Big Data Analytics Techniques and Applications
Homework IV

0556562 陳鴻君

1. I used regression decision tree as framework to build prediction model. To predict “WeatherDelay” is numeric, therefore regression analysis is used. About the features, I’ve already tried many sets, such as 'ArrTime', 'CRSArrTime', 'FlightNum', 'ActualElapsedTime', 'CRSElapsedTime', 'AirTime', 'ArrDelay', 'DepDelay', 'Distance', 'Cancelled', 'NASDelay', 'SecurityDelay', 'LateAircraftDelay', 'WeatherDelay'. But MSE is still around 60.
2. I used Holdout validation to validate training model. (Training set : Test set) = (7:3).
3. Framework:
    1. Load 2003 to 2008 data into dataframes.
    2. Union 2003 to 2007’s dataframes.
    3. Select columns and replace ‘NA’ with ‘0’.
    4. Translate datafrme into labeled point.
    5. Split 0.7 and 0.3 randomly.
    6. Put training set into Decision Tree Regressor training model with property impurity='variance', maxDepth=8, maxBins=256.
    7. Predict test data.
    8. Calculate MAE and RMSE.
    9. Print Regression Tree model, MAE and RMSE.

	Validation	Test 2008
MAE	0.983698446598	7.47653332258
RMSE	1.06372900555	8.25008737581

Comment:

```
spark-submit --packages com.databricks:spark-csv_2.10:1.5.0 --conf
"spark.default.parallelism=50" --conf
"spark.yarn.driver.memoryOverhead=400" --conf
“spark.yarn.executor.memoryOverhead=2048" dataProcessor.py
```