Big Data Analytic Homework4 - Airline delay prediction

0556555 林哲宇

1. Program WorkFlow
   a. Set Spark Config (executor memory overhead , memory)
   b. Load in Data with csv format
   c. Put Carriers into Dictionary for categorical computing
   d. Union all years data
   e. Select the features, switch NA -> 0, using user defined function to switch UniqueCarrier value, then drop null rows
   f. Change the rdd format data into LabeledPoint form
   g. Split data into 70% training, 30% evaluation
   h. Training and evaluate the performance of the model

2. Execution Command
   spark-submit --packages com.databricks:spark-csv_2.10:1.5.0 --conf "spark.default.parallelism=100" hw4.py > output.txt

3. Answers
   Q1.Explain the predictive framework you designed.
   A1. RandomForest Regression with depth = 8 , 4 trees.
   The feature I selected are "Month","UniqueCarrier","Diverted","ArrDelay","DepDelay","Distance","NASDelay"
   I simply chose the weather-related and time-related feature to calculate the weather delay. And of course different carriers and if it is diverted should influence the data too.
   And of course “WeatherDelay" is the ground truth.

   Q2.Explain how method you use to validate your model when training.
   A2. I chose holdout to validate my model. With 70% training data and 30% testing data.

   Q3.Show the evaluation results of validation in training and prediction in testing by following those evaluation metric: average MAE and average RMSE.
   A3.

	Random Forest 4 trees, Validation	Random Forest 4 trees, Testing	Average
MSE	59.1098	73.0523	66.0811
RMSE	7.6883	8.5471	8.1177
MAE	1.0550	1.1659	1.1105

4. Discussion

   As you can see, the MSE and MAE is rather high for this model. Actually I've put several combination of features into training, but all of them didn't work well. By far this combination is relevantly good in some condition. I think the most possible reason is that these data couldn't represent the occurrence of weather. If we added the weather data of each Origin and Destination, it should be more accurate in theory.

   At first, I chose Origin and Dest as features, but in order to treat the columns as categorical index in random forest, which are 3400 indexes, the executor would fail calculating in these case. So I abandoned these two features at last.

   All in all, these data couldn't be good representation for weather delay. I think we'll need to combine the daily weather data to get better result.