MSCI: 6110 Fall 2019 Big Data Management and Analytics Homework 4

Due 10/31/2019 6:00PM. Submit on ICON Dropbox

Total points: 100

Instructions:

1. Please submit a zip file with a .R script with your code and figure images generated for some of the questions. Mark questions and your explanations using R comments in your R script.

2. Your code should be able to run correctly in SparkR. Load all the libraries needed. No need to include the install.packages() functions.

Q1 (50 pts). Random Forest. There are several steps in this question:

(1) Load all the trips on Jan 1st (pickup time) between 10am and 11 am in the NYC_Taxi_Jan table you previously created. Save these trips into a Spark DataFrame. Filter the records and only keep the records with CSH or CRD payment types. Remove any record with missing values. You may either issue sql() commands or use spark DataFrame manipulation functions. Save this DataFrame as "day1_training".

(2) Train a random forest model using day1_training as the input dataset. The random forest should be trained to predict the payment type (CSH or CRD) using the following ~~9~~ columns: ~~pickup_datetime, dropoff_datetime~~, passenger_count, trip_distance, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, fare_amount. Use 20 trees with maximum tree depth = 5. Save the model as a variable.

(3) Load the trips on Jan 2nd (pickup time) between 10am and 11 am from the NYC_taxi_Jan table and **repeat the steps in (1)** to generate a new DataFrame called "day2_testing". Do not train a new model. Apply the model learned in (2) on this testing dataset to do prediction. Print the top 20 rows of your prediction output.

(4) Calculate the accuracy of your model on the testing set. The accuracy is calculated as: number of rows with correct prediction/number of rows in the testing set. If the "prediction" column has the same value as the "payment_type" column, then this row is predicted correctly. **Print this number and also report this number using an R comment.**

(5) Treat CRD as the positive class ("1") and CSH as the negative class ("0"). Calculate the precision and recall of the prediction result in (4). **Print these two numbers and also report them using R comment lines in your code.**

Q2 (35 pts). Linear Regression.

(1) (20 pts) Use the "day1_training" DataFrame obtained in Q1(1) to train a linear regression model. Train the model to predict the "fare_amount" column using trip_distance, pickup_longitude, pickup_latitude, dropoff_latitude, and dropoff_longitude.

(2) (15 pts) Apply the model learned in Q2(1) on the "day2_testing" dataframe to do prediction and calculate the mean squared error (MSE) using the output. Print the first 20 rows of prediction result and the MSE value. **Report the MSE in your code using an R comment.**

Q3. (15 pts) Clustering.

(1) Use the kmeans function in SparkR to cluster the trips in the day1_training dataframe into 5 clusters. You should only use the following attributes for clustering: pickup_latitude, pickup_longitude, dropoff_latitude, dropoff_longitude. Fit the learned model on the day1_training dataframe (same dataset for training and testing). Show the first 20 rows of the clustering results.