STA 35C: Homework 9

Instructor: Akira Horiguchi

Student name: ABCDE FGHIJ; Student ID: 123456789

Nov 26, 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.

There might be more questions to come. I am posting these just in case you want to get started early or have already started these problems.

1. K-means clustering

Recall the K-means clustering algorithm. 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

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.

2. K-means clustering (code)

ISLR Chapter 12, exercise 10 (a), (c), (d), and (e). In this problem, you will generate simulated data, and then perform K-means clustering on the data.

(i) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.

Hint: There are a number of functions in R that you can use to generate data. One example is the `rnorm()` function; `runif()` is another option. Be sure to add a mean shift to the observations in each class so that there are three distinct classes.

(ii) Perform K-means clustering of the observations with $K = 3$. How well do the clusters that you obtained in K-means clustering compare to the true class labels?

Hint: You can use the `table()` function in R to compare the true class labels to the class labels obtained by clustering. Be careful how you interpret the results: K-means clustering will arbitrarily number the clusters, so you cannot simply check whether the true class labels and clustering labels are the same.

(iii) Perform K-means clustering with $K = 2$. Describe your results.

(iv) Now perform K-means clustering with $K = 4$, and describe your results.