Unsupervised-Learning

Guy Van-Dam

January 2021

1 Abstract

2 Introduction

When building a model, we need to experiment with different Machine-Learning methods, as different methods are built for different tasks. When we do not have a lot of information about our data, this step is crucial if not necessary for building an effective and efficient model. Trying different algorithms and analyzing and comparing their performances against each other will guide us towards building the right model for the task at hand.

When is comes to Unsupervised Machine-Learning, where we rarely have external labels, and we do not have a lot, if any information about our data. we have to try different clustering algorithms and compare their result for classifying our data-set in the most accurate way.

3 Methods

All of the code for this **paper** was written in *Python* using external libraries for faster and more accurate calculations and for visualization of the results. These external libraries include *Matplotlib* for plotting and visualizing results and clusters, *Numpy* for fast mathematical calculations. *Pandas* for reading and manipulating the data-sets. *Scikit-Learn* for the clustering algorithms implementations, normalization and dimension reduction and *Scipy* for the statistical tests. Due to memory and run-time constraints, data-sets 2 and 3 needed to be sub-sampled to 14000 points, While data-set 1 could stay whole at about 12000 points.

Many of the functions I used are randomly initialized or have a **certain randomness in them**, so a fixed random state was required for reproducibility. I used 42 for the random seed.

Our data-sets were all above 3 dimensions, so a dimension reduction algorithm needed to be used. I used the Principal Component Analysis, or PCA algorithm with a fixed random seed, **as it does not assume a lot about that data, is fast and intuitive.** PCA needs the input data to be normalized. I normalized the data using the *Standard Scalar* class and *Normalize* method from *Scikit-Learn*.

The data was passed through the PCA algorithm, before clustering for faster computation. reducing the data size to an average of 10%.

I ran the PCA algorithm with 2 principal components and my predetermined random seed. As do not look at the algorithm performance.

The clustering algorithms ran with the same parameters for all of the datasets. The clustering algorithms I used were:

- The individual entries are indicated with a black dot, a so-called bullet.
- The text in the entries may be of any length.

4 Results

5 Discussion

6 References