Changelog: 12 Dec 2016

# Advantages & Disadvantages of k-Means and Hierarchical clustering (Unsupervised Learning)

Machine Learning for Language Technology
ML4LT (2016)

*Marina Santini*
Department of Linguistics and Philology
Uppsala University

# Outline

- k-Means: Advantages and Disadvantages
- Hierarchical Clustering: Advantages and Disadvantages

# k-Means: Advantages and Disadvantages

**Advantages**

- Easy to implement
- With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
- k-Means may produce tighter clusters than hierarchical clustering
- An instance can change cluster (move to another cluster) when the centroids are re-computed.

**Disavantages**

- Difficult to predict the number of clusters (K-Value)
- Initial seeds have a strong impact on the final results
- The order of the data has an impact on the final results
- Sensitive to scale: rescaling your datasets (normalization or standardization) will completely change results. While this itself is not bad, not realizing that *you have to spend extra attention to scaling your data* might be bad.

# Hierarchical Clustering: Advantages and Disadvantages

**Advantages**

- *Hierarchical clustering* outputs a hierarchy, ie a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram (see suggestion on how to cut a dendrogram in lab8).
- Easy to implement

**Disavantages**

- It is not possible to undo the previous step: once the instances have been assigned to a cluster, they can no longer be moved around.
- Time complexity: not suitable for large datasets
- Initial seeds have a strong impact on the final results
- The order of the data has an impact on the final results
- Very sensitive to outliers

# The end