Laboratory 6 – Clustering

1. Classification vs. Clustering:
   a. Classification – Supervised learning. Learns a metod from prelabeled data.
   b. Clustering – Unsupervised learning. Finds natural grouping of instances given unlabeled data. The process of grouping physical or abstract objects into classes of similar objects.
2. A good clustering method will produce high quality clusters, which are characterized by high intra-class similarity and low inter-class similarity.
3. Similarity is expressed in terms of a distance function, which is typically metric d(i,j). There exist different measures for interval-scaled, Boolean, categorical, ordinal and ratio variables.
4. The most commonly used distance measures is Minkowski distance $L = \sqrt[q]{\sum_{i=1}^{k}(|x_i - y_i|)^q}$, especially:
   a. L_1 norm - Manhattan distance $L_1 = \sum_{i=1}^{k} |x_i - y_i|$
   b. L_2 norm - Euclidean distance $L_2 = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$
   c. L_∞ norm - Czebyszew norm $L_\infty = max_{i=1}^{k}|x_i - y_i|$
5. If the values of attributes are in different units then it is likely that some of them will take very large values and hence the distance between two cases on this variable can be a big number. Other attributes may take small values – in this case the distance will be smaller. The attributes with high variability or range will dominate the distance metric. To overcome this problem standardization or normalization of values is performed.
6. There exist many clustering algorithms, which can be divided with respect to different features. The main categories of clustering methods are:
   a. Partitioning algorithms – construct various partitions and then evaluate them by some criterion. Example of algorithms: k-means, k-medoids, PAM.
   b. Hierarchical algorithms – create a hierarchical decomposition of the set of data (or objects) using some criterion. Example of algorithms: AHC.
7. K-means:
   a. Pick a number of cluster centers – centroids.
   b. Assign every item to its nearest cluster center.
   c. Move each cluster center to the mean of its assigned items.
   d. Repeat steps b and c until centroids stop moving or change is less than defined threshold.
8. AHC builds a tree-based taxonomy (dendrogram) from the set of unlabeled instances. It is a bottom up (agglomerative) approach. It starts with all instances in their own cluster. Until there is only one cluster determine two clusters that are the most similar and merge them into a one cluster. The distance between clusters can be calculated using different approaches:
   a. Single linkage – minimum distance
   b. Complete linkage – maximum distance
   c. Mean distance
   d. Average distance
9. Dendrogram shows how the clusters were merged (bottom up) or decomposed (top down). Clustering of the data objects is obtained by cutting the dendrogram at the desired level. In this case each connected component forms a cluster.