STA 141A - Spring 2025 - Homework 5

Instructor: Dr. Akira Horiguchi

Student name: ABCDE FGHIJ; Student ID: 123456789

Due date: May 14, 2025 at 9 PM (PT)

The assignment must be done in an R Markdown or Quarto document. The assignment must be submitted by the due date above by uploading:

1. a .pdf file in GRADESCOPE (if you can knit/compile your .rmd to a .html file only, please save the created .html file as a .pdf file (by opening the .html file -> print -> save to .pdf)).

Email submissions will not be accepted.

Each answer has to be based on `R` code that shows how the result was obtained. The code has to answer the question or solve the task. For example, if you are asked to find the largest entry of a vector, the code has to return the largest element of the vector. If the code just prints all values of the vector, and you determine the largest element by hand, this will not be accepted as an answer. No points will be given for answers that are not based on `R`. This homework already contains chunks for your solution (you can also create additional chunks for each solution if needed, but it must be clear to which tasks your chunks belong).

There are many possible ways to write `R` code that is needed to answer the questions or do the tasks, but for some of the questions or tasks you might have to use something that has not been discussed during the lectures or the discussion sessions. You will have to come up with a solution on your own. Try to understand what you need to do to complete the task or to answer the question, feel free to search the Internet for possible solutions, and discuss possible solutions with other students. It is perfectly fine to ask what kind of an approach or a function other students use. However, you are not allowed to share your code or your answers with other students. Everyone has to write the code, do the tasks and answer the questions on their own.

During the discussion sessions, you may be asked to present and share your solutions.

1. Logistic Regression with the `titanic` dataset

Use the following codes to download the `titanic` data, which provide information on the fate of passengers of Titanic.

```
install.packages("titanic")  # run just once, then comment out
library(titanic)  # run every time
```

(a) Fit a logistic regression model using the `titanic_train` dataset, with `Survived` as the response variable. Use passenger class, sex, age, and fare as predictors. Which predictors are significant, with significance level $\alpha = 0.05$?

```
### Your Solution (Code)
```

Your Solution (Text)

(b) Find the confusion matrix of the train data. Calculate the accuracy.

```
### Your Solution (Code)
```

(c) Predict the survival status (binary) using the `titanic_test` dataset. What portion of passengers are predicted to survive?

```
### Your Solution (Code)
```

2. Linear Discrimant Analysis

We now perform Linear Discrimant Analysis (LDA) on a subset of the `iris` dataset. The `iris` data set is a data frame with 150 samples (rows) and 5 variables (columns) named `Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width`, and `Species`.

(a) Create a subset of the `iris` data frame, where the rows for the species `virginica` and the columns `Sepal.Length` and `Petal.Width` are excluded.

```
### Your Solution (Code)
```

(b) Create a scatter plot, where `Petal.Length` is plotted against `Sepal.Width`, and where each dot is colored by its corresponding species.

```
### Your Solution (Code)
```

(c) Perform LDA using the created subset, and print the fitted model. Please explain the values of `Prior probabilities of groups`, `Group means`, `Coefficients of linear discriminants`. (Hint: The package `MASS` is required for the `lda()` function)

```
### Your Solution (Code)
```

(d) Create the plot in (b) and add the LDA line that discriminates between the two classes.

```
### Your Solution (Code)
```

(e) Obtain the predicted class using the fitted model on the created training dataset. Further, print the confusion matrix. What is the train accuracy for this classifier? (`help(predict)`)

```
### Your Solution (Code)
```

3. Logistic Regression by hand

One is interested in predicting individuals with a certain balance to the class "default" or to "no default" by using logistic regression. We calculated the estimates $\hat{\beta}_0 = -10.4$ and $\hat{\beta}_1 = 0.007$ for the true, but unknown parameters β_0 and β_1 in the logistic regression model, respectively. To which class would you assign individuals with balances $X = \$1,000$ and $X = \$1,750$, given that we conservatively assign individuals with balance $X = x$ to the class "default" if $P(X = x) \geq 0.2$.

4. Discriminant functions by hand

One is interested to analyze the effect of the number of hours regularly working a day in front of a laptop (modeled by X) on concentration (modeled by Y). We distinguish between full concentration ability (class "1"), mild concentration difficulties (class "2") and severe concentration difficulties (class "3"). We have the following data:

$$(10,2),\ (6,2),\ (5,3),\ (6,1),\ (11,2),\ (12,3),\ (13,2),\ (9,2),\ (3,1),\ (2,1),\ (1,1),\ (11,3).$$

a. Calculate the discriminant functions for each of the three classes.

b. Explain to which class you would assign people who regularly work $x = 4, x = 7, x = 14$ hours per day in front of a laptop.

5. Confusion matrix by hand

We assigned 80 individuals to a certain class based on their numbers of hours they regularly do sports a week. Afterwards, we asked them if they feel down or balanced, in other words, if they belong to class “I” (Null) or class “II” (Non-null). We obtained the confusion matrix:

Predicted / True	I	II	Total
I	35	8	43
II	3	34	37
Total	38	42	80

: Confusion matrix

Calculate the accuracy, the false positive rate, the true positive rate, the positive predictive value, and the negative predictive value.