Northeastern University, Data Mining Techniques - CS6220
Fall 2017

Solutions to Homework 2, Part 3

Nakul Camasamudram

October 25, 2017

3 Evaluation

3.1 Cophenetic Correlation Coefficient

Solution:
Table 8.7 from TSK:

Table 8.7. Cophenetic distance matrix for single link and data in table 8.3

Point	P1	P2	P3	P4	P5	P6
P1	0	0.222	0.222	0.222	0.222	0.222
P2	0.222	0	0.148	0.151	0.139	0.148
P3	0.222	0.148	0	0.151	0.148	0.110
P4	0.222	0.151	0.151	0	0.151	0.151
P5	0.222	0.139	0.148	0.151	0	0.148
P6	0.222	0.148	0.110	0.151	0.148	0

a. The cophenetic distance between two clusters is the smallest distance between the two clusters when they were initially merged together.

- Clusters 3 and 6 were the first to be merged in and when they were, the distance between them was 0.11
- Clusters 2 and 5 were merged together next and the distance between them when they were merged was 0.139
- and similarly for P3/P5 and P2/P6

b.
The cophenetic correlation can be computed as follows

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}.$$

- $x(i,j) = |X_i - X_j|$, the ordinary Euclidean distance between the i^{th} and j^{th} observations
- $t(i,j) =$ the cophenetic distance between the points T_i and T_j

Computation:

Datapoints	x(i, j)	t(i, j)	x(i, j) - Mean[x(i, j)]	t(i, j) - Mean[t(i, j)]	Column E * Column F	(Column E)^2	(Column F)^2
p1, p2	0.24	0.222	0.001333333	0.051666667	6.88889E-05	1.77778E-06	0.002669444
p1,p3	0.22	0.222	-0.018666667	0.051666667	-0.000964444	0.000348444	0.002669444
p1,p4	0.37	0.222	0.131333333	0.051666667	0.006785556	0.017248444	0.002669444
p1,p5	0.34	0.222	0.101333333	0.051666667	0.005235556	0.010268444	0.002669444
p1,p6	0.23	0.222	-0.008666667	0.051666667	-0.000447778	7.51111E-05	0.002669444
p2,p3	0.15	0.148	-0.088666667	-0.022333333	0.001980222	0.007861778	0.000498778
p2,p4	0.2	0.151	-0.038666667	-0.019333333	0.000747556	0.001495111	0.000373778
p2,p5	0.14	0.139	-0.098666667	-0.031333333	0.003091556	0.009735111	0.000981778
p2,p6	0.25	0.148	0.011333333	-0.022333333	-0.000253111	0.000128444	0.000498778
p3,p4	0.15	0.151	-0.088666667	-0.019333333	0.001714222	0.007861778	0.000373778
p3,p5	0.28	0.148	0.041333333	-0.022333333	-0.000923111	0.001708444	0.000498778
p3,p6	0.11	0.11	-0.128666667	-0.060333333	0.007762889	0.016555111	0.003640111
p4,p5	0.29	0.151	0.051333333	-0.019333333	-0.000992444	0.002635111	0.000373778
p4,p6	0.22	0.151	-0.018666667	-0.019333333	0.000360889	0.000348444	0.000373778
p5,p6	0.39	0.148	0.151333333	-0.022333333	-0.003379778	0.022901778	0.000498778
	Mean[x(I, j)]	Mean[t(i, j)]			Sum(Column E * Column F)	Sum((Column E)^2)	Sum((Column F)^2)
	0.238666667	0.1703333			0.020786667	0.099173333	0.021459333

Using the above equation and computations from the table,

$$
\begin{aligned}
c &= \frac{Sum(ColumnE * ColumnF)}{\sqrt{Sum((ColumnE)^2) \cdot Sum((ColumnF)^2)}} \\
&= \frac{0.020786667}{\sqrt{0.099173333 \times 0.021459333}} \\
&= .450587616
\end{aligned}
$$

3.2 Purity

Solution:

Table 8.9 from TSK:

Table 8.9. K-means clustering results for the *LA Times* document data set.

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

a. Calculation: Let p_i be the purity of the i^{th} cluster, m_i be the number of objects in the i_{th} cluster and m_{ij} be the number of objects of class j that under cluster i.

- $p_1 = \frac{1}{m_1} \cdot max_j(m_{1j}) = \frac{1}{677} \cdot 506 = 0.7474$
- $p_2 = \frac{1}{m_2} \cdot max_j(m_{2j}) = \frac{1}{361} \cdot 280 = 0.7756$
- $p_3 = \frac{1}{m_3} \cdot max_j(m_{3j}) = \frac{1}{685} \cdot 671 = 0.9796$
- $p_4 = \frac{1}{m_4} \cdot max_j(m_{4j}) = \frac{1}{369} \cdot 162 = 0.4390$
- $p_5 = \frac{1}{m_5} \cdot max_j(m_{5j}) = \frac{1}{464} \cdot 331 = 0.7134$
- $p_6 = \frac{1}{m_6} \cdot max_j(m_{6j}) = \frac{1}{648} \cdot 358 = 0.5525$

Overall purity of the clustering $= \sum_{i=1}^{6} \frac{m_i}{m} \cdot p_i = \frac{1}{[354+555+341+943+273+738]} \cdot [506 + 280 + 671 + 162 + 331 + 358] = 0.7203$

b. Purity is a measure of the extent to which each cluster contains only objects from a single class. So, ideally, the overall purity of the clustering should be 1. In the above case, since it is 0.72034, the clustering is good.

By this metric, Cluster 3 is particularly good as it contains objects almost entirely from the "Sports" class.