

Richard Gresham
CPSC 375-01
November 17, 2022

Homework 8

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 should submit to Canvas. Please mention the group members name in the PDF.

Due: check on Canvas.

1. What would the 1-Nearest Neighbor approach predict is the class when the feature set is {Soymilk, Apple, Eggs} and the following data is used as the training data (same table as the classwork)? Use the Jaccard similarity to calculate neighbors. Show your calculations. (Note: do NOT write any code for this problem. The answers are to be computed by hand.)

	<i>Feature set</i>	<i>Class</i>
A	{Apple, Banana, Soymilk, Yogurt}	Vegetarian
B	{Apple, Peanuts, Yogurt}	Vegetarian
C	{Tomatoes, Potatoes, Yogurt}	Non-Vegetarian
D	{Apple, Tomatoes, Potatoes}	Non-Vegetarian
E	{Soymilk, Apple, Eggs}	? Vegetarian: class(A)

$$J(E,A) = 2/5 = .4 \quad J(E,C) = 0/6 = 0$$

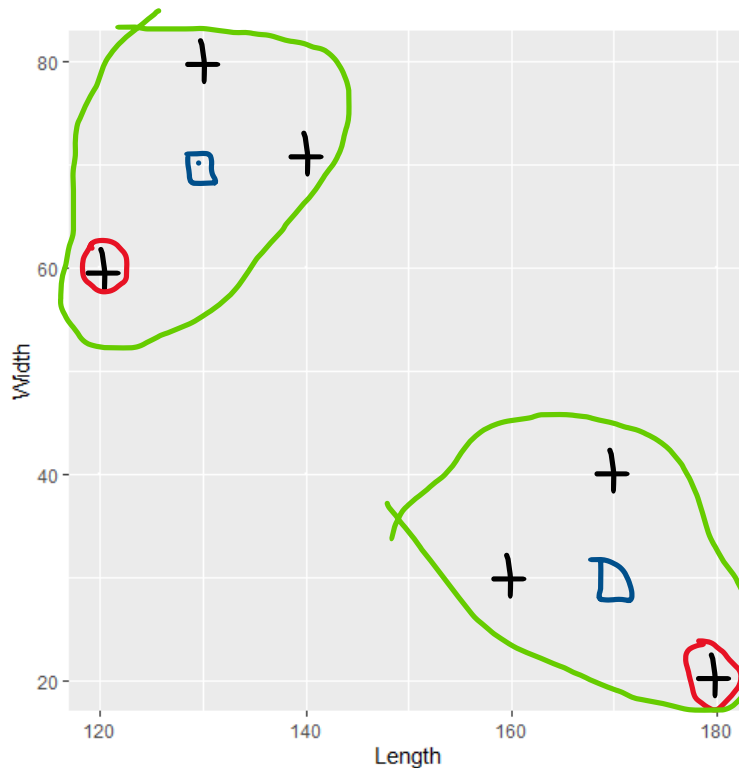
$$J(E,B) = 1/5 = .2 \quad J(E,D) = 1/5 = .2$$

So, by 1NN for jaccard similarity will be with class A and be Vegetarian.

2. Consider the following dataset. (Note: do NOT write any code for this problem. The answers are to be computed by hand and marked on the graph. You can visually guess some of the answers.)

Length	120	140	130	170	160	180
Width	60	70	80	40	30	20

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=120, Width=60). 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: (130, 70)

New center 2: (170, 30)

I can visually tell what clusters will hold what point so now we will add all values divided by the number as shown below.

center 1: $((120 + 130 + 140) / 3, (60 + 80 + 70) / 3) = (130, 70)$

center 2: $((160 + 170 + 180) / 3, (30 + 40 + 20) / 3) = (170, 30)$

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.*

Ovals are drawn above.

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

The k-means algorithm at this point will terminate. This is because since in this new iteration all of the original points in the cluster remain the same, there will be no difference when calculating a new center and thus the algorithm will terminate.

f) If a new point (Length=140, Width=60) is given, to which cluster will it belong?

At this length it is visually closer to the 1st cluster at the top and thus belong to it.

$$\text{sqrt}((140 - 130)^2 + (60 - 70)^2) = \text{sqrt}(-10^2 + -10^2) = \text{sqrt}(200) = \mathbf{14.1421}$$

$$\text{sqrt}((140 - 170)^2 + (70 - 30)^2) = \text{sqrt}(-30^2 + 40^2) = \text{sqrt}(900 + 1600) = \text{sqrt}(2500) =$$

50

Thus, the first cluster will own it.

3. Consider the file `breast-cancer-wisconsin.csv` (in the Datasets module on Canvas) which contains “Features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.”¹ The goal is to cluster the data based on the features to distinguish Benign and Malignant cases.

- a. Read the data from the file into an object called “mydata”. 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.

```
cancer <- read_csv("breast-cancer-wisconsin.csv")
```

```
cancer %>% nrow()
```

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

```
cancer %>% filter(Class == "B") %>% nrow()
```

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

```
cancer %>% filter(Class == "M") %>% nrow()
```

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

- b. Run k-means clustering using **all the rows** and **only the following features**: ClumpThickness, CellSize, and Nuclei. Use `nstart=10`.

- i. What should be the value of k? k = 2

This is because there are only two cases B or M for our dataset.

- ii. Give R code:

```
km <- kmeans(cancer[,c(2,3,7)],centers = 2, nstart = 10)
```

```
km$cluster
```

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

- i. What is the contingency table of your clustering? (Hint: use R’s `table()` function. You can arbitrarily assign cluster 1/2 to Benign/Malignant)

	Cluster 1 Benign	Cluster 2 Malignant
Benign	437	7
Malignant	20	219

- ii. Give R code:

```
km$cluster
```

```
table(cancer$Class, km$cluster)
```

```
1 2
```

¹Original dataset from Breast Cancer Wisconsin (Diagnostic) Data Set

[https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+\(Diagnostic\)](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))

B 437 7
M 20 219

4. Using the contingency table that you obtained from the previous problem (3.c), calculate the following metrics (consider Malignant as the Positive class):

1. Accuracy: $(437 + 219) / (20 + 7 + 437 + 219) = .9604685$
2. Error: $(20 + 7) / 683 = .03953148$
3. Precision: $219 / (219 + 7) = 0.9690265$
4. Recall: $(219) / (219 + 20) = 0.916318$
5. F-score: $(2 * .9690265 * .916318) / (.9690265 + .916318) = 0.9419355$

Consider a “silly” classifier for this problem that makes every prediction as Malignant. Calculate the metrics for this “silly” classifier.

	Benign	Malignant
Benign	0	444
Malignant	0	239

1. Accuracy : $TP + TN / ALL = 0 + 239 / (444 + 239) = .3499268$
2. Error : $(444 + 0) / 683 = 0.6500732$
3. Precision : $TP = 239 / (0 + 239) = 1$
4. Recall: $239 / (239 + 0) = 1$
5. F-score: $(2 \times 1 \times 1) / 2 = 1$