Summary on the topic of Maximization of Mutual Information for Information Clustering

Tilman Hinnerichs

February 13, 2020

1 Clustering

Clustering algorithms are a group of algorithms for the exploration of similarity structures among datasets. It aims on identifying groups within the data, that are somehow similar to each other. As no additional data is used, these are classified as unsupervised learning methods. Crucially, these so called clusters do not have any label, are only based on the underlying information of each data point. In general, whether a clustering is good or bad can only be determined by an expert, or ground truth. Alternatively, quality of clusters can be measured by various dataset independent measures, e.g., average radius of the clusters and distance between different clusters. The various algorithms differ in their interpretation of similarity.

2 k-means-clustering

k-means is one of the simplest and widely known clustering algorithms and shall thus be explained here briefly. k hereby defines the number of clusters.
The steps are the following:

1. Randomly initialize k points as initial means
2. Assign each data point to the closest of the k means
3. For each of the sets of data points assigned to one of the k means, calculate the mean among each group of assignments

Repeat steps 2 and 3 until the means are not changing anymore or within a certain margin.