Development of unsupervised learning transformations through supervised learning methods.

Author: Patricia Cortajarena Sauca

Ponente: Carlos Roberto del Blanco Adán

Tutor: Pedro Morales

Trabajo Fin de Grado

ETSIT UPM

Madrid. January, 2018

Abstract

The aim of this project is

Acknowledgements

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Working with large datasets and high-dimensional data in nowadays' problems has encouraged the use of dimensionality reduction algorithms which try to preserve as much information as possible even decreasing the number of features needed to describe that same dataset. Thus, time and memory in huge implementations can be saved.
Taking into account that this turns into a difficult task, we can find that numerous approaches have been proposed.
Although they look forward to achieve more or less the same performance, they differ from one another and we can not reassure which would suite for a specific problem or even if the behaviour of the algorithm is going to reach the results we expected or needed.
The first point to take into account is the existence of parametric and non parametric algorithms, and secondly, in both of them we can find different models proposed depending on what to optimize, yet not everything is going to be preserved as well as in the original dataset, so we need to prioritize some aspects.
So our decision of which to implement depends on the previous study of our data, the performance requirements and the later purpose and usage of the reduced data.

We propose the research and then base our study in the next dimensionality reduction algorithms:

- PCA (Principal Component Analysis)
- MDS (Multidimensional Scaling)
- TSNE (T-Stochastic Neighbour Embedding)

1.1 PCA: Principal Component Analysis

Principal Component Analysis algorithm is based on reducing the number of features by processing the correlations between the features of the datapoints. The aim is to eliminate this correlations by transforming the matrix $\mathbf{X} \in \Re^{mxn}$ (with m being the number of data points and n de number of features) into an orthogonal basis. By omiting the correlation between columns of the matrix $\mathbf{X}$ we are capable of doing away with redundancies.
The model starts by computing de covariance matrix, which results in a $\Re^{nxn}$ symmetric matrix. We obtain it by using the next expresion:

$$\text{cov}(\mathbf{X}) = \frac{1}{m-1}\,\mathbf{X}^T\mathbf{X}$$

Because the aim of the PCA is to eliminate the correlations, the covariance matrix of the result $\mathbf{Y}$ should be a diagonal matrix with just the variances of the columns.
PCA is famous because of a great advantage: we can find a linear transformation ($\mathbf{Y = XP}$), which makes this a parametric model, easy to reuse and quite computationally simple because of some covariance matrix calculation approaches.
For symmetric matrices ($\mathbf{X}$) we can find eigenvalue decomposition with a diagonal matrix ($\mathbf{Y}$), matching exactly with our linear problem with $\mathbf{X}$ and $\mathbf{Y}$.

$$\mathbf{Y = XP}$$
$$\text{cov}(\mathbf{Y}) = \tfrac{1}{m-1}\,\mathbf{Y}^T\mathbf{Y} = \tfrac{1}{m-1}\,(\mathbf{XP})^T\mathbf{XP} = \mathbf{P}^T\text{cov}(\mathbf{X})\mathbf{P}$$
$$\mathbf{D} = \mathbf{V}^T\mathbf{AV}$$
$$\mathbf{A} = \text{cov}(\mathbf{X});\ \mathbf{P} = \mathbf{V}^T;\ \mathbf{D} = \text{cov}(\mathbf{Y})$$

With the previous expresions we get to the point that computing the eigenvectors of the covariance matrix $\mathbf{X}$ we can get a linear transformation from space $\mathbf{X}$ to space $\mathbf{Y}$. The eigenvalues matrix obtained ($\text{cov}(\mathbf{Y})$) sorted decreasingly would be the orthogonal basis values. Choosing the $\mathbf{N}$ first values of this matrix, being $\mathbf{N}$ the desired output dimension, and computing the product between this $\mathbf{N}$ eigenvalues and our datapoints, we would obtain our reduced dimensionally points.

1.2 MDS: Multidimensional Scaling

Bibliography

Code