2018 Spring, INE5008 Data Mining, by Kichun Lee

HW 1

Due 1 pm, April 20, 2017

**PART #1**

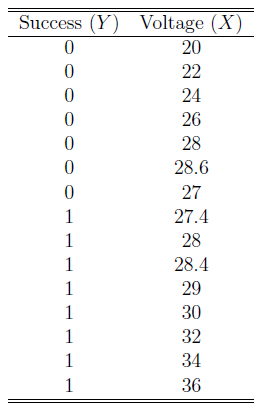
Refer to the attached *HWtrain.data.txt* and *HWtest.data.txt* in a comma-separated form. The file *HWtrain.data.txt* contains 173 observations (rows) of wine quality. The first value in a row represents the wine quality of the row in terms of 1, 2, and 3. The values from the 2nd column and 14th column are some descriptive values for the wine, as discussed in class. The file *HWtest.data.txt* contains five future-observations that we would like to estimate the wine qualities (marked by X).

We would like to apply principle component analysis after normalizing the descriptive values in *HWtrain.data.txt*

1. When we use the first two principle components, how much information of the total do we use? 55.3%
2. Find the top-three original variables among the 13 that contributed to the construction of the first principle component.
3. V2: 0.4227611
4. V12: 0.3934570
5. V7 : 0.3793968
6. Draw a plot of the 1st scores (on the X axis) and the 2nd scores (on the Y axis) of the descriptive values in *HWtrain.data.txt* together with the wine qualities and similarly transformed score values of the five future observations in *HWtest.data.txt.*  Then try to predict the wine quality of each of the five future observations.

**PART #2**

Refer to the following data (HW2\_data\_2.xlsx):



Normalize , where is the minimum and is the maximum of all .

Looking at the trend, the following predictor is suggested:

Construct the confusion matrix and compute the accuracy, sensitivity, specificity, and precision for each cut-off value 0.4, 0.5, and 0.6.

**PART #3**

Refer to the attached *Sham.data* and *PAB.data*. The file, *Sham.data,* contains 25 observations of blood pressure at a certain position of *normal* rats. Similarly, the file, *PAB.data,* contains 15 observations of blood pressure at the same position of *irregular* rats. Irregular rats are shot by a certain medicine and under the influence of it. For your information, the sample average of blood pressure values for the *Sham.data* data is 0.380, and that for the *PAB.data* data is 1.153. We will consider the null (the status quo) to be the normal-rat case (*Sham.data*) and the alternative to be the irregular-rat case (*PAB.data*).

We devised a simple classification rule as follows:

“If the blood pressure value is less than a threshold value, *t*, then we predict the case to be of the normal-rat case (*Sham.data*).

Otherwise, it is of the irregular-rat case (*PAB.data*).”

Draw a ROC curve by varying the threshold value, *t*, and find the best threshold value which generates the closest point to the best separation case (0,1) in the ROC curve.

Note: there are many possible ways and implementations to accomplish this; however, a partial implementation is in *HW2\_Prob\_3\_Rscript\_part.txt* for your reference.