CS 5644: Assignment 3

1. (50 points) **Regression;** Consider the Bike Share dataset from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset). The dataset contains three files, viz. *day.csv, hour.csv,* and *Readme.txt.* Both the datasets (data.csv and hour.csv) contain a combination of integer-valued (e.g., season, weekday or not) and real-valued features (e.g., temperature, windspeed). Details of the dataset are described in the README (*Readme.txt*) in the *Data Characteristics* section. Just like the previous assignment, spend some time understanding the structure of the dataset, how the instances are organized, how the features are organized, what the various features mean (info in README), what features are useful for the task at hand, and so on. Do not attempt to run any machine learning algorithm before understanding the structure of the dataset.

   Note, in particular, the last three fields in the data, viz. ***casual*** *(denotes casual riders),* ***registered*** *(registered riders), and* ***cnt*** *(total ridership count).*

   a. Using only hour.csv, implement regression algorithms (both linear and k-nearest neighbors) to predict the hourly values for:
      i. the number of casual riders
      ii. the number of registered riders
      iii. total ridership count □

   b. Using only *day.csv*, implement regression algorithms (both linear and k-nearest neighbors) to predict the daily values for:
      i. the number of casual riders
      ii. the number of registered riders
      iii. total ridership count

   Therefore, for each dataset, you are reporting 3 models for linear regression and 3 models for KNN regression.

   Note: Remember that using one of the target values (as a feature) in predicting the outcome of any of the counts – casual, registered, or total would defeat the purpose of the learning algorithm. It will make the problem too easy. That is, the number of ***casual*** riders cannot be used as a feature to predict the number of **registered** riders or **total** ridership. Similarly, the number of **registered** riders cannot be used to predict the other two values, and so on. Only features like season, temperature, real-feel, etc. have to be used by the learning algorithm.

   For instance, suppose you are using *hourly.csv* dataset and you want to predict the

number of **casual** riders. You have to remove the columns related to the number of **registered** riders and **total** ridership first and then start training/testing your model. Similarly, when you are making the prediction model for **total** ridership, you have to remove the columns related to **casual** rides and **registered** riders first and then start training/testing your model.

As before, you will need to separate the data into training set and test set (decide on the proportion of splits yourself). Evaluate the performance of your regression using suitable measures. Report on the performance results and which model(s) worked best (and why in your opinion).

2. (50 points) **Clustering;** Consider the Seeds data set from the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/seeds). The dataset comprises of features from three different types of wheat kernels. There are seven features (area, perimeter, compactness, length, width, asymmetry coefficient, and length of kernel groove) that describe each data point. (Note that the dataset has an eighth column (*class information with labels 1, 2, and 3*), which we will use as ground truth to verify our clustering results.)

   Using the ***k-means*** algorithm cluster this dataset into three clusters based on the seven features at your disposal. Demonstrate the effectiveness of your implementation by comparing the results against the ground truth. Follow the steps in the k-means demo video from the lectures.

   Also note that the default label values in scikit learn start from ***0***, whereas the dataset here starts labels with ***1***. While evaluating your implementation's effectiveness, ensure to account for this discrepancy.

   As a performance measure, compare the clusters identified by k-means w.r.t. the ground truth data and make observations.

What to submit:
A zipped file containing:

1. a PDF document summarizing answers to questions 1 and 2. Instead distill your lessons and experiences succinctly.
2. Either hyperlinks to or actual attachments of your data files and your iPython notebook(s).