TDT4300 Datavarehus og datagruvedrift - Spring 2013
Assignment 4: Clustering

Summary

In this assignment we will repeat some of the more exam-relevant theory parts of the book and work with clustering.

1 Repetition: Apriori Algorithm

Given the shopping basket in Table 1, use the apriori algorithm to generate all possible association rules (for minimum support .5 and minimum confidence .8). Provide step-by-step notes on how you reach your result.

2 Clustering

2.1 k-Means Clustering

You are given the one-dimensional data set shown in Table 2. Perform k-Means clustering on this data.

Perform the clustering for two initial centroids: 2 and 5; and for three initial centroids: 2, 6, 8. Document your observations.

ID	*Transaction*
1	A,B,C
2	A,C
4	A,D
5	B,E,F

Table 1: Shopping Basket.

ID	*X*
P1	3
P2	1
P3	2
P4	4
P5	7
P6	9
P7	6
P8	9
P9	6
P10	8

Table 2: Data for k-means clustering.

X	Y
1	11
1	9
1	5
1	2
6	7
11	7

Table 3: Data for Hierarchical clustering.

Note Use Euclidean distance (the distance calculations are easier with one dimension only $d(q,p) = \sqrt{(p_1 - q_1)^2}$, e.g. to compute the distance between points P2 and P1 is $\sqrt{(1-2)^2} = \sqrt{1}$. Hint, this is equivalent to the absolute value of the difference, i.e. $d(q,p) = |p_1 - p_2|$.

2.2 Hierarchical Agglomerative Clustering HAC

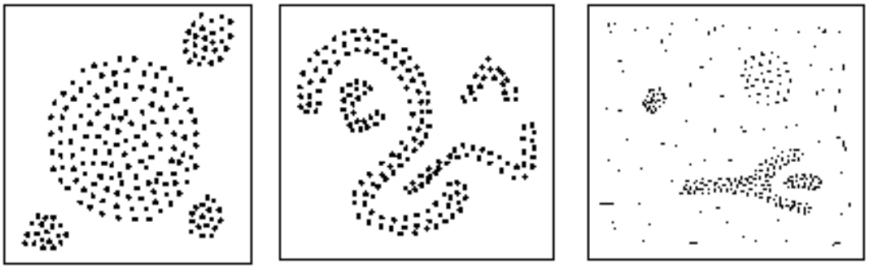
You are given the two-dimensional data points in Table 3.

1. Explain the hierarchical clustering and the difference between MIN-link and MAX-link.

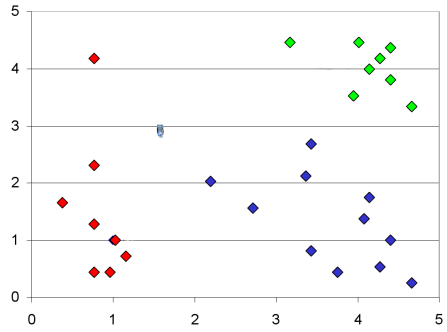
2. Perform hierarchical agglomerative clustering on the dataset of Table 3 and show the resulting dendrogram. Perform both MIN-link and MAX-link. **All calculations performed should be written in the report**.

2.3 Clustering Methods

Given the following three descriptions of datasets, decide what clustering algorithm to use and argue why (i.e. for each of the datasets choose one of k-means, HAC, or DBSCAN:



(a)



(b)

Figure 1: Plots of different data sets.

1. Text collection (100.000 documents, 30.000 dimensions, i.e. 30.000 distinct words)
2. Noisy data collection (200 instances, 3 dimensions)
3. Data collection with only little noise, with taxonomy-like relations in between some of the instances (ca 400 instances, around 20 dimensions)

Further, assign one best-matching clustering algorithm to each data plot in Figure 1. Explain what the advantages and disadvantages of the different algorithms are and why they fit the different kinds of data.

Notes

Your submission in its learning is a **pdf** file with your report. This is a manual "pen and paper kind" of exercise. There's no need for programming. Write down step for step of your solutions (**including distance calculations in all cases that calculations are needed**).