Cloud Fog Computing and big data analysis

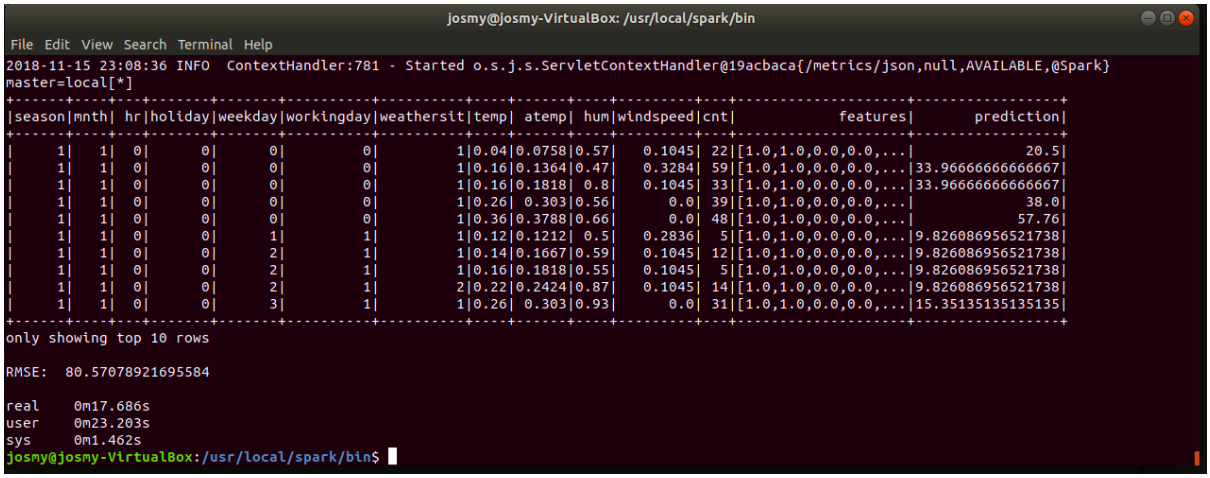
0516251 費蓋德 (Gueter Josmy Faure)

Homework 4: Spark MLlib & ML Pipelines

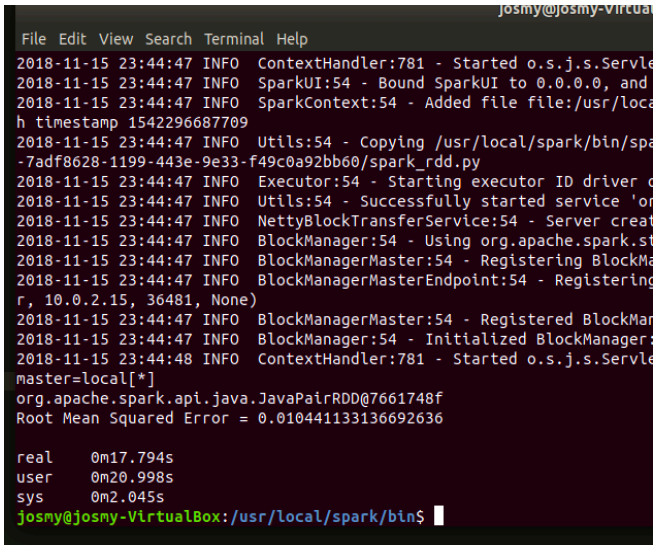
Part I: Spark MLlib Decision Tree - Regression version

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:



RDD version:



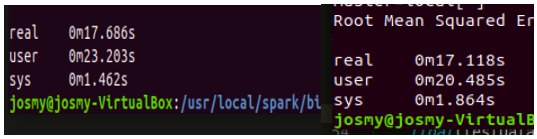
Dataset version(missing):

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

Report

Comparing execution time between RDD version code and DataFrame:

Dataframe time RDD time



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.

I 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 !