REI502M - Introduction to Data Mining

Solutions to homework 7

Elías Snorrason November 12, 2019

Problem 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 ≤ K ≤ 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?

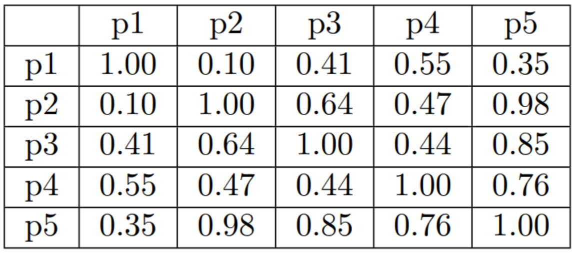
Typically, empty cluster indicate that K, the number of clusters, is greater than the number of points. If this is happening for 100 points in the given range for K, *then all the points in the data set must be identical*, giving only a single non-empty cluster.

Since all points are identical, the similarity/distance matrix would be a matrix of ones/zeroes. For single link agglomerative clustering, clusters can merge with the points in the order that they appear in the data set.
Given that the distance matrix is zero, all points lie within a single circle with an arbitrary radius ε. *With the appropriate value of MinPts, all points will be core points.* Otherwise they will all be considered noise.

Problem 16

Use the similarity matrix in Table 7.13 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.

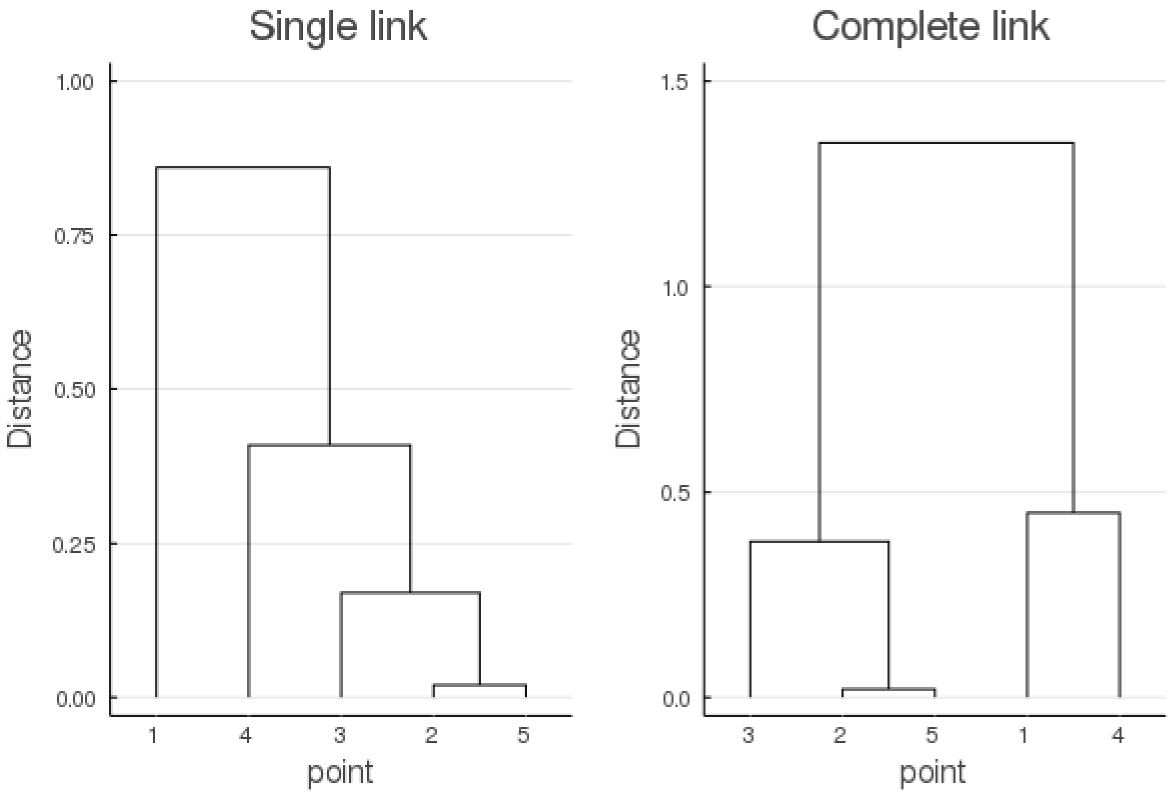
Table 7.13. Similarity matrix for Exercise 16.



	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

In both cases, clusters which minimize distances are merged. Hence, the similarity matrix is changed to a distance/dissimilarity matrix $D = 1 - S$.

```
using StatsPlots, Clustering

S = [1.00 0.10 0.41 0.55 0.35;
     0.10 1.00 0.64 0.47 0.98;
     0.41 0.64 1.00 0.44 0.85;
     0.55 0.47 0.44 1.00 0.76;
     0.35 0.98 0.85 0.76 1.00]

D = 1 .- S

hcsingle = hclust(D,linkage=:single);
pltsingle = plot(hcsingle,xlabel="point", ylabel="Distance", yaxis =[0.0,1.0],title="Single link")

hccomplete = hclust(D,linkage=:complete)
pltcomplete = plot(hccomplete,xlabel="point",ylabel="Distance",yaxis=[0.0,1.5],title="Complete link")

plot(pltsingle,pltsomplete)
```



Problem 17

Hierarchical clustering is sometimes used to generate K clusters, K > 1 by taking the clusters at the Kth 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 total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

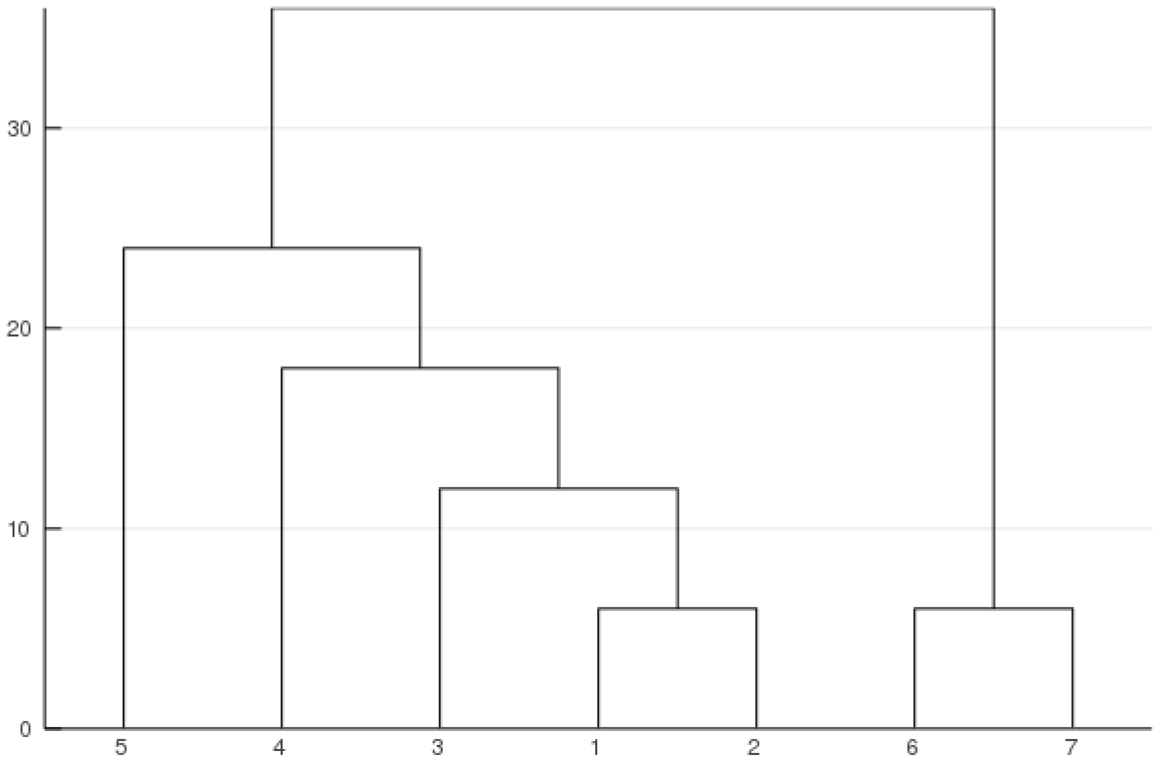
- $\{18, 45\}$
  Clusters are $\{6, 12, 18, 24, 30\}$ and $\{42, 48\}$
  with total square error $(12^2 + 6^2 + 0^2 + 6^2 + 12^2) + (3^2 + 3^2) = 360 + 18 = 378$.
- $\{15, 40\}$
  Clusters are $\{6, 12, 18, 24\}$ and $\{30, 42, 48\}$
  with total square error $(9^2 + 3^2 + 3^2 + 9^2) + (10^2 + 2^2 + 8^2) = 180 + 168 = 348$.

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?

Incidentally, the initial centroids remain unchanged after clustering. So the clusters are stable in both cases.

C. What are the two clusters produced by single link?

Here is the dendrogram produced by single-linking:



The last two elements are in their own cluster. 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.)

K-means selects the second clustering from the first part, while single link selects the first clustering in the first part. Of the two, single link cluster produces more natural clusters (whose centroids are further apart).

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

Single linked clustering is biased towards more dense clusters, which tend to be more contiguous.

F. What well-known characteristic of the K-means algorithm explains the previous behavior?

The natural clusters have different sizes, which K-means handle's poorly. Instead, it will break up the larger cluster and move the centroids together.

Problem 20

Consider the following four faces shown in Figure 7.39. Again, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points

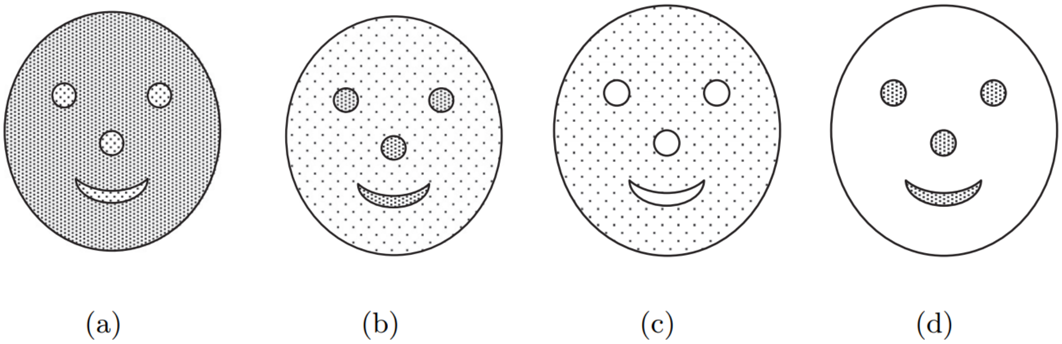


Figure 7.39. Figure for Exercise 20.

A. For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain.

The nose, eyes and mouth would have to be relatively denser than the surrounding areas in order for single-link to work properly. This corresponds to *figures b and d*.

B. For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain

The same applies for K-means since the features are almost globular (except the mouth). With the correct number of clusters (4), K-means would find all features in *figures b and d*, although b will include some lower density portions.

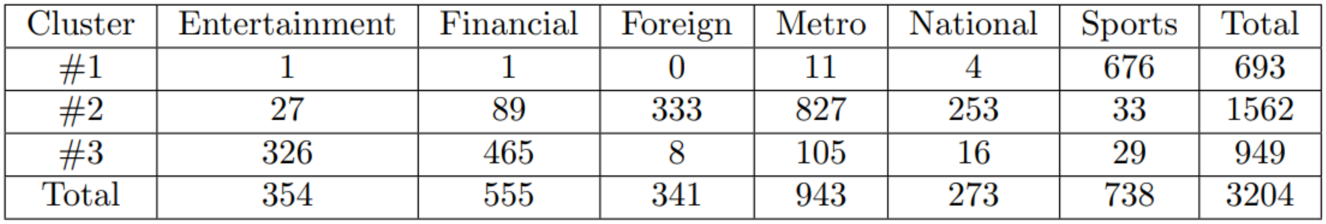
C. What limitation does clustering have in detecting all the patterns formed by the points in Figure 7.39(c)?

The patterns can't be empty sets, otherwise the clustering algorithm will priorotize areas in the figure with higher densities.

Problem 21

Compute the entropy and purity for the confusion matrix in Table 7.14

Table 7.14. Confusion matrix for Exercise 21.



Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Total
#1	1	1	0	11	4	676	693
#2	27	89	333	827	253	33	1562
#3	326	465	8	105	16	29	949
Total	354	555	341	943	273	738	3204

Calculating the entropy for each cluster:

$$\begin{aligned}\text{Entropy}\,(\#\ 1) = &-\frac{1}{693}\log_2\left(\frac{1}{693}\right)-\frac{1}{693}\log_2\left(\frac{1}{693}\right)\\&-\cancelto{0}{\frac{0}{693}\log_2\left(\frac{0}{693}\right)}-\frac{11}{693}\log_2\left(\frac{11}{693}\right)\\&-\frac{4}{693}\log_2\left(\frac{4}{693}\right)-\frac{676}{693}\log_2\left(\frac{676}{693}\right)\\&=0.200\end{aligned}$$

$$\begin{aligned}\text{Entropy}\,(\#\ 2) = &-\frac{27}{1562}\log_2\left(\frac{27}{1562}\right)-\frac{89}{1562}\log_2\left(\frac{89}{1562}\right)\\&-\frac{333}{1562}\log_2\left(\frac{333}{1562}\right)-\frac{872}{1562}\log_2\left(\frac{872}{1562}\right)\\&-\frac{253}{1562}\log_2\left(\frac{253}{1562}\right)-\frac{33}{1562}\log_2\left(\frac{33}{1562}\right)\\&=1.841\end{aligned}$$

$$\begin{aligned}\text{Entropy}\,(\#\ 3) = &-\frac{326}{949}\log_2\left(\frac{326}{949}\right)-\frac{465}{949}\log_2\left(\frac{465}{949}\right)\\&-\frac{8}{949}\log_2\left(\frac{8}{949}\right)-\frac{105}{949}\log_2\left(\frac{105}{949}\right)\\&-\frac{16}{949}\log_2\left(\frac{16}{949}\right)-\frac{29}{949}\log_2\left(\frac{29}{949}\right)\\&=1.696\end{aligned}$$

Purity:

$$\text{Purity}\,(\#\ 1)=\frac{676}{693}=0.975$$

$$\text{Purity}\,(\#\ 2)=\frac{827}{1562}=0.529$$

$$\text{Purity}\,(\#\ 3)=\frac{465}{949}=0.490$$

Total entropy and total purity:

$$\text{Entropy}=\frac{1}{3204}\left(693\cdot 0.200+1562\cdot 1.841+949\cdot 1.696\right)=1.443$$

$$\text{Purity}=\frac{1}{3204}\left(693\cdot 0.975+1562\cdot 0.529+949\cdot 0.490\right)=0.614$$

Problem 24

Given the set of cluster labels and similarity matrix shown in Tables 7.15 and 7.16, respectively, compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ijth entry is 1 if two objects belong to the same cluster, and 0 otherwise.

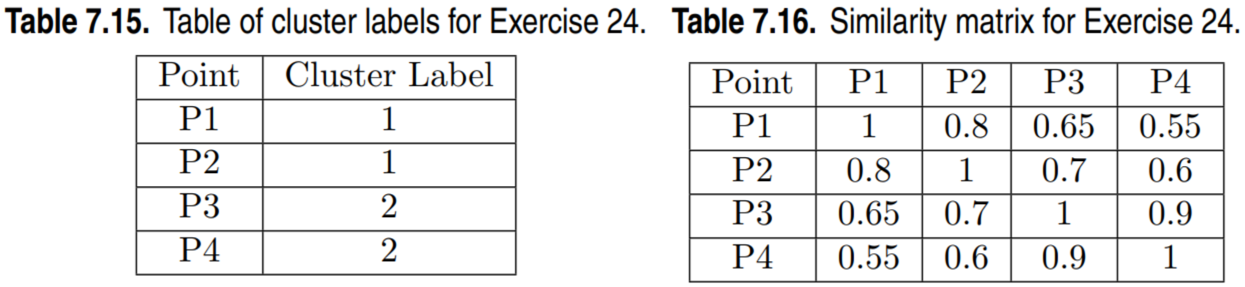


Table 7.15. Table of cluster labels for Exercise 24.

Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 7.16. Similarity matrix for Exercise 24.

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

From the textbook:

> Because the actual and ideal similarity matrices are symmetric, the correlation is calculated only among the $n(n-1)/2$ entries below or above the diagonal of the matrices.

The ideal similarity matrix is a block diagonal matrix (with ones on the block diagonals). We can vectorize the upper triangular parts of both matrices and compare:

$$\mathbf{a} = \{0.8, 0.65, 0.7, 0.55, 0.6, 0.9\} \quad \text{and} \quad \mathbf{b} = \{1, 0, 0, 0, 0, 1\}$$

In Julia:

```
using Statistics

a = [0.8, 0.65, 0.7, 0.55, 0.6, 0.9];
b = [1, 0, 0, 0, 0, 1];

# Standard deviations
std(a) # 0.1304
std(b) # 0.5164

# Covariance
cov(a,b) # 0.0600

# Correlation
cor(a,b) # 0.8911

# Test
cov(a,b) /(std(a)*std(b)) == cor(a,b) # true
```

So the correlation is:

$$\text{Corr}\left[\mathbf{a}, \mathbf{b}\right] = 0.8911$$