COSI 126A Homework 1

Jie Tang

FEBRUARY 17, 2020

Section I: Clustering Problem

Problem 1

Many partitional clustering algorithms that automatically determine the number of clusters claim that this is an advantage. List two situations in which this is not the case.

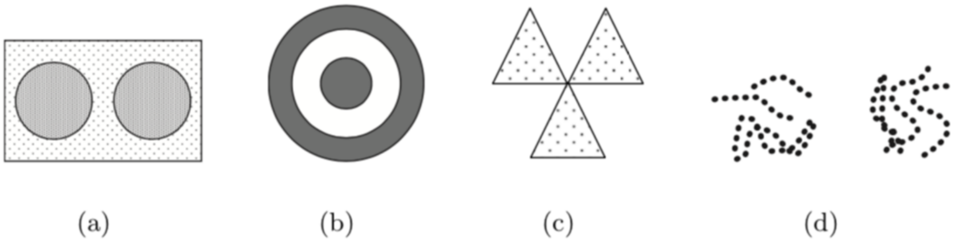
Answer:

(1) One situation that is not an advantage for automatic clustering is when the number of clusters calculated is greater than the system can handle.

(2) The second situation that is not an advantage for automatic clustering when the data set is known and therefore running the algorithm doesn't not return any additional information.

Problem 2

Identify the clusters below using the center-, contiguity-, and density- based definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means K-means, contiguity-based means single link, and density-based means DBSCAN.



Answer:

Figure (a) has three clusters. Because the density-based definition clustering defines the clustering type with the main two clusters have a high density with the remaining surrounding area has a lower density within the figure.

Figure (b) the density-based definition clustering defines the cluster type with the larger outer ring and the inner circle clusters. The ring just outlining the center circle has a lower density which would fit into the center-based definition.

Figure (c) has 3 clusters. And the center-based definition defines the cluster type with each triangle defined as an individual cluster. Center-based definition is chosen due to each triangle, cluster, contains the sample center point.

Figure (d) the continuity-based definition defines the cluster with two clusters. The two clusters that are within the figure is divided by the lower density area between the two sections.

Problem 3

Hierarchical clustering algorithms require O(m2log(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 √ m points are sampledfrom 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 Kth 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 − √ m remaining points can be assigned to the cluster associated with the closest centroid.

For each of the following types of data or clusters, discuss briefly if (1) sampling will cause problems for this approach and (2) what those problems are. Assume that the sampling technique randomly chooses points from the total set of m points and that any unmentioned characteristics of the data or clusters are as optimal as possible. In other words, focus only on problems caused by the particular characteristic mentioned. Finally, assume that K is very much less than m.

(a) Data with very different sized clusters.
Answer: In this case, sampling will cause problems. For example, when there are two clusters, one with 10 points, and the other with 1000 points. Sampling 100 points from the data can be biased.

(b) High-dimensional data.
Answer: In this case, sampling is not a problem. Because when we do sampling, we do randomly select the rows, not the columns.

(c) Data with outliers, i.e., atypical points.
Answer: In this case, sampling might not be problematic. Because the outliers are only a little and accounts few parts of the whole data, it’s unlikely to get many outliers when sampling.

(d) Data with highly irregular regions.
Answer: In this case, sampling is a problem. Because in irregular regions, when we do sampling on points, the outlines of regions might lose.

(e) Data with globular clusters.
Answer: It’s not a problem.

(f) Data with widely different densities.
Answer: In this case, sampling will cause problems. For example, when there are two clusters, cluster one with 10 points, and cluster two with 1000 points. Sampling might choose most of the points from cluster two, and the result can be biased.
(g) Data with a small percentage of noise points.
Answer: In this case, sampling might not be problematic. Because the noise points are only a little and accounts few parts of the whole data, it’s unlikely to get many noise points when sampling.

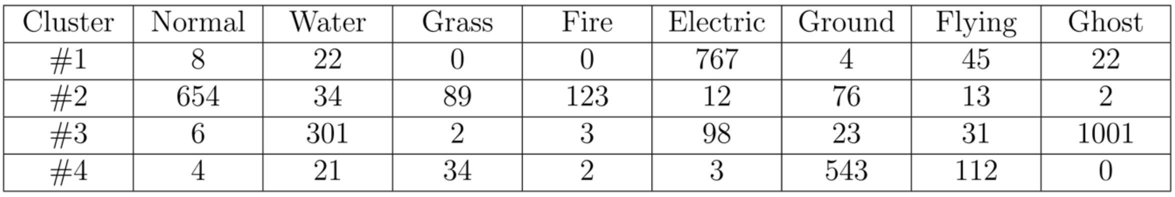
(h) Non-Euclidean data.
Answer: Not a problem.

(i) Euclidean data.
Answer: Not a problem.

(j) Data with many and mixed attribute types.
Answer: In this case, sampling can cause problems. When Data have many and mixed attribute types, the data can be very sparse and the density of data can be very low. Sampling can be biased.

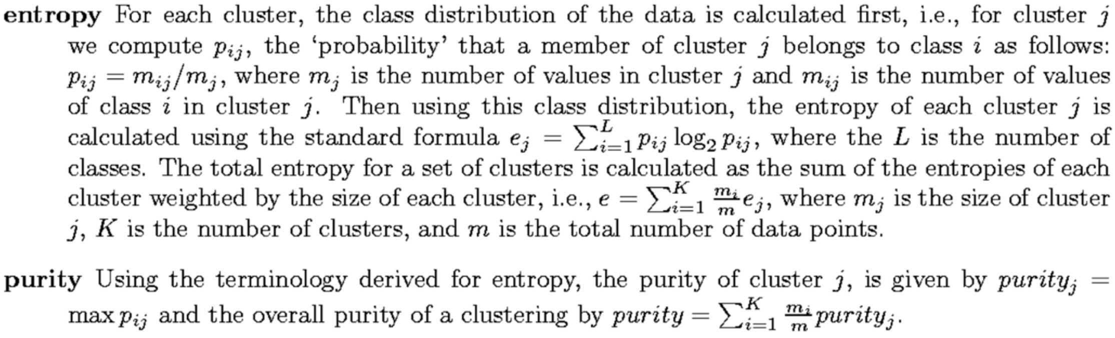
Problem 4

(a) Compute the entropy and purity for each cluster in the confusion matrix below.
(b) Compute the total entropy and total purity.
(c) Compute the following F-measure: F(#3,Water)



Cluster	Normal	Water	Grass	Fire	Electric	Ground	Flying	Ghost
#1	8	22	0	0	767	4	45	22
#2	654	34	89	123	12	76	13	2
#3	6	301	2	3	98	23	31	1001
#4	4	21	34	2	3	543	112	0

Answer:
we know that the formulas of entropy, purity and F-measure are as follows:



entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

(a): Cluster1:

Total points in cluster = 8+22+0+0+767+4+45+22 = 868

E = $\frac{-8}{868}$*$\log_2\frac{8}{868}$+ $\frac{-22}{868}$*$\log_2\frac{22}{868}$ + $\frac{-767}{868}$*$\log_2\frac{767}{868}$ + $\frac{-4}{868}$*$\log_2\frac{4}{868}$ + $\frac{-45}{868}$*$\log_2\frac{45}{868}$ + $\frac{-22}{868}$*$\log_2\frac{822}{868}$ = 0.746

P = $\frac{767}{868}$ = 0.884

For other clusters, we use the same methods and can get the following table result:

Cluster	Entropy	Purity
1	0.746	0.884
2	1.707	0.652
3	1.381	0.683
4	1.180	0.755

(b) Compute the total entropy and total purity.

Answer:

According to the definition of total entropy and total purity, we can see that, they are the weighted entropy and purity of every rows.

Total num = 868 +1003 +1465+ 719 = 4055

total entropy = 868/4055 * 0.746 + 1003/4055*1.707 + 1465/4055*1.381+719/4055*1.180
= 1.290

**total purity = ** 868/4055 * 0.884+ 1003/4055*0.652+ 1465/4055*0.683+719/4055*0.755
= 0.731

(c) Compute the following F-measure: F(#3,Water)

Answer:

$$F = \frac{(\alpha^2 + 1)P * R}{\alpha^2 (P + R)}$$

From the above equation, we can know that F-Measure is the weight of Precision and Recall.

The most common F-measure is when a = 1. F-measure = $\frac{2*P*R}{P+R}$

Precise = 301/1465 = 0.206

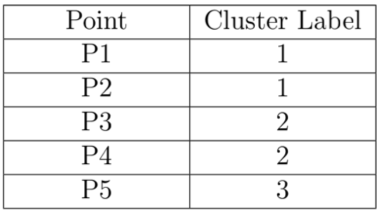
Recall = 301/378 = 0.796

P(water)	R(water)	F(water)
0.206	0.796	0.322

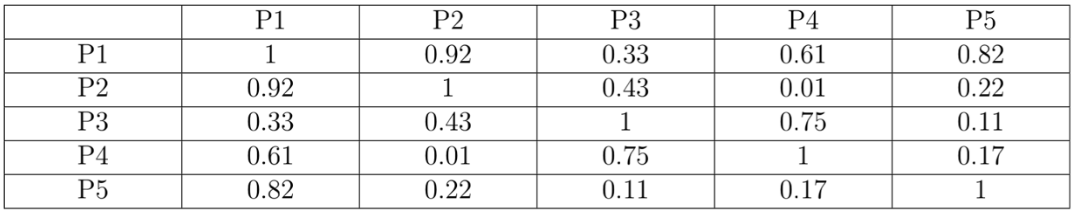
Problem 5

(a) Given the set of cluster labels and similarity matrix shown below, 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.

(b) Compute the silhouette coefficient for each point, each of the three clusters, and the overall clustering.



Point	Cluster Label
P1	1
P2	1
P3	2
P4	2
P5	3



	P1	P2	P3	P4	P5
P1	1	0.92	0.33	0.61	0.82
P2	0.92	1	0.43	0.01	0.22
P3	0.33	0.43	1	0.75	0.11
P4	0.61	0.01	0.75	1	0.17
P5	0.82	0.22	0.11	0.17	1

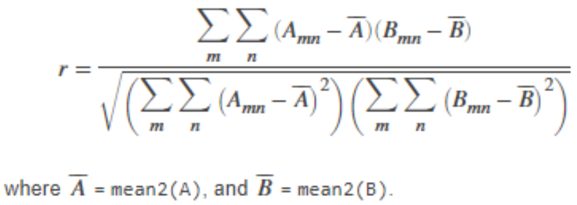
Answer:

(a):

The ideal similarity matrix is shown below:

	P1	P2	P3	P4	P5
P1	1	1	0	0	0
P2	1	1	0	0	0
P3	0	0	1	1	0
P4	0	0	1	1	0
P5	0	0	0	0	1

And the following formula shows how to calculate the correlation between two matrix:



$$r = \frac{\sum_m \sum_n (A_{mn} - \overline{A})(B_{mn} - \overline{B})}{\sqrt{\left(\sum_m \sum_n (A_{mn} - \overline{A})^2\right)\left(\sum_m \sum_n (B_{mn} - \overline{B})^2\right)}}$$

where $\overline{A}$ = `mean2(A)`, and $\overline{B}$ = `mean2(B)`.

we can get the correlation in MATLAB and it is **0.65**.

(b):

Discuss: the way we compute the silhouette coefficient

For an individual point, i

Calculate a = average distance of i to the points in its cluster

Calculate b = min (average distance of i to points in another cluster)

The silhouette coefficient for a point is then given by

s = 1 – a/b if a < b, (or s = b/a - 1 if a ⩾ b, not the usual case)

Typically between 0 and 1. The closer to 1 the better.

We need to get the distance matrix first:

	P1	P2	P3	P4	P5
P1	0	0.08	0.67	0.39	0.18
P2	0.08	0	0.57	0.99	0.78
P3	0.67	0.57	0	0.25	0.89
P4	0.39	0.99	0.25	0	0.83
P5	0.18	0.78	0.89	0.83	0

For each point:

Let S1, S2, S3, S4, S5 stands for silhouette coefficients of P1, P2, P3, P4, P5 respectively.

a1 = (0+0.08)/2 = 0.04, b1 = min{(0.67+0.39)/2, 0.18} = min{0.53, 0.18} = 0.18
S1 = 1 – 0.04/0.18 = 0.78

a2 = (0+0.08)/2 = 0.04, b2 = min{(0.57 +0.99)/2, 0.78} = min{0.78, 0.78} = 0.78
S2 = 1 – 0.04/0.78 = 0.95

a3 = (0+0.25)/2 =0.125, b3 = min{(0.67+0.57)/2, 0.89} = min{0.62, 0.89} = 0.62
S3 = 1 – 0.125/0.62 = 0.80

a4 = (0+0.25)/2 =0.125, b4 = min{(0.99+0.39)/2, 0.17} = min{0.69, 0.83} = 0.69
S4 = 1-0.125/0.69 = 0.82

a5 = 0, b5 = min{(0.18+0.78)/2, (0.89+0.83)/2} = min{0.48, 0.86} = 0.48
S5 = 1 – 0/0.48 = 1
So the result should be:

Point	Sihouette
P1	0.78
P2	0.95
P3	0.80
P4	0.82
P5	1

For each clusters:
Cluster1: S11 = (S1 + S2)/2 = (0.78+0.95)/2 = 0.825
Cluster2: S22 = (S3 + S4)/2 = (0.80+0.82)/2 = 0.81
Cluster3: S33 = S5 = 1

Cluster	Silhouette
1	0.825
2	0.81
3	1

Overall clustering:
S = (S11 + S22 + S33)/3 = (0.825+0.81+1)/3 = 0.878

Section II: Programming

Note: Please see the other file

Bonus Points:
1. My DBSCAN is less than 40 lines (5 points)

2. When we run global k-means and do sampling, why we choose 59 samples so that at least one of them is located in the top 5% with 95% confidence interval? (5 points)

Answer: we assume that there are **n** points in the dataset, and we choose **x** points so that all of them are below 95%, and the probability is 95%. As a result, we can get the following equation:

1 – 0.95^{x} >= 0.95
And the least number of x is 59.