**Data Science HW2 (Due on 10/16 24:00)**

\*\*\*\*\*Only factor data can be used in association rule mining// only numeric data can be used in clustering

1. Based on the winequality-red dataset, remove the last column (quality) and treat the remaining columns as input features. Then, respectively perform *K* means and *K* medoids to conduct clustering. Use NbClust package to help you decide the optimal number of clusters (from 2 clusters to 9 clusters). Finally, indicate the centroids and calculate the average quality within each group.
2. Redo the regression problem using the winequality-white dataset (response is quality). Construct multiple linear regression (MLR) to identify the significant predictors. Based on these significant predictors, conduct K nearest neighbors (KNN) to conduct forecasting for quality. Please compare their performances (MLR and KNN) in RMSE, MAE, and MAPE.
3. Redo the classification problem using the winequality-white dataset (response is rating with only two levels: poor (quality=3, 4 ,5) or good (quality=6, 7 ,8)). Construct KNN to conduct forecasting for rating. Demonstrate the performances in overall accuracy and respective accuracies for two levels of rating (poor or good).
4. Based on pima dataset, remove zeros in the following columns, Glucose, BloodPressure, SkinThickness, Insulin, and BMI. Then, remove the last column (Outcome) and conduct fuzzy *C* means clustering. Show the best-fitted number of clusters (based on XB, FS indices), associated centroids, and a corresponding Table between Outcome labels (Yes/No means with or without diabetes) and grouping labels. Calculate the Euclidean distances between various pairs of centroids.
5. Based on pima dataset, remove zeros in the following columns, Glucose, BloodPressure, SkinThickness, Insulin, and BMI. Then, remove the last column (Outcome) and conduct Gaussian mixture clustering (GMC). Show the best-fitted number of clusters, associated centroids, covariances, and a corresponding Table between Outcome labels (Yes/No means with or without diabetes) and grouping labels. Calculate the Mahalanbis distances between various pairs of centroids.
6. Based on gender\_outlier dataset, apply hierarchical clustering (single link, complete link, group average, Ward’s method) to perform clustering. Use DB index to choose the best number of clusters (from 2 clusters until 9 clusters) and specify the majority class (male or female) in each group.