**Intelligent Data Analysis (CS5152/6052)**

**Fall 2018**

**Homework #4**

**Due Date: Nov. 29th, 2018, 9PM.**

1. Consider the wine datasets given at this [LINK.](https://archive.ics.uci.edu/ml/datasets/wine+quality) One dataset at this site is for red wines and the other is for white wines. Data has eleven features and the twelfth column is the target attribute: “wine-quality”. The values of wine-quality range between 0 and 10. The goal of this task is to learn a regression model that predicts “wine-quality” for a given set of features. Consider the white wine dataset and perform the following tasks with this dataset. Use Matlab or Python SKLearn libraries to perform these tasks.
   1. (8) Do linear regression to learn the single-feature regression models, one model for each of the 11 features. Find the R2 and AIC values for each of these models. Report these values for the models.
   2. (8) Select the model with the highest R2 value, combine with its feature other features, one at a time, and thus generate all bivariate regression models (models containing two features). One of these two features is from the selected single-feature model and the other is from one of the remaining 10 features. Report the R2 and AIC values for all the bivariate regression models.
   3. (8) Select the bivariate model with the highest R2 value as the Best model at this stage. Combine a third feature from the remaining nine features with this selected bivariate model to build (and then select the best) 3-feature regression models. Report the R2 and AIC values of all these models.
   4. (14)Repeat the steps above to generate (k+1)-feature models from the k-feature models until the following situation arises: all the (k+1)-feature models have an AIC value higher than the AIC value of the k-feature model from which they are being generated. Stop the process and report the k-feature model found as being the best regression model for this data. Report the features included, their coefficients, and p-values for the coefficients. Comment on the magnitudes of the p-values.
   5. (7) Find the five wines that have the largest magnitudes of difference between the predicted and the actual wine-quality values. Look at the regression model, the rest of the data, and comment on why you think these wines are outliers.
2. Consider the data-file attached with this homework containing 6600 data points on a 2-D plane. You will need to use the BIC metric to determine the quality of a clustering. This is computed here as: BIC = n\*log(SSE/n) + log(n)\*c\*(d+1) where n is the number of data points, c is the number of clusters, and d is the number of features (dimensionality of the data). Remember to use the sum of SSEs for all the clusters in any clustering.
   1. (3) Plot the data on a 2-D scatter plot and mark by hand the boundaries of the ideal clusters that you would like discovered in this dataset.
   2. (12) Run the k-means algorithm for k = 3, 5, 7, 9, 11, 13, 15, 17 and 19. Plot the total SSE and BIC values for the above values of k. What is the best number of clusters for this dataset? How did you find the best number of clusters, briefly explain.
   3. (8) For the best number of clusters selected above, plot the scatter plot of the data showing the points of each cluster with a different color/symbol. Mark the points on the scatter plot that belong to clusters other than what your intuition says. Why did k-means algorithm place them in these different clusters – explain very briefly.
   4. (5) Plot the silhouette diagram for the best clustering you have selected. Comment on the characteristics of the silhouette diagram that you think are informative about this clustering. Comment using the cluster numbers and their plots on the silhouette diagram.
   5. (12) Perform single-linkage hierarchical clustering for this data and cut the dendrogram to obtain 11 clusters. There are options/parameters in most toolboxes to generate a given number of clusters. Plot the 2-D scatter plot of the dataset showing data points of each of the 11 clusters with different color/symbol.
   6. (5) Mark any data points on this scatter plot that are clustered differently from your intuitive view of the correct clusters. Explain why Single-linkage clustering may have placed them in counter-intuitive clusters.