STA 6923 Homework 5, Fall 2021

Due date: Tuesday, November 2, 11:59 pm

Instruction: Answer your questions item by item. Write your conclusions clearly and concisely. You need to provide two files as follows.

- Your .Rmd code file, with comments on your code
- Your full homework report, written by R markdown pdf (or html) file. Before each output, you should keep your code line as well by using "Echo = TRUE", or you can define it in the beginning.

---

*From text book (ISL)** (Each problem worths 10 points)

Chapter 4, page 168:

- Conceptual problems:
  - 1, 3, 4, 6, 8
- Applied problem:

– 10 (In part (i), be concise; only describe and provide the output of your best prediction.)

Not from the text book

Problem Breast cancer is the most common malignancy among women, accounting for nearly 1 in 3 cancers diagnosed among women in the United States, and it is the second leading cause of cancer death among women. Breast Cancer occurs as a results of abnormal growth of cells in the breast tissue, commonly referred to as a Tumor. A tumor does not mean cancer - tumors can be benign (not cancerous), pre-malignant (pre-cancerous), or malignant (cancerous). Tests such as MRI, mammogram, ultrasound and biopsy are commonly used to diagnose breast cancer performed.

Data sources **Breast Cancer Wisconsin Data Sets** were created by Dr. William H. Wolberg, physician at the University Of Wisconsin Hospital at Madison, Wisconsin,USA. To create the dataset Dr. Wolberg used fluid samples, taken from patients with solid breast masses and an easy-to-use graphical computer program called Xcyt, which is capable of perform the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, than it calculates the mean value, extreme value and standard error of each feature for the image, returning a 30 real-valuated vector.

You can get the data from UCI repository webpage. The data set can be found by clicking Data Folder in the Breast Cancer Wisconsin (Diagnostic) Data Set page. On the other hand, you may just use the following code to extract the data (note that the first column is the class with two categories: “B” for benign and “M” for malignant).

```
library(RCurl)
fileURL = "https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data"
BC_data = read.csv(fileURL, header = FALSE, sep = ",")
names(BC_data) <- c('id_number', 'diagnosis', 'radius_mean',
         'texture_mean', 'perimeter_mean', 'area_mean',
         'smoothness_mean', 'compactness_mean',
         'concavity_mean','concave_points_mean',
         'symmetry_mean', 'fractal_dimension_mean',
         'radius_se', 'texture_se', 'perimeter_se',
         'area_se', 'smoothness_se', 'compactness_se',
         'concavity_se', 'concave_points_se',
         'symmetry_se', 'fractal_dimension_se',
         'radius_worst', 'texture_worst',
         'perimeter_worst', 'area_worst',
         'smoothness_worst', 'compactness_worst',
         'concavity_worst', 'concave_points_worst',
         'symmetry_worst', 'fractal_dimension_worst')

BC_data$id_number <- NULL
```

We then remove all the "_worst" variables, i.e., you will only have 21 variables in your data frame. We want to use this data to do some classification analyses.

1. Do certain exploratory analysis first.

(a) Check the scatterplots as well as correlations between the predictors. It would be reasonable only to compare those predictors with their own metrics, i.e., means, standard errors, and worst cases.

(b) Check whether there are strong multicollinearity effects among the predictors

2. Split data by using the following commands

```
library(caret)
set.seed(12)
tr.ind = createDataPartition(BC_data$diagnosis, p = 0.7, list = F)
BC.tr = BC_data[tr.ind,]
BC.te = BC_data[-tr.ind,]
```

Do you still see multicollinearity problems for the training set?

3. Run a logistic regression model for the data by using the **diagnosis** column on all the predictors. Report confusion matrices for the test data as well as training data predictions. Display ROC curve for the test data prediction, along with reporting the AUC.

4. Redo part **3.** for LDA and QDA, as well as Naive Bayes methods, respectively.

5. Do part **3.** by using the KNN classification method for k=1, and k=7.

6. Since KNN uses Euclidean distances to calculate the *distance*, different scales in predictors may affect each other. Scale all the predictors for the original data, i.e., combined with training and test data. Then split the date by “tr.ind”. Redo part **5.** for this scaled data.

7. Comments all the classification results you’ve done above.