

CS534 - HW 3

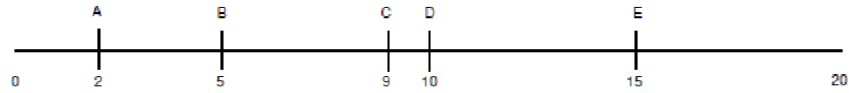
Keith Chester

Due date: July 10th 2022

Problem 1

In this problem we are exploring K -means clustering.

We have a set of data representing five customer ratings on a new car, with a scale from 0 to 20. The image below represents that data:



Part A

Assuming that $K = 2$, and the two initial centroids are 3 and 4, we will try to utilize the K-means algorithm to show all computational steps.

$$K = [2, 5, 9, 10, 15]; C_0 = 3; C_1 = 4 \quad (1)$$

Next we find the distance (simple, since we're working with a singular dimension) from each point in order to assign it.

Data point	Distance from C_0	Distance from C_1
2	1	2
5	2	1
9	6	5
10	7	6
15	12	11

Here we've bolded the closer distance to the associated centroid. We will now recompute the resulting centroid position by taking the mean value of all points assigned.

$$C_i = \frac{1}{n} \sum_{x_i \rightarrow C_i} x_i \quad (2)$$

$$C_0 = \frac{1}{n} \sum_{x_i \rightarrow C_0} x_i = \frac{1}{1} \sum (2) = 2 \quad (3)$$

$$C_1 = \frac{1}{n} \sum_{x_i \rightarrow C_1} x_i = \frac{1}{4} \sum (5, 9, 10, 15) = 9.75 \quad (4)$$

This gives us two new centroids at $C_0 = 2$ and $C_1 = 9.75$. We continue this process until we no longer have any points change centroid assignments.

Data point	Distance from C_0	Distance from C_1
2	0	7.75
5	3	4.75
9	7	0.75
10	8	0.25
15	13	5.25

...since we have a new point on the first centroid, we run the calculations to get the new centroid averages again:

$$C_0 = \frac{1}{n} \sum_{x_i \rightarrow C_0} x_i = \frac{1}{2} \sum (2, 5) = 3.50 \quad (5)$$

$$C_1 = \frac{1}{n} \sum_{x_i \rightarrow C_1} x_i = \frac{1}{3} \sum (9, 10, 15) = 11.33 \quad (6)$$

...leaving us with $C_0 = 3.50$ and $C_1 = 11.33$. Continuing the process:

Data point	Distance from C_0	Distance from C_1
2	0	9.33
5	3	6.333
9	5.50	2.33
10	6.50	0.67
15	11.50	3.67

Since we see no point changes, so our centroids are finalized at $C_0 = 3.50$ and $C_1 = 11.33$.

To find the Silhouette Coefficient Index, we take a given cluster C_i , select a point x_i within it, and find the average distance between it and all other points within that cluster, which we'll mark as a_1 . Then for each other cluster C_j , find the *minimum* average distance (a_j) between this point and the average distance from each point in the other clusters.

...choosing $x_i = x_0 = 2$ for our point

$$a_1 = 3 \quad (7)$$

/noindent ...we need to test both x_0 and x_1 to find the minimum average distance to utilize for our calculation.

$$a_2 = \frac{1}{n} \sum d(x_0, C_1); = \frac{1}{3} (7, 8, 13) = 9.33 \quad (8)$$

$$a_2 = \frac{1}{n} \sum d(x_1, C_1); = \frac{1}{3} (4, 5, 10) = 6.33 \quad (9)$$

...which, since we have only the one other cluster, we can stop here. We'll use the smaller 6.33 average distance. If we had more clusters, we would take the minimum of the average distances from each point in other clusters, but we can stop here. We can now calculate the index S_1 :

$$S_1 = 1 - \frac{a_1}{a_2} = 1 - \frac{3}{6.33} = 0.53 \quad (10)$$

...and then we do the same for the next cluster C_2 . We use $x_i = x_3 = 9$.

$$a_1 = \frac{1}{n} \sum d(x_2, C_1); = \frac{1}{2} (1, 6) = 3.5 \quad (11)$$

...we need to test both x_2 , x_3 and x_4 to find the minimum average distance to utilize for our calculation.

$$a_2 = \frac{1}{n} \sum d(x_2, C_0); = \frac{1}{2} (7, 4) = 5.5 \quad (12)$$

$$a_2 = \frac{1}{n} \sum d(x_3, C_0); = \frac{1}{2} (8, 5) = 6.5 \quad (13)$$

$$a_2 = \frac{1}{n} \sum d(x_4, C_0); = \frac{1}{2} (12, 10) = 11 \quad (14)$$

...using $a_2 = 5.5$ as our minimum.

$$S_2 = 1 - \frac{a_1}{a_2} = 1 - \frac{3.5}{5.5} = 0.37 \quad (15)$$

...so the Silhouette Coefficient Index for our clusters C_0 and C_1 are $S_1 = 0.76$ and $S_2 = 0.37$, respectively.

We are now going to calculate the Davies-Bouldin Index. To find this, we will be first finding the average distance from a given cluster's C_i 's points to its centroid (s_i), the distance between the centroids of two given clusters (d_{ij}), and then the max $R_{i,j}$ where...

$$R_{i,j} = \frac{s_1 + s_2}{d_{ij}} \quad (16)$$

...then we find the max $R_{i,j}$ (which will be the only one we calculate as we only have two clusters), and then take the average of those maxes - will again be just R_{ij} given our setup.

$$D_i = \max(R_{i,j}) \quad (17)$$

$$DB_{index} = \frac{1}{N} \sum_{i=1}^N D_i \quad (18)$$

...and now with our values:

$$s_1 = \frac{1}{2} \sum (1.5, 1.5) = 1.5 \quad (19)$$

$$s_2 = \frac{1}{3} \sum (2.33, 1.33, 3.67) = 2.44 \quad (20)$$

$$R_{0,1} = \frac{1.5 + 2.44}{7.83} = 0.50 \quad (21)$$

Finally we are moving onto the Calinski-Harabasz Index. To calculate this, we will be utilizing the following equation: