STA 141A - Spring 2025 - Homework 6

Instructor: Dr. Akira Horiguchi

Student name: ABCDE FGHIJ; Student ID: 123456789

Due date: May 21, 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 two files:

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. Cross-validation

We perform cross-validation on a simulated data set.

```
set.seed(1)
x = runif(100) # 100 values being uniformly distributed (btw 0 and 1) are generated
y = 1 + x -  x^2  + rnorm(100, 0, 0.1)
```

(a) Create a scatterplot where `y` is plotted against `x`. Describe your findings.

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

Your Solution (Text)

(b) Use `lm()` to fit the three models below. Print the summary tables for the three fitted models and comment your findings.

- Model I: $Y = \beta_0 + \beta_1 X + \varepsilon$
- Model II: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$
- Model III: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

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

Your Solution (Text)

(c) Calculate the leave-one-out-cross-validation mean squared error for each model I-III.

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

(d) Calculate the k-fold cross-validation mean squared error for each model I-III for $k = 10$.

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

(e) Which model has the smallest cross-validation error based on your results in (c) and (d)? Briefly explain why.

Your Solution (Text)

(f) Explain the individual concepts and the relationship between the validation set approach, leave-one-out cross-validation and k-fold cross-validation in about 1/2 page (maximum one page).

Your Solution (Text)

2. Bootstrap

Herein, the `iris` data set is used.

(a) Provide an estimate for the population mean of `Sepal.Width`, which is hereafter denoted by $\hat{\mu}$.

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

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result. (Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.)

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

(c) Estimate the standard error of $\hat{\mu}$ by using the bootstrap. How does this compare to your answer from (b)?

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

Your Solution (Text)

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of `Sepal.Width`. Compare it to the results obtained by using `t.test(iris$Sepal.Width)`. *Hint*: You can approximate a 95% confidence interval using the formula

$$\left(\hat{\mu} - 2SE(\hat{\mu}),\, \hat{\mu} + 2SE(\hat{\mu})\right)$$

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

3. K-means clustering

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

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

(a) Let $K = 2$ and consider the clustering induced by using `tag` 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.

Your Solution (Text)

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

(b) Prove the following identity:

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

- $\bar{x}_{C,j} = \frac{1}{|C|}\sum_{i\in C} x_{i,j}$ 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. (The left-hand side of either panel above is exactly the within-cluster variation from Sec 9, page 6 of the lecture notes.) *Hint*: you might get more insight by writing Equation (1) in vector notation.

Your Solution (Text)

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

Your Solution (Text)

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 $p = 3$). 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.