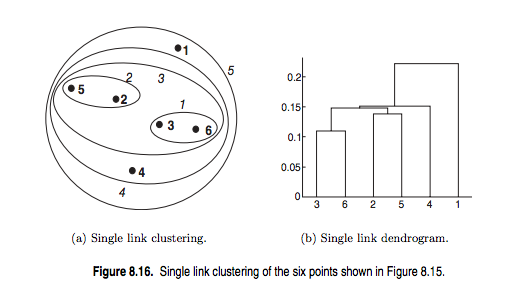
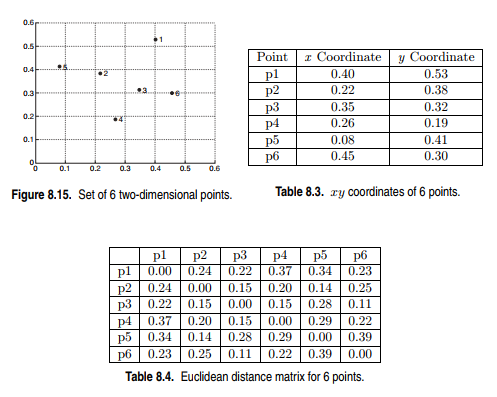
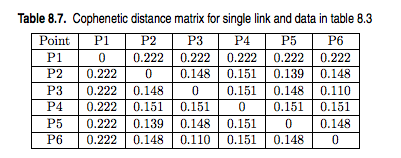
3. Evaluation





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.

**Solution:**

The distance table is provided in Table 8.4

P3/P6 = dist({3, 6}) = 0.11

P2/P5 = dist({2, 5}) = 0.139

P3/P5 = dist({3, 5}) = min(dist(3, 2), dist(6, 5), dist(6, 2), dist(3, 5) = min(0.148, 0.39, 0.25, 0.28) = 0.148

P2/P6 = dist({2, 6}) = min(dist(3, 2), dist(6, 5), dist(6, 2), dist(3, 5) = min(0.148, 0.39, 0.25, 0.28) = 0.148

b. Based upon Tables 8.4 and 8.7, show all work to compute the Cophenetic Correlation Coefficient. (Note that individual additions/subtractions are not necessary, but results of sums and products are – make sure to demonstrate the process.)

**Solution:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Distance** | **CP** | **(x- Mu\_x)** | **(y-Mu\_y)** | **(x-Mu\_x)^2** | | **(y-Mu\_y)^2** |
| 0.24 | 0.222 | 0 | 0.052 | 0 | 0.002704 | |
| 0.22 | 0.222 | -0.02 | 0.052 | 0.0004 | 0.002704 | |
| 0.37 | 0.222 | 0.13 | 0.052 | 0.0169 | 0.002704 | |
| 0.34 | 0.222 | 0.1 | 0.052 | 0.01 | 0.002704 | |
| 0.23 | 0.222 | -0.01 | 0.052 | 1E-04 | 0.002704 | |
| 0.15 | 0.148 | -0.09 | -0.022 | 0.0081 | 0.000484 | |
| 0.2 | 0.151 | -0.04 | -0.019 | 0.0016 | 0.000361 | |
| 0.14 | 0.139 | -0.1 | -0.031 | 0.01 | 0.000961 | |
| 0.25 | 0.148 | 0.01 | -0.022 | 0.0001 | 0.000484 | |
| 0.15 | 0.151 | -0.09 | -0.019 | 0.0081 | 0.000361 | |
| 0.28 | 0.148 | 0.04 | -0.022 | 0.0016 | 0.000484 | |
| 0.11 | 0.11 | -0.13 | -0.06 | 0.0169 | 0.0036 | |
| 0.29 | 0.151 | 0.05 | -0.019 | 0.0025 | 0.000361 | |
| 0.22 | 0.151 | -0.02 | -0.019 | 0.0004 | 0.000361 | |
| 0.39 | 0.148 | 0.15 | -0.022 | 0.0225 | 0.000484 | |

Mu\_x = avg(distance) = 0.2387

Mu\_y = avg(CP) = 0.1703

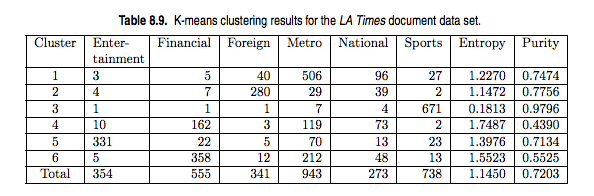
Covariance(distance, CP) = 0.00148

Stddev(distance) = 0.0842

Stddev(CP) = 0.0392

Single link = Stddev(distance) \* Stddev(CP) / Covariance(distance, CP) = 0.45

3.2 Purity



While not robust to increasing K, purity is a simple measure accounting for the extent to which clusters contain a single class.

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

**Solution:**

To calculate Purity: Loop through each cluster (LA Times document). Then for each cluster, select the maximum value from each row, sum them together and finally divide by the total number of data points

For example: for cluster 1

506/(3+5+40+506+96+27=677) = 0.7474

2: 280/361 = 0.7756

3: 671/685 = 0.9796

4: 162/369 = 0.4390

5: 331/464 = 0.7134

6: 358/648 = 0.5525

1. 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.

**Solution:**

Not really, as you can see from the purity value, most of them is about 0.7. Cluster 3 is good though because it’s mostly sports.