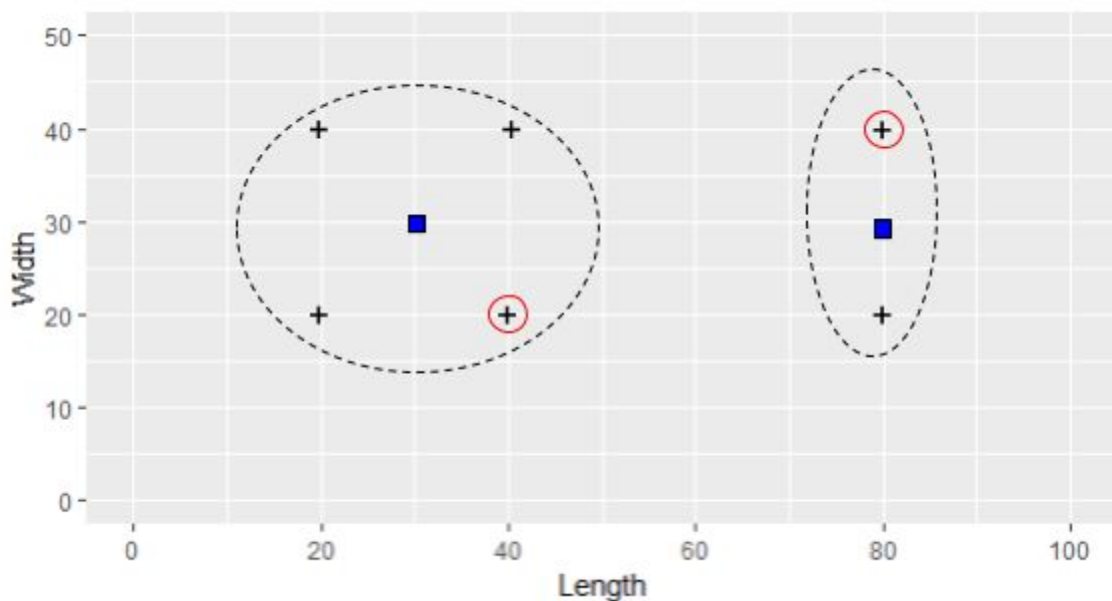# Homework 3 Solution

1. Consider the following dataset. (Note: do NOT write any code for this problem. The answers are to be computed by hand.)

Length	20	40	80	20	40	80
Width	20	20	20	40	40	40

a) Mark the data points on the graph below (use '+' to indicate each point).

b) Let k=2. Let one of the two initial centers be (Length=40, Width=20). Select the second center using the **Farthest Distance Heuristic**. Indicate the two centers on the graph (circle the centers).

c) Recompute the centers after the first iteration of the k-means algorithm.
New center 1:([20+40+20+40]/4, [20+20+40+40]/4) = (30,30)
New center 2:([80+80]/2, [20+40]/2) = (80,30)
Indicate the two new centers on the graph (mark new centers with squares).

d) What are the two clusters after this first iteration? Draw two ovals, each containing all the points in one cluster in the graph above.
e) Will the k-means algorithm terminate after this first iteration or will it continue? Answer in 1-2 sentences.

Since there is no change in which points where grouped together, the cluster centers will not change and the algorithm will terminate.

   f) If a new point (Length=50, Width=25) is given, to which cluster will it belong?

Distance between (50,25) and (30,30) = sqrt($20^2 + 5^2$) =sqrt(425)

Distance between (50,25) and (45,62) = sqrt($5^2 + 37^2$)

Since (30,30) is closer, the new point will belong to its cluster.

2. Consider the attached file `breast-cancer-wisconsin.csv` which contains "Features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass."[1] The goal is to automatically cluster the data based on the features to distinguish Benign and Malignant cases.

   a. Load the data. Column 1 ("Code") is the anonymized subject code and will not be used here. Columns 2-10 are the 9 features. Column 11 is the diagnosis: [B]enign or [M]alignant.

       i.   How many total cases are there in the data?: 683

       ii.   How many [B]enign cases are there in the data?: 444

       iii.   How many [M]alignant cases are there in the data?: 239

   b. Run the k-means clustering algorithm using all the rows and all the 9 features. Use k=2, nstart=10.

       i.   ~~What should be the value of k? k =____~~ (already given)

       ii.   Give R code:

   km <- kmeans(cancer[,2:10], centers=2, nstart=10)

   c. Evaluation: Compare the resulting clusters with the known diagnosis .

       i.   What is the contingency table of your clustering? (Note: you can arbitrarily assign cluster 1/2 to Benign/Malignant)

> table(cancer$Class, km$cluster)

```
    1   2
B   9 435
M 221  18
```

	Cluster 1	Cluster 2
Benign	9	435
Malignant	221	18

---

[1]Original dataset from Breast Cancer Wisconsin (Diagnostic) Data Set
https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

3. The goal is to evaluate three classifiers intending to identify gender (male/female) given the height and weight. The evaluation is to be based on the following dataset with 5 instances:

Gender	Height (cm)	Weight (kg)
Male	148	60
Male	149	66
Female	150	60
Male	151	62
Male	161	72

The three classifiers to be evaluated are:

C1: "*Everyone is male*"

C2: "*Anyone with height over 150cm is male; all others are female*"

C3: Classify using a 1-nearest neighbor classifier trained on the following dataset:

Male	148	65
Female	149	61
Male	160	70

Calculate the following metrics for the classifiers
- Accuracy
- Error rate
- Precision of identifying males
- Recall of identifying males
- F1-score of identifying males

Complete the following table with your answers.

Classifier	Accuracy	Error rate	Precision	Recall	F1-score
C1	4/5 =0.8	1-0.8=0.2	4/5 =0.8	4/4 = 1.0	2*0.8/(1.8)=0.88
C2	3/5=0.6	1-0.6=0.4	2/2 = 1	2/4 = 0.5	2*1*0.5/(1.5)=0.67
C3	3/5=0.6	1-0.6=0.4	2/2 = 1	2/4 = 0.5	2*1*0.5/(1.5)=0.67

Hint: It is easiest to first complete a table with predictions:

Gender	Height (cm)	Weight (kg)	Predicted (C1)	Predicted (C2)	Predicted (C3)
Male	148	60	Male	Female	Female
Male	149	66	Male	Female	Male
Female	150	60	Male	Female	Female
Male	151	62	Male	Male	Female
Male	161	72	Male	Male	Male