CSE 351 - Introduction to Data Science (Summer 2020)
Prof. Praveen Tripathi
Homework 2: Modeling

This homework will investigate model building in IPython. It revolves around predicting the fare of a taxi ride given information including a pickup and a drop off location. Our goal is to explore the data and uncover interesting observations about the New York Taxi operations.

Data downloading

Please find the dataset "taxi_fare.csv" attached on Blackboard under "Assignment 2". Below is a description of the data:

- key - Unique `string` identifying each row in both the training and test sets. Use it as a unique ID field.
- pickup_datetime - `timestamp` value indicating when the taxi ride started.
- pickup_longitude - `float` for longitude coordinate of where the taxi ride started.
- pickup_latitude - `float` for latitude coordinate of where the taxi ride started.
- dropoff_longitude - `float` for longitude coordinate of where the taxi ride ended.
- dropoff_latitude - `float` for latitude coordinate of where the taxi ride ended.
- passenger_count - `integer` indicating the number of passengers in the taxi ride.
- fare_amount - `float` dollar amount of the cost of the taxi ride. You are predicting this value for the test set.

Python in Google Colab

Please use the notebook template "cse351_hw2_template.ipynb" attached on Blackboard to complete this assignment. The template provides a layout of the required tasks. Open the notebook in Google Colab, insert code cells to run code and text cells to write explanations and answers. Feel free to use any libraries in Python to complete the tasks.

The scikit-learn library is useful for this assignment. Please find its documentation at the following link: https://scikit-learn.org/stable/

Tasks(50 points)

1. Take a look at the data. There may be anomalies in the data that you may need to factor in before you start on the other tasks. Clean the data first to handle these issues. Explain what you did to clean the data. **(5 points)**

2. Split the data into training set and test set. Save them as "taxi_fare_train.csv" and "taxi_fare_test.csv". Explain how you divided the dataset. **(5 points)**

3. For the training set, compute the Pearson correlation between the following: **(6 points)**
    - Euclidean distance of the ride and the taxi fare
    - time of day and distance traveled
    - time of day and the taxi fare

    Which has the highest correlation?

4. For each subtask of (3), create a plot visualizing the relation between the variables. Comment on whether you see non-linear or any other interesting relations. **(9 points)**

5. Set up a simple linear regression model to train on “taxi_fare_train.csv”, and then predict taxi fares for instances in file “taxi_fare_test.csv” **(Note that you need to drop the “fare_amount” column in the test set first)**. Use your generated features from the previous task if applicable. How well/badly does the model work? **(Evaluate the correctness of your predictions based on the original “fare_amount” column)** What are the coefficients for your features? Which variable(s) are the most important one? **(10 points)**

6. Now, try to build a better prediction model that works harder to solve the task. Perhaps it will still use linear regression but with new features. Perhaps it will preprocess features better (e.g. normalize or scale the input vector, convert non-numerical value into float, or do a special treatment of missing values). Perhaps it will use a different machine learning approach (e.g. nearest neighbors, random forests, etc). Explain what you did differently here versus the simple model. How does the model perform? **(10 points)**

7. Visual Appeal and Layout - For all the tasks above, please include an explanation wherever asked and make sure that your procedure is documented (suitable comments) as well as you can. Don’t forget to **label** all plots and include legends wherever necessary as this is key to making good visualizations! Ensure that the plots are visible enough by playing with size parameters. Be sure to use appropriate color schemes wherever possible to maximize the ease of understandability. All output must be laid out and clearly shown in a python notebook(.ipynb). **(5 points)**

Submission

1. This assignment must be done **individually** by every student. Your code will be checked thoroughly to detect copying/plagiarism. Do your own work!
2. Please use Piazza to ask any questions.
3. Submit everything through Blackboard. You will need to upload:
    1. The Jupyter notebook all your work is in (.ipynb file)
    2. Python file (export the notebook as .py)
    3. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse351_hw2_*lastname_firstname_sbuid*.ipynb
2. cse351_hw2_*lastname_firstname_sbuid*.py
3. cse351_hw2_*lastname_firstname_sbuid*.pdf

References

1. Part of the instructions are courtesy of Professor Steven Skiena (CSE 519)