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

Written by: Minh Duc Vu
Audited by: Agrawal Naman

1. True or False

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

Solution:

(a) Predicted Value: A

(b) Residual: B

(c) Line of Best Fit: D

(d) Intercept: Not Pictured

(e) Observed Value: C

3. Anesthesia Study:

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. Logistic regression model (called $M1$) for response Y.

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. Regressors that are not significant at significant level 0.1.

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

3. Coefficient of the variable D, duration of surgery, in model $M1$

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. Coefficient of the variable T, the type of device used to secure the airway, in model $M1$.

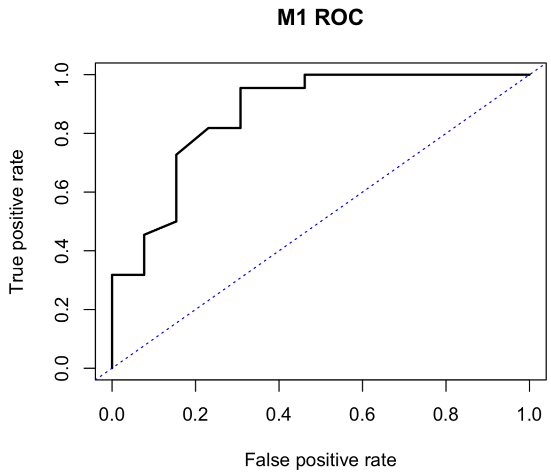
Solution: T is a significant variable in this model (at significance level of 0.1). T is a binary categorical variable, and T = 0 is the reference level
Level T = 1 is indicated by the Indicator variable $I(T = 1)$ with coefficient -1.666
This means that given the same condition in D, compared to T = 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. ROC curve of model M1 and 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
```

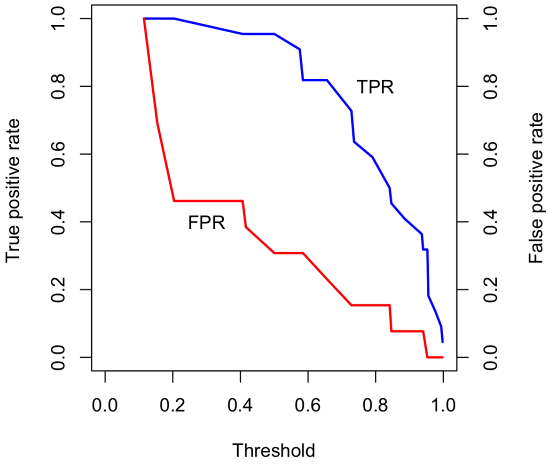


6. Let δ denote the threshold used for a classifier based on the probability derived from model $M1$. How the TPR and the FPR of the classifier change when the threshold δ changes.

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)

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. 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. Code to predict the probability that the patient will experience sore throat upon waking up from the surgery.

Solution:

```r
# Patient A: D = 80, T = laryngeal mask airway:
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. Naive Bayes classifier ($M2$).

Solution:

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

10. Accuracy of $M2$ on the given dataset.

Solution:

```
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. Probability of having sore throat after surgery.

Solution:

```
# 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. Decision tree model (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. Most important feature:

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. 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. Predict the status of having sore throat after surgery.

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. Iris Data Set

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

1. K-means clustering method to cluster all the flowers into k groups and WSS:

Solution:

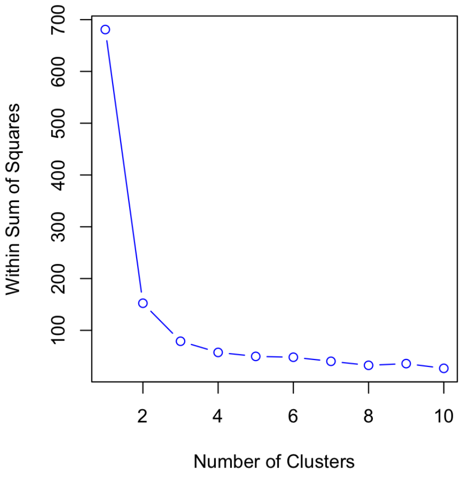
```
K = 10
wss <- numeric(K)
for (k in 1:K) {
  wss[k] <- sum(kmeans(data2, centers = k)$withinss)
}
```

2. Plot of WSS against k

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. Centroids of all 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.
```