**Problem 1**

1) C1 -> {12.0,33.0}

C2 -> {18.0,27.0}

Initial Centroids à Instance 2(12.0,33.0) & 4(18.0,27.0)

First Iteration

|  |  |  |  |
| --- | --- | --- | --- |
| Instances | Point | C1-> {12.0,33.0} | C2 -> {18.0,27.0} |
| 1 | {12,15} | 18 | 13.416 |
| 2 | {12,13} | 0 | 15.23 |
| 3 | {18,15} | 18.91 | 12 |
| 4 | {18,27} | 8.48 | 0 |
| 5 | {24,21} | 16.97 | 8.48 |
| 6 | {36,42} | 25.63 | 23.43 |

After first iteration, C1->{2} & C2->{1,3,5,6,4}

New Centroids à (12.0,33.0) & (21.6,24.0)

Second Iteration

|  |  |  |  |
| --- | --- | --- | --- |
| Instances | Point | C1-> {12.0,33.0} | C2 -> {21.6,24.0} |
| 1 | {12,15} | 18 | 13.15 |
| 2 | {12,13} | 0 | 14.6 |
| 3 | {18,15} | 18.97 | 9.69 |
| 4 | {18,27} | 8.48 | 4.68 |
| 5 | {24,21} | 16.97 | 3.84 |
| 6 | {36,42} | 25.63 | 23.05 |

After Second iteration, C1->{2} & C2->{1,3,5,6,4}

More iterations are not required to get the final clusters because in 1st iteration we get clusters C1->{2} & C2->{1,3,5,6,4} and in second iteration the clusters obtained are C1->{2} & C2->{1,3,5,6,4} and the centroid remains the same for every iteration after completing the second iteration.

2)Usual shape is spherical (with the radius equal to the distance between the centroid and the farthest point).

Ex: It cannot handle non globular clusters or clusters of different sizes and densities.

K means -> updates the cluster by recomputing the centroid and the presence of outliers may push the centroid near or closer to the outlier.

3)In the above given image k-means doesn’t perform well when groups are non-spherical because the k-means tends to pick spherical groups. One pre-processing technique is to use many clusters.

**Problem 2** (<https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>)

Single-link and complete-link clustering reduce the assessment of cluster quality to a single similarity between a pair of documents: the two most similar documents in single-link clustering and the two most dissimilar documents in complete-link clustering. A measurement based on one pair cannot fully reflect the distribution of documents in a cluster. It is therefore not surprising that both algorithms often produce undesirable clusters. Single-link clustering can produce straggling clusters as shown in Figure 2.1. Since the merge criterion is strictly local, a chain of points can be extended for long distances without regard to the overall shape of the emerging cluster. This effect is called *chaining*.

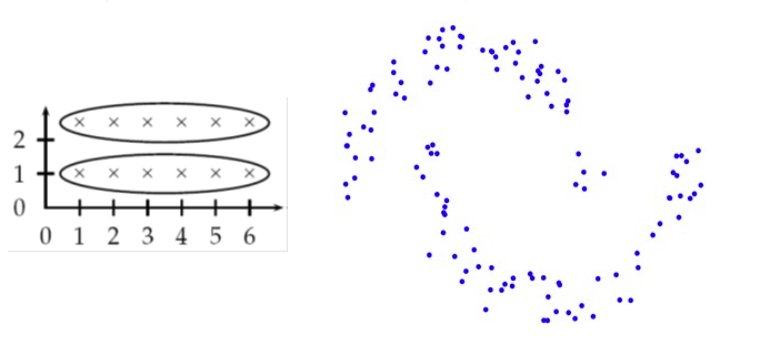


Figure 2.1 Chaining in single-link clustering. The local criteria in single-link clustering can cause undesirable elongated clusters

The chaining effect is also apparent in Figure 2.2. The last eleven merges of the single-link clustering add on single documents or pairs of documents, corresponding to a chain.

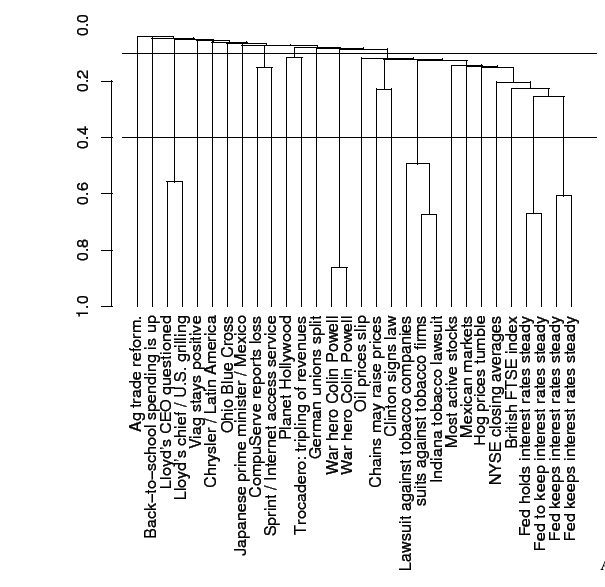
.

Figure 2.2 A dendrogram of a single-link clustering of 30 documents from Reuters-RCV1. Two possible cuts of the dendrogram are shown: at 0.4 into 24 clusters and at 0.1 into 12 clusters.

The complete-link clustering has a more useful organization of the data than a clustering with chains.However, complete-link clustering suffers from a different problem. It pays too much attention to outliers, points that do not fit well into the global structure of the cluster. In the example in Figure 2.3 the four documents d2,d3,d4,d5 are split because of the outlier d1 at the left edge . Complete-link clustering does not find the most intuitive cluster structure .

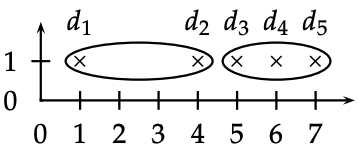


Figure 2.3 Outliers in complete-link clustering. The five documents have the x-coordinates 1 + 2ǫ, 4, 5 + 2ǫ, 6 and 7 − ǫ. Complete-link clustering creates the two clusters shown as ellipses. The most intuitive two-cluster clustering is {{d1}, {d2, d3, d4 , d5}}, but in complete-link clustering, the outlier d1 splits {d2, d3, d4 , d5} as shown.