← Go Back to Making Sense of Unstructured Data

Course Content

Dimensionality Reduction (PCA & tSNE)

In real-world situations, we often deal with **high-dimensional** data, with lots of columns or "features" that represent the information collected about each observation. High-dimensional data is disadvantageous for a couple of reasons:

- It is generally difficult to analyze or visualize high-dimensional data and identify hidden patterns.
- Not all the features or dimensions of the data are equally important.

Therefore, we need to reduce the dimensionality of the dataset in such a way that by losing only a minimal amount of information, we can visualize the data and identify patterns more easily with a smaller number of features.

Two important techniques that we can use for dimensionality reduction are:

- PCA
- t-SNE

PCA

The main idea of principal component analysis (PCA) is to reduce the dimensionality of a dataset consisting of many variables correlated with each other, either heavily or lightly, while retaining the variation present in the dataset.

This is done by transforming the variables to a new coordinate space of variables, which are known as principal components (or simply, the PCs), and are orthogonal to each other.

The selection of principal components is such that the retention of variation present in the original variables is the maximum for the first principal component and decreases as we move down in order. The principal components are the eigenvectors of the covariance matrix, and hence they are orthogonal.

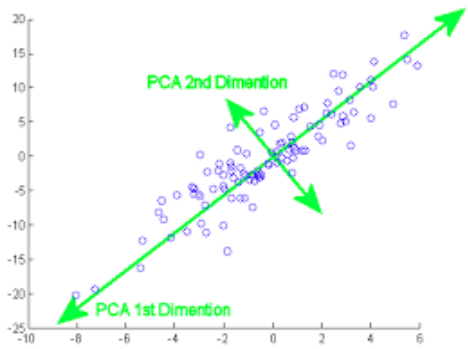

Image Source (Links to an external site.)Links to an external site.

t-SNE

The t-SNE algorithm calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space. This is done by setting the probabilities from the low-dimensional space similar to those of the high-dimensional space. We measure the difference between the probability distributions of the two-dimensional spaces using Kullback-Leibler divergence and try to optimize it.

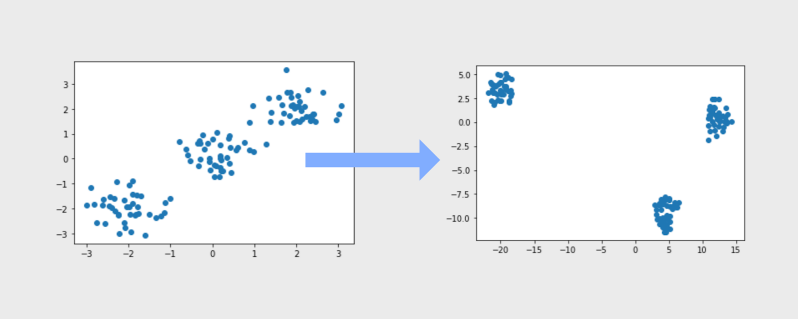

Image Source (Links to an external site.)Links to an external site.

Previous

Next