General Report

Baochen Hu
Ning Zhang

June 20, 2015

Abstract

This report aims to answer the questions in Assignment1.

1 Question 1

What performance metrics did you implement and use to evaluate Classification algorithms?

Please see

Assignment_Part1_Plain_Python.pdf
Assignment1_Part2&3_Spark.pdf.

2 Question 2

What performance metrics did you implement and use to evaluate Classification algorithms?

Answer:
For the Plain Python, we used Mean Absolute Error to evaluate most of the performances. For the Spark MLlib, we used Mean Squared Error to evaluate most of the performances.

3 Question 3

What performance metrics did you implement and use to evaluate Regression algorithms?

Answer:
For the Plain Python, we used Mean Absolute Error to evaluate most of the performances. For the Spark MLlib, we used Mean Squared Error to evaluate most of the performances.

4 Question 4

Compare and contrast using Just Scipy libraries vs using Apache Spark. How did the results vary? When would you use just Python libraries and when would you use Apache Spark?

Answer:
For the small datasets, Scipy libraries seems to be more efficient than spark. But we believe for large datasets, the spark will more efficient. Besides, for the SGD, large scale data set can only be applied to spark. Since our spark is running on single machine instead of cluster, which will surely influence the performance. With regard to accuracy, the two platform's algorithm can get very similar results.