04_Model_Training

April 13, 2022

0.1 Load required libraries

[2]:
```
import boto3
import sagemaker
from sagemaker.xgboost.estimator import XGBoost
from sagemaker.session import Session

sess = sagemaker.Session()
role = sagemaker.get_execution_role()
region = boto3.Session().region_name
```

0.2 Create an XGBoost estimator

[2]:
```
# Construct a SageMaker estimator that calls the xgboost-container

from sagemaker.debugger import Rule, rule_configs
from sagemaker import image_uris

bucket = "ads508-team4-xgboost"
prefix = "models"
s3_output_location='s3://{}/{}/{}'.format(bucket, prefix, 'xgboost')

# Set up container

container = sagemaker.image_uris.retrieve("xgboost", region, "1.2-1")
print(container)

xgb_model = sagemaker.estimator.Estimator(
    image_uri = container,
    role = role,
    instance_count = 1,
    instance_type = 'ml.m5.large',
    volume_size = 5,
    output_path = s3_output_location,
    sagemaker_session = sagemaker.Session(),
    rules = [Rule.sagemaker(rule_configs.create_xgboost_report())]
)
```

683313688378.dkr.ecr.us-east-1.amazonaws.com/sagemaker-xgboost:1.2-1

0.3 Set hyperparameters for xgboost

[3]:
```
xgb_model.set_hyperparameters(
    max_depth = 5,
    eta = 0.2,
    gamma = 4,
    min_child_weight = 6,
    subsample = 0.7,
    objective = "multi:softmax",
    num_round = 20,
    num_class = 8
)
```

0.4 Set path for input files

[4]:
```
from sagemaker.session import TrainingInput

content_type = "csv"

train_input = TrainingInput('s3://ads508-team4-split/train/df_train.
 ↪csv',content_type = content_type)
validation_input = TrainingInput('s3://ads508-team4-split/validation/
 ↪df_validation.csv', content_type = content_type)
test_input = TrainingInput('s3://ads508-team4-split/test/df_test.csv',␣
 ↪content_type = content_type)
```

0.5 Start Training

[5]:
```
xgb_model.fit({"train": train_input, "validation": validation_input}, wait=True)
```

```
2022-03-30 19:55:33 Starting - Starting the training job...
2022-03-30 19:55:57 Starting - Preparing the instances for
trainingCreateXgboostReport: InProgress
ProfilerReport-1648670133: InProgress
...
2022-03-30 19:57:31 Downloading - Downloading input data...
2022-03-30 19:58:31 Training - Downloading the training image...
2022-03-30 19:59:31 Training - Training image download completed. Training in
progress.[2022-03-30 19:59:24.315 ip-10-0-83-31.ec2.internal:1 INFO

utils.py:27] RULE_JOB_STOP_SIGNAL_FILENAME: None
INFO:sagemaker-containers:Imported framework

sagemaker_xgboost_container.training
INFO:sagemaker-containers:Failed to parse hyperparameter objective value

multi:softmax to Json.
```

```
Returning the value itself
INFO:sagemaker-containers:No GPUs detected (normal if no gpus

installed)
INFO:sagemaker_xgboost_container.training:Running XGBoost Sagemaker in

algorithm mode
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Determined delimiter of CSV input is ','
INFO:root:Single node training.
[2022-03-30 19:59:25.698 ip-10-0-83-31.ec2.internal:1 INFO

json_config.py:91] Creating hook from json_config at

/opt/ml/input/config/debughookconfig.json.
[2022-03-30 19:59:25.699 ip-10-0-83-31.ec2.internal:1 INFO hook.py:199]

tensorboard_dir has not been set for the hook. SMDebug will not be exporting

tensorboard summaries.
[2022-03-30 19:59:25.700 ip-10-0-83-31.ec2.internal:1 INFO

profiler_config_parser.py:102] User has disabled profiler.
[2022-03-30 19:59:25.701 ip-10-0-83-31.ec2.internal:1 INFO hook.py:253]

Saving to /opt/ml/output/tensors
[2022-03-30 19:59:25.701 ip-10-0-83-31.ec2.internal:1 INFO

state_store.py:77] The checkpoint config file

/opt/ml/input/config/checkpointconfig.json does not exist.
INFO:root:Debug hook created from config
INFO:root:Train matrix has 192332 rows and 76 columns
INFO:root:Validation matrix has 41214 rows
[0]#011train-merror:0.63710#011validation-merror:0.64303
[2022-03-30 19:59:31.917 ip-10-0-83-31.ec2.internal:1 INFO hook.py:413]

Monitoring the collections: feature_importance, metrics, predictions, labels,

hyperparameters
[2022-03-30 19:59:31.919 ip-10-0-83-31.ec2.internal:1 INFO hook.py:476]

Hook is writing from the hook with pid: 1
[1]#011train-merror:0.62081#011validation-merror:0.62515
[2]#011train-merror:0.61823#011validation-merror:0.62338
[3]#011train-merror:0.61492#011validation-merror:0.61916
[4]#011train-merror:0.60763#011validation-merror:0.61220
[5]#011train-merror:0.60080#011validation-merror:0.60581
[6]#011train-merror:0.59533#011validation-merror:0.60120
[7]#011train-merror:0.59256#011validation-merror:0.59754
[8]#011train-merror:0.59230#011validation-merror:0.59696
```

```
[9]#011train-merror:0.59390#011validation-merror:0.59856
[10]#011train-merror:0.58636#011validation-merror:0.59155
[11]#011train-merror:0.58443#011validation-merror:0.58929
[12]#011train-merror:0.58244#011validation-merror:0.58783
[13]#011train-merror:0.57860#011validation-merror:0.58335
[14]#011train-merror:0.57612#011validation-merror:0.58036
[15]#011train-merror:0.57434#011validation-merror:0.57905
[16]#011train-merror:0.57119#011validation-merror:0.57551
[17]#011train-merror:0.57040#011validation-merror:0.57493
[18]#011train-merror:0.56768#011validation-merror:0.57211
[19]#011train-merror:0.56533#011validation-merror:0.56995

2022-03-30 20:01:37 Uploading - Uploading generated training model
2022-03-30 20:01:37 Completed - Training job completed
Training seconds: 258
Billable seconds: 258
```

0.6 Show the name of the training job

[6]:
```
training_job_name = xgb_model.latest_training_job.name
print("Training Job Name:  {}".format(training_job_name))
```

```
Training Job Name:  sagemaker-xgboost-2022-03-30-19-55-33-695
```

0.7 Show training job metrics

[7]:
```
xgb_model.training_job_analytics.dataframe()
```

```
Warning: No metrics called train:mae found
Warning: No metrics called validation:mae found
Warning: No metrics called train:rmse found
Warning: No metrics called validation:accuracy found
Warning: No metrics called train:mlogloss found
Warning: No metrics called validation:balanced_accuracy found
Warning: No metrics called train:cox-nloglik found
Warning: No metrics called validation:f1 found
Warning: No metrics called validation:rmse found
Warning: No metrics called validation:cox-nloglik found
Warning: No metrics called validation:mse found
Warning: No metrics called validation:ndcg found
Warning: No metrics called train:accuracy found
Warning: No metrics called train:mse found
Warning: No metrics called train:f1 found
Warning: No metrics called validation:mlogloss found
Warning: No metrics called train:ndcg found
Warning: No metrics called train:map found
Warning: No metrics called validation:map found
```

[7]:

	timestamp	metric_name	value
0	0.0	train:merror	0.616582
1	60.0	train:merror	0.585638
2	120.0	train:merror	0.568650
3	0.0	validation:merror	0.621455
4	60.0	validation:merror	0.590569
5	120.0	validation:merror	0.573125

1 Deploy the model to a real-time endpoint

[8]:
```
xgb_predictor = xgb_model.deploy(initial_instance_count = 1, instance_type = 'ml.
↪m5.xlarge')
```

```
------!
```

To send it in an HTTP POST request, we'll serialize it as a CSV string and then decode the resulting CSV.

[14]:
```
!pip install boto3 --upgrade
from sagemaker.serializers import CSVSerializer


xgb_predictor.serializers = sagemaker.serializers.CSVSerializer()
```

```
/opt/conda/lib/python3.7/site-packages/secretstorage/dhcrypto.py:16:
CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes
instead
  from cryptography.utils import int_from_bytes
/opt/conda/lib/python3.7/site-packages/secretstorage/util.py:25:
CryptographyDeprecationWarning: int_from_bytes is deprecated, use int.from_bytes
instead
  from cryptography.utils import int_from_bytes
Requirement already satisfied: boto3 in /opt/conda/lib/python3.7/site-packages
(1.21.29)
Requirement already satisfied: s3transfer<0.6.0,>=0.5.0 in
/opt/conda/lib/python3.7/site-packages (from boto3) (0.5.0)
Requirement already satisfied: botocore<1.25.0,>=1.24.29 in
/opt/conda/lib/python3.7/site-packages (from boto3) (1.24.29)
Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in
/opt/conda/lib/python3.7/site-packages (from boto3) (0.10.0)
Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in
/opt/conda/lib/python3.7/site-packages (from botocore<1.25.0,>=1.24.29->boto3)
(2.8.1)
Requirement already satisfied: urllib3<1.27,>=1.25.4 in
/opt/conda/lib/python3.7/site-packages (from botocore<1.25.0,>=1.24.29->boto3)
(1.26.7)
Requirement already satisfied: six>=1.5 in /opt/conda/lib/python3.7/site-
packages (from python-dateutil<3.0.0,>=2.1->botocore<1.25.0,>=1.24.29->boto3)
```

```
(1.14.0)
WARNING: Running pip as the 'root' user can result in broken permissions
and conflicting behaviour with the system package manager. It is recommended to
use a virtual environment instead: https://pip.pypa.io/warnings/venv
WARNING: You are using pip version 21.3.1; however, version 22.0.4 is
available.
You should consider upgrading via the '/opt/conda/bin/python -m pip install
--upgrade pip' command.
```

2 Download our test file and saved it to local instance

[16]:
```
import pandas as pd
import numpy as np
import csv

!aws s3 cp 's3://ads508-team4-split/test/df_test.csv' ./data/

df_test = pd.read_csv(
    "./data/df_test.csv",
    delimiter=",",
    quoting=csv.QUOTE_NONE,
)
df_test.head()
```

```
download: s3://ads508-team4-split/test/df_test.csv to data/df_test.csv
```

[16]:

	4	115	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	...	0.63	1.2	0.64	0.65	\
0	6	0	1	0	0	0	0	0	0	1	...	0	0	0	0	
1	0	84	0	0	0	0	1	0	0	0	...	0	1	0	0	
2	2	77	1	0	0	0	0	0	0	1	...	0	0	0	0	
3	5	44	1	0	0	0	0	0	0	1	...	0	0	0	0	
4	1	79	0	0	0	0	0	0	0	0	...	0	0	0	0	

	0.66	0.67	0.68	0.69	1.3	0.70
0	0	0	0	0	1	0
1	0	0	0	0	1	0
2	0	0	0	0	1	0
3	0	0	0	0	1	0
4	0	0	0	0	0	0

[5 rows x 77 columns]