Northeastern University
CS6220 – Data Mining Techniques
October 2017, Emily Dutile

Homework 2, Part 3

3.1 Cophenetic Correlation Coefficient

a. Examine Table 8.7 in the TSK text. Explain how the following cells of the table were computed: P3/P6, P2/P5, P3/P5, P2/P6

Table 8.3: x y coordinates from TSK

Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Table 8.4: Euclidean distance matrix from TSK

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

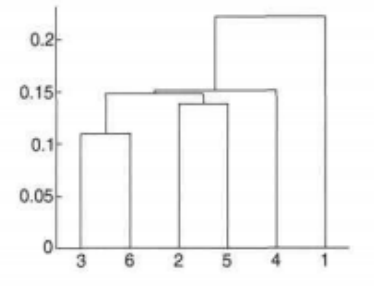
Table 8.7: Cophentetic distance matrix from TSK

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

When applying the Single Link, we should be getting the results that we see in Table 8.4. However, there looks to be an error in the textbook but we will continue to use the textbook values to show the calculations. Below is the sample calculation for the cophenetic distance between P2 and P3. Seen in the image below, the height of the dendrogram should be the distance between cluster P3, P6 and P2, P5:

{P3, P6} and {P2, P5} = $\sqrt{(0.22-0.35)^2+(0.38-0.32)^2} = 0.143$
P3/P6 = dist(3,6) = $\sqrt{(0.35-0.45)^2+(0.32-0.3)^2} = 0.11$ (should be .101)
P2/P5 = dist(2,5) = $\sqrt{(0.22-0.08)^2+(0.38-0.41)^2} = 0.139$ (should be .143)
P3/P5 = dist(3,5) = min(dist(3,2), dist(3,5), dist(6,5), dist(6,2) = min(0.148, 0.28, 0.39, 0.25) = 0.148
P2/P6 = dist(2,6) = min(dist(3, 2), dist(3,5), dist(6,5), dist(6,2) = min(0.148, 0.28, 0.39, 0.25) = 0.148

As we saw from TSK, we get the Single Link dendogram below.



b. Based upon Tables 8.4 and 8.7, show all work to compute the Cophenetic Correlation Coefficient.

$$c = \frac{\sum_{i<j}(d_{ij}-\bar{d})(c_{ij}-\bar{c})}{\sqrt{[\sum_{i<j}(d_{ij}-\bar{d})^2][\sum_{i<j}(c_{ij}-\bar{c})^2]}}$$

Distance	CP
0.24	0.222
0.22	0.222
0.37	0.222
0.34	0.222
0.23	0.222
0.15	0.148
0.2	0.151
0.14	0.139
0.25	0.148
0.15	0.151
0.28	0.148
0.11	0.11
0.29	0.151
0.22	0.151
0.39	0.148

Using the single link agglomerative clustering hierarchy clustering technique, we get the following:

Distance	CP	(x-x_bar)	(y-y_bar)	(x-x_bar)*(y-y_bar)	(x-x_bar)^2	(y-y_bar)^2
0.24	0.222	0.001	0.052	0.000052	0.000001	0.002704
0.22	0.222	-0.019	0.052	-0.000988	0.000361	0.002704
0.37	0.222	0.131	0.052	0.006812	0.017161	0.002704
0.34	0.222	0.101	0.052	0.005252	0.010201	0.002704
0.23	0.222	-0.009	0.052	-0.000468	8.1E-05	0.002704
0.15	0.148	-0.089	-0.022	0.001958	0.007921	0.000484
0.2	0.151	-0.039	-0.019	0.000741	0.001521	0.000361
0.14	0.139	-0.099	-0.031	0.003069	0.009801	0.000961
0.25	0.148	0.011	-0.022	-0.000242	0.000121	0.000484
0.15	0.151	-0.089	-0.019	0.001691	0.007921	0.000361
0.28	0.148	0.041	-0.022	-0.000902	0.001681	0.000484
0.11	0.11	-0.129	-0.06	0.00774	0.016641	0.0036
0.29	0.151	0.051	-0.019	-0.000969	0.002601	0.000361
0.22	0.151	-0.019	-0.019	0.000361	0.000361	0.000361
0.39	0.148	0.151	-0.022	-0.003322	0.022801	0.000484
				0.020785	**0.099175**	**0.021461**

x_bar	y_bar
0.23866667	0.17033333
0.239	0.17

SUM(x-x_bar)^2 * SUM(y-y_bar)^2	C
0.002128395	0.45053024

3.2 Purity

a. Examine Table 8.9 in the TSK text. Show each step to compute the values in the Purity column of the table

The calculation (from class slides):
For each cluster, count the number of data points from the most common class in the cluster
Sum over all clusters and divide by the total number of data points

Sum the values of each row (N). Take the max value in the row (maxv). We get, (1/N)*maxv

Table 8.9 from TSK

Cluster	Enter-tainment	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

i.e.: Cluster 1 row sum=$3 + 5 + 40 + 506 + 96 + 27 = 677$
MAX(3,5,40,506,96,27)=506

Cluster 1: $(1/677) * 506 = 0.7474$
Cluster 2: $(1/361) * 280 = 0.7756$
Cluster 3: $(1/685) * 671 = 0.9796$
Cluster 4: $(1/369) * 162 = 0.4390$
Cluster 5: $(1/464) * 331 = 0.7134$

Cluster 6: $(1/648) * 358 = 0.5525$
Totals: $(506 + 280 + 671 + 162 + 331 + 358)/(354 + 555 + 341 + 943 + 273 + 738) = 0.7203$

b. Based upon the purity metric, is this a good clustering? Which of the clusters is particularly good via this metric provide a reasonable explanation as to why this might be true.

The purity range is from 0 to 1, with 1 being perfect. Cluster 3 looks particularly good via this metric which is .9796. The number of items that is linked with this item (sports) is very high. Looking at the number of clusters and the number of data points all together, there are relatively few clusters in comparison to the number of data points. A smaller number of items will acquire high purity values.