# Association analysis - Lab 2

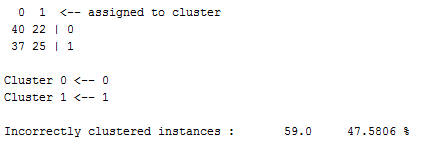
The data set used in the lab, monk1, consists of six different attributes and a class variable. In the first step is two different clustering algorithms tested to examine if the true classes can be found by standard clustering algorithms. Thereafter is association analysis also tested to see if a set of rules can be found that classifies the observations correctly.

The two different clustering algorithms tested are SimpleKMeans and Hierarchical clustering. Two different number of clusters is tested for the SimpleKMeans clustering and two different links are tested for the hierarchical clustering.

## SimpleKMeans

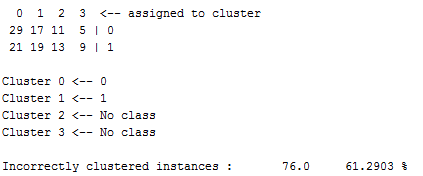
### K = 2 and seed = 10

This clustering performed just marginally better than random guessing with 47.6 % of the observations being misclassified.



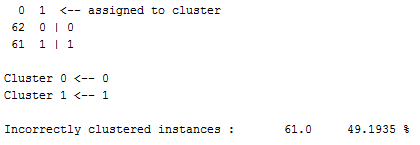
### K = 4 and seed = 10

The clustering with k=4 also performs badly and have clear problems with finding the existing class division. In almost every cluster around half of the observations has *0* as their true class and the other half has class *1* as their true class. The misclassification rate is over 50 % so the classification performed by the clustering is worse than random guessing.

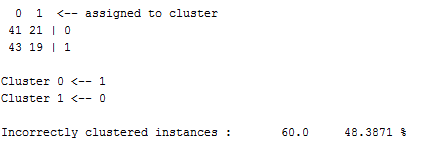


## Hierarchical clustering with k = 2 and single link

The hierarchical clustering with two clusters and single link as distance measure also fails with the objective to find the existing class division. All observations but one are in the same cluster and the misclassification rate is almost 50 %.



## Hierarchical clustering with k = 2 and complete link

The hierarchical clustering with complete link instead of single link as distance measure produces a different result, but with a just slightly lower misclassification rate. In the obtained clusters are almost half of the observations belonging to either of the true classes. Hence, also for this clustering is the conclusion that it not was able to find the existing class division. 

## Association analysis

### Monk's problem

For the target concept associated with the Monk-1 problem either of the following has to be true to perfectly describe class 1:

* *a1=a2*
* *a5=1*

If either of the statements above are true the class is equal to 1.

If either of the statements above not is true the class is equal to 0.

### The rules

The rules that are searched for then is the rules where *a1=a2* or *a5=1*.

Of the 19 rules are then the following rules the ones of interest.

* attribute#5=1 29 ==> class=1 29 conf:(1)
* attribute#1=3 attribute#2=3 17 ==> class=1 17 conf:(1)
* attribute#1=2 attribute#2=2 15 ==> class=1 15 conf:(1)
* attribute#1=1 attribute#2=1 9 ==> class=1 9 conf:(1)

The three rules containing *a1* and *a2* are unique and the three rules together has a support count of 41. The attribute *a5* can be combined with the three other rules, so the support count of the first rule is not unique.

If the rules are applied on the data set, the class 1 will be accurately predicted. Since the class 1 is perfectly descried, so is also class 0 because the ones that not are classified as 1s then has to belong to class 0.

## Why can the clustering algorithms not find a clustering that matches the class division in the database?

The tested clustering algorithms may have troubles to form clusters that are capable to satisfy conditions like *a1=a2*. If *a1=1* and *a2=1* the observation should be classified to class 1, but if *a2=2* it should not. This dependence between the variables *a1* and *a2* explains why the clustering algorithms not did find a clustering that matches the true underlying classification of the observations.

The centroid-based clustering algorithms, like SimpleKMeans, uses the distance from the centroids for clustering the observations. This algorithm does not take the dependence between *a1* and *a2* into consideration. An effect of this is that points who satisfies *a1=a2* will be clustered together with similar points which not necessary is points that also satisfies *a1=a2*.

For the hierarchical clustering performed is the problem similar. The distance between the observations are the measure used for creating the clusters, not the dependency between the attributes. Therefore, also this clustering algorithm fails to find the classes.

The clustering uses all algorithms. Can be seen by the rules found with the association analysis that for perfectly describing class 1 it is enough to use rules that only contains the attributes *a1*, *a2* and *a5*. As was given by the target concept associated with the data set.

## Finally, would you say that the clustering algorithms fail or perform poorly for the monk1 dataset? Why or why not?

Yes, I would say that the clustering algorithms performs poorly for the monk1 dataset. The misclassification rates for the performed clustering's are only slightly better than random guessing.