

CS 7301

Sol<sup>n</sup> for HW#3 (Part-of-Samples)

14. You are given a data set with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values of  $K$ ,  $1 \leq K \leq 100$ , the K-means algorithm returns only one non-empty cluster. You then apply an incremental version of K-means, but obtain exactly the same result. How is this possible? How would single link or DBSCAN handle such data?

- (a) The data consists completely of duplicates of one object.
- (b) Single link (and many of the other agglomerative hierarchical schemes) would produce a hierarchical clustering, but which points appear in which cluster would depend on the ordering of the points and the exact algorithm. However, if the dendrogram were plotted showing the proximity at which each object is merged, then it would be obvious that the data consisted of duplicates. DBSCAN would find that all points were core points connected to one another and produce a single cluster.

15. Traditional agglomerative hierarchical clustering routines merge two clusters at each step. Does it seem likely that such an approach accurately captures the (nested) cluster structure of a set of data points? If not, explain how you might postprocess the data to obtain a more accurate view of the cluster structure.

- (a) Such an approach does not accurately capture the nested cluster structure of the data. For example, consider a set of three clusters, each of which has two, three, and four subclusters, respectively. An ideal hierarchical clustering would have three branches from the root—one to each of the three main clusters—and then two, three, and four branches from each of these clusters, respectively. A traditional agglomerative approach cannot produce such a structure.
- (b) The simplest type of postprocessing would attempt to flatten the hierarchical clustering by moving clusters up the tree.

16. Use the similarity matrix in Table 8.1 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

The solutions are shown in Figures 8.6(a) and 8.6(b).

17. Hierarchical clustering is sometimes used to generate  $K$  clusters,  $K > 1$  by taking the clusters at the  $K^{th}$  level of the dendrogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: {6, 12, 18, 24, 30, 42, 48}.

- (a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the

Table 8.1. Similarity matrix for Exercise 16.

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

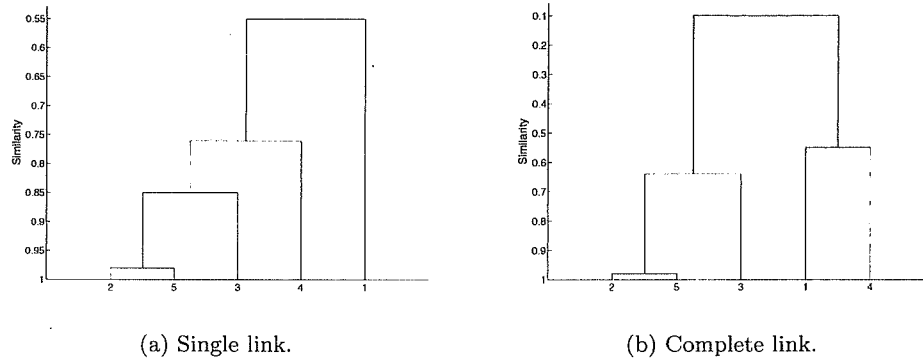


Figure 8.6. Dendrograms for Exercise 16.

total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

- i. {18, 45}
  - First cluster is 6, 12, 18, 24, 30.
  - Error = 360.
  - Second cluster is 42, 48.
  - Error = 18.
  - Total Error = 378
- ii. {15, 40} First cluster is 6, 12, 18, 24 .
  - Error = 180.
  - Second cluster is 30, 42, 48.
  - Error = 168.
  - Total Error = 348.

(b) Do both sets of centroids represent stable solutions; i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

Yes, both centroids are stable solutions.

- (c) What are the two clusters produced by single link?

The two clusters are {6, 12, 18, 24, 30} and {42, 48}.

- (d) Which technique, K-means or single link, seems to produce the “most natural” clustering in this situation? (For K-means, take the clustering with the lowest squared error.)

MIN produces the most natural clustering.

- (e) What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.)

MIN produces contiguous clusters. However, density is also an acceptable answer. Even center-based is acceptable, since one set of centers gives the desired clusters.

- (f) What well-known characteristic of the K-means algorithm explains the previous behavior?

K-means is not good at finding clusters of different sizes, at least when they are not well separated. The reason for this is that the objective of minimizing squared error causes it to “break” the larger cluster. Thus, in this problem, the low error clustering solution is the “unnatural” one.

18. Suppose we find  $K$  clusters using Ward’s method, bisecting K-means, and ordinary K-means. Which of these solutions represents a local or global minimum? Explain.

Although Ward’s method picks a pair of clusters to merge based on minimizing SSE, there is no refinement step as in regular K-means. Likewise, bisecting K-means has no overall refinement step. Thus, unless such a refinement step is added, neither Ward’s method nor bisecting K-means produces a local minimum. Ordinary K-means produces a local minimum, but like the other two algorithms, it is not guaranteed to produce a global minimum.

19. Hierarchical clustering algorithms require  $O(m^2 \log(m))$  time, and consequently, are impractical to use directly on larger data sets. One possible technique for reducing the time required is to sample the data set. For example, if  $K$  clusters are desired and  $\sqrt{m}$  points are sampled from the  $m$  points, then a hierarchical clustering algorithm will produce a hierarchical clustering in roughly  $O(m)$  time.  $K$  clusters can be extracted from this hierarchical clustering by taking the clusters on the  $K^{\text{th}}$  level of the dendrogram. The remaining points can then be assigned to a cluster in linear time, by using various strategies. To give a specific example, the centroids of the  $K$  clusters can be computed, and then each of the  $m - \sqrt{m}$  remaining points can be assigned to the cluster associated with the closest centroid.

means would find the nose, eyes, and mouth straightforwardly as long as the number of clusters was set to 4.

- (c) What limitation does clustering have in detecting all the patterns formed by the points in Figure 8.7(c)?

Clustering techniques can only find patterns of points, not of empty spaces.

21. Compute the entropy and purity for the confusion matrix in Table 8.2.

**Table 8.2.** Confusion matrix for Exercise 21.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total	Entropy	Purity
#1	1	1	0	11	4	676	693	0.20	0.98
#2	27	89	333	827	253	33	1562	1.84	0.53
#3	326	465	8	105	16	29	949	1.70	0.49
Total	354	555	341	943	273	738	3204	1.44	0.61

22. You are given two sets of 100 points that fall within the unit square. One set of points is arranged so that the points are uniformly spaced. The other set of points is generated from a uniform distribution over the unit square.

- (a) Is there a difference between the two sets of points?

Yes. The random points will have regions of lesser or greater density, while the uniformly distributed points will, of course, have uniform density throughout the unit square.

- (b) If so, which set of points will typically have a smaller SSE for  $K=10$  clusters?

The random set of points will have a lower SSE.

- (c) What will be the behavior of DBSCAN on the uniform data set? The random data set?

DBSCAN will merge all points in the uniform data set into one cluster or classify them all as noise, depending on the threshold. There might be some boundary issues for points at the edge of the region. However, DBSCAN can often find clusters in the random data, since it does have some variation in density.

23. Using the data in Exercise 24, compute the silhouette coefficient for each point, each of the two clusters, and the overall clustering.

Cluster 1 contains {P1, P2}, Cluster 2 contains {P3, P4}. The dissimilarity matrix that we obtain from the similarity matrix is the following:

**Table 8.3.** Table of distances for Exercise 23

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Let  $a$  indicate the average distance of a point to other points in its cluster.  
Let  $b$  indicate the minimum of the average distance of a point to points in another cluster.

$$\text{Point P1: SC} = 1 - a/b = 1 - 0.1/((0.65+0.55)/2) = 5/6 = 0.833$$

$$\text{Point P2: SC} = 1 - a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846$$

$$\text{Point P3: SC} = 1 - a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556$$

$$\text{Point P4: SC} = 1 - a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478$$

$$\text{Cluster 1 Average SC} = (0.833+0.846)/2 = 0.84$$

$$\text{Cluster 2 Average SC} = (0.556+0.478)/2 = 0.52$$

$$\text{Overall Average SC} = (0.840+0.517)/2 = 0.68$$

24. Given the set of cluster labels and similarity matrix shown in Tables 8.4 and 8.5, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose  $ij^{th}$  entry is 1 if two objects belong to the same cluster, and 0 otherwise.

**Table 8.4.** Table of cluster labels for Exercise 24. **Table 8.5.** Similarity matrix for Exercise 24.

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

We need to compute the correlation between the vector  $\mathbf{x} = \langle 1, 0, 0, 0, 0, 1 \rangle$  and the vector  $\mathbf{y} = \langle 0.8, 0.65, 0.55, 0.7, 0.6, 0.3 \rangle$ , which is the correlation between the off-diagonal elements of the distance matrix and the ideal similarity matrix.

We get:

$$\text{Standard deviation of the vector } \mathbf{x} : \sigma_x = 0.5164$$

$$\text{Standard deviation of the vector } \mathbf{y} : \sigma_y = 0.1703$$

$$\text{Covariance of } \mathbf{x} \text{ and } \mathbf{y} : \text{cov}(\mathbf{x}, \mathbf{y}) = -0.200$$

27

$$\begin{aligned}
\frac{1}{2|C_i|} \sum_{x \in C_i} \sum_{y \in C_i} (x - y)^2 &= \frac{1}{2|C_i|} \sum_{x \in C_i} \sum_{y \in C_i} ((x - c_i) - (y - c_i))^2 \\
&= \frac{1}{2|C_i|} \left( \sum_{x \in C_i} \sum_{y \in C_i} (x - c_i)^2 - 2 \sum_{x \in C_i} \sum_{y \in C_i} (x - c_i)(y - c_i) \right. \\
&\quad \left. + \sum_{x \in C_i} \sum_{y \in C_i} (y - c_i)^2 \right) \\
&= \frac{1}{2|C_i|} \left( \sum_{x \in C_i} \sum_{y \in C_i} (x - c_i)^2 + \sum_{x \in C_i} \sum_{y \in C_i} (y - c_i)^2 \right) \\
&= \frac{1}{|C_i|} \sum_{x \in C_i} |C_i|(x - c_i)^2 \\
&= \text{SSE}
\end{aligned}$$

The cross term  $\sum_{x \in C_i} \sum_{y \in C_i} (x - c_i)(y - c_i)$  is 0.

28. Prove Equation 8.15.

$$\begin{aligned}
\frac{1}{K} \sum_{i=1}^K \sum_{j=1}^K |C_i|(c_j - c_i)^2 &= \frac{1}{2K} \sum_{i=1}^K \sum_{j=1}^K |C_i|((m - c_i) - (m - c_j))^2 \\
&= \frac{1}{2K} \left( \sum_{i=1}^K \sum_{j=1}^K |C_i|(m - c_i)^2 - 2 \sum_{i=1}^K \sum_{j=1}^K |C_i|(m - c_i)(m - c_j) \right. \\
&\quad \left. + \sum_{i=1}^K \sum_{j=1}^K |C_i|(m - c_j)^2 \right) \\
&= \frac{1}{2K} \left( \sum_{i=1}^K \sum_{j=1}^K |C_i|(m - c_i)^2 + \sum_{i=1}^K \sum_{j=1}^K |C_i|(m - c_j)^2 \right) \\
&= \frac{1}{K} \sum_{i=1}^K K|C_i|(m - c_i)^2 \\
&= \text{SSB}
\end{aligned}$$

Again, the cross term cancels.

29. Prove that  $\sum_{i=1}^K \sum_{x \in C_i} (x - m_i)(m - m_i) = 0$ . This fact was used in the proof that  $\text{TSS} = \text{SSE} + \text{SSB}$  on page 557.