Project 2: Structure in compressed spaces (Unsupervised learning

January 24, 2019

1 Description

Clustering tends to operate nicely if there is clear structure in the data - which however is both hard to define and hard to discover. In this project we will use neural networks to learn good features for clustering by training autoencoders.

We will try a novel approach based on a a novel autoencoder structure. We will train an end-to-end autoencoder, with a feature layer that uses softmax neurons, followed by normal reconstruction layers; the goal here is to force the network to learn a representation of the data that forces it to "bin" data into different groups. We will take the maximum of the feature vector as the potential cluster assignment.

2 Tasks

1. Select 3 different reasonably-sized datasets from UCB archive, Kaggle, etc., load and inspect them
2. Cluster all of them using standard methods (e.g., pick some of the methods from [3]) and evaluate (for example using the completeness Score [4]).
3. Train an auto-encoder on your data [2] and use the autoencoder features as inputs to clustering algorithms.
4. Train the same auto-encoders, but now change the middle layer to use softmax features.
5. Evaluate your method using different cluster sizes and regularisation strengths (you are free to experiment here).

3 References

1. Xie, Junyuan, Ross Girshick, and Ali Farhadi. "Unsupervised deep embedding for clustering analysis." International conference on machine learning. 2016..
2. Building autoencoders in Keras
3. Scikit-learn clustering
4. Completeness score

4 Dataset examples

1. https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones
2. https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008
3. https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions