ISyE 7406: Data Mining & Statistical Learning
HW#3

Classification. In this problem, you are asked to write a report to summarize your analysis of the popular "Auto MPG" data set in the literature. Much research has been done to analyze this data set, and here the objective of our analysis is to predict whether a given car gets high or low gas mileage based 7 car attributes such as cylinders, displacement, horsepower, weight, acceleration, model year and origin.

(a) The "Auto MPG" data set is available at UCI Machine Learning (ML) Repository:

`https://archive.ics.uci.edu/ml/datasets/Auto+MPG`

Download the data file "auto-mpg.data" from UCI ML Repository or from Canvas, and use Excel or Notepad to see the data (this is a .txt file).

There are 398 rows (i.e., 398 different kinds of cars), and 9 columns (the car attributes and name). Before we do any analysis, we need to clean the raw data. In particular, some values are missing for this dataset. Many statistical methods have been proposed to deal with missing values, and please conduct literature research by yourself. For the purpose of simplicity in this homework, here we adopt a simple though inefficient method to remove those rows with missing values. Also we remove the last column of car names, which is text/string and may cause trouble in our numerical analysis. These two deletions lead to a new cleaned data set of 392 observations and 8 columns. To save your time, you can also directly download the cleaned data from the file "Auto.csv" from Canvas.

(b) Create a binary variable, `mpg01`, that contains a 1 if `mpg` contains a value above its median, and a 0 if `mpg` contains a value below its median. This binary variable will be the response variable in this homework. Note that you need to first compute the median value of the `mpg` variable in the data set.

(c) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

(d) Split the data into a training set and a test set. Any reasonable splitting is acceptable, as long as you clearly explain how you split and why you think it is reasonable. For your convenience, you can either randomly split, or save every fifth (or tenth) observations as testing data.

(e) For the purpose of this homework, perform the following classification methods on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (c). What is the test error of the model obtained?

(1) *LDA* (2) *QDA* (3) *Naive Bayes* (4) *Logistic Regression*

(5) *KNN* with **several** values of K. Use only the variables that seemed most associated with `mpg01` in (c). Which value of K seems to perform the best on this data set?

(6) (Optional) **PCA-KNN**. The Principal Component Analysis (PCA) or other dimension reduction methods can easily be combined with other data mining methods. Recall that the essence of the PC-reduction is to replace the p-dimensional explanatory variable $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$, for $i = 1, \ldots, n$, with a new p-dimensional explanatory variable $\mathbf{u}_i = (u_{i1}, \ldots, u_{ip})$, where $\mathbf{u}_i = \mathbf{A}_{p\times p}\mathbf{x}_i$. Then we can apply standard data mining methods such as **KNN** to the first $r(\leq p)$ entries of the $\mathbf{u}_i$'s, $(u_{i1}, \ldots, u_{ir})$, to predict Y_i's. Find the testing errors when the **KNN** with different values of K (neighbors) is applied to the PCA-dimension-reduced data for different $r = p-1, p-2, \ldots, 1$.

(7) (Optional) *Any other classification methods* you want to propose or use.

Write a report to summarize your findings, e.g., what is the best method and how to use your results in the context of guiding to manufacture or buy high gas mileage cars. The report should include (i) **Introduction,** (ii) **Exploratory (or preliminary) Data Analysis**, (iii) **Methods**, (iv) **Results** and (v) **Findings.** Please attach your computing code for R, Python, or other statistical software (without, or with limited, output) in the appendix of your report, and please do not just dump the computer output in the body of the report. It is important to summarize and interpret your computer output results.

Remarks: (a) From now on, we might no longer ask you to conduct cross-validation (CV) explicitly as in the previous HWs, but you might need to do it on your own if you want your results more convincing.

(b) In this HW, you are asked to use those explanatory variables that are seemed most associated with `mpg01`. This approach often occurs in manufacturing or biomedical applications where one wants to combine certain domain knowledge to improve the performance. Note that this might not be applicable in other applications of machine learning such as deep neural networks or random forest, where it is okay to use all explanatory variables.

(c) Below some sample R code if you wnat (please feel free to use python or matlab or other software):

```
## Suppose that you save the datafile in the local folder of your computer, say, ‘‘C:/Temp":
Auto1 <- read.table(file = "C:/Temp/Auto.csv", sep = ",", header=T);


#### (b)
# Note you may find it helpful to use the {\tt data.frame()} function to create
# a single data set containing both {\tt mpg01} and the other {\tt Auto} variables.

mpg01 = I(Auto1$mpg >= median(Auto1$mpg))
Auto  = data.frame(mpg01, Auto1[,-1]);  ## replace column "mpg" by "mpg01".

### END####
```