyword argument values.

```
repo_name = 'pca'
```

```
repo_version = 1
```

```
DEFAULT_MINI_BATCH_SIZE = 500
```

```
create_model()
```

Return a `PCAModel` referencing the latest s3 model data produced by this Estimator.

```
classmethod attach(training_job_name, sagemaker_session=None)
```

Attach to an existing training job.

Create an Estimator bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching, if the training job has a Complete status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, attach will block and display log messages from the training job, until the training job completes.

#### Parameters

- **training\_job\_name** (*str*) – The name of the training job to attach to.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

#### Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling Estimator Class with the attached training job.

```
data_location
```

```
delete_endpoint()
```

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

```
deploy(initial_instance_count, instance_type, endpoint_name=None, **kwargs)
```

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

#### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, ‘ml.c4.xlarge’.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model()`. Implementations may customize `create_model()` to accept **\*\*kwargs** to customize model creation during deploy. For more, see the implementation docs.

### Returns

A predictor that provides a **predict()** method, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** *sagemaker.predictor.RealTimePredictor*

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

*records* should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set()` to construct a `RecordSet` object from `ndarray` arrays.

### Parameters

- **records** (`RecordSet`) – The records to train this Estimator on
- **mini\_batch\_size** (*int* or *None*) – The size of each mini-batch to use when training. If *None*, a default value will be used.
- **wait** (*bool*) – Whether the call should wait until the job completes (default: `True`).
- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).
- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

**hyperparameters** ()

Return the hyperparameters as a dictionary to use for training.

The `fit()` method, which trains the model, calls this method to find the hyperparameters.

**Returns** The hyperparameters.

**Return type** `dict[str, str]`

**model\_data**

The model location in S3. Only set if Estimator has been `fit()`.

**Type** `str`



**record\_set** (*train*, *labels=None*, *channel='train'*)

Build a RecordSet from a numpy ndarray matrix and label vector.

For the 2D ndarray *train*, each row is converted to a Record object. The vector is stored in the “values” entry of the *features* property of each Record. If *labels* is not None, each corresponding label is assigned to the “values” entry of the *labels* property of each Record.

The collection of Record objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the *train\_instance\_count* property on this Estimator. One S3 object is created per training instance.

#### Parameters

- **train** (*numpy.ndarray*) – A 2D numpy array of training data.
- **labels** (*numpy.ndarray*) – A 1D numpy array of labels. Its length must be equal to the number of rows in *train*.
- **channel** (*str*) – The SageMaker TrainingJob channel this RecordSet should be assigned to.

**Returns** A RecordSet referencing the encoded, uploading training and label data.

**Return type** RecordSet

**train\_image** ()

Return the Docker image to use for training.

The *fit* () method, which does the model training, calls this method to find the image to use for model training.

**Returns** The URI of the Docker image.

**Return type** str

**training\_job\_analytics**

Return a TrainingJobAnalytics object for the current training job.

**transformer** (*instance\_count*, *instance\_type*, *strategy=None*, *assemble\_with=None*, *output\_path=None*, *output\_kms\_key=None*, *accept=None*, *env=None*, *max\_concurrent\_transforms=None*, *max\_payload=None*, *tags=None*, *role=None*, *volume\_kms\_key=None*)

Return a Transformer that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

#### Parameters

- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).

- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the `Model`, which is also used during transform jobs. If not specified, the role from the `Estimator` will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

**class** `sagemaker.PCAModel` (*model\_data*, *role*, *sagemaker\_session=None*)

Bases: `sagemaker.model.Model`

Reference PCA s3 model data. Calling `deploy()` creates an Endpoint and return a Predictor that transforms vectors to a lower-dimensional representation.

**class** `sagemaker.PCAPredictor` (*endpoint*, *sagemaker\_session=None*)

Bases: `sagemaker.predictor.RealTimePredictor`

Transforms input vectors to lower-dimensional representations.

The implementation of `predict()` in this `RealTimePredictor` requires a `numpy ndarray` as input. The array should contain the same number of columns as the feature-dimension of the data used to fit the model this Predictor performs inference on.

`predict()` returns a list of `Record` objects, one for each row in the input `ndarray`. The lower dimension vector result is stored in the `projection` key of the `Record.label` field.

## 4.3 LinearLearner

The Amazon SageMaker LinearLearner algorithm.

```

class sagemaker.LinearLearner(role, train_instance_count, train_instance_type, predic-
tor_type, binary_classifier_model_selection_criteria=None,
target_recall=None, target_precision=None, posi-
tive_example_weight_mult=None, epochs=None, use_bias=None,
num_models=None, num_calibration_samples=None,
init_method=None, init_scale=None, init_sigma=None,
init_bias=None, optimizer=None, loss=None, wd=None,
l1=None, momentum=None, learning_rate=None,
beta_1=None, beta_2=None, bias_lr_mult=None,
bias_wd_mult=None, use_lr_scheduler=None,
lr_scheduler_step=None, lr_scheduler_factor=None,
lr_scheduler_minimum_lr=None, normalize_data=None, normal-
ize_label=None, unbias_data=None, unbias_label=None,
num_point_for_scaler=None, margin=None, quan-
tile=None, loss_insensitivity=None, huber_delta=None,
early_stopping_patience=None, early_stopping_tolerance=None,
num_classes=None, accuracy_top_k=None, f_beta=None,
balance_multiclass_weights=None, **kwargs)

```

Bases: `sagemaker.amazon.amazon_estimator.AmazonAlgorithmEstimatorBase`

An Estimator for binary classification and regression.

Amazon SageMaker Linear Learner provides a solution for both classification and regression problems, allowing for exploring different training objectives simultaneously and choosing the best solution from a validation set. It allows the user to explore a large number of models and choose the best, which optimizes either continuous objectives such as mean square error, cross entropy loss, absolute error, etc., or discrete objectives suited for classification such as F1 measure, `precision@recall`, accuracy. The implementation provides a significant speedup over naive hyperparameter optimization techniques and an added convenience, when compared with solutions providing a solution only to continuous objectives.

This Estimator may be fit via calls to `fit_ndarray()` or `fit()`. The former allows a LinearLearner model to be fit on a 2-dimensional numpy array. The latter requires Amazon Record protobuf serialized data to be stored in S3.

To learn more about the Amazon protobuf Record class and how to prepare bulk data in this format, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

After this Estimator is fit, model data is stored in S3. The model may be deployed to an Amazon SageMaker Endpoint by invoking `deploy()`. As well as deploying an Endpoint, `deploy` returns a `LinearLearnerPredictor` object that can be used to make class or regression predictions, using the trained model.

LinearLearner Estimators can be configured by setting hyperparameters. The available hyperparameters for LinearLearner are documented below. For further information on the AWS LinearLearner algorithm, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/linear-learner.html>

#### Parameters

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if accessing AWS resource.
- **train\_instance\_count** (*int*) – Number of Amazon EC2 instances to use for training.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.

- **predictor\_type** (*str*) – The type of predictor to learn. Either “binary\_classifier” or
- or **"regressor"**. (*"multiclass\_classifier"*) –
- **binary\_classifier\_model\_selection\_criteria** (*str*) – One of ‘accuracy’, ‘f1’, ‘f\_beta’,
- **'recall\_at\_target\_precision', 'cross\_entropy\_loss', 'loss\_function' ('precision\_at\_target\_recall',)** –
- **target\_recall** (*float*) – Target recall. Only applicable if `binary_classifier_model_selection_criteria` is `precision_at_target_recall`.
- **target\_precision** (*float*) – Target precision. Only applicable if `binary_classifier_model_selection_criteria` is `recall_at_target_precision`.
- **positive\_example\_weight\_mult** (*float*) – The importance weight of positive examples is multiplied by this constant. Useful for skewed datasets. Only applies for classification tasks.
- **epochs** (*int*) – The maximum number of passes to make over the training data.
- **use\_bias** (*bool*) – Whether to include a bias field
- **num\_models** (*int*) – Number of models to train in parallel. If not set, the number of parallel models to train will be decided by the algorithm itself. One model will be trained according to the given training
- **parameter** (*regularization, optimizer, loss*) –
- **num\_calibration\_samples** (*int*) – Number of observations to use from validation dataset for doing model
- **calibration** (*finding the best threshold*) –
- **init\_method** (*str*) – Function to use to set the initial model weights. One of “uniform” or “normal”
- **init\_scale** (*float*) – For “uniform” init, the range of values.
- **init\_sigma** (*float*) – For “normal” init, the standard-deviation.
- **init\_bias** (*float*) – Initial weight for bias term
- **optimizer** (*str*) – One of ‘sgd’, ‘adam’, ‘rmsprop’ or ‘auto’
- **loss** (*str*) – One of ‘logistic’, ‘squared\_loss’, ‘absolute\_loss’, ‘hinge\_loss’,
- **'eps\_insensitive\_absolute\_loss', 'quantile\_loss', 'huber\_loss' or ('eps\_insensitive\_squared\_loss',)** –
- or **'auto'**. (*'softmax\_loss'*) –
- **wd** (*float*) – L2 regularization parameter i.e. the weight decay parameter. Use 0 for no L2 regularization.
- **l1** (*float*) – L1 regularization parameter. Use 0 for no L1 regularization.
- **momentum** (*float*) – Momentum parameter of sgd optimizer.
- **learning\_rate** (*float*) – The SGD learning rate
- **beta\_1** (*float*) – Exponential decay rate for first moment estimates. Only applies for adam optimizer.

- **beta\_2** (*float*) – Exponential decay rate for second moment estimates. Only applies for adam optimizer.
- **bias\_lr\_mult** (*float*) – Allows different learning rate for the bias term. The actual learning rate for the
- **is learning rate times bias\_lr\_mult.** (*bias*) –
- **bias\_wd\_mult** (*float*) – Allows different regularization for the bias term. The actual L2 regularization weight
- **the bias is wd times bias\_wd\_mult. By default there is no regularization on the bias term.** (*for*) –
- **use\_lr\_scheduler** (*bool*) – If true, we use a scheduler for the learning rate.
- **lr\_scheduler\_step** (*int*) – The number of steps between decreases of the learning rate. Only applies to learning rate scheduler.
- **lr\_scheduler\_factor** (*float*) – Every lr\_scheduler\_step the learning rate will decrease by this quantity. Only applies for learning rate scheduler.
- **lr\_scheduler\_minimum\_lr** (*float*) – The learning rate will never decrease to a value lower than this.
- **lr\_scheduler\_minimum\_lr** – Only applies for learning rate scheduler.
- **normalize\_data** (*bool*) – Normalizes the features before training to have standard deviation of 1.0.
- **normalize\_label** (*bool*) – Normalizes the regression label to have a standard deviation of 1.0. If set for classification, it will be ignored.
- **unbias\_data** (*bool*) – If true, features are modified to have mean 0.0.
- **ubias\_label** (*bool*) – If true, labels are modified to have mean 0.0.
- **num\_point\_for\_scaler** (*int*) – The number of data points to use for calculating the normalizing and unbiasing terms.
- **margin** (*float*) – the margin for hinge\_loss.
- **quantile** (*float*) – Quantile for quantile loss. For quantile q, the model will attempt to produce
- **such that true\_label < prediction with probability q.** (*predictions*) –
- **loss\_insensitivity** (*float*) – Parameter for epsilon insensitive loss type. During training and metric
- **any error smaller than this is considered to be zero.** (*evaluation,*) –
- **huber\_delta** (*float*) – Parameter for Huber loss. During training and metric evaluation, compute L2 loss for
- **smaller than delta and L1 loss for errors larger than delta.** (*errors*) –
- **early\_stopping\_patience** (*int*) – the number of epochs to wait before ending training if no improvement is
- **The improvement is training loss if validation data is not provided, or else it is the validation** (*made.*) –

- or the binary classification model selection criteria like accuracy, f1-score etc. To disable early (*loss*)-
- set `early_stopping_patience` to a value larger than epochs. (*stopping*,-)
- `early_stopping_tolerance` (*float*) – Relative tolerance to measure an improvement in loss. If the ratio of
- improvement in loss divided by the previous best loss is smaller than this value, early stopping will (*the*)-
- the improvement to be zero. (*consider*)-
- `num_classes` (*int*)- The number of classes for the response variable. Required when `predictor_type` is
- and ignored otherwise. The classes are assumed to be labeled 0, ..., `num_classes - 1`. (*multiclass\_classifier*)-
- `accuracy_top_k` (*int*)- The value of k when computing the Top K Accuracy metric for multiclass
- An example is scored as correct if the model assigns one of the top k scores to the true (*classification*.)-
- `label`. -
- `f_beta` (*float*) – The value of beta to use when calculating F score metrics for binary or multiclass
- Also used if `binary_classifier_model_selection_criteria` is `f_beta`. (*classification*.)-
- `balance_multiclass_weights` (*bool*) – Whether to use class weights which give each class equal importance in
- loss function. Only used when `predictor_type` is `multiclass_classifier`. (*the*)-
- **\*\*kwargs** – base class keyword argument values.

```
repo_name = 'linear-learner'
```

```
repo_version = 1
```

```
DEFAULT_MINI_BATCH_SIZE = 1000
```

```
classmethod attach (training_job_name, sagemaker_session=None)
```

Attach to an existing training job.

Create an Estimator bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching, if the training job has a Complete status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, attach will block and display log messages from the training job, until the training job completes.

#### Parameters

- `training_job_name` (*str*) – The name of the training job to attach to.

- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

## Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling `Estimator` Class with the attached training job.

**data\_location**

**delete\_endpoint()**

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

**deploy** (*initial\_instance\_count*, *instance\_type*, *endpoint\_name=None*, *\*\*kwargs*)

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, ‘ml.c4.xlarge’.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model()`. Implementations may customize `create_model()` to accept `**kwargs` to customize model creation during deploy. For more, see the implementation docs.

### Returns

**A predictor that provides a `predict()` method**, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** `sagemaker.predictor.RealTimePredictor`

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

`records` should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set()` to construct a `RecordSet` object from `ndarray` arrays.

#### Parameters

- **records** (`RecordSet`) – The records to train this `Estimator` on
- **mini\_batch\_size** (`int` or `None`) – The size of each mini-batch to use when training. If `None`, a default value will be used.
- **wait** (`bool`) – Whether the call should wait until the job completes (default: `True`).
- **logs** (`bool`) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).
- **job\_name** (`str`) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

#### `hyperparameters()`

Return the hyperparameters as a dictionary to use for training.

The `fit()` method, which trains the model, calls this method to find the hyperparameters.

**Returns** The hyperparameters.

**Return type** `dict[str, str]`

#### `model_data`

The model location in S3. Only set if `Estimator` has been `fit()`.

**Type** `str`

#### `record_set(train, labels=None, channel='train')`

Build a `RecordSet` from a `numpy ndarray` matrix and label vector.

For the 2D `ndarray train`, each row is converted to a `Record` object. The vector is stored in the “values” entry of the `features` property of each `Record`. If `labels` is not `None`, each corresponding label is assigned to the “values” entry of the `labels` property of each `Record`.

The collection of `Record` objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the `train_instance_count` property on this `Estimator`. One S3 object is created per training instance.

#### Parameters

- **train** (`numpy.ndarray`) – A 2D `numpy` array of training data.
- **labels** (`numpy.ndarray`) – A 1D `numpy` array of labels. Its length must be equal to the number of rows in `train`.
- **channel** (`str`) – The SageMaker TrainingJob channel this `RecordSet` should be assigned to.

**Returns** A `RecordSet` referencing the encoded, uploading training and label data.

**Return type** `RecordSet`

#### `train_image()`

Return the Docker image to use for training.

The `fit()` method, which does the model training, calls this method to find the image to use for model training.



**Returns** The URI of the Docker image.

**Return type** `str`

**training\_job\_analytics**

Return a `TrainingJobAnalytics` object for the current training job.

**transformer** (*instance\_count*, *instance\_type*, *strategy=None*, *assemble\_with=None*, *output\_path=None*, *output\_kms\_key=None*, *accept=None*, *env=None*, *max\_concurrent\_transforms=None*, *max\_payload=None*, *tags=None*, *role=None*, *volume\_kms\_key=None*)

Return a `Transformer` that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

**Parameters**

- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The `ExecutionRoleArn` IAM Role ARN for the `Model`, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

**normalize\_data**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**normalize\_label**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**unbias\_data**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**unbias\_label**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**num\_point\_for\_scaler**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**margin**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**quantile**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**loss\_insensitivity**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**huber\_delta**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**early\_stopping\_patience**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**early\_stopping\_tolerance**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**num\_classes**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**accuracy\_top\_k**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**f\_beta**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**balance\_multiclass\_weights**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**create\_model()**

Return a `LinearLearnerModel` referencing the latest s3 model data produced by this Estimator.

**class** `sagemaker.LinearLearnerModel` (*model\_data*, *role*, *sagemaker\_session=None*)

Bases: `sagemaker.model.Model`

Reference `LinearLearner` s3 model data. Calling `deploy()` creates an `Endpoint` and returns a `LinearLearnerPredictor`

**class** `sagemaker.LinearLearnerPredictor` (*endpoint*, *sagemaker\_session=None*)

Bases: `sagemaker.predictor.RealTimePredictor`

Performs binary-classification or regression prediction from input vectors.

The implementation of `predict()` in this `RealTimePredictor` requires a `numpy ndarray` as input. The array should contain the same number of columns as the feature-dimension of the data used to fit the model this Predictor performs inference on.

`predict()` returns a list of `Record` objects, one for each row in the input `ndarray`. The prediction is stored in the "predicted\_label" key of the `Record.label` field.

## 4.4 Amazon Estimators

Base class for Amazon Estimator implementations

```
class sagemaker.amazon.amazon_estimator.AmazonAlgorithmEstimatorBase (role,
                                                                    train_instance_count,
                                                                    train_instance_type,
                                                                    data_location=None,
                                                                    **kwargs)
```

Bases: `sagemaker.estimator.EstimatorBase`

Base class for Amazon first-party Estimator implementations. This class isn't intended to be instantiated directly.

Initialize an `AmazonAlgorithmEstimatorBase`.

**Parameters** `data_location` (*str* or *None*) – The s3 prefix to upload RecordSet objects to, expressed as an S3 url. For example “s3://example-bucket/some-key-prefix/”. Objects will be saved in a unique sub-directory of the specified location. If *None*, a default data location will be used.

**feature\_dim**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**mini\_batch\_size**

An algorithm hyperparameter with optional validation. Implemented as a python descriptor object.

**train\_image ()**

Return the Docker image to use for training.

The `fit ()` method, which does the model training, calls this method to find the image to use for model training.

**Returns** The URI of the Docker image.

**Return type** *str*

**hyperparameters ()**

Return the hyperparameters as a dictionary to use for training.

The `fit ()` method, which trains the model, calls this method to find the hyperparameters.

**Returns** The hyperparameters.

**Return type** `dict[str, str]`

**data\_location**

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

`records` should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set ()` to construct a `RecordSet` object from `ndarray` arrays.

**Parameters**

- **records** (`RecordSet`) – The records to train this Estimator on
- **mini\_batch\_size** (*int* or *None*) – The size of each mini-batch to use when training. If *None*, a default value will be used.
- **wait** (*bool*) – Whether the call should wait until the job completes (default: *True*).

- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).
- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

**record\_set** (*train*, *labels=None*, *channel='train'*)

Build a `RecordSet` from a `numpy ndarray` matrix and label vector.

For the 2D `numpy ndarray train`, each row is converted to a `Record` object. The vector is stored in the “values” entry of the `features` property of each `Record`. If `labels` is not `None`, each corresponding label is assigned to the “values” entry of the `labels` property of each `Record`.

The collection of `Record` objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the `train_instance_count` property on this Estimator. One S3 object is created per training instance.

#### Parameters

- **train** (*numpy.ndarray*) – A 2D `numpy` array of training data.
- **labels** (*numpy.ndarray*) – A 1D `numpy` array of labels. Its length must be equal to the number of rows in `train`.
- **channel** (*str*) – The SageMaker TrainingJob channel this `RecordSet` should be assigned to.

**Returns** A `RecordSet` referencing the encoded, uploading training and label data.

**Return type** `RecordSet`

## 4.5 FactorizationMachines

The Amazon SageMaker Factorization Machines algorithm.

```
class sagemaker.FactorizationMachines (role, train_instance_count, train_instance_type,
                                     num_factors, predictor_type, epochs=None,
                                     clip_gradient=None, eps=None, rescale_grad=None,
                                     bias_lr=None, linear_lr=None, factors_lr=None,
                                     bias_wd=None, linear_wd=None, factors_wd=None,
                                     bias_init_method=None, bias_init_scale=None,
                                     bias_init_sigma=None, bias_init_value=None,
                                     linear_init_method=None, linear_init_scale=None,
                                     linear_init_sigma=None, linear_init_value=None,
                                     factors_init_method=None, factors_init_scale=None,
                                     factors_init_sigma=None, factors_init_value=None,
                                     **kwargs)
```

Bases: `sagemaker.amazon.amazon_estimator.AmazonAlgorithmEstimatorBase`

Factorization Machines is Estimator for general-purpose supervised learning.

Amazon SageMaker Factorization Machines is a general-purpose supervised learning algorithm that you can use for both classification and regression tasks. It is an extension of a linear model that is designed to parsimoniously capture interactions between features within high dimensional sparse datasets.

This Estimator may be fit via calls to `fit()`. It requires Amazon `Record` protobuf serialized data to be stored in S3. There is an utility `record_set()` that can be used to upload data to S3 and creates `RecordSet` to be passed to the `fit` call.

To learn more about the Amazon protobuf Record class and how to prepare bulk data in this format, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

After this Estimator is fit, model data is stored in S3. The model may be deployed to an Amazon SageMaker Endpoint by invoking `deploy()`. As well as deploying an Endpoint, `deploy` returns a `FactorizationMachinesPredictor` object that can be used for inference calls using the trained model hosted in the SageMaker Endpoint.

FactorizationMachines Estimators can be configured by setting hyperparameters. The available hyperparameters for FactorizationMachines are documented below.

For further information on the AWS FactorizationMachines algorithm, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/fact-machines.html>

### Parameters

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if accessing AWS resource.
- **train\_instance\_count** (*int*) – Number of Amazon EC2 instances to use for training.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, ‘ml.c4.xlarge’.
- **num\_factors** (*int*) – Dimensionality of factorization.
- **predictor\_type** (*str*) – Type of predictor ‘binary\_classifier’ or ‘regressor’.
- **epochs** (*int*) – Number of training epochs to run.
- **clip\_gradient** (*float*) – Optimizer parameter. Clip the gradient by projecting onto the box [-clip\_gradient, +clip\_gradient]
- **eps** (*float*) – Optimizer parameter. Small value to avoid division by 0.
- **rescale\_grad** (*float*) – Optimizer parameter. If set, multiplies the gradient with `rescale_grad` before updating. Often choose to be `1.0/batch_size`.
- **bias\_lr** (*float*) – Non-negative learning rate for the bias term.
- **linear\_lr** (*float*) – Non-negative learning rate for linear terms.
- **factors\_lr** (*float*) – Non-negative learning rate for factorization terms.
- **bias\_wd** (*float*) – Non-negative weight decay for the bias term.
- **linear\_wd** (*float*) – Non-negative weight decay for linear terms.
- **factors\_wd** (*float*) – Non-negative weight decay for factorization terms.
- **bias\_init\_method** (*string*) – Initialization method for the bias term: ‘normal’, ‘uniform’ or ‘constant’.
- **bias\_init\_scale** (*float*) – Non-negative range for initialization of the bias term that takes effect when `bias_init_method` parameter is ‘uniform’
- **bias\_init\_sigma** (*float*) – Non-negative standard deviation for initialization of the bias term that takes effect when `bias_init_method` parameter is ‘normal’.
- **bias\_init\_value** (*float*) – Initial value of the bias term that takes effect when `bias_init_method` parameter is ‘constant’.

- **linear\_init\_method** (*string*) – Initialization method for linear term: ‘normal’, ‘uniform’ or ‘constant’.
- **linear\_init\_scale** (*float*) – Non-negative range for initialization of linear terms that takes effect when linear\_init\_method parameter is ‘uniform’.
- **linear\_init\_sigma** (*float*) – Non-negative standard deviation for initialization of linear terms that takes effect when linear\_init\_method parameter is ‘normal’.
- **linear\_init\_value** (*float*) – Initial value of linear terms that takes effect when linear\_init\_method parameter is ‘constant’.
- **factors\_init\_method** (*string*) – Initialization method for factorization term: ‘normal’, ‘uniform’ or ‘constant’.
- **factors\_init\_scale** (*float*) – Non-negative range for initialization of factorization terms that takes effect when factors\_init\_method parameter is ‘uniform’.
- **factors\_init\_sigma** (*float*) – Non-negative standard deviation for initialization of factorization terms that takes effect when factors\_init\_method parameter is ‘normal’.
- **factors\_init\_value** (*float*) – Initial value of factorization terms that takes effect when factors\_init\_method parameter is ‘constant’.
- **\*\*kwargs** – base class keyword argument values.

```
repo_name = 'factorization-machines'
```

```
repo_version = 1
```

```
classmethod attach(training_job_name, sagemaker_session=None)
```

Attach to an existing training job.

Create an Estimator bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching, if the training job has a Complete status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, attach will block and display log messages from the training job, until the training job completes.

#### Parameters

- **training\_job\_name** (*str*) – The name of the training job to attach to.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

#### Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling `Estimator` Class with the attached training job.

**data\_location**

**delete\_endpoint()**

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

**deploy** (*initial\_instance\_count*, *instance\_type*, *endpoint\_name=None*, *\*\*kwargs*)

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

#### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, 'ml.c4.xlarge'.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model()`. Implementations may customize `create_model()` to accept **\*\*kwargs** to customize model creation during deploy. For more, see the implementation docs.

#### Returns

A predictor that provides a **predict()** method, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** `sagemaker.predictor.RealTimePredictor`

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

`records` should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set()` to construct a `RecordSet` object from `ndarray` arrays.

#### Parameters

- **records** (`RecordSet`) – The records to train this Estimator on
- **mini\_batch\_size** (*int* or *None*) – The size of each mini-batch to use when training. If *None*, a default value will be used.
- **wait** (*bool*) – Whether the call should wait until the job completes (default: `True`).
- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).
- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

**hyperparameters ()**

Return the hyperparameters as a dictionary to use for training.

The `fit ()` method, which trains the model, calls this method to find the hyperparameters.

**Returns** The hyperparameters.

**Return type** `dict[str, str]`

**model\_data**

The model location in S3. Only set if Estimator has been `fit ()`.

**Type** `str`

**record\_set (train, labels=None, channel='train')**

Build a `RecordSet` from a `numpy ndarray` matrix and label vector.

For the 2D `ndarray train`, each row is converted to a `Record` object. The vector is stored in the “values” entry of the `features` property of each `Record`. If `labels` is not `None`, each corresponding label is assigned to the “values” entry of the `labels` property of each `Record`.

The collection of `Record` objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the `train_instance_count` property on this Estimator. One S3 object is created per training instance.

**Parameters**

- **train** (`numpy.ndarray`) – A 2D `numpy` array of training data.
- **labels** (`numpy.ndarray`) – A 1D `numpy` array of labels. Its length must be equal to the number of rows in `train`.
- **channel** (`str`) – The SageMaker TrainingJob channel this `RecordSet` should be assigned to.

**Returns** A `RecordSet` referencing the encoded, uploading training and label data.

**Return type** `RecordSet`

**train\_image ()**

Return the Docker image to use for training.

The `fit ()` method, which does the model training, calls this method to find the image to use for model training.

**Returns** The URI of the Docker image.

**Return type** `str`

**training\_job\_analytics**

Return a `TrainingJobAnalytics` object for the current training job.

**transformer (instance\_count, instance\_type, strategy=None, assemble\_with=None, output\_path=None, output\_kms\_key=None, accept=None, env=None, max\_concurrent\_transforms=None, max\_payload=None, tags=None, role=None, volume\_kms\_key=None)**

Return a `Transformer` that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

**Parameters**

- **instance\_count** (`int`) – Number of EC2 instances to use.
- **instance\_type** (`str`) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.



- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The ExecutionRoleArn IAM Role ARN for the Model, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

**create\_model** ()

Return a FactorizationMachinesModel referencing the latest s3 model data produced by this Estimator.

**class** sagemaker.**FactorizationMachinesModel** (*model\_data, role, sagemaker\_session=None*)

Bases: *sagemaker.model.Model*

Reference S3 model data created by FactorizationMachines estimator. Calling *deploy()* creates an Endpoint and returns *FactorizationMachinesPredictor*.

**class** sagemaker.**FactorizationMachinesPredictor** (*endpoint, sagemaker\_session=None*)

Bases: *sagemaker.predictor.RealTimePredictor*

Performs binary-classification or regression prediction from input vectors.

The implementation of *predict()* in this *RealTimePredictor* requires a numpy ndarray as input. The array should contain the same number of columns as the feature-dimension of the data used to fit the model this Predictor performs inference on.

*predict()* returns a list of Record objects, one for each row in the input ndarray. The prediction is stored in the "score" key of the Record.label field. Please refer to the formats details described: <https://docs.aws.amazon.com/sagemaker/latest/dg/fm-in-formats.html>

## 4.6 LDA

The Amazon SageMaker LDA algorithm.

```
class sagemaker.LDA(role, train_instance_type, num_topics, alpha0=None, max_restarts=None,
                    max_iterations=None, tol=None, **kwargs)
Bases: sagemaker.amazon.amazon_estimator.AmazonAlgorithmEstimatorBase
```

Latent Dirichlet Allocation (LDA) is `Estimator` used for unsupervised learning.

Amazon SageMaker Latent Dirichlet Allocation is an unsupervised learning algorithm that attempts to describe a set of observations as a mixture of distinct categories. LDA is most commonly used to discover a user-specified number of topics shared by documents within a text corpus. Here each observation is a document, the features are the presence (or occurrence count) of each word, and the categories are the topics.

This Estimator may be fit via calls to `fit()`. It requires Amazon Record protobuf serialized data to be stored in S3. There is an utility `record_set()` that can be used to upload data to S3 and creates `RecordSet` to be passed to the `fit` call.

To learn more about the Amazon protobuf Record class and how to prepare bulk data in this format, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

After this Estimator is fit, model data is stored in S3. The model may be deployed to an Amazon SageMaker Endpoint by invoking `deploy()`. As well as deploying an Endpoint, `deploy` returns a `LDAPredictor` object that can be used for inference calls using the trained model hosted in the SageMaker Endpoint.

LDA Estimators can be configured by setting hyperparameters. The available hyperparameters for LDA are documented below.

For further information on the AWS LDA algorithm, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/lda.html>

### Parameters

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if accessing AWS resource.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.
- **num\_topics** (*int*) – The number of topics for LDA to find within the data.
- **alpha0** (*float*) – Optional. Initial guess for the concentration parameter
- **max\_restarts** (*int*) – Optional. The number of restarts to perform during the Alternating Least Squares (ALS) spectral decomposition phase of the algorithm.
- **max\_iterations** (*int*) – Optional. The maximum number of iterations to perform during the ALS phase of the algorithm.
- **tol** (*float*) – Optional. Target error tolerance for the ALS phase of the algorithm.
- **\*\*kwargs** – base class keyword argument values.

```
repo_name = 'lda'
```

```
repo_version = 1
```

```
create_model()
```

Return a `LDAModel` referencing the latest s3 model data produced by this Estimator.

```
classmethod attach(training_job_name, sagemaker_session=None)
```

Attach to an existing training job.

Create an Estimator bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching,

if the training job has a Complete status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, `attach` will block and display log messages from the training job, until the training job completes.

#### Parameters

- **training\_job\_name** (*str*) – The name of the training job to attach to.
- **sagemaker\_session** (*sagemaker.session.Session*) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

#### Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling `Estimator` Class with the attached training job.

#### `data_location`

#### `delete_endpoint()`

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

#### `deploy(initial_instance_count, instance_type, endpoint_name=None, **kwargs)`

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

#### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, 'ml.c4.xlarge'.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model()`. Implementations may customize `create_model()` to accept `**kwargs` to customize model creation during deploy. For more, see the implementation docs.

#### Returns

A predictor that provides a `predict()` method, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** *sagemaker.predictor.RealTimePredictor*

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

*records* should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set()` to construct a `RecordSet` object from `ndarray` arrays.

#### Parameters

- **records** (`RecordSet`) – The records to train this Estimator on
- **mini\_batch\_size** (*int* or *None*) – The size of each mini-batch to use when training. If *None*, a default value will be used.
- **wait** (*bool*) – Whether the call should wait until the job completes (default: *True*).
- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when *wait* is *True* (default: *True*).
- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

**hyperparameters** ()

Return the hyperparameters as a dictionary to use for training.

The `fit()` method, which trains the model, calls this method to find the hyperparameters.

**Returns** The hyperparameters.

**Return type** `dict[str, str]`

**model\_data**

The model location in S3. Only set if Estimator has been `fit()`.

**Type** `str`

**record\_set** (*train*, *labels=None*, *channel='train'*)

Build a `RecordSet` from a `numpy ndarray` matrix and label vector.

For the 2D `ndarray train`, each row is converted to a `Record` object. The vector is stored in the “values” entry of the `features` property of each `Record`. If *labels* is not *None*, each corresponding label is assigned to the “values” entry of the `labels` property of each `Record`.

The collection of `Record` objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the `train_instance_count` property on this Estimator. One S3 object is created per training instance.

#### Parameters

- **train** (*numpy.ndarray*) – A 2D `numpy` array of training data.
- **labels** (*numpy.ndarray*) – A 1D `numpy` array of labels. Its length must be equal to the number of rows in *train*.
- **channel** (*str*) – The SageMaker TrainingJob channel this `RecordSet` should be assigned to.

**Returns** A RecordSet referencing the encoded, uploading training and label data.

**Return type** RecordSet

**train\_image()**

Return the Docker image to use for training.

The `fit()` method, which does the model training, calls this method to find the image to use for model training.

**Returns** The URI of the Docker image.

**Return type** str

**training\_job\_analytics**

Return a TrainingJobAnalytics object for the current training job.

**transformer** (*instance\_count*, *instance\_type*, *strategy=None*, *assemble\_with=None*, *output\_path=None*, *output\_kms\_key=None*, *accept=None*, *env=None*, *max\_concurrent\_transforms=None*, *max\_payload=None*, *tags=None*, *role=None*, *volume\_kms\_key=None*)

Return a Transformer that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

#### Parameters

- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The ExecutionRoleArn IAM Role ARN for the Model, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

**class** sagemaker.LDAModel (model\_data, role, sagemaker\_session=None)

Bases: *sagemaker.model.Model*

Reference LDA s3 model data. Calling *deploy()* creates an Endpoint and return a Predictor that transforms vectors to a lower-dimensional representation.

**class** sagemaker.LDAPredictor (endpoint, sagemaker\_session=None)

Bases: *sagemaker.predictor.RealTimePredictor*

Transforms input vectors to lower-dimesional representations.

The implementation of *predict()* in this *RealTimePredictor* requires a numpy ndarray as input. The array should contain the same number of columns as the feature-dimension of the data used to fit the model this Predictor performs inference on.

*predict()* returns a list of Record objects, one for each row in the input ndarray. The lower dimension vector result is stored in the *projection* key of the Record.*label* field.

## 4.7 NTM

The Amazon SageMaker NTM algorithm.

**class** sagemaker.NTM (role, train\_instance\_count, train\_instance\_type, num\_topics, encoder\_layers=None, epochs=None, encoder\_layers\_activation=None, optimizer=None, tolerance=None, num\_patience\_epochs=None, batch\_norm=None, rescale\_gradient=None, clip\_gradient=None, weight\_decay=None, learning\_rate=None, \*\*kwargs)

Bases: *sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase*

Neural Topic Model (NTM) is Estimator used for unsupervised learning.

This Estimator may be fit via calls to *fit()*. It requires Amazon Record protobuf serialized data to be stored in S3. There is an utility *record\_set()* that can be used to upload data to S3 and creates RecordSet to be passed to the *fit* call.

To learn more about the Amazon protobuf Record class and how to prepare bulk data in this format, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

After this Estimator is fit, model data is stored in S3. The model may be deployed to an Amazon SageMaker Endpoint by invoking *deploy()*. As well as deploying an Endpoint, *deploy* returns a NTMPredictor object that can be used for inference calls using the trained model hosted in the SageMaker Endpoint.

NTM Estimators can be configured by setting hyperparameters. The available hyperparameters for NTM are documented below.

For further information on the AWS NTM algorithm, please consult AWS technical documentation: <https://docs.aws.amazon.com/sagemaker/latest/dg/ntm.html>

### Parameters

- **role** (*str*) – An AWS IAM role (either name or full ARN). The Amazon SageMaker training jobs and APIs that create Amazon SageMaker endpoints use this role to access training data and model artifacts. After the endpoint is created, the inference code might use the IAM role, if accessing AWS resource.
- **train\_instance\_type** (*str*) – Type of EC2 instance to use for training, for example, 'ml.c4.xlarge'.
- **num\_topics** (*int*) – Required. The number of topics for NTM to find within the data.

- **encoder\_layers** (*list*) – Optional. Represents number of layers in the encoder and the output size of each layer.
- **epochs** (*int*) – Optional. Maximum number of passes over the training data.
- **encoder\_layers\_activation** (*str*) – Optional. Activation function to use in the encoder layers.
- **optimizer** (*str*) – Optional. Optimizer to use for training.
- **tolerance** (*float*) – Optional. Maximum relative change in the loss function within the last `num_patience_epochs` number of epochs below which early stopping is triggered.
- **num\_patience\_epochs** (*int*) – Optional. Number of successive epochs over which early stopping criterion is evaluated.
- **batch\_norm** (*bool*) – Optional. Whether to use batch normalization during training.
- **rescale\_gradient** (*float*) – Optional. Rescale factor for gradient.
- **clip\_gradient** (*float*) – Optional. Maximum magnitude for each gradient component.
- **weight\_decay** (*float*) – Optional. Weight decay coefficient. Adds L2 regularization.
- **learning\_rate** (*float*) – Optional. Learning rate for the optimizer.
- **\*\*kwargs** – base class keyword argument values.

```
repo_name = 'ntm'
```

```
repo_version = 1
```

```
classmethod attach (training_job_name, sagemaker_session=None)
```

Attach to an existing training job.

Create an Estimator bound to an existing training job, each subclass is responsible to implement `_prepare_init_params_from_job_description()` as this method delegates the actual conversion of a training job description to the arguments that the class constructor expects. After attaching, if the training job has a Complete status, it can be `deploy()` ed to create a SageMaker Endpoint and return a `Predictor`.

If the training job is in progress, attach will block and display log messages from the training job, until the training job completes.

#### Parameters

- **training\_job\_name** (*str*) – The name of the training job to attach to.
- **sagemaker\_session** (`sagemaker.session.Session`) – Session object which manages interactions with Amazon SageMaker APIs and any other AWS services needed. If not specified, the estimator creates one using the default AWS configuration chain.

#### Examples

```
>>> my_estimator.fit(wait=False)
>>> training_job_name = my_estimator.latest_training_job.name
Later on:
>>> attached_estimator = Estimator.attach(training_job_name)
>>> attached_estimator.deploy()
```

**Returns** Instance of the calling `Estimator` Class with the attached training job.

**data\_location**

**delete\_endpoint()**

Delete an Amazon SageMaker Endpoint.

**Raises** `ValueError` – If the endpoint does not exist.

**deploy** (*initial\_instance\_count*, *instance\_type*, *endpoint\_name=None*, *\*\*kwargs*)

Deploy the trained model to an Amazon SageMaker endpoint and return a `sagemaker.RealTimePredictor` object.

More information: <http://docs.aws.amazon.com/sagemaker/latest/dg/how-it-works-training.html>

#### Parameters

- **initial\_instance\_count** (*int*) – Minimum number of EC2 instances to deploy to an endpoint for prediction.
- **instance\_type** (*str*) – Type of EC2 instance to deploy to an endpoint for prediction, for example, ‘ml.c4.xlarge’.
- **endpoint\_name** (*str*) – Name to use for creating an Amazon SageMaker endpoint. If not specified, the name of the training job is used.
- **\*\*kwargs** – Passed to invocation of `create_model()`. Implementations may customize `create_model()` to accept **\*\*kwargs** to customize model creation during deploy. For more, see the implementation docs.

#### Returns

A predictor that provides a **predict()** method, which can be used to send requests to the Amazon SageMaker endpoint and obtain inferences.

**Return type** `sagemaker.predictor.RealTimePredictor`

**fit** (*records*, *mini\_batch\_size=None*, *wait=True*, *logs=True*, *job\_name=None*)

Fit this Estimator on serialized Record objects, stored in S3.

`records` should be an instance of `RecordSet`. This defines a collection of S3 data files to train this Estimator on.

Training data is expected to be encoded as dense or sparse vectors in the “values” feature on each Record. If the data is labeled, the label is expected to be encoded as a list of scalars in the “values” feature of the Record label.

More information on the Amazon Record format is available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/cdf-training.html>

See `record_set()` to construct a `RecordSet` object from `ndarray` arrays.

#### Parameters

- **records** (`RecordSet`) – The records to train this Estimator on
- **mini\_batch\_size** (*int* or *None*) – The size of each mini-batch to use when training. If *None*, a default value will be used.
- **wait** (*bool*) – Whether the call should wait until the job completes (default: `True`).
- **logs** (*bool*) – Whether to show the logs produced by the job. Only meaningful when `wait` is `True` (default: `True`).



- **job\_name** (*str*) – Training job name. If not specified, the estimator generates a default job name, based on the training image name and current timestamp.

#### **hyperparameters** ()

Return the hyperparameters as a dictionary to use for training.

The `fit()` method, which trains the model, calls this method to find the hyperparameters.

**Returns** The hyperparameters.

**Return type** `dict[str, str]`

#### **model\_data**

The model location in S3. Only set if Estimator has been `fit()`.

**Type** `str`

#### **record\_set** (*train, labels=None, channel='train'*)

Build a `RecordSet` from a `numpy ndarray` matrix and label vector.

For the 2D `ndarray train`, each row is converted to a `Record` object. The vector is stored in the “values” entry of the `features` property of each `Record`. If `labels` is not `None`, each corresponding label is assigned to the “values” entry of the `labels` property of each `Record`.

The collection of `Record` objects are protobuf serialized and uploaded to new S3 locations. A manifest file is generated containing the list of objects created and also stored in S3.

The number of S3 objects created is controlled by the `train_instance_count` property on this Estimator. One S3 object is created per training instance.

#### **Parameters**

- **train** (*numpy.ndarray*) – A 2D `numpy` array of training data.
- **labels** (*numpy.ndarray*) – A 1D `numpy` array of labels. Its length must be equal to the number of rows in `train`.
- **channel** (*str*) – The SageMaker TrainingJob channel this `RecordSet` should be assigned to.

**Returns** A `RecordSet` referencing the encoded, uploading training and label data.

**Return type** `RecordSet`

#### **train\_image** ()

Return the Docker image to use for training.

The `fit()` method, which does the model training, calls this method to find the image to use for model training.

**Returns** The URI of the Docker image.

**Return type** `str`

#### **training\_job\_analytics**

Return a `TrainingJobAnalytics` object for the current training job.

#### **transformer** (*instance\_count, instance\_type, strategy=None, assemble\_with=None, output\_path=None, output\_kms\_key=None, accept=None, env=None, max\_concurrent\_transforms=None, max\_payload=None, tags=None, role=None, volume\_kms\_key=None*)

Return a `Transformer` that uses a SageMaker Model based on the training job. It reuses the SageMaker Session and base job name used by the Estimator.

#### **Parameters**

- **instance\_count** (*int*) – Number of EC2 instances to use.
- **instance\_type** (*str*) – Type of EC2 instance to use, for example, ‘ml.c4.xlarge’.
- **strategy** (*str*) – The strategy used to decide how to batch records in a single request (default: None). Valid values: ‘MULTI\_RECORD’ and ‘SINGLE\_RECORD’.
- **assemble\_with** (*str*) – How the output is assembled (default: None). Valid values: ‘Line’ or ‘None’.
- **output\_path** (*str*) – S3 location for saving the transform result. If not specified, results are stored to a default bucket.
- **output\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the transform output (default: None).
- **accept** (*str*) – The content type accepted by the endpoint deployed during the transform job.
- **env** (*dict*) – Environment variables to be set for use during the transform job (default: None).
- **max\_concurrent\_transforms** (*int*) – The maximum number of HTTP requests to be made to each individual transform container at one time.
- **max\_payload** (*int*) – Maximum size of the payload in a single HTTP request to the container in MB.
- **tags** (*list[dict]*) – List of tags for labeling a transform job. If none specified, then the tags used for the training job are used for the transform job.
- **role** (*str*) – The ExecutionRoleArn IAM Role ARN for the Model, which is also used during transform jobs. If not specified, the role from the Estimator will be used.
- **volume\_kms\_key** (*str*) – Optional. KMS key ID for encrypting the volume attached to the ML compute instance (default: None).

**create\_model** ()

Return a NTMModel referencing the latest s3 model data produced by this Estimator.

**class** sagemaker.NTMModel (*model\_data, role, sagemaker\_session=None*)

Bases: *sagemaker.model.Model*

Reference NTM s3 model data. Calling *deploy()* creates an Endpoint and return a Predictor that transforms vectors to a lower-dimensional representation.

**class** sagemaker.NTMPredictor (*endpoint, sagemaker\_session=None*)

Bases: *sagemaker.predictor.RealTimePredictor*

Transforms input vectors to lower-dimensional representations.

The implementation of *predict()* in this *RealTimePredictor* requires a numpy ndarray as input. The array should contain the same number of columns as the feature-dimension of the data used to fit the model this Predictor performs inference on.

*predict()* returns a list of Record objects, one for each row in the input ndarray. The lower dimension vector result is stored in the *projection* key of the Record.label field.

**S**

`sagemaker.session`, 15



## A

accuracy\_top\_k (*sagemaker.LinearLearner* attribute), 54

AmazonAlgorithmEstimatorBase (*class in sagemaker.amazon.amazon\_estimator*), 54

analytics() (*sagemaker.tuner.HyperparameterTuner* method), 11

AnalyticsMetricsBase (*class in sagemaker.analytics*), 26

as\_json\_range() (*sagemaker.tuner.CategoricalParameter* method), 12

as\_tuning\_range() (*sagemaker.tuner.CategoricalParameter* method), 11

attach() (*sagemaker.estimator.Estimator* class method), 5

attach() (*sagemaker.FactorizationMachines* class method), 58

attach() (*sagemaker.KMeans* class method), 38

attach() (*sagemaker.LDA* class method), 62

attach() (*sagemaker.LinearLearner* class method), 50

attach() (*sagemaker.NTM* class method), 67

attach() (*sagemaker.PCA* class method), 43

attach() (*sagemaker.transformer.Transformer* class method), 14

attach() (*sagemaker.tuner.HyperparameterTuner* class method), 9

## B

balance\_multiclass\_weights (*sagemaker.LinearLearner* attribute), 54

best\_training\_job() (*sagemaker.tuner.HyperparameterTuner* method), 10

boto\_region\_name (*sagemaker.session.Session* attribute), 15

## C

CategoricalParameter (*class in sagemaker.tuner*), 11

clear\_cache() (*sagemaker.analytics.AnalyticsMetricsBase* method), 26

clear\_cache() (*sagemaker.analytics.HyperparameterTuningJobAnalytics* method), 26

clear\_cache() (*sagemaker.analytics.TrainingJobAnalytics* method), 27

CLOUDWATCH\_NAMESPACE (*sagemaker.analytics.TrainingJobAnalytics* attribute), 27

COMPLETE (*sagemaker.session.LogState* attribute), 15

config (*sagemaker.session.s3\_input* attribute), 24

container\_def() (*in module sagemaker.session*), 23

ContinuousParameter (*class in sagemaker.tuner*), 11

create\_endpoint() (*sagemaker.session.Session* method), 20

create\_endpoint\_config() (*sagemaker.session.Session* method), 19

create\_model() (*sagemaker.estimator.Estimator* method), 4

create\_model() (*sagemaker.FactorizationMachines* method), 61

create\_model() (*sagemaker.KMeans* method), 41

create\_model() (*sagemaker.LDA* method), 62

create\_model() (*sagemaker.LinearLearner* method), 54

create\_model() (*sagemaker.mxnet.estimator.MXNet* method), 30

create\_model() (*sagemaker.NTM* method), 70

create\_model() (*sagemaker.PCA* method), 43

create\_model() (*sagemaker.session.Session* method), 19

create\_model() (*sage-*

- maker.tensorflow.estimator.TensorFlow method*), 34
- `create_model_from_job()` (*sagemaker.session.Session method*), 19
- ## D
- `data_location` (*sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase attribute*), 55
- `data_location` (*sagemaker.FactorizationMachines attribute*), 58
- `data_location` (*sagemaker.KMeans attribute*), 39
- `data_location` (*sagemaker.LDA attribute*), 63
- `data_location` (*sagemaker.LinearLearner attribute*), 51
- `data_location` (*sagemaker.NTM attribute*), 68
- `data_location` (*sagemaker.PCA attribute*), 43
- `dataframe()` (*sagemaker.analytics.AnalyticsMetricsBase method*), 26
- `default_bucket()` (*sagemaker.session.Session method*), 16
- `DEFAULT_ESTIMATOR_CLS_NAME` (*sagemaker.tuner.HyperparameterTuner attribute*), 8
- `DEFAULT_ESTIMATOR_MODULE` (*sagemaker.tuner.HyperparameterTuner attribute*), 8
- `DEFAULT_MINI_BATCH_SIZE` (*sagemaker.LinearLearner attribute*), 50
- `DEFAULT_MINI_BATCH_SIZE` (*sagemaker.PCA attribute*), 43
- `delete_endpoint()` (*sagemaker.estimator.Estimator method*), 6
- `delete_endpoint()` (*sagemaker.FactorizationMachines method*), 59
- `delete_endpoint()` (*sagemaker.KMeans method*), 39
- `delete_endpoint()` (*sagemaker.LDA method*), 63
- `delete_endpoint()` (*sagemaker.LinearLearner method*), 51
- `delete_endpoint()` (*sagemaker.NTM method*), 68
- `delete_endpoint()` (*sagemaker.PCA method*), 43
- `delete_endpoint()` (*sagemaker.predictor.RealTimePredictor method*), 13
- `delete_endpoint()` (*sagemaker.session.Session method*), 20
- `delete_endpoint()` (*sagemaker.tuner.HyperparameterTuner method*), 10
- `deploy()` (*sagemaker.estimator.Estimator method*), 6
- `deploy()` (*sagemaker.FactorizationMachines method*), 59
- `deploy()` (*sagemaker.KMeans method*), 39
- `deploy()` (*sagemaker.LDA method*), 63
- `deploy()` (*sagemaker.LinearLearner method*), 51
- `deploy()` (*sagemaker.model.Model method*), 25
- `deploy()` (*sagemaker.NTM method*), 68
- `deploy()` (*sagemaker.PCA method*), 43
- `deploy()` (*sagemaker.tuner.HyperparameterTuner method*), 10
- `description()` (*sagemaker.analytics.HyperparameterTuningJobAnalytics method*), 26
- ## E
- `early_stopping_patience` (*sagemaker.LinearLearner attribute*), 54
- `early_stopping_tolerance` (*sagemaker.LinearLearner attribute*), 54
- `endpoint_from_job()` (*sagemaker.session.Session method*), 21
- `endpoint_from_model_data()` (*sagemaker.session.Session method*), 22
- `endpoint_from_production_variants()` (*sagemaker.session.Session method*), 22
- `Estimator` (*class in sagemaker.estimator*), 3
- `eval_metrics` (*sagemaker.KMeans attribute*), 41
- `expand_role()` (*sagemaker.session.Session method*), 22
- `export_csv()` (*sagemaker.analytics.AnalyticsMetricsBase method*), 26
- ## F
- `f_beta` (*sagemaker.LinearLearner attribute*), 54
- `FactorizationMachines` (*class in sagemaker*), 56
- `FactorizationMachinesModel` (*class in sagemaker*), 61
- `FactorizationMachinesPredictor` (*class in sagemaker*), 61
- `feature_dim` (*sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase attribute*), 55
- `fit()` (*sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimator method*), 55
- `fit()` (*sagemaker.estimator.Estimator method*), 6
- `fit()` (*sagemaker.FactorizationMachines method*), 59
- `fit()` (*sagemaker.KMeans method*), 39
- `fit()` (*sagemaker.LDA method*), 63
- `fit()` (*sagemaker.LinearLearner method*), 51
- `fit()` (*sagemaker.NTM method*), 68
- `fit()` (*sagemaker.PCA method*), 44
- `fit()` (*sagemaker.tensorflow.estimator.TensorFlow method*), 34
- `fit()` (*sagemaker.tuner.HyperparameterTuner method*), 9

## G

`get_caller_identity_arn()` (*sagemaker.session.Session* method), 23

`get_execution_role()` (in module *sagemaker.session*), 24

## H

`huber_delta` (*sagemaker.LinearLearner* attribute), 54

`hyperparameter_ranges()` (*sagemaker.tuner.HyperparameterTuner* method), 11

`hyperparameters()` (*sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase* method), 55

`hyperparameters()` (*sagemaker.estimator.Estimator* method), 4

`hyperparameters()` (*sagemaker.FactorizationMachines* method), 59

`hyperparameters()` (*sagemaker.KMeans* method), 41

`hyperparameters()` (*sagemaker.LDA* method), 64

`hyperparameters()` (*sagemaker.LinearLearner* method), 52

`hyperparameters()` (*sagemaker.NTM* method), 69

`hyperparameters()` (*sagemaker.PCA* method), 44

`hyperparameters()` (*sagemaker.tensorflow.estimator.TensorFlow* method), 35

*HyperparameterTuner* (class in *sagemaker.tuner*), 8

*HyperparameterTuningJobAnalytics* (class in *sagemaker.analytics*), 26

## I

*IntegerParameter* (class in *sagemaker.tuner*), 11

## J

`JOB_COMPLETE` (*sagemaker.session.LogState* attribute), 15

## K

*KMeans* (class in *sagemaker*), 37

*KMeansModel* (class in *sagemaker*), 41

*KMeansPredictor* (class in *sagemaker*), 41

## L

*LDA* (class in *sagemaker*), 61

*LDAModel* (class in *sagemaker*), 65

*LDAPredictor* (class in *sagemaker*), 66

*LinearLearner* (class in *sagemaker*), 46

*LinearLearnerModel* (class in *sagemaker*), 54

*LinearLearnerPredictor* (class in *sagemaker*), 54

`logs_for_job()` (*sagemaker.session.Session* method), 23

*LogState* (class in *sagemaker.session*), 15

`loss_insensitivity` (*sagemaker.LinearLearner* attribute), 54

## M

`margin` (*sagemaker.LinearLearner* attribute), 54

`mini_batch_size` (*sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase* attribute), 55

*Model* (class in *sagemaker.model*), 24

*model\_data* (*sagemaker.estimator.Estimator* attribute), 6

`model_data` (*sagemaker.FactorizationMachines* attribute), 60

`model_data` (*sagemaker.KMeans* attribute), 40

`model_data` (*sagemaker.LDA* attribute), 64

`model_data` (*sagemaker.LinearLearner* attribute), 52

`model_data` (*sagemaker.NTM* attribute), 69

`model_data` (*sagemaker.PCA* attribute), 44

*MXNet* (class in *sagemaker.mxnet.estimator*), 29

*MXNetModel* (class in *sagemaker.mxnet.model*), 30

*MXNetPredictor* (class in *sagemaker.mxnet.model*), 31

## N

`name` (*sagemaker.analytics.HyperparameterTuningJobAnalytics* attribute), 26

`name` (*sagemaker.analytics.TrainingJobAnalytics* attribute), 27

`normalize_data` (*sagemaker.LinearLearner* attribute), 53

`normalize_label` (*sagemaker.LinearLearner* attribute), 53

*NTM* (class in *sagemaker*), 66

*NTMModel* (class in *sagemaker*), 70

*NTMPredictor* (class in *sagemaker*), 70

`num_classes` (*sagemaker.LinearLearner* attribute), 54

`num_point_for_scaler` (*sagemaker.LinearLearner* attribute), 54

## P

*PCA* (class in *sagemaker*), 42

*PCAModel* (class in *sagemaker*), 46

*PCAPredictor* (class in *sagemaker*), 46

`predict()` (*sagemaker.predictor.RealTimePredictor* method), 12

`prepare_container_def()` (*sagemaker.model.Model* method), 25

`prepare_container_def()` (*sagemaker.mxnet.model.MXNetModel* method), 31

prepare\_container\_def() (sagemaker.tensorflow.model.TensorFlowModel method), 35  
 production\_variant() (in module sagemaker.session), 23

## Q

quantile (sagemaker.LinearLearner attribute), 54

## R

RealTimePredictor (class in sagemaker.predictor), 12  
 record\_set() (sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase method), 56  
 record\_set() (sagemaker.FactorizationMachines method), 60  
 record\_set() (sagemaker.KMeans method), 40  
 record\_set() (sagemaker.LDA method), 64  
 record\_set() (sagemaker.LinearLearner method), 52  
 record\_set() (sagemaker.NTM method), 69  
 record\_set() (sagemaker.PCA method), 44  
 repo\_name (sagemaker.FactorizationMachines attribute), 58  
 repo\_name (sagemaker.KMeans attribute), 38  
 repo\_name (sagemaker.LDA attribute), 62  
 repo\_name (sagemaker.LinearLearner attribute), 50  
 repo\_name (sagemaker.NTM attribute), 67  
 repo\_name (sagemaker.PCA attribute), 43  
 repo\_version (sagemaker.FactorizationMachines attribute), 58  
 repo\_version (sagemaker.KMeans attribute), 38  
 repo\_version (sagemaker.LDA attribute), 62  
 repo\_version (sagemaker.LinearLearner attribute), 50  
 repo\_version (sagemaker.NTM attribute), 67  
 repo\_version (sagemaker.PCA attribute), 43

## S

s3\_input (class in sagemaker.session), 24  
 sagemaker.session (module), 15  
 SAGEMAKER\_ESTIMATOR\_CLASS\_NAME (sagemaker.tuner.HyperparameterTuner attribute), 8  
 SAGEMAKER\_ESTIMATOR\_MODULE (sagemaker.tuner.HyperparameterTuner attribute), 8  
 sagemaker\_session (sagemaker.tuner.HyperparameterTuner attribute), 11  
 Session (class in sagemaker.session), 15  
 set\_hyperparameters() (sagemaker.estimator.Estimator method), 4

STARTING (sagemaker.session.LogState attribute), 15  
 stop\_tuning\_job() (sagemaker.session.Session method), 18  
 stop\_tuning\_job() (sagemaker.tuner.HyperparameterTuner method), 10

## T

TAILING (sagemaker.session.LogState attribute), 15  
 TensorFlow (class in sagemaker.tensorflow.estimator), 33  
 TensorFlowModel (class in sagemaker.tensorflow.model), 35  
 TensorFlowBasePredictor (class in sagemaker.tensorflow.model), 36  
 train() (sagemaker.session.Session method), 16  
 train\_image() (sagemaker.amazon.amazon\_estimator.AmazonAlgorithmEstimatorBase method), 55  
 train\_image() (sagemaker.estimator.Estimator method), 4  
 train\_image() (sagemaker.FactorizationMachines method), 60  
 train\_image() (sagemaker.KMeans method), 40  
 train\_image() (sagemaker.LDA method), 65  
 train\_image() (sagemaker.LinearLearner method), 52  
 train\_image() (sagemaker.NTM method), 69  
 train\_image() (sagemaker.PCA method), 45  
 training\_job\_analytics (sagemaker.estimator.Estimator attribute), 7  
 training\_job\_analytics (sagemaker.FactorizationMachines attribute), 60  
 training\_job\_analytics (sagemaker.KMeans attribute), 40  
 training\_job\_analytics (sagemaker.LDA attribute), 65  
 training\_job\_analytics (sagemaker.LinearLearner attribute), 53  
 training\_job\_analytics (sagemaker.NTM attribute), 69  
 training\_job\_analytics (sagemaker.PCA attribute), 45  
 training\_job\_summaries() (sagemaker.analytics.HyperparameterTuningJobAnalytics method), 27  
 TrainingJobAnalytics (class in sagemaker.analytics), 27  
 transform() (sagemaker.session.Session method), 18  
 transform() (sagemaker.transformer.Transformer method), 14  
 Transformer (class in sagemaker.transformer), 13



`transformer()` (*sagemaker.estimator.Estimator method*), 7  
`transformer()` (*sagemaker.FactorizationMachines method*), 60  
`transformer()` (*sagemaker.KMeans method*), 40  
`transformer()` (*sagemaker.LDA method*), 65  
`transformer()` (*sagemaker.LinearLearner method*), 53  
`transformer()` (*sagemaker.NTM method*), 69  
`transformer()` (*sagemaker.PCA method*), 45  
`tune()` (*sagemaker.session.Session method*), 17  
`TUNING_JOB_NAME_MAX_LENGTH` (*sagemaker.tuner.HyperparameterTuner attribute*), 8  
`tuning_ranges` (*sagemaker.analytics.HyperparameterTuningJobAnalytics attribute*), 26

## U

`unbias_data` (*sagemaker.LinearLearner attribute*), 53  
`unbias_label` (*sagemaker.LinearLearner attribute*), 53  
`upload_data()` (*sagemaker.session.Session method*), 15

## W

`wait()` (*sagemaker.transformer.Transformer method*), 14  
`wait()` (*sagemaker.tuner.HyperparameterTuner method*), 10  
`wait_for_endpoint()` (*sagemaker.session.Session method*), 21  
`wait_for_job()` (*sagemaker.session.Session method*), 20  
`wait_for_transform_job()` (*sagemaker.session.Session method*), 21  
`wait_for_tuning_job()` (*sagemaker.session.Session method*), 20  
`WAIT_IN_PROGRESS` (*sagemaker.session.LogState attribute*), 15