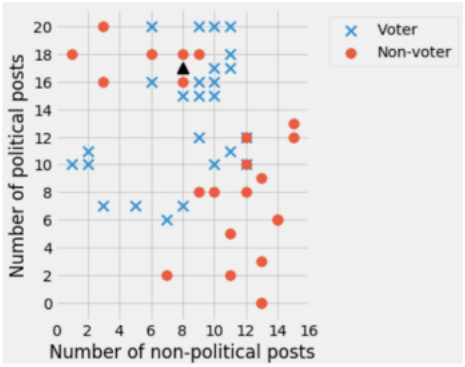
DSA1101 - Introduction to Data Science Suggested Solutions (Semester 1: AY2023/24)

Written by: Minh Duc Vu
Audited by: Agrawal Naman

1. True or False
   (a) If we use linear regression to predict response y based on regressor x for n observations, the average of our residuals, $(1/n)\Sigma_{i=1}^{n} e_i$ will always be zero.
   (b) In order to build a k-nearest neighbors classifier, you do not need to know the class label (response) of any of the training observations.
   (c) A classifier is considered to be overfitting if it performs very well on the training set, but not very well on the test set.
   (d) If I am using House Price and Household Monthly Income (both in Singapore dollars) as two features for my KNN classifier, I do not need to standardize them since they have the same units.
   (e) Candidate A decides to train a classifier to predict whether people will vote in the 2020 U.S. election or not. He gathered data on voting records from the 2018 U.S. election and decides to use two features: the number of political and non-political posts on social media that a person made in the month leading up to the election. A scatter plot of his data is shown below. The candidate is trying to classify the point at (8, 17) shown as a triangle on the graph. If he uses a 3-nearest neighbor classifier, the classification will be "voter".



Solution:

(a) True, assuming that the intercept is present. For simple regression, it's easy to verify: $e_i = y_i - \hat{y}_i = y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}) \implies \sum e_i = \sum(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})) = \sum y_i - n\bar{y} - \hat{\beta}_1(\sum x_i - n\bar{x}) = 0$. In fact, this is also true for a multivaraiate regression model.

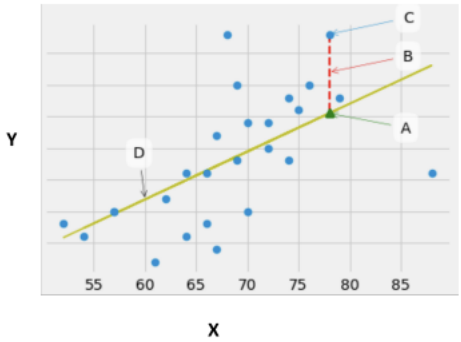
(b) False

(c) True

(d) False, since the scale of the values may still be very different.

(e) False

2. Multiple Choice Questions
A linear regression model was built using a dataset where Y is the response and X is the regressor. The visualization with certain aspects labeled A, B, C, or D is given below. Match each term below to its label on the graph, or "Not Pictured". Some letters may be used multiple times or not at all.



(a) Predicted value
◯ A ◯ B ◯ C ◯ D ◯ Not Pictured

(b) Residual
◯ A ◯ B ◯ C ◯ D ◯ Not Pictured

(c) Line of Best Fit
◯ A ◯ B ◯ C ◯ D ◯ Not Pictured

(d) Intercept
◯ A ◯ B ◯ C ◯ D ◯ Not Pictured

(e) Observed Value
◯ A ◯ B ◯ C ◯ D ◯ Not Pictured

Solution:

(a) Predicted Value: A

(b) Residual: B

(c) Line of Best Fit: D

(d) Intercept: Not Pictured

(e) Observed Value: C

3. A research study investigated Y = whether a patient having surgery with general anesthesia experienced a sore throat on waking (response variable) as a function of D = the duration of the surgery (in minutes) and T = the type of device used to secure the airway.

Variable	Description
Patient	patient's identity number
Y	0=no; 1=yes
T	0=laryngeal mask airway; 1=tracheal tube
D	duration of the surgery (in minutes)

Part I: Logistic Regression Model

```
data1 = read.csv("data1-finals.csv", header = TRUE)

names(data1)[2] = "D" # Change the name of "Duration" variable to "D"
# Y and T are categorical variables
data1$Y = as.factor(data1$Y)
data1$T = as.factor(data1$T)
```

1. Write code to form a logistic regression model (called $M1$) for response Y. Write down the fitted model and explain in detail any notations used.

Solution:

```
M1 <- glm(Y ~ T + D, family = binomial(link = "logit"), data1)
summary(M1)
```

Fitted Model: $log(\hat{p}/(1-\hat{p})) = -1.401 - 1.666 * I(T = 1) + 0.068 * D$
With $\hat{p}$ is the predicted probability of $Y = 1$, or the patient experienced a sore throat upon waking up and Indicator variable I(T = 1) will return 1 if $T = 1$ else 0

2. Report any regressor that is not significant at significant level 0.1.

Solution: At significant level 0.1, there are no insignificant regressors

3. Report the coefficient of the variable D, duration of surgery, in model $M1$. Interpret it in the context of this study.

Solution: D is a very significant variable in this model with coefficient = 0.068.
This means that given the same condition in T, when D, the duration of the surgery increase by 1 minute, then the log-odds of $Y = 1$ will increase by 0.068 or the odds of $Y = 1$ will increase by $e^{0.068} = 1.07$ times

4. Report the coefficient of the variable T, the type of device used to secure the airway, in model $M1$. Interpret it in the context of this study.

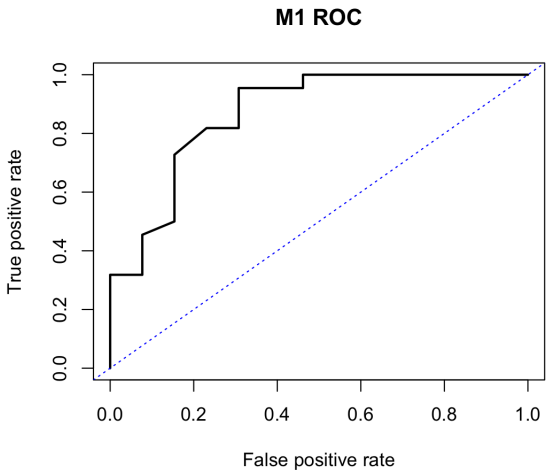
Solution: T is a significant variable in this model (at significance level of 0.1). T is a binary categorical variable, and T = 0 is the reference level
Level T = 1 is indicated by the Indicator variable $I(T = 1)$ with coefficient -1.666
This means that given the same condition in D, compared to T = 0 (laryngeal mask airway) the log-odds of $Y = 1$ for $T = 1$ (tracheal tube) is less than by 1.666; or the odds of $Y = 1$ for $T = 1$ will be less than that of T = 0 by $e^{1.666} = 5.291$ times

5. Write code to plot the ROC curve of model M1. Derive and report the value of AUC.

Solution:

```
M1_predict <- predict(M1, data1, type = "response")
actual_class <- data1$Y == 1
M1_prediction_obj <- prediction(M1_predict, actual_class)

M1_roc = performance(M1_prediction_obj, "tpr", "fpr")
plot (M1_roc, lwd = 2, main = 'M1 ROC', xlim = c(0, 1), ylim = c(0, 1))
abline (a=0, b=1, col ="blue", lty = 3)
M1_auc <- performance(M1_prediction_obj, "auc")@y.values[[1]]
M1_auc # 0.869
# The AUC value of Model M1 is 0.869
```

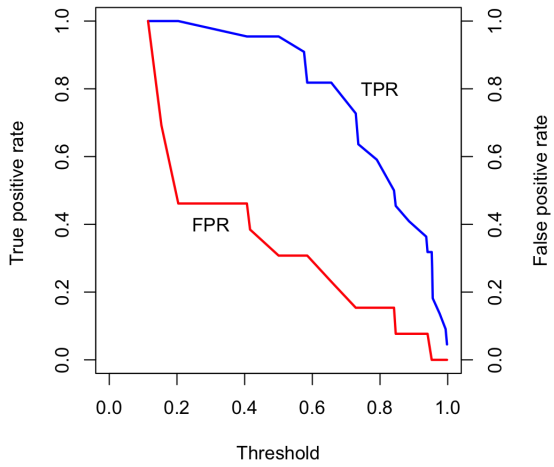


6. Let δ denote the threshold used for a classifier based on the probability derived from model $M1$. Write code to plot a figure to show how the TPR and the FPR of the classifier change when the threshold δ changes.

Solution:

```
alpha <- round(as.numeric(unlist(M1_roc@alpha.values)) ,4)
fpr <- round(as.numeric(unlist(M1_roc@x.values)) ,4)
tpr <- round(as.numeric(unlist(M1_roc@y.values)) ,4)
```

```r
par(mar = c(5, 5, 2 ,5))
plot(alpha, tpr, xlab = "Threshold", xlim = c(0 ,1), ylab = "True positive rate ",
     type ="l", col = "blue", lwd = 2)
par(new = "True")
plot(alpha, fpr, xlab = "", ylab = "", axes = F, xlim = c(0, 1),
     type = "l", col = "red", lwd = 2)
axis(side = 4)
mtext(side = 4, line = 3, "False positive rate")
text(0.3, 0.4, "FPR")
text(0.8, 0.8, "TPR")
```



7. Assume TPR is prioritized over FPR, find the value of δ that gives the best TPR as long as FPR is not larger than 0.5.

Solution:

```r
tpr_fpr = cbind(alpha, tpr, fpr)
tpr_fpr[fpr <= 0.5, ]
# The best threshold is 0.204 with TPR = 1 and FPR = 0.462
```

8. Two patients will have surgery with the estimated time for duration (D) and the type of device used to secure the airway (T) listed below. Write code to predict the probability that the patient will experience sore throat upon waking up from the surgery.
Patient A: D = 80, T = laryngeal mask airway;
Patient B: D = 125, T = tracheal tube.

Solution:

```r
# Patient A: D = 80, T = laryngeal mask airway:
```

```r
predict(M1, data.frame("D" = 80, "T" = "0"), type = 'response')
# 0.983
# Patient B: D = 125, T = tracheal tube:
predict(M1, data.frame("D" = 125, "T" = "1"), type = 'response')
# 0.996
```

Part II: Naive Bayes Classifier

9. We now use the naive Bayes classifier for the dataset given. Write code to form the classifier, named as $M2$.

Solution:

```r
M2 <- naiveBayes(Y ~ T + D, data1)
```

10. Calculate the accuracy of $M2$ on the given dataset.

Solution:

```r
M2_predict = predict(M2, data1, type = "class")
confusion_matrix = table(M2_predict, data1$Y)
M2_acuracy = (confusion_matrix[1, 1] + confusion_matrix[2, 2]) / sum(confusion_matrix)
# M2 Model accuracy is 0.829
```

11. Using the classifier $M2$, predict the probability of having sore throat after surgery for each patient listed in Question 8. Report the probabilities.

Solution:

```r
# Patient A: D = 80, T = laryngeal mask airway:
predict(M2, data.frame("D" = 80, "T" = "0"), type = 'raw')
# 0.997
# Patient B: D = 125, T = tracheal tube:
predict(M2, data.frame("D" = 125, "T" = "1"), type = 'raw')
# 1.0
```

Part III: Decision Trees

12. Write code to form a decision tree (called M3) to predict if a patient has sore throat upon waking up after a surgery, with minsplit $= 4$, where variable selection and split points are based on information gain.

Solution:

```
M3 <- rpart(Y ~ T + D, method = 'class',
            data1, control = rpart.control(minsplit = 4),
            parms = list(split = 'information'))
```

13. Among the features used to form the tree, which one is the most important?

Solution:

```
rpart.plot(M3, type=4, extra=2, varlen=0, faclen=0, clip.right.labs=FALSE)
# Alt, may look at variable importances:
M3$variable.importance
```

Among features used to train the model, D is the most important and significant

14. Calculate the accuracy of M3 on the given dataset.

Solution:

```
M3_predict = predict(M3, data1, type = "class")
confusion_matrix = table(M3_predict, data1$Y) # confusion matrix
M3_acuracy = (confusion_matrix[1, 1] + confusion_matrix[2, 2]) / sum(confusion_matrix)
# M3 Model accuracy is 0.914
```

15. Using the classifier M3, predict the status of having sore throat after surgery for each patient listed in Question 8. Report the results.

Solution:

```
# Patient A: D = 80, T = laryngeal mask airway:
predict(M3, data.frame("D" = 80, "T" = "0"), type = 'class')
# 1
# Patient B: D = 125, T = tracheal tube:
predict(M3, data.frame("D" = 125, "T" = "1"), type = 'class')
# 1
```

4. Consider the famous Iris Flower Data set which was first introduced in 1936 by the famous statistician Ronald Fisher. This data set consists of observations from flowers of Iris species. For each observation, four features were measured: the flower's length and width of the sepals and petals (in cm).

```
data2 = read.csv("data2-finals.csv", header = TRUE)
```

1. Use K-means clustering method to cluster all the flowers into k groups where k = 1, 2, 3, ..., 10, where for each value of k the value of W SS - the within sum of squares is obtained

Solution:

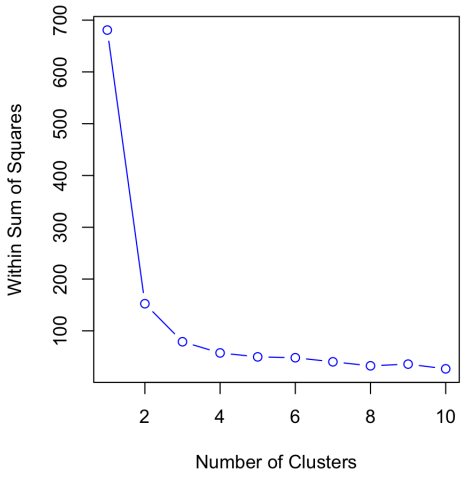
```
K = 10
wss <- numeric(K)
for (k in 1:K) {
  wss[k] <- sum(kmeans(data2, centers = k)$withinss)
}
```

2. Write code to obtain the plot of W SS against k. Which value of k would you choose as the number of clusters for all the observations in the data set? Explain.

Solution:

```
plot(1:K, wss, col = "blue", type="b",
    xlab="Number of Clusters", ylab="Within Sum of Squares")
```

WSS is greatly reduced when k from 1 to 2.
Another substantial reduction in WSS occurs at k from 2 to 3.
However, the improvement in WSS is fairly linear for k ¿ 3.
Therefore, k = 3 will be chosen for the k-means clustering analysis.



3. With the value of k chosen above, report the centroids of all the clusters and the number of the observations in each cluster.

Solution:

```
kout <- kmeans(data2, centers = 3)
kout$centers
kout$size
# Centroids of 3 clusters
# sepal.length sepal.width petal.length petal.width
# 5.006        3.418       1.464        0.244
# 6.85         3.074       5.742        2.071
# 5.902        2.748       4.394        1.434
# Number of points in each clusters: 50, 38, 62 (respectively to the centroids above)
# Note: The values may be different for every run due to random initilization of centroids.
```