Dimensionality Reduction

Introduction

In many applications of data science it is often the case that we would like to represent objects in a shared space in such a way as to be able to realize their differences and highlight their similarities. This type of an representation is often required in instances where the original representations of the data points that we are comparing doesn't provide sufficient information to make the distinctions and/or the representation is highly dimensional and therefore difficult to work with (e.g. comparing images in their original binary representation). In addition, on many data science tasks, such as classification and regression, we are also faced with the dilemma on the types of features that we would like to select to represent the data in order to achieve better results. One approach for solving these problems is to use dimensionality reduction techniques which perform mapping from the original representation space to a lower dimensional space by automatically selecting and extracting features using a particular transformation approach.

Project Description

In this project we are going to learn and explore several dimensionality reduction techniques such as Principle Component Analysis (PCA), Latent Semantic Analysis (LSA), both of which are based on Singular Value Decomposition (SVD). We'll also learn about Independent Component Analysis (ICA), Neighborhood Component Analysis (NCA) and Linear Discriminant Analysis (LDA). Using the data representations of the different dimensionality reduction approaches we'll perform exploratory data analysis over a particular data collection. Doing so we'll perform data visualization of the low dimensional projects. In addition we'll use the generated representations of the data points as features in order to perform classification using a single layer neural network. Finally we'll analyze the classification performances across the different dimensionality reduction techniques in terms of the size of the low-dimensional space.

Datasets Used

We encourage students to suggest datasets where dimensionality reduction techniques may be useful on a particular task or as means to perform exploratory data analysis. We'll provide two types of data collections:

1. Olivetti Faces. This is a collection of 400 grayscale face images.
2. MNIST Handwritten Digits. A collection of handwritten digits.