**Quiz 4**

*Content: KNN*

*Functions in R: knn(), sample().*

**PART I**

Consider data set <Smarket.csv>. One would want to split this data set into two parts randomly: one part is used to form a model, and other part is used to test the goodness of fit of the model. R code is given in the photo.

A screenshot of a computer

Description automatically generated

**Q1.** Is the train set and the test set are selected randomly from the original data set?

Ans: NO.

Explanation: The train set is selected which includes all the data point in Year 2002 to Year 2004; while the test set include the data points in Year 2005.

Note: In this code, the original dataset was NOT split randomly. For this code, we focus more on how the algorithm of KNN works. The part on how to split the data randomly is detailed in the later part.

**Q2.** What is the object “prediction” defined by the code in line 20?

The real outcome for data points in the test set, based on KNN classifier with k = 5.

The predicted outcome for data points in the original data set, on KNN classifier with k = 5.

The predicted outcome for data points in the train set, on KNN classifier with k = 5.

The predicted outcome for data points in the test set, based on KNN classifier with k = 5.

Ans: The predicted outcome for data points in the test set, based on KNN classifier with k = 5.

**Q3.** Object “confusion.matrix.1” created by the code in line 22 helps to check the goodness of fit of the KNN classifier with k = 5 on

Test data set

Train data set

Original data set (whole)

ANS: Test data set

Explanation: The code to form “prediction” is: *knn(train[,X],* ***test[,X]****, train[,Y],k=5)*

Which wants to predict the outcome for test set.

The code in line 22: *confusion.matrix.1 = table(prediction, test[ , Y])*

Is to form a table for the predicted outcome (prediction) and the real outcome of the test set, test[ , Y].

**Q4.** Object “pred” is formed using the code in line 24. What is the length of “pred”?

ANS: 1250

Explanation: The classifier here uses the train set as the original data set, *knn(market[,X], market[,X], market[,Y], k=5)*

And the test set is also the original data set, *knn(market[,X],* ***market[,X]****, market[,Y], k=5)*

So, the prediction of outcome is for the test set with 1250 points.

**Q5.** The train set and the test set used to form the KNN classifier in “pred” are the same which is all the data points given in the original data set. True or False?

ANS: True.

**PART II**

There are 9 students with the names are: A, B, C, D, E, F, G, H and I. We would want to form 3 groups where each group has 2 students by a random way. The R code is given in the photo below.

A screenshot of a computer

Description automatically generated

**Q6.** From the R code given, the three groups are formed randomly. True or False?

ANS: True

There should be 3 groups, labeled as 1, 2, and 3.

Each name should be matched randomly to a label. Each label should have 3 names.

Vector <y> is a vector that has labels and it has been mixed up well by the function sample().

For example:

If vector y is: 3 1 1 2 2 3 3 2 1

then the 1st name <A> should be in group 3

the 2nd name <B> should be in group 1

the 3rd name <C> should be in group 1

The 4th name <D> should be in group 2

...

The 9th name <I> should be in group 1.

**Q7.** Ms Pham runs this part of code for the first time and got students D and E for group 1, B and C for group 2, A and F for group 3. However, Valerie Lim also runs the same code, but she got different list as: B and C for group 1; A and D for group 2; E and F for group 3.

It is impossible for Ms Pham and Valerie Lim to get different answers when they run the same code above. True or False?

ANS: False.

Explanation: It is possible for two of them to get different answers. This is due to the sample() function in line 30.  
Line 28 helps to form a set of group labels. 9 students, each one has a group label, 1 or 2 or 3.

Line 30 helps to mix the labels well, not in any order.

Line 34 help to assign students to group 1. Similarly, line 36 and 38 helps to assign students to group 2 and 3, respectively.