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

Tilman Hinnerichs

February 14, 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 r adius 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.

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

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.

- maximizes information between spatially-preserved features and compact features
- incorporate knowledge about locality in the input significantly improves performance in downstream tasks
- DIM simultaneously estimates and maximizes the mutual information between input data and learned high-level representations

The authors present various techniques to estimate the MI, eventually deciding for the Noise-Contrastive Estimation (infoNCE).

Deep Infomax consists of three parts

1. global, and
2. local MI maximization, and
3. prior matching

→ Formula for goal and MI

Just use Figure 2 for explanation of the setup.
Setup:

- Convolution to obtain best global feature
- First global discriminator tries to distinguish between global + $M \times M$ matrix fake/real
- Second local discriminator tries to distinguish between local + $M \times M$ matrix fake/real for each segment each!!
- Third discriminator was trained to estimate the divergence from a prior distribution with some property

Drawbacks:

- Only applicable on images
- Needs preprocessing creating the images features
- needs k-means after representation calculation -¿ prone to degeneration
- calculates over continuous random variables, which requires complex estimators
- DeepInfomax estimator reduces to entropy

3.2 On Mutual Information Maximization for Representation Learning

3.3 Invariant Information Clustering

Benefits:

- Only reliant on the information itself
- Range is discrete → exact calculation of MI

Two issues:

- Degeneracy of clustering → e.g. as with k-means
- Noisy data with unknown or distractor classes (e.g. STL10)

Content:

- semantic clustering vs. representation learning
- theoretical foundations of IIC in statistical learning
- Set state of the art performance on various dataset

Let $x, x' \in \mathcal{X}$ be paired data sample from a joint probability distribution $P(x, x')$, e.g., two images containing the same image. We now want to maximize the mutual information between representation of similar entities.
Goal is to make representations of paired samples the same, which is not the same as minimizing representation distance.
Representation space: $\mathcal{Y} = \{1, \dots, C\}$
Introduce notation: $P(z = c|x) = E_{\psi,c} \in [0, 1]^C$
Now consider a pair of such cluster assignment variables z, z' for two inputs x, x'. Their conditional joint distribution is given by $P(z = c, z' = c'|x, x') = E_{\psi,c}(x) \cdot E_{\psi,c'}(x)$ (This states that z, z' are independent given inputs x, x', which is generally not the case)
Build matrix $\mathbf{P}$
Why degenerate solutions are avoided: $I(z, z') = H(z) - H(z|z')$ ($H(z)$ maximized when all equally likely to happen, $H(z|z')$ is 0 when assignments are exactly predictable from each other. This encourages deterministic one-hot predictions
For image clustering no labels available. Thus, objective becomes: $max_\psi \mathcal{I}(E_\psi(X), E_\psi(gx)))$