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.

3 Discussed papers

When trying to find a suitable representation, we are trying to have as much information from the input within the output, i.e. loose as little information from the input as possible. Thus, the *transinformation* or *mutual information* (MI) has to be maximized also known as the infomax principle (Linsker, 1988). Hence, making MI the objective function for representation learning seems to be a suitable approach. However, mutual information is generally quite difficult to compute, especially in continuous and high dimensional settings.
Let $\mathcal{X}, \mathcal{Y}$ be the domain and the range of a continuous and (almost everywhere) differentiable parametric function $E_\psi : \mathcal{X} \to \mathcal{Y}$ with parameters ψ. Additionally, let $\mathcal{E}_\Psi = \{E_\psi\}_{\psi \in \Psi}$ over Ψ.
Find the set of parameters ψ, such that the mutual information, $\mathcal{I}(X; E_\psi(X))$ is maximized. Ideally the marginal distribution induced by forwarding samples from the empirical distribution for a given ψ should match a prior distribution, thus encouraging the output of the encoder to have desired characteristics, e.g., independence.

3.1 Learning Deep Representations by mututal Inforamtion Estimation and Maximization – Deep InfoMax (DIM)

- incorporate knowledge about locality in the input significantly improves performance in downstream tasks
- DIM simultaneously estimates and maximiizes the mutual inforamtion between input data and learned high-level representations

3.2 On Mutual Information Maximization for Representation Learning

3.3 Invariant Information Clustering