Homework #4: k-Means Clustering
Jana Obšteter, **jana.obsteter@kis.si**

We have run classification tree, logistic regression and k-mean clustering on gene expression data GDS4168.

1. The classification tree works very well on the data with AUC of 0.988 and classification accuracy of 0.981 (even a "stump" tree with maximum depth of 1). The reason for its success is it finds an informative feature (gene ASAP1) and classifies upon that – it cancels out all the others that represent noise for classification. Also logistic regression performs well with AUC of 0.925 and classification accuracy of 0.981. This means that the data can be separated with a single hyperplane.

2. k-means clustering does identify two clusters, but assigns erroneous membership to the data points. Cluster 1 contains 31 cases and 9 controls and cluster 2 contains 20 cases and 2 controls. Also, the silhouette score for the clustering with 2 clusters is low, i.e. 0.211, although still the best one of the 2 – 8 clusters. This means that the separation of clusters is not good, i.e. there is not a big distinction between matching the data to its own cluster and the neighbouring cluster.

3. *If not: what does it tell you about the data? How can you have excellent predictive models but clustering is not able to re-discover the two classes as subgroups in the data?*
k-means makes some assumptions about the shape and the distribution of the data. If the k-means fail to correctly assign the classes via clustering, it means, that the assumptions are broken i.e. data does not follow the assumed shape and distribution. However, some of the predictive model might still work well, since they do no make the same assumptions and are also able to learn from the data.

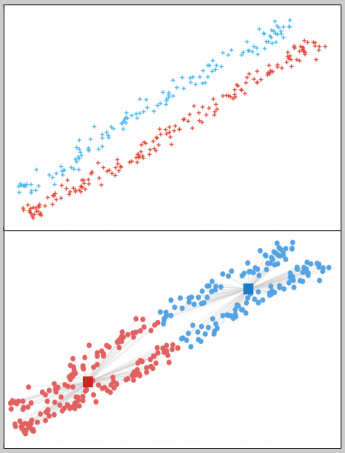

Figure 1: Data classes (above) and cluster membership (below). Example of non-spherical classes.

The first assumption of the k-means is the assumption of the spherical clusters. If the clusters are non-spherical, e.g. disk-shaped, and near to each other, k-means will fail to correctly identify the clusters. However, if the clusters are non-spherical but far apart, the k-means might still succeed. Therefore, one possible explanation for this specific example is that the classes are not spherical and are near to each other. k-means tries to minimise sum of squared error (or the variance) within the clusters and maximise the one between the clusters. So if the distances within the classes are larger than the distances between the classes (variance within > variance between) the k-means clustering will fail in correctly assigning the classes. An examples is shown in Figure 1. The logistic regression would work well in this situation, since the classes can be separated with a hyperplane (AUC = 1). All the cases / controls and above / below a certain threshold. The classes are not spherical and near – consequently k-means fails.

Further more, k-means clustering makes another assumptions of the roughly equal sizes of the clusters. If the classes have uneven number of instances, such as in this dataset (41 cases and 11 controls), k-means might fail. An example of such is shown in Figure 2. Again, logistic regression would perform well since the data can be separated with a hyperplane (AUC = 1).

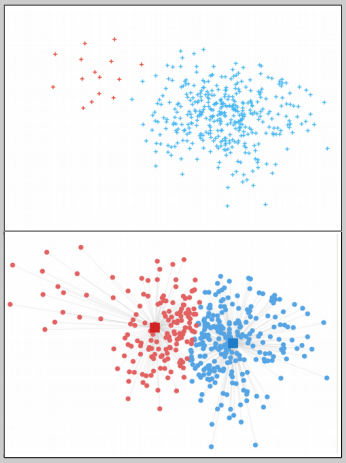

Figure 2: Data classes (above) and cluster membership (below). Example of uneven class sizes.

Moreover, k-means clustering is unable to cancel out the noise. All the attributes are equally important. If the data contains a lot of noise (uninformative attributes), k-means will perform poorly. This is where trees work well, since they can identify the most informative attribute to split the data through the learning process.

The clustering problems stated above get even worse in higher dimensions (curse of dimensionality). As already mentioned, other predictive models are able to deal with these problems since they are supervised and can learn from the data – e.g. trees can identify informative attributes and logistic regression can identify a hyperplane to split the data and give higher probabilities to more informative attributes. k-means is unsupervised algorithm therefore does not learn from the data. However, each algorithm makes its own assumptions and we must know them and the limitation when choosing the appropriate model for our data.