Hierarchical Clustering

Mads Møller

LinkedIn: https://www.linkedin.com/in/madsmoeller1/

Mail: mads.moeller@outlook.com

This paper is the sixth in the series about machine learning algorithms. The Hierarchical Clustering algorithm is used for *clustering* problems. Hierarchical Clustering is also our first unsupervised machine learning algorithm, meaning that our data do not need to be labelled. So you do not need to know your target classes in order to use the Hierarchical Clustering algorithm.

1 Intuition

In many ways clustering can relate to the task of nearest-neighbor algorithms. We would like to group observations that are similar given a dissimilarity. Each observation is in a group and different groups do not overlap. A group is often in unsupervised learning referred to as a **cluster**. Within each group we have *inter homogeneity*, meaning that observations within a cluster should be similar. We call the collection of clusters a *clustering*. Within clustering we have *inter heterogeneity*, meaning that clusters should be dissimilar to other clusters.

Often clustering methods are used for *customer segmentation*, *grouping of products*, *recommender systems* (segmentation of customers and products). So indeed clustering methods are useful in the real-world as well.

2 Hierarchical Clustering

There are different methods of clustering. In this paper we will inspect the clustering type called Hierarchical Clustering. In the next paper we will look at *Partitioning Clustering*.

2.1 Dissimilarities (between observations)

The basic way of measuring a dissimilarity is through a distance. We already looked into dissimilarities in the paper about K Nearest Neighbors (KNN). Therefore, we will not go too much into dissimilarities again, but some of the ways you could measure dissimilarities will be listed here:

Euclidean distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2} \tag{1}$$

Manhattan distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ik} - x_{jk}| \tag{2}$$

Maximum distance:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \max\{|x_{i1} - x_{j1}|, |x_{i2} - x_{j2}|, \ldots, |x_{ik} - x_{jk}|\} \tag{3}$$

The general formulation of l_p-distance (**Minkowski**):

$$d(\mathbf{x}_i, \mathbf{x}_j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \ldots + |x_{ik} - x_{jk}|^p)^{1/p} \tag{4}$$

Hamming distance (categorical variables):

$$d(\mathbf{x}_i, \mathbf{x}_j) = \text{cardinality}\{s : x_{is} \neq x_{js}\} \tag{5}$$

2.2 Dissimilarities (between clusters)

Now that we want to do clustering we also need to be able to calculate the distance between two clusters. Like there are different methods to calculate the distance between two observations there are also different methods to calculate the distance between two sets of observations (clusters). In clustering the distance between two clusters are also called the **linkage** between two sets. Let us define two clusters as S_1 and S_2. We will look into three different methods.

Average Linkage:

$$L(S_1, S_2) = \frac{1}{|S_1| \times |S_2|} \sum_{x_1 \in S_1; x_2 \in S_2} d(x_1, x_2)$$

This is the average linkage or average dissimilarity. The equation calculates the distance between each observation in each cluster $x_1 \in S_1$ and $x_2 \in S_2$ and then divides with the number of combinations $|S_1| \times |S_2|$. Therefore we get the average distance between observations in the two clusters.

Single Linkage:

$$L(S_1, S_2) = \min_{x_1 \in S_1; x_2 \in S_2} d(x_1, x_2)$$

In the single linkage formulation we are only interested in the distance between the two closets observations between the two clusters. So we calculate each distance combination between observations in each cluster and then only chooses the smallest.

Complete Linkage:

$$L(S_1, S_2) = \max_{x_1 \in S_1; x_2 \in S_2} d(x_1, x_2)$$

In the complete linkage formulation we are only interested in the distance between the two observations that are has the greatest distance between the two clusters.

2.3 Hierarchical Clustering

In Hierarchical Clustering we have all the observations together in the beginning (at the top of figure 1). In the end we have each observation into a cluster by itself (in the bottom of figure 1). We can cluster observations in this way by making horizontal cuts. Different horizontal cuts give different clustering.

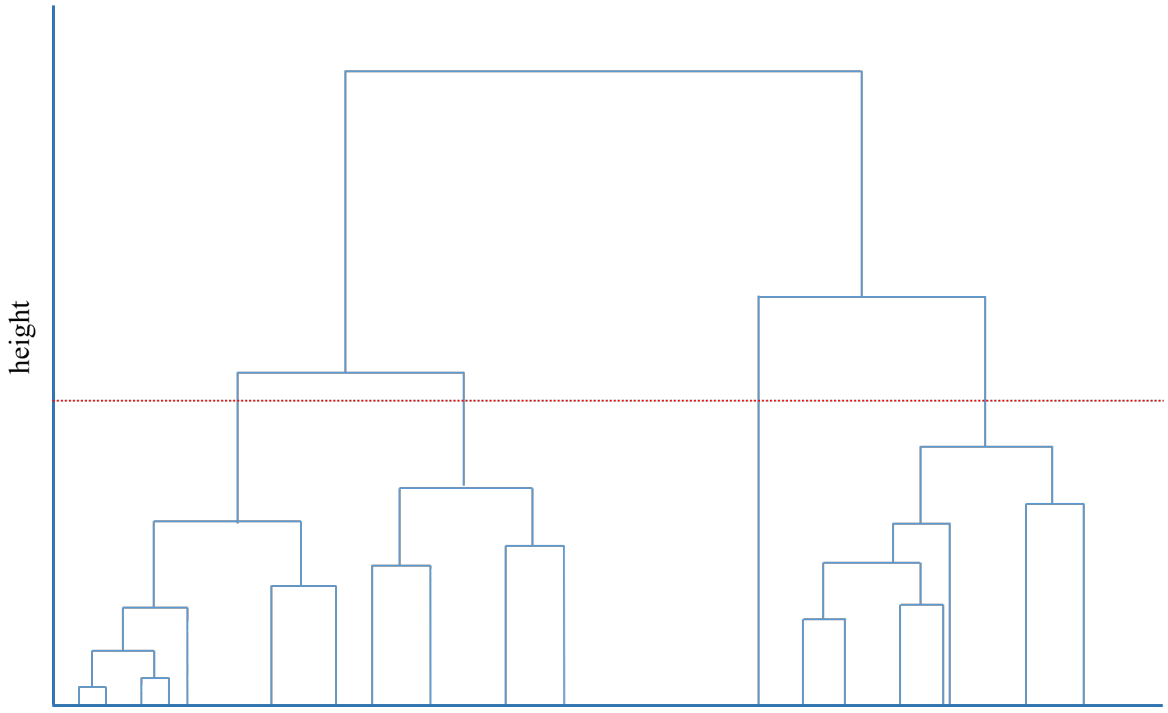

Figure 1: Cluster Dendrogram

If we inspect the red line on the **cluster dendrogram** at figure 1. It symbolizes a horizontal cut. If we made this cut our observations would be clustered into four categories (four branches at the cut). Therefore the number of clusters are highly dependent of the cut-off value for the horizontal cut. Any Hierarchical Clustering software should be able to make a cluster dendrogram. The *height* parameter of the cluster dendrogram is a parameter who says something about the similarity between clusters. Hierarchical Clustering is often appealing for users, since it is visual (like decision trees). We build a whole tree of partitions.

2.4 Hyperparameters

Like every other machine learning algorithm we have hyperparameters to tune. Both the linkage method and the distance method to measure similarity are hyperparameters. The best way is to investigate different combinations and see if the clusters makes sense with the observations within it.