STA 141A: Homework 6

Instructor: Akira Horiguchi

Student name: ABCDE FGHIJ; Student ID: 123456789

Nov 19, 2025 (Wednesday), 22:59 PST

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

- 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. Linear Discriminant Analysis

We now perform Linear Discriminant 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. 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)
```

2. 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.

3. K-means clustering

Recall the K-means clustering algorithm (lecture notes, Sec 9, page 8). Consider the following dataset where $n = 6$ and $p = 2$:

init	x1	x2
1	1	4
1	0	3
2	0	4
1	5	2
2	6	2
2	6	0

(a) Consider the clustering induced by using `init` as the cluster labels. Using this clustering as step 1 of the algorithm, perform each iteration of step 2 of the algorithm until the induced clusters stop changing.

Problems (b) and (c) are from Problem 12.6.1 of ISLR2, and involve the K-means clustering algorithm.

(b) Prove the following identity (*Note*: The left-hand side of Equation (1) is exactly the within-cluster variation from Sec 9, page 6 of the lecture notes):

$$\frac{1}{|C|}\sum_{i,i'\in C}\sum_{j=1}^{p}(x_{i,j}-x_{i',j})^2 = 2\sum_{i\in C}\sum_{j=1}^{p}(x_{i,j}-\bar{x}_{C,j})^2 \tag{1}$$

- $\bar{x}_{C,j} = \frac{1}{|C|}\sum_{i\in C} x_{i,j}$ is the mean of the j-th feature of the points in cluster C,
- $|C|$ is the number of points in cluster C,
- $\|\cdot\|_2$ is the usual Euclidean norm.

Hint: you might get more insight by writing Equation (1) in vector notation.

(c) Denote the value in Equation (1) as $W(C)$. Suppose someone has already chosen a positive integer K. On the basis of the above identity, argue that each iteration of Step 2 of the K-means clustering algorithm (lecture notes, Sec 9, page 8) decreases the objective $\sum_{k=1}^{K} W(C_k)$, where C_k is the cluster $k \in \{1,\dots,K\}$ of K clusters.

4. Principal Component Analysis

Consider the real-valued random variables X_1, X_2, X_3. Suppose the random variable X_1 is independent of the random variable $X_2 + X_3$. Also suppose that the correlation between X_2 and X_3 is 0.5. Suppose we measure X_1, X_2, X_3 on $n = 100$ observations (so here $p = 3$). Furthermore, suppose $Var(X_1) = 5$, and $Var(X_2) = Var(X_3) = 1$. For this data, what are reasonable directions for the first two principal components? (You can write each direction as a unit vector pointing to that direction.) Explain your reasoning.