ComS 474
Homework 7

Sean Gordon

Nov 20, 2020

1 K-means

1) $X_i \in cluster1$ if $dist(X_i, C_1) < dist(X_i, C_2)$, meaning $||X_i - C_1||^2 < ||X_i - C_2||^2$, else $X_i \in cluster2$

2) Step 1: For a point x, calculate the euclidean distance d_i to each centroid C_i.
Step 2: Select the smallest d_i and record the index i.
Step 3: Assign x to the cluster of the respective centroid C_i.
Step 4: Repeat steps 1-3 for all points in X.

Using the steps above, it can be found that each cluster contains the points below:
Cluster 1: [0, 4, 5], Cluster 2: [1, 2, 3]

These values were calculated using the CLOSEST(X, C) function in kmeans.py.

3) Step 1: For a centroid C, calculate the arithmetic mean M for all points in its cluster.
Step 2: Assign C the calculated value M.
Step 3: Repeat this for all centroids C.

$C_1 = \frac{(-0.57,0.87,-0.89)+(-0.28,0.25,-1.54)+(-1.18,1.26,-0.33)}{3} \Rightarrow$
C_1 = (-0.6767, 0.7933, -0.92)

$C_1 = \frac{(0.04,-0.76,0.41)+(0.55,-0.38,0.56)+(-0.65,-1.66,0.35)}{3} \Rightarrow$
C_1 = (-0.02, -0.9333, 0.44)

---

2 Single-linkage clustering

4)

	(1)	(2)	(3)	(4)	(5)	(6)
(1)	0	2.17232594	2.21797205	2.8186699	0.94392796	0.91531415
(2)	2.17232594	0	0.65345237	1.13564959	2.2192341	2.47313566
(3)	2.21797205	0.65345237	0	1.7670597	2.34431227	2.54452353
(4)	2.8186699	1.13564959	1.7670597	0	2.71239746	3.0446182
(5)	0.94392796	2.2192341	2.34431227	2.71239746	0	1.81499311
(6)	0.91531415	2.47313566	2.54452353	3.0446182	1.81499311	0

5) As clusters 2 and 3 have the shortest distance between them (0.6534), they should be merged.

6)

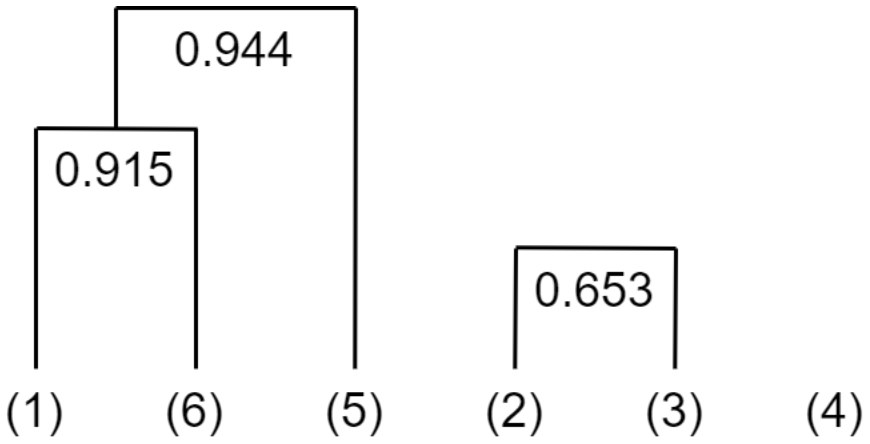
	(1)	(2)	(3)	(4)	(5)	(6)
(1)	0	2.17232594	2.21797205	2.8186699	0.94392796	0.91531415
(2)	2.17232594	0	**0.65345237**	1.13564959	2.2192341	2.47313566
(3)	2.21797205	**0.65345237**	0	1.7670597	2.34431227	2.54452353
(4)	2.8186699	1.13564959	1.7670597	0	2.71239746	3.0446182
(5)	0.94392796	2.2192341	2.34431227	2.71239746	0	1.81499311
(6)	0.91531415	2.47313566	2.54452353	3.0446182	1.81499311	0

	(1)	(2, 3)	(4)	(5)	(6)
(1)	0	2.17232594	2.8186699	0.94392796	**0.91531415**
(2, 3)	2.17232594	0	1.13564959	2.2192341	2.47313566
(4)	2.8186699	1.13564959	0	2.71239746	3.0446182
(5)	0.94392796	2.2192341	2.71239746	0	1.81499311
(6)	**0.91531415**	2.47313566	3.0446182	1.81499311	0

	(1, 6)	(2, 3)	(4)	(5)
(1, 6)	0	2.17232594	2.8186699	**0.94392796**
(2, 3)	2.17232594	0	1.13564959	2.2192341
(4)	2.8186699	1.13564959	0	2.71239746
(5)	**0.94392796**	2.2192341	2.71239746	0

	(1, 5, 6)	(2, 3)	(4)
(1, 6)	0	2.17232594	2.71239746
(2, 3)	2.17232594	0	1.13564959
(4)	2.71239746	1.13564959	0

Dendrogram:



---

3 DBSCAN

7) Neighbors = {B, C, D, G}.

8) A sample is a core point if its number of neighbors > T.

B has 1 neighbor A. C has 3 neighbors A, E, F.
D has 1 neighbor A. G has 4 neighbors A, H, I, J.

As the only neighbor of A that has > 3 neigbors is G, G is the only neigbor of A that is a core point.

9) C={A}

C={A,B,C,D,G}	Added from A
C={A,B,C,D,G,H,I,J}	Added from G
C={A,B,C,D,G,H,I,J,K}	Added from I
C={A,B,C,D,G,H,I,J,K,L,N}	Added from K

10)

Algorithm 1: Shortened DBSCAN Pseudocode.

Data: X: samples, T:, a threshold

```
1  Initialize cluster index i ← 1;
2  foreach sample x ∈ X do
3      if x are NOT assigned to a cluster then
4          Seed set of cluster i: S ← x;
5          while S ≠ ∅ do
6              y ← one element of S;
7              if |N(y)| > T then
8                  Assign y to cluster i;
9                  S ← S ∪ N(y);
10             end
11             Remove y from S;
12         end
13         i++;
14     end
15 end
```