**ST4061 – Computer Intensive Statistical Analytics II**

2020-2021

In-class test 1

**NAME AND SURNAME: AMIT SOMNATH SAMBREKAR**

**STUDENT NUMBER: 120220153**

**PROGRAM: MSC in DATA SCIENCE AND ANALYTICS**

**INTRUCTIONS**

* Provide your answers in this document, after each question.
* Paste the R code you used for each question item.
* **Save your files regularly.**

Your Word document will be copied directly from your account for assessment.

**Question 1**

Load the following libraries and dataset into your R session, and split the dataset into a training set x.train (all observations prior to 2019) and a test set x.test (all 2019 observations) as follows:

library(class)

library(MASS)

x = read.csv(file="CA1\_data.csv", stringsAsFactors=TRUE)

itrain = (x$Year<2019)

x.train = x[itrain,]

x.test = x[-itrain,]

This dataset contains variables relating to a company’s commercial activity between 2015 and 2019. The response variable of interest in this question is x$Increase, which indicates whether the stock for the company increases on that day. All covariates in the dataset except for Year are used as potential predictors.

1. Fit a logistic regression model using all predictors except for Year. Provide the corresponding confusion matrix obtained for the test set x.test.

**Your answer:**

**y.test**

**glm\_pred No Yes**

**FALSE 28 4**

**TRUE 4 17**

**R code for (1):**

**rm(list =ls())**

**library(class)**

**library(MASS)**

**x = read.csv(file="CA1\_data.csv", stringsAsFactors=TRUE)**

**y = x$Increase**

**itrain = (x$Year<2019)**

**x.train = x[itrain,]**

**x.test = x[itrain == FALSE,]**

**y.train = y[itrain ==FALSE]**

**x.train$Year = NULL**

**y.test = x.test$Increase**

**length(y.test)**

**glm\_xtrain = glm(Increase ~ ., data = x.train, family = binomial(link = logit))**

**summary(glm\_xtrain)**

**glm\_pred = predict(glm\_xtrain, newdata = x.test[,-c(1,2)], type = "response") > 0.5**

**cofusion\_mat = table(glm\_pred, y.test)**

1. Perform linear discriminant analysis using all predictors except for Year. Provide the corresponding confusion matrix obtained for the test set x.test.

**Your answer:**

**> (tb = table(lda.p$class,x.test$Increase))**

**No Yes**

**No 26 5**

**Yes 6 16**

**> sum(diag(tb))/sum(tb)**

**[1] 0.7924528**

**R code for (2):**

**# LDA:**

**lda.o = lda(Increase~., data=x.train)**

**lda.p = predict(lda.o, newdata=x.test)**

**names(lda.p)**

**(tb = table(lda.p$class,x.test$Increase))**

**sum(diag(tb))/sum(tb)**

1. Perform quadratic discriminant analysis using all predictors except for Year. Provide the corresponding confusion matrix obtained for the test set x.test.

**Your answer:**

**> (tb = table(qda.p$class, x.test$Increase))**

**No Yes**

**No 27 5**

**Yes 5 16**

**> sum(diag(tb))/sum(tb)**

**[1] 0.8113208**

**R code for (2):**

qda.o = qda(Increase~., data=x.train)

qda.p = predict(qda.o, newdata=x.test)

(tb = table(qda.p$class, x.test$Increase))

sum(diag(tb))/sum(tb)

1. Compare the classifiers obtained in (2) and (3) and explain, using relevant output:

* How they differ;.
* *(Note: If you did not manage to make your code work, you may still answer what you would expect to find for (a) and (b).)*

**Your answer:**

LDA assumes equality of covariances among the predictor variables X across each all levels of Y. This assumption is relaxed with the QDA model; hence the accuracy is maximum in QDA (81.11%) compared to the accuracy obtained in the LDA model.

LDA (79.24%) is a much less flexible classifier than QDA, and so has substantially lower variance.

QDA is recommended if the training set is very large, so that the variance of the classifier is not a major concern.

**Question 2**

A student is analyzing a dataset in R named dat, whose top 6 rows are shown below. This dataset comprises of 150 observations summarized by 7 covariates x1, …, x7, and a 3-level categorical response variable named type:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **x1** | **x2** | **x3** | **x4** | **x5** | **x6** | **x7** | **type** |
| 5 | 86 | 68 | 28 | 30.2 | 0.364 | 24 | class3 |
| 7 | 195 | 70 | 33 | 25.1 | 0.163 | 55 | class2 |
| 5 | 77 | 82 | 41 | 35.8 | 0.156 | 35 | class3 |
| 0 | 165 | 76 | 43 | 47.9 | 0.259 | 26 | class2 |
| 0 | 107 | 60 | 25 | 26.4 | 0.133 | 23 | class1 |
| 5 | 97 | 76 | 27 | 35.6 | 0.378 | 52 | class3 |

1. The student is applying the kth-nearest neighbours classifier to the above dataset, using the cross-validated value k=5, with the below R instruction:

set.seed(1)

itrain = sample(1:150, 100)

dat.train = dat[itrain, -8]

Y.train = dat[itrain, 8]

dat.test = dat[-itrain, -8]

knn.out = knn(dat.train, dat.test, Y.train, k=5)

What is the student doing wrong when calling knn()? Briefly explain your answer.

**Ans**

Scaling is absent from this data collection. We need to apply scaling as columns of various sizes and KNN does not work well on non-scaled results.

The KNN functions by measuring the nearest point, so it is easier to deal with scaled points rather than unscaled where the distance between the columns is not uniform.

1. Briefly discuss whether logistic regression could be applied to this dataset, and why.

**Your answer:** Logistic regression is a kind of regression analysis used for

predicting the outcome of dependent variable based on one

or more independent variables. A dependent variable can take only two values.