CS5785 Homework 1

Due date: Sept. 29th, 2014

The homework is generally split into programming exercises and written exercises. Due dates will be given for each homework. You should turn in an electronic copy of your solutions to the homework. Please submit your homework to CMS. You are responsible for submitting clear, organized answers to the questions. Please include all relevant information for a question, including text response, equations, figures, graphs, etc. Please pay attention to the discussion board for relevant information regarding updates, tips, and policy changes. You are required to work in groups of 2 (unless given an exemption from the course staff). Version 3.

PROGRAMMING EXERCISES

1. **Crazy Taxi.** The main goal of this problem will be to create a model which can predict travel time given information such as pick up location, drop off location, pick up time, and trip distance.

   (a) Download the zip file provided with HW1 assignment. Inside should be a file called example_data.csv with 9999 examples. Each data point contains information from a single taxi cab trip.

   (b) Use the script provided in the zip file to calculate distances between pickup and dropoff locations. Calculate the minimum and maximum distance and provide a method to eliminate outliers. Please provide an explanation of this method in the writeup. Use this outlier filtering method to clean any data used in further subproblems.

   (c) The main variables we are interested in are trip distance, pick up time, and distance between pickup and dropoff. Please create three scatterplots which compare these variables against trip time in secs.

   (d) Separate the example data into a training and test set by assigning every fourth point to the test set. Using least squares fitting on the training data, create a linear model which predicts trip time given trip distance. Calculate both the ordinary least squares accuracy (OLS) and total least squares accuracy (TLS) on the test set.

   (e) Make a significant change your outlier filtering method as described in part (b) and try part (d) again; are your results better or worse? Be sure to describe your new filtering method.

   (f) Go the website (`http://www.andresmh.com/nyctaxitrips/`) and download the file named `trip_data_2.csv.zip`. Remember to use your outlier filter on this data set. Use all of the example data to train a new linear model and evaluate this model on trip_data_2. Report the OLS, TLS, and the correlation coefficient.

   (g) Consider the parameter pick up time. Currently, this value is cyclical and has a highly nonlinear relationship with trip time. Think of a way we can simplify this variable so that it will have a linear relationship and write down one such technique. Hint: consider transforming it into a binary variable.

(h) Create a feature vector which includes trip distance, distance between pick up location and drop off location, pick up latitude, pick up longitude, drop off latitude, drop off longitude, and pick up time. Normalize each element of your feature vector. Write the formula used for the normalization of trip distance.

(i) Create a linear model which predicts trip time given the feature vector described above. Train on trip_data_1.csv file on the website. Using trip_data_2.csv, calculate TLS, OLS, and correlation coefficient. Remember to use your outlier filter on these data sets.

(j) Is this model more accurate than the simple linear model which compared trip time to trip distance? Explain why or why not.

2. **The Titanic Disaster**

(a) Join the Titanic: Machine Learning From Disaster competition on Kaggle. Kaggle is a platform for data prediction competitions and contains datasets and benchmark models that will may use throughout the course. Download the training and test data.

(b) Using logistic regression, try to predict whether a passenger survived the disaster. You can choose the features (or combinations of features) you would like to use or ignore, provided you justify your reasoning.

(c) Train your classifier using all of the training data, and test it using the testing data. Submit your results to Kaggle and include them in the writeup.

WRITTEN EXERCISES

1. In this problem, we will explore certain values of Pearson's correlation coefficient. For each problem, label the axes with logical dimensions. Briefly explain why the dimensions you use make sense. Please do not copy your examples from the lecture notes or any other source, and do not reuse dimensions.

(a) Draw a graph with a correlation coefficient value of close to 0.

(b) Draw a graph with a correlation coefficient value of close to 1.

(c) Draw a graph with a correlation coefficient value of close to -1.

2. In this problem, we will explore the difference between using the L1 and L2 norms to calculate error rate. Note that least squares is the L2 norm minimization problem and least absolute deviations (aka least absolute errors) is the L1 norm minimization problem.

(a) Consider a graph with most data lying along a line but with one large outlier. Briefly describe what the regression line will look like under the two different norms.

(b) Draw an example where linear regression with the L2 norm leads to a single optimal solution but L1 norm leads has multiple optimal solutions with the same error rate.

(c) In order for L1 to be as useful as L2, what properties would it need to have?