Applied Data Mining: Homework #7

Due on Fill-in this please

Instructor: Hasan Kurban

Student Name

November 27, 2017

Problem 1

In this problem, you are asked to use SVM to predict whether a given car gets high or low gas mileage based on the Auto data set. The data set can be obtained as follows:

```
> library(ISLR)
> View(Auto)
```

1.1 Create a binary variable that takes on a 1 for cars with gas mileage above the median, and a 0 for cars with gas mileage below the median. Add this variable to the data as a new variable and name it as "mpglevel" (mpglevel is the response variable for questions 1.2 and 1.3).

R Code

Listing 1: Sample R Script With Highlighting

```
%% You provide code here %%
```

1.2 Fit a linear support vector classifier to the data with various values of cost (cost = c(0.01, 0.1, 1, 5, 10, 100)), in order to predict whether a car gets high or low gas mileage. Report the cross-validation errors associated with different values of this parameter. Comment on your results, i.e., what is the cost value for the model that has the lowest cross-validation error?

R Code

Listing 2: Sample R Script With Highlighting

```
%% You provide code here %%
```

Cross-validation Errors and Discussion of the Results

Answer here ...

1.3 Now repeat (1.2), this time using SVMs with radial and polynomial basis kernels, with different values of gamma (c(0.01, 0.1, 1, 5, 10, 100)) and degree (c(2, 3, 4)) and cost (c(0.1, 1, 5, 10)). Use the cost and degree parameters values for polynomial kernels. The cost and gamma parameters values are given for radial basis kernels. Comment on your results, i.e., what are the parameters values (cost, degree, gamma) for the model that has the lowest cross-validation error?

R Code

Listing 3: Sample R Script With Highlighting

```
%% You provide code here %%
```

Discussion of Results

Answer here ...

Problem 2

Load the Caravan data set as follows and answer the questions below.

```
> library(ISLR)
> View(Caravan)
```

2.1 Create a training set consisting of the first 1,000 observations, and a test set consisting of the remaining observations. The class variable is "Purchase" whose values are "No" and "Yes". Transform "No" to "0" "Yes" to 1. Place the R code below.

R Code

Listing 4: Sample R Script With Highlighting

```
%% You provide code here %%
```

2.2 Fit a boosting model to the training set with Purchase as the response and the other variables as predictors. Use 1,000 trees, and a shrinkage value of 0.01. Which predictors appear to be the most important? (R package for boosting: "gbm".)

R Code

Listing 5: Sample R Script With Highlighting

```
%% You provide code here %%
```

Most Important Predictors

Answer here ...

2.3 Use the boosting model to predict the response on the test data. Predict that a person will make a purchase if the estimated probability of purchase is greater than 20%. Form a confusion matrix. What fraction of the people predicted to make a purchase do in fact make one?

R Code

Listing 6: Sample R Script With Highlighting

```
%% You provide code here %%
```

Results

Answer here (Confusion matrix, what fraction of the people predicted to make a purchase do in fact make one?) ...

Problem 3

In this question, you are asked to compare Naive Bayes and K-nearest Neighbors (KNN) algorithms over Ionosphere Data Set. Here is the beginning of an R session that allows us to read this data from the web into our local R session:

```
> library(data.table)
> library("curl")
> mydata <- fread("https://archive.ics.uci.edu/ml/machine-learning-databases/
                   ionosphere/ionosphere.data")
> mydata <- as.data.frame(mydata)
> mydata <- mydata[,-2] #remove the second variable
```

3.1 Create a training data set containing a random sample of 300 data points and a test set containing the remaining observations. Name the training data and test data as mydata.training and mydata.testing, respectively. Place the R code below. You will use mydata.training and mydata.testing to answer rest of the questions. Thus, create them once and use mydata.training to train the models (classifiers) and mydata.testing to test the models. The last variable variable (35th variable in the data) is the response and the other variables are predictors.

R Code

Listing 7: Sample R Script With Highlighting

```
%% You provide code here %%
```

3.2 Train a naive bayes classifier using 10-fold cross-validation over mydata.training. Use this model to predict the observations in mydata.testing. Form a confusion matrix and report the error rate of the classifier over mydata.testing.

R Code

Listing 8: Sample R Script With Highlighting

```
%% You provide code here %%
```

Confusion Matrix and Error Rate

Answer here ...

3.3 Perform KNN on mydata.training, with several values of k (k= (2,5,10,50), in order to classify radar returns from the ionosphere. What test errors do you obtain over my mydata.testing? Report the confusion matrices. Which value of k seems to perform the best on the test data.

R Code

Listing 9: Sample R Script With Highlighting

```
%% You provide code here %%
```

Discussion and Results

Answer here ...

3.4 Compare Naive Bayes classifier with KNN, i.e., Which one performed better?

Comparison of the Algorithms

Answer here ...