Homework 3 - Jared, Kurt, Khoa

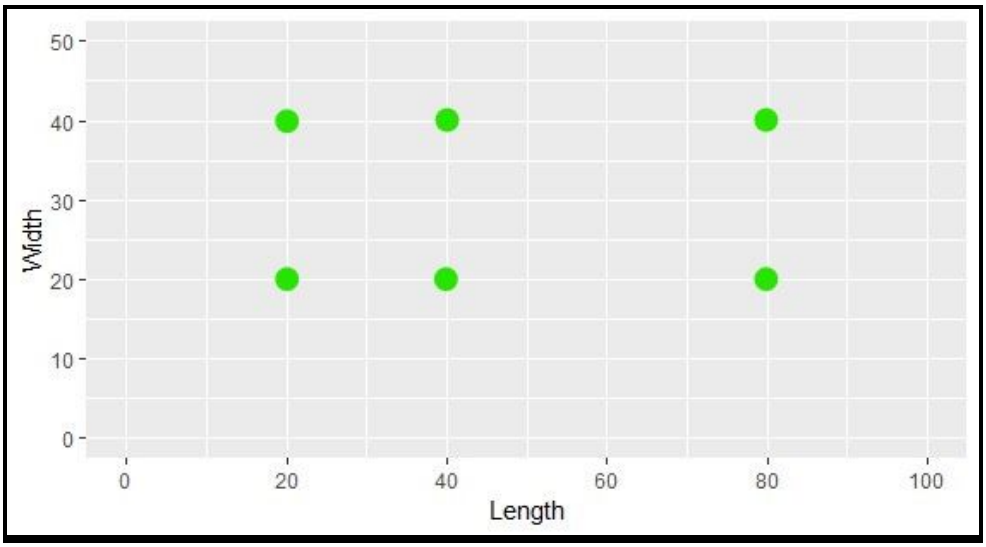
Jared Castaneda
Kurt Prutsman
Khoa Do

Due: Tuesday 10/22, 5:30pm on Titanium. Prepare your answers as a single PDF file.
Group work: You may work in groups of 1-3. Include all group member names in the PDF file. Only one person in the group needs to submit to Titanium.

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).

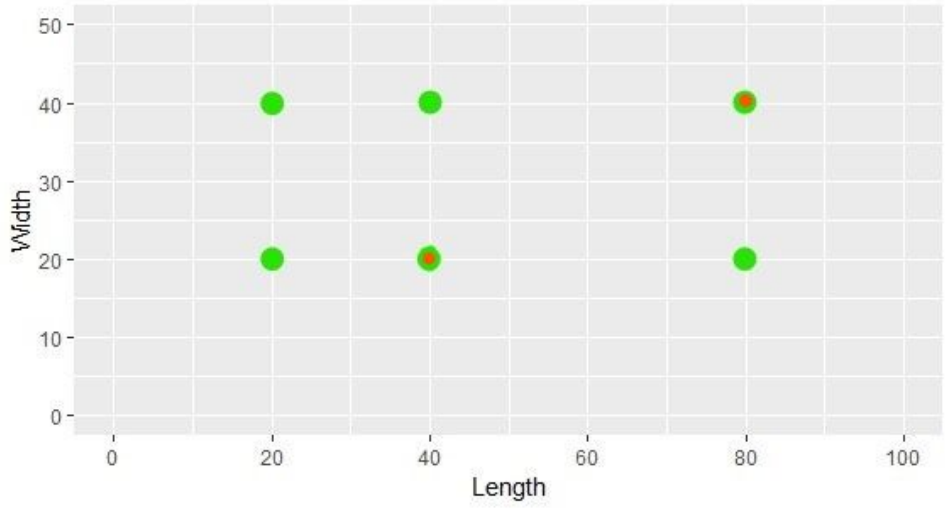
Farthest Distance Heuristic Option 1
Center 1 (40, 20)
Center 2 (80,40)

$$\sqrt{(40-80)^2 + (20-40)^2} = 20\sqrt{5} = 44.721$$

Farthest Distance Heuristic Option 2
Center 1(40,20)
Center 2 (80,20)

$$\sqrt{(40-80)^2 + (20-20)^2} = 40$$

Option 1 has a greater Euclidean Distance, so that will be the second center.



c) Recompute the centers after the first iteration of the k-means algorithm.

$$p(a,b) = |x_2 - x_1| + |y_2 - y_1|$$
$$p(Cluster\ 1,\ P1) = |40-20| + |20-20| = 20$$
$$p(Cluster\ 1,\ P2) = |40-40| + |20-20| = 0$$
$$p(Cluster\ 1,\ P3) = |40-80| + |20-20| = 40$$
$$p(Cluster\ 1,\ P4) = |40-20| + |20-40| = 40$$
$$p(Cluster\ 1,\ P5) = |40-40| + |20-40| = 20$$
$$p(Cluster\ 1,\ P6) = |40-80| + |20-40| = 60$$

$$p(Cluster\ 2,\ P1) = |80-20| + |40-20| = 80$$
$$p(Cluster\ 2,\ P2) = |80-40| + |40-20| = 60$$
$$p(Cluster\ 2,\ P3) = |80-80| + |40-20| = 20$$
$$p(Cluster\ 2,\ P4) = |80-20| + |40-40| = 60$$
$$p(Cluster\ 2,\ P5) = |80-40| + |40-40| = 40$$
$$p(Cluster\ 2,\ P6) = |80-80| + |40-40| = 0$$

	Cluster 1 (40,20)	Cluster 2 (80,40)	Cluster
P1 (20,20)	20	80	Cluster 1
P2 (40,20)	0	60	Cluster 1
P3(80,20)	40	20	Cluster 2
P4 (20,40)	40	60	Cluster 1
P5(40,40)	20	40	Cluster 1
P6(80,40)	60	0	Cluster 2

$$Cluster\ 1\ \{P1(20,20),\ P2(40,20),\ P4(20,40),\ P5(40,40)\}$$
$$Cluster\ 2\ \{P3(80,20), P6(80,40)\}$$

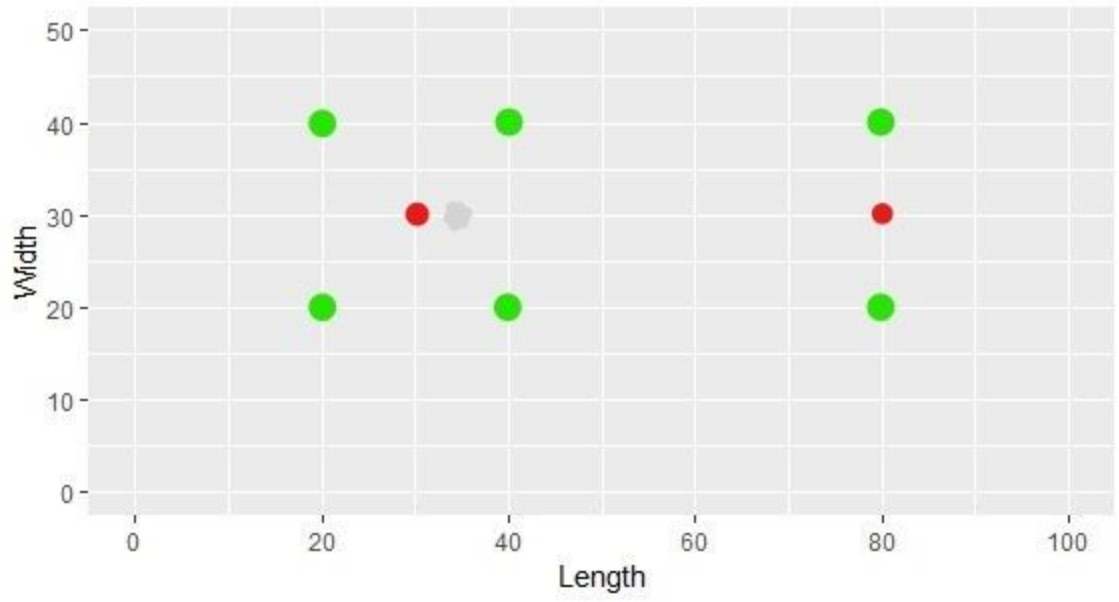
Cluster 1 $\frac{(20+40+20+40)}{4} = 30$ $\frac{(20+20+40+40)}{4} = 30$

Cluster 2 $\frac{(80+80)}{2} = 80$ $\frac{(20+40)}{2} = 30$

New center 1: **(30, 30)**

New center 2: **(80, 30)**

Indicate the two new centers on the graph (mark new centers with squares).



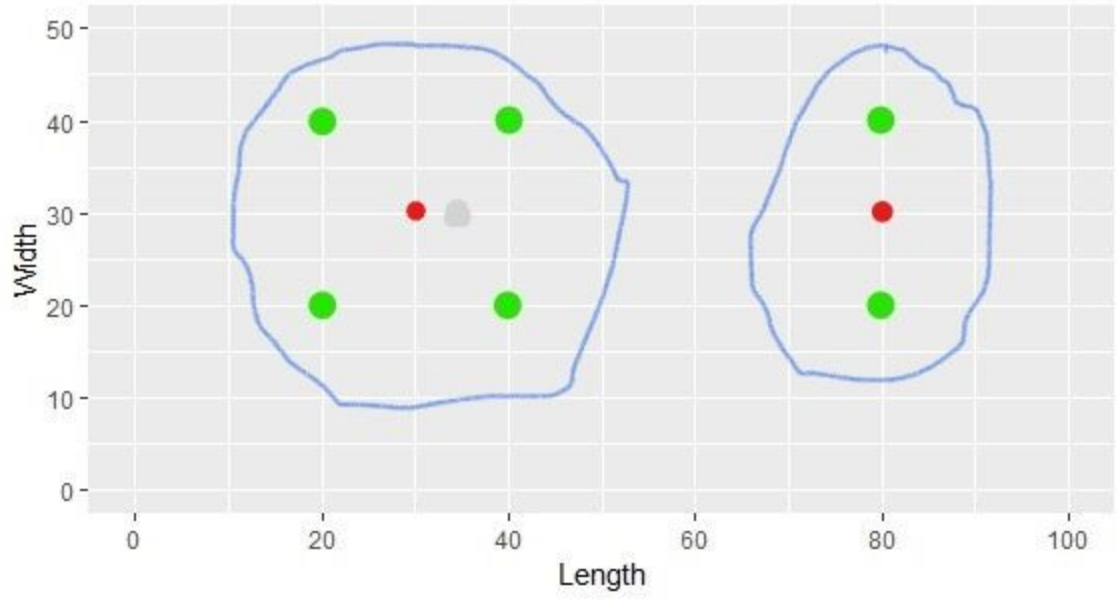
d) What are the two clusters after this first iteration? Draw two ovals, each containing all the points in one cluster in the graph above.

$$p(a,b) = |x_2 - x_1| + |y_2 - y_1|$$

$$p(Cluster\ 1,\ P1) = |30-20| + |30-20| = 20$$
$$p(Cluster\ 1,\ P2) = |30-40| + |30-20| = 20$$
$$p(Cluster\ 1,\ P3) = |30-80| + |30-20| = 60$$
$$p(Cluster\ 1, P4) = |30-20| + |30-40| = 20$$
$$p(Cluster\ 1,\ P5) = |30-40| + |30-40| = 20$$
$$p(Cluster 1,\ P6) = |30-80| + |30-40| = 60$$

$$p(Cluster\ 2,\ P1) = |80-20| + |30-20| = 70$$
$$p(Cluster\ 2,\ P2) = |80-40| + |30-20| = 50$$
$$p(Cluster\ 2,\ P3) = |80-80| + |30-20| = 10$$
$$p(Cluster\ 2,\ P4) = |80-20| + |30-40| = 70$$
$$p(Cluster\ 2,\ P5) = |80-40| + |30-40| = 50$$
$$p(Cluster\ 2,\ P6) = |80-80| + |30-40| = 10$$

	Cluster 1 (35,30)	Cluster 2 (80,30)	Cluster
P1 (20,20)	20	70	Cluster 1
P2 (40,20)	20	50	Cluster 1
P3(80,20)	60	10	Cluster 2
P4 (20,40)	20	70	Cluster 1
P5(40,40)	20	50	Cluster 1
P6(80,40)	60	10	Cluster 2



e) Will the k-means algorithm terminate after this first iteration or will it continue? Answer in 1-2 sentences.

The algorithm wouldn't terminate <u>unless</u> there's no change in which points were grouped together and the clusters change.

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

$$p(Cluster\ 1,\ P7) = |35 - 50| + |30 - 25| = 20$$
$$p(Cluster\ 2,\ P7) = |80 - 50| + |30 - 25| = 35$$

	Cluster 1 (35,30)	Cluster 2 (80,30)	Cluster
P7 (50,25)	20	35	Cluster 1

A new point of (50, 25) would go to Cluster 1.

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___

   R code:

```
nrow(data)
nrow(data %>% filter(Class == "B"))
nrow(data %>% filter(Class == "M"))
```

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 = n/a___
   ii. Give R code:

```
km <- kmeans(data[,2:10], centers = 2, nstart = 10)
data$cluster <- km$cluster
data$cluster <- as.factor(data$cluster)
```

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)

contingencyTable

Filter

	Predicted/Cluster1	Predicted/Cluster2
Actual/M	18	221
Actual/B	435	9

R code:

```
kM1 <- nrow(data %>% filter(Class == "M" & cluster == 1))
kB1 <- nrow(data %>% filter(Class == "B" & cluster == 1))

kM2 <- nrow(data %>% filter(Class == "M" & cluster == 2))
kB2 <- nrow(data %>% filter(Class == "B" & cluster == 2))

contingencyTable <- matrix(c(c(kM1, kB1), c(kM2, kB2)), nrow = 2, ncol = 2)
colnames(contingencyTable) <- paste("Predicted/Cluster", sep = "", c(1, 2))
rownames(contingencyTable) <- paste("Actual/", sep = "", c("M", "B"))
```

[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

Note: do NOT write any code for this problem. The answers are to be computed by hand.

Complete the following table with your answers.

Classifier	Accuracy	Error rate	Precision	Recall	F1-score
C1	.8	.2	.8	1	.89
C2	.6	.4	1	.5	.67
C3	.6	.4	1	.5	.67

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
			All

Accuracy = (TP + TN)/All

Error rate: *1 – accuracy*, or

Error rate = (FP + FN)/All

Precision = $\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$

Recall = $\mathrm{TP}/\mathrm{P}=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$

F = $(2\times\mathrm{precision}\times\mathrm{recall})/(\mathrm{precision}+\mathrm{recall})$

Classifier	Accuracy	Error rate	Precision	Recall	F1-score
C1	(4+0)/5 = .8	1-.8 = .2	4/(4+1) = .8	4/(4+0) = 1	(2*.8*1)/(.8+1) = .89
C2	(2+1)/5 = .6	1 - .6 = .4	2/(2+0) = 1	2/(2+2) = .5	(2*1*.5)/(1+.5) = .67
C3	(2+1)/5 = .6	1 - .6 = .4	2/(2+0) = 1	2/(2+2) = .5	(2*1*.5)/(1+.5) = .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	M	F	F
Male	149	66	M	F	M
Female	150	60	M	F	F
Male	151	62	M	M	F
Male	161	72	M	M	M

Calculating Predicted C1 and C2 is straightforward.

Calculating Predicted C3 can be done by the following rule:

For each test measure loop

For each train measure loop

take the Euclidean Distance between train and test:

$d(m_{test}, m_{train}) = \sqrt{(x_{test} - x_{train})^2 + (y_{test} - y_{train})^2}$

Please note: we can take a shortcut:

$p(m_{test}, m_{train}) = |x_{test} - x_{train}| + |y_{test} - y_{train}|$

end loop

$p(m_{test}, m_{train})_{min} = \min(p(m_{test}, m_{train})_1, p(m_{test}, m_{train})_2, p(m_{test}, m_{train})_3, \ldots)$

$C3_{p\_test\_train}$ = label of $p(m_{test}, m_{train})_{min}$

end loop

Test$_1$

Height (cm)	Weight (kg)	Predicted
148	60	Female

Class/Label	Height (cm)	Weight (kg)	p
Male	148	65	$\|148-148\|+\|65-60\|=5$
Female	149	61	$\|149-148\|+\|61-60\|=2$
Male	160	70	$\|160-148\|+\|70-60\|=22$

Test$_2$

Height (cm)	Weight (kg)	Predicted
149	66	Male

Class/Label	Height (cm)	Weight (kg)	p
Male	148	65	$\|148-149\|+\|65-66\|=$ 2
Female	149	61	$\|149-149\|+\|61-66\|=5$
Male	160	70	$\|160-149\|+\|70-66\|=$ 11+4=15

Test$_3$

Height (cm)	Weight (kg)	Predicted
150	60	Female

Class/Label	Height (cm)	Weight (kg)	p
Male	148	65	$\|148-150\|+\|65-60\|=7$
Female	149	61	$\|149-150\|+\|61-60\|=2$
Male	160	70	$\|160-150\|+\|70-60\|=20$

Test$_4$

Height (cm)	Weight (kg)	Predicted
151	62	Female

Class/Label	Height (cm)	Weight (kg)	p
Male	148	65	$\|148-151\|+\|65-62\|=6$
Female	149	61	$\|149-151\|+\|61-62\|=3$
Male	160	70	$\|160-151\|+\|70-62\|=17$

Test$_5$

Height (cm)	Weight (kg)	Predicted
161	72	Male

Class/Label	Height (cm)	Weight (kg)	p
Male	148	65	$\|148-161\|+\|65-72\|=13+7=20$
Female	149	61	$|149-161|+|61-72|=$12+11=23
Male	160	70	$\|160-161\|+\|70-72\|=3$

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

Contingency Tables

C1:

Actual\Predicted	Male	Female
Male	4	0
Female	1	0

C2:

Actual\Predicted	Male	Female
Male	2	2
Female	0	1

C3:

Actual\Predicted	Male	Female
Male	2	2
Female	0	1