TDDD41 - Lab 3

Martin Estgren, Alice Reinaudo

March 9, 2017

1 Clustering

For k-Means, we only have the options of Euclidean or Manhattan distance. Both are concepts of distance that are very intuitive for humans, however they seem to fail with the *monk*1 dataset. For both measurements, it turns out that 47.5806% of the instances are incorrectly clustered. When using MakeDensityBasedClusterer, we get 45.9677%, not much of an improvement. For both algorithms we used the default parameters.

When plotting attributes against classification, it can be seen that there is considerable overlap and no well-defined boundary between the two classes. Under these conditions it is understandably quite hard to cluster, since similar points are supposed to lie close to each other in order for the algorithms to succeed.

2 Association analysis

Association analysis does not require points that have the same classification to be close together. Therefore it may be easier to find rules that correctly classify points, so long as there is a pattern to be found.

We remove those rules where the antecedent is a superset of the antecedent of another rule, because then it is redundant. This is not the case in general but only when the more general rules have 100% confidence. This is because if the confidence was lower, then the rules with a superset-antecedent would provide additional information and therefore it would not be wise to discard them. For cluster 1, we end up with the following rules.

Table 1: Association rules

Parameters	Cluster	Occurences	Confidience
attribute 5 = 1	1	29	1
attribute 1 = 3 attribute 2 = 3	1	17	1
attribute 1 = 2 attribute 2 = 2	1	15	1
attribute 1 = 1 attribute 2 = 1	1	9	1

These rules then encompass both clusters since everything that does not belong to cluster 1 must belong to cluster 0.