Lab7-Subscriber

April 4, 2022

[1]:
```
import os
os.environ['PYSPARK_SUBMIT_ARGS'] = '--packages org.apache.spark:
 ↪spark-sql-kafka-0-10_2.12:3.1.2 pyspark-shell'
```

[2]:
```
from pyspark.sql.types import StructType, StringType, FloatType
from pyspark.sql import SparkSession
from pyspark.sql.functions import explode
from pyspark.sql.functions import split
from pyspark.ml.classification import RandomForestClassificationModel
from pyspark.ml.evaluation import MulticlassClassificationEvaluator
from pyspark.ml.linalg import Vectors
from pyspark.sql import Row
import pyspark.sql.functions as f
from pyspark.ml import PipelineModel
from itertools import chain

spark = SparkSession\
    .builder\
    .appName("Iris-Prediction")\
    .config("spark.driver.extraClassPath", "/home/ubuntu/jars/
 ↪spark-sql-kafka-0-10_2.12-3.1.2.jar,/home/ubuntu/jars/commons-pool2-2.11.0.
 ↪jar")\
    .getOrCreate()

spark.sparkContext.setLogLevel('WARN')
```

```
:: loading settings :: url = jar:file:/usr/lib/spark/jars/ivy-2.4.0.jar!/org/apa
che/ivy/core/settings/ivysettings.xml

Ivy Default Cache set to: /root/.ivy2/cache
The jars for the packages stored in: /root/.ivy2/jars
org.apache.spark#spark-sql-kafka-0-10_2.12 added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-
parent-b4e4886c-b385-40e0-8941-cde23bf18d8d;1.0
        confs: [default]
        found org.apache.spark#spark-sql-kafka-0-10_2.12;3.1.2 in central
        found org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.1.2 in
central
```

```
        found org.apache.kafka#kafka-clients;2.6.0 in central
        found com.github.luben#zstd-jni;1.4.8-1 in central
        found org.lz4#lz4-java;1.7.1 in central
        found org.xerial.snappy#snappy-java;1.1.8.2 in central
        found org.slf4j#slf4j-api;1.7.30 in central
        found org.spark-project.spark#unused;1.0.0 in central
        found org.apache.commons#commons-pool2;2.6.2 in central
:: resolution report :: resolve 650ms :: artifacts dl 12ms
        :: modules in use:
        com.github.luben#zstd-jni;1.4.8-1 from central in [default]
        org.apache.commons#commons-pool2;2.6.2 from central in [default]
        org.apache.kafka#kafka-clients;2.6.0 from central in [default]
        org.apache.spark#spark-sql-kafka-0-10_2.12;3.1.2 from central in
[default]
        org.apache.spark#spark-token-provider-kafka-0-10_2.12;3.1.2 from central
in [default]
        org.lz4#lz4-java;1.7.1 from central in [default]
        org.slf4j#slf4j-api;1.7.30 from central in [default]
        org.spark-project.spark#unused;1.0.0 from central in [default]
        org.xerial.snappy#snappy-java;1.1.8.2 from central in [default]
        ---------------------------------------------------------------------
        |                  |            modules            ||   artifacts   |
        |       conf       | number| search|dwnlded|evicted|| number|dwnlded|
        ---------------------------------------------------------------------
        |      default     |   9   |   0   |   0   |   0   ||   9   |   0   |
        ---------------------------------------------------------------------
:: retrieving :: org.apache.spark#spark-submit-
parent-b4e4886c-b385-40e0-8941-cde23bf18d8d
        confs: [default]
        0 artifacts copied, 9 already retrieved (0kB/11ms)
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use
setLogLevel(newLevel).
22/04/04 16:44:33 INFO org.apache.spark.SparkEnv: Registering MapOutputTracker
22/04/04 16:44:33 INFO org.apache.spark.SparkEnv: Registering BlockManagerMaster
22/04/04 16:44:33 INFO org.apache.spark.SparkEnv: Registering
BlockManagerMasterHeartbeat
22/04/04 16:44:33 INFO org.apache.spark.SparkEnv: Registering
OutputCommitCoordinator
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.apache.spark_spark-sql-kafka-0-10_2.12-3.1.2.jar
added multiple times to distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.apache.spark_spark-token-provider-
kafka-0-10_2.12-3.1.2.jar added multiple times to distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.apache.kafka_kafka-clients-2.6.0.jar added multiple
times to distributed cache.
```

```
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.apache.commons_commons-pool2-2.6.2.jar added
multiple times to distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.spark-project.spark_unused-1.0.0.jar added multiple
times to distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/com.github.luben_zstd-jni-1.4.8-1.jar added multiple
times to distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.lz4_lz4-java-1.7.1.jar added multiple times to
distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.xerial.snappy_snappy-java-1.1.8.2.jar added multiple
times to distributed cache.
22/04/04 16:44:36 WARN org.apache.spark.deploy.yarn.Client: Same path resource
file:///root/.ivy2/jars/org.slf4j_slf4j-api-1.7.30.jar added multiple times to
distributed cache.
```

[3]:

```python
df = spark.readStream.format('kafka').option('kafka.bootstrap.servers', '10.188.
 ↪0.2:9092').option("startingOffsets", "earliest").option('subscribe',␣
 ↪'iris-data').option("failOnDataLoss", "false").load()
df = df.selectExpr("CAST(value AS STRING)")

schema = StructType()\
    .add("sepal_length", FloatType())\
    .add("sepal_width", FloatType())\
    .add("petal_length", FloatType())\
    .add("petal_width", FloatType())\
    .add("species", StringType())

print(df.isStreaming)

df.printSchema()

df = df.select(f.from_json(f.decode(df.value, 'utf-8'), schema=schema).
 ↪alias("input"))
df = df.select("input.*")
```

```
True
root
 |-- value: string (nullable = true)
```

[4]:

```python
query3 = df.writeStream.outputMode('update').format('console').start()
```

```
22/04/04 16:44:44 WARN org.apache.spark.sql.streaming.StreamingQueryManager:
Temporary checkpoint location created which is deleted normally when the query
```

```
didn't fail: /tmp/temporary-e61cb8ee-a68b-490f-b8b2-6e5c31d7beb1. If it's
required to delete it under any circumstances, please set
spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know
deleting temp checkpoint folder is best effort.
22/04/04 16:44:44 WARN org.apache.spark.sql.streaming.StreamingQueryManager:
spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and
will be disabled.
```

[5]:

```
model_path = 'gs://big-data-lab-madhurj/Ass7/Pipeline_Model'
model = PipelineModel.load(model_path)
print('Model Loaded....')

predictions = model.transform(df)

mapping = dict(zip([0.0,1.0,2.0],␣
 ↪['Iris-setosa','Iris-versicolor','Iris-virginica']))
mapping_expr = f.create_map([f.lit(x) for x in chain(*mapping.items())])
output_df = predictions.withColumn('prediction', mapping_expr[f.
 ↪col("prediction")])[['prediction','species']]

output_df = output_df.withColumn('correct', f.when((f.
 ↪col('prediction')=='Iris-setosa') & (f.col('species')=='Iris-setosa'),1).
 ↪when((f.col('prediction')=='Iris-versicolor') & (f.
 ↪col('species')=='Iris-versicolor'),1).when((f.
 ↪col('prediction')=='Iris-virginica') & (f.
 ↪col('species')=='Iris-virginica'),1).otherwise(0))

df_acc = output_df.select(f.format_number(f.avg('correct')*100,2).
 ↪alias('accuracy'))

output_df2 = output_df[['prediction','species','correct']]
output_df2.createOrReplaceTempView('output')
```

```
22/04/04 16:44:45 WARN org.apache.hadoop.util.concurrent.ExecutorHelper: Thread
(Thread[GetFileInfo #1,5,main]) interrupted:
java.lang.InterruptedException
        at
com.google.common.util.concurrent.AbstractFuture.get(AbstractFuture.java:510)
        at com.google.common.util.concurrent.FluentFuture$TrustedFuture.get(Flue
ntFuture.java:88)
        at org.apache.hadoop.util.concurrent.ExecutorHelper.logThrowableFromAfte
rExecute(ExecutorHelper.java:48)
        at org.apache.hadoop.util.concurrent.HadoopThreadPoolExecutor.afterExecu
te(HadoopThreadPoolExecutor.java:90)
        at
java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1157)
        at
```

```
java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:624)
        at java.lang.Thread.run(Thread.java:750)
```

```
-------------------------------------------
Batch: 0
-------------------------------------------
+------------+-----------+------------+-----------+-----------+
|sepal_length|sepal_width|petal_length|petal_width|    species|
+------------+-----------+------------+-----------+-----------+
|         4.9|        3.0|         1.4|        0.2|Iris-setosa|
|         4.7|        3.2|         1.3|        0.2|Iris-setosa|
|         4.6|        3.1|         1.5|        0.2|Iris-setosa|
|         5.0|        3.6|         1.4|        0.2|Iris-setosa|
|         5.4|        3.9|         1.7|        0.4|Iris-setosa|
|         4.6|        3.4|         1.4|        0.3|Iris-setosa|
|         5.0|        3.4|         1.5|        0.2|Iris-setosa|
|         4.4|        2.9|         1.4|        0.2|Iris-setosa|
|         4.9|        3.1|         1.5|        0.1|Iris-setosa|
|         5.4|        3.7|         1.5|        0.2|Iris-setosa|
|         4.8|        3.4|         1.6|        0.2|Iris-setosa|
|         4.8|        3.0|         1.4|        0.1|Iris-setosa|
|         4.3|        3.0|         1.1|        0.1|Iris-setosa|
|         5.8|        4.0|         1.2|        0.2|Iris-setosa|
|         5.7|        4.4|         1.5|        0.4|Iris-setosa|
|         5.4|        3.9|         1.3|        0.4|Iris-setosa|
|         5.1|        3.5|         1.4|        0.3|Iris-setosa|
|         5.7|        3.8|         1.7|        0.3|Iris-setosa|
|         5.1|        3.8|         1.5|        0.3|Iris-setosa|
|         5.4|        3.4|         1.7|        0.2|Iris-setosa|
+------------+-----------+------------+-----------+-----------+
only showing top 20 rows
```

```
Model Loaded…
```

```
22/04/04 16:44:59 WARN org.apache.spark.ml.feature.StringIndexerModel: Input
column class does not exist during transformation. Skip StringIndexerModel for
this column.
```

[ ]:

```
query1 = output_df2.writeStream.queryName("output").outputMode('update').
 ↪format('console').start()
query2 = df_acc.writeStream.outputMode('update').format('console').start()

query1.awaitTermination()
query2.awaitTermination()
```

```
22/04/04 16:44:59 WARN org.apache.spark.sql.streaming.StreamingQueryManager:
Temporary checkpoint location created which is deleted normally when the query
```

```
didn't fail: /tmp/temporary-8745b3c2-deb6-466c-8cb1-3e71326cd594. If it's
required to delete it under any circumstances, please set
spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know
deleting temp checkpoint folder is best effort.
22/04/04 16:44:59 WARN org.apache.spark.sql.streaming.StreamingQueryManager:
spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and
will be disabled.
22/04/04 16:44:59 WARN org.apache.spark.sql.streaming.StreamingQueryManager:
Temporary checkpoint location created which is deleted normally when the query
didn't fail: /tmp/temporary-10ef7f80-0603-4e4e-aa26-71dd181266d4. If it's
required to delete it under any circumstances, please set
spark.sql.streaming.forceDeleteTempCheckpointLocation to true. Important to know
deleting temp checkpoint folder is best effort.
22/04/04 16:44:59 WARN org.apache.spark.sql.streaming.StreamingQueryManager:
spark.sql.adaptive.enabled is not supported in streaming DataFrames/Datasets and
will be disabled.

-------------------------------------------
Batch: 0
-------------------------------------------
+--------+
|accuracy|
+--------+
|   83.89|
+--------+

-------------------------------------------
Batch: 0
-------------------------------------------
+-----------+-----------+-------+
| prediction|    species|correct|
+-----------+-----------+-------+
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
|Iris-setosa|Iris-setosa|      1|
+-----------+-----------+-------+
only showing top 20 rows
```