

Cloud Fog Computing and big data analysis

0516251 費蓋德 (Gueter Josmy Faure)

Homework 4: Spark MLlib & ML Pipelines

Part I: Spark MLlib Decision Tree - Regression version

Note: I was already very far in my assignment when I realize that my id should be in the command line, sorry I did not respect this requirement (but my name is on it).

Dataframe version:

```
Josmy@Josmy-VirtualBox: /usr/local/spark/bin
File Edit View Search Terminal Help
2018-11-15 23:08:36 INFO ContextHandler:781 - Started o.s.j.s.ServletContextHandler@19acbac{ /metrics/json,null,AVAILABLE,@Spark}
master=local[*]
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|season|mnth|hr|holiday|weekday|workingday|weathersit|temp|atemp|hum|windspeed|cnt|features|prediction|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|1|1|1|0|0|0|0|1|0.04|0.0758|0.57|0.1045|22|[1.0,1.0,0.0,0.0,...]|20.5|
|1|1|1|0|0|0|0|1|0.16|0.1364|0.47|0.3284|59|[1.0,1.0,0.0,0.0,...]|33.96666666666667|
|1|1|1|0|0|0|0|1|0.16|0.1818|0.8|0.1045|33|[1.0,1.0,0.0,0.0,...]|33.96666666666667|
|1|1|1|0|0|0|0|1|0.26|0.303|0.56|0.0|39|[1.0,1.0,0.0,0.0,...]|38.0|
|1|1|1|0|0|0|0|1|0.36|0.3788|0.66|0.0|48|[1.0,1.0,0.0,0.0,...]|57.76|
|1|1|1|0|0|1|1|1|0.12|0.1212|0.5|0.2836|5|[1.0,1.0,0.0,0.0,...]|9.826086956521738|
|1|1|1|0|0|2|1|1|0.14|0.1667|0.59|0.1045|12|[1.0,1.0,0.0,0.0,...]|9.826086956521738|
|1|1|1|0|0|2|1|1|0.16|0.1818|0.55|0.1045|5|[1.0,1.0,0.0,0.0,...]|9.826086956521738|
|1|1|1|0|0|2|1|1|0.22|0.2424|0.87|0.1045|14|[1.0,1.0,0.0,0.0,...]|9.826086956521738|
|1|1|1|0|0|3|1|1|0.26|0.303|0.93|0.0|31|[1.0,1.0,0.0,0.0,...]|15.35135135135135|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 10 rows
RMSE: 80.57078921695584
real 0m17.686s
user 0m23.203s
sys 0m1.462s
Josmy@Josmy-VirtualBox: /usr/local/spark/bin$
```

RDD version:

```
Josmy@Josmy-VirtualBox: /usr/local/spark/bin
File Edit View Search Terminal Help
2018-11-15 23:44:47 INFO ContextHandler:781 - Started o.s.j.s.ServletContextHandler@19acbac{ /metrics/json,null,AVAILABLE,@Spark}
2018-11-15 23:44:47 INFO SparkUI:54 - Bound SparkUI to 0.0.0.0, and SparkContext:54 - Added file file:/usr/local/spark/bin/spark-2.4.0-bin-h288.jar to the classpath of the driver
2018-11-15 23:44:47 INFO h timestamp 1542296687709
2018-11-15 23:44:47 INFO Utils:54 - Copying /usr/local/spark/bin/spark-2.4.0-bin-h288.jar to 0.0.0.0:4040
2018-11-15 23:44:47 INFO Executor:54 - Starting executor ID driver on 0.0.0.0:4040
2018-11-15 23:44:47 INFO Utils:54 - Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on 0.0.0.0:4040
2018-11-15 23:44:47 INFO NettyBlockTransferService:54 - Server created on 0.0.0.0:4040
2018-11-15 23:44:47 INFO BlockManager:54 - Using org.apache.spark.storage.BlockManagerMasterEndpoint for block management
2018-11-15 23:44:47 INFO BlockManagerMaster:54 - Registering BlockManagerMasterEndpoint on 0.0.0.0:4040
2018-11-15 23:44:47 INFO BlockManagerMasterEndpoint:54 - Registering BlockManagerMaster on 0.0.0.0:4040
2018-11-15 23:44:47 INFO BlockManagerMaster:54 - Registered BlockManagerMaster on 0.0.0.0:4040
2018-11-15 23:44:47 INFO BlockManager:54 - Initialized BlockManager
2018-11-15 23:44:48 INFO ContextHandler:781 - Started o.s.j.s.ServletContextHandler@19acbac{ /metrics/json,null,AVAILABLE,@Spark}
master=local[*]
org.apache.spark.api.java.JavaPairRDD@7661748f
Root Mean Squared Error = 0.010441133136692636
real 0m17.794s
user 0m20.998s
sys 0m2.045s
Josmy@Josmy-VirtualBox: /usr/local/spark/bin$
```

Dataset version(missing):

Failed to convert dataframe to dataset (not familiar with scala)

Report

Comparing execution time between RDD version code and DataFrame:

Dataframe time

RDD time

```
real    0m17.686s
user    0m23.203s
sys     0m1.462s

josny@josny-VirtualBox: /usr/local/spark/bin
josny@josny-VirtualBox: ~$
```

Slightly different execution time. For the real execution time, RDD is faster, for user it's still faster but not for system

Why do we use Decision Tree - the regression version in this homework, instead of using Decision Tree - the classification version?

In short, because we are dealing with continuous features.

Explanation:

Regression: the output variable takes continuous values.

Classification: the output variable takes class labels.

In our case, the output feature was “cnt” and in this column the values are numbers (continuous values) thus we use regression. If we had **categories instead of numbers** in ‘cnt’, using classification would be a better idea.

What is the main difference between DataSet and DataFrame?

A Dataset is a strongly typed collection of domain-specific objects that can be transformed in parallel using functional or relational operations. By this definition, we can deduce that at Dataframe is just an untyped view of a dataset or simply a Dataset of row.

What I have learned, what problems I encountered, and how the problems were resolved

I have encountered quite a few problems and consequently had to email the TA frequently.

Firstly, I had problems with spark-submit because I was not in the right directory (easy to solve).

While writing the Dataframe part of the homework, I was unaware about which one should be the output features. As a result, my RMSE was very large. I fixed it after knowing the right output.

Using pyspark, I had no idea how to open file in RDD format. I went on stackoverflow and found how to transform a dataframe to RDD and checked the official spark documentation to find how to perform decision tree on RDD.

Thank you !