Homework Assignment #9
KNN
DATS 6101 - Spring 2020

Pima Dataset

This exercise uses the Pima.te and Pima.tr dataset from the MASS package, which includes measurement on a population of women of Pima Indian heritage living near Phoenix, Arizona. The women were tested for diabetes according to World Health Organization criteria.

The variables in the dataset are:

- npreg : number of pregnancies.
- glu : plasma glucose concentration
- bp : diastolic blood pressure (mm Hg) skin triceps skinfold thickness (mm)
- bmi : body mass index
- ped : diabetes pedigree function
- age : age in years
- type : Yes or No: diabetic by WHO criteria

Question 1
In the MASS library, combine the two datasets Pima.te and Pima.tr back into one complete dataset, call it pima. (Try function rbind().) How many observations are there?

Question 2
Obtain some basic summary data for pima. (You can try the xkablesummary() function introduced in the previous HW7 solution.)

Question 3
Another quick EDA to perform, you can plot the pairs(). The plot function can handle both numerical and categorical variable type. See also other suggestions in the sample RMD.

Question 4
In order to perform KNN analysis, we need to separate the X-variables and the y-labels. (Which should be our y-variable?) Before we separate them out, create a vector/array of 1 and 2 to create a train-test split in the ratio of 3:1. (Set a constant seed value so that we can duplicate the results.) So eventually, you will get the training Xs as a dataframe, training y-label (a vector), as well as the test sets together in four groups. Make sure the train-X and train-y are not mixed up in the ordering during the process. Same for test-X and test-y.

Question 5
Perform the KNN analysis, with different k values. You do not need to show all the results from different k, but please include the one with the best (total) accuracy in your submission. How does the accuracy compared to the percentages of being T/F in the dataset?

Question 6
Compare to the best logistic regression you can get. (Use the full model with all variables, since that is what we have for KNN.) How is the accuracy (assumes the standard cutoff of 0.5) compared to KNN?

Question 7
What is the score for the logit model using ROC-AUC? We should be able to compute the ROC-AUC value for the KNN model the same way. Can you compare them?