The Wang Yanan Institute for Studies in Economics (WISE)

Dimensional Reduction

Yufeng Gu & Huajin Yu

April 20, 2019

Abstract

In this assignment, we talk about high-dimensional data. When the dimension becomes increasing high, there will be a problem called dimensionality disaster. In order to solve this kind of problem, we need to reduce the dimension of data to a relatively low one.

1 High-Dimensional Data

High-dimensional data means that the dimension of the data is very large, even much larger than the number of samples. The obvious performance of high-dimensional data is that the data is very sparse in space and the sample size is always very small compared with the dimension of space.

The biggest problem encountered in the analysis of high-dimensional data is the expansion of dimension, which is usually called the "dimension disaster" problem. The results show that with the increase of dimension, the number of spatial samples will increase exponentially.

2 Dimensional Disaster

As shown below, when the dimension of the data space increases from 1 to 3, the most obvious change is the increase in the sample requirement; in other words, when the sample size is determined, the sample density will decrease and the sample will become sparse.

Assuming that the sample size is $n = 12$ and the width of a single dimension is 3, then in one-dimensional space, the sample density is $12/3 = 4$, and in two-dimensional space, the size of the sample distribution space is 3×3, then the sample density is $12/9 = 1.33$, In three-dimensional space, the sample density is $12/27 = 0.44$.

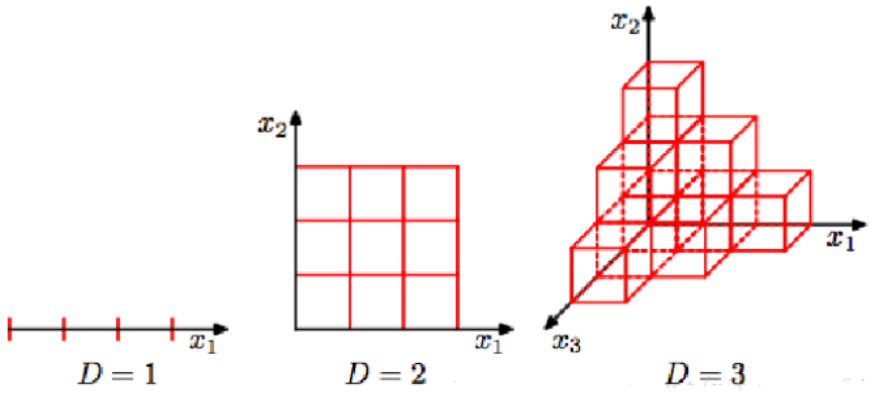

Figure 1: Expansion.

Imagine, when the data space is a higher dimension, $X = [x_1, x_2, .., x_n]$, what will happen?

1. More samples are needed, and the sample increases exponentially with the increase of the dimension of the data;
2. Data becomes more sparse, leading to data disaster;
3. In high-dimensional data spaces, prediction will no longer be easy;
4. Cause model over-fitting.

3 Dimensional Reduction

The dimensional reduction in the field of machine learning refers to mapping the data points in the original high-dimensional space to the low-dimensional space by some

mapping method. The essence of dimensional reduction is to learn a mapping function $f : x \to y$, where x is the representation of the original data points. y is a low-dimensional vector representation after data point mapping.

For high-dimensional data, the solution to the over-fitting problem caused by dimension disaster is:

1. increasing sample size;
2. reducing sample characteristics.

As the lecture shows, lasso is one the method to reduce the dimension of data. Next we will introduce some other methods to solve this problem.

4 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is the most commonly used linear dimensional reduction method. Its goal is to map the high dimensional data to the low dimensional space by some kind of linear projection, and expect the maximum variance of the data in the projected dimension. In this way, fewer data dimensions are used while retaining the properties of more original data points.

Let n-dimensional vector w be an axis direction (called mapping vector) of the target subspace, maximizing the variance after data mapping, such as:

$$\max_{w} \frac{1}{m-1} \sum_{i=1}^{m} (w^T(x_i - \bar{x}))^2 \tag{1}$$

Where m is the number of data instances, x_i is the vector representation of data instance i, and $\bar{x}$ is the average vector of all data instances. w is defined as a matrix containing all mapping vectors as column vectors. Through linear algebraic transformation, the following optimization objective functions can be obtained:

$$\min_{W} tr(W^T A W), \ s.t. W^T W = I \tag{2}$$

$$A = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})(x_i - \bar{x})^T \tag{3}$$

Where tr represents the trace of the matrix and A is the data covariance matrix. The

optimal W is composed of the eigenvectors corresponding to the first k largest eigenvalues of the data covariance matrix as column vectors. These eigenvectors form a set of orthogonal bases and best retain the information in the data. The output of PCA is $Y = W'X$, which is reduced from the original dimension of X to the k dimension.

PCA seeks to maximize the intrinsic information of the data after dimensional reduction and to measure the importance of that direction by measuring the size of the variance in the projection direction. However, this projection does not distinguish the data greatly, even it may make the data points indistinguishable from each other. This is also the biggest problem with PCA, which leads to poor classification using PCA in many cases. As shown in the following figure, when you project a data point onto a one-dimensional space using PCA, PCA selects a axis-2, which makes it impossible for the two clusters to be distinguished. If you select the axis-1, you will get a good distinction.

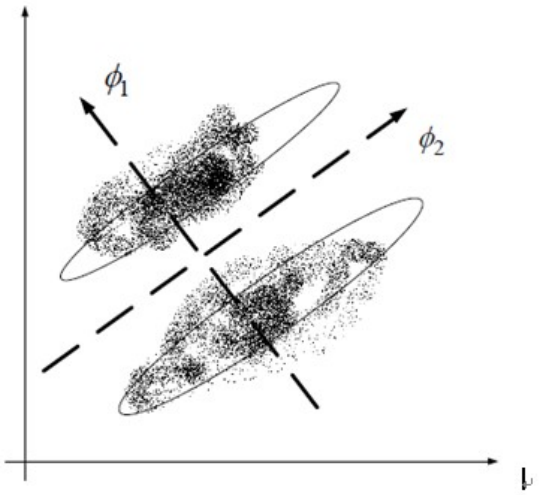

Figure 2: PCA Selection.

5 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (also called Fisher Linear Discriminant) is one kind of supervised linear dimensional reduction algorithm. Unlike PCA, which maintains data information, LDA is designed to make reduced data points as easily distinguishable as possible.

Suppose the raw data is represented as X (a $m \times n$ matrix, m is the dimension, n is the number of sample). Since it is linear, it is hoped that the mapping vector a will be

found so that the data points after $a'X$ can maintain the following two properties:

1. Data points of the same kind are as close as possible (within class);
2. Data points of different classes are as separate as possible (between class).

6 Summary

PCA seeks to maximize the intrinsic information of the data after dimensional reduction and to measure the importance of that direction by measuring the size of the variance in the projection direction.

While the core idea of LDA is to project to the normal vector of the linear discriminant hyperplane.The purpose of LDA is to make the reduced data points as easily to distinguish as possible. These two are the most commonly used linear dimensional reduction methods.

Reference

[1] CSDN, "A detailed explanation of dimensional reduction method in Machine Learning."

[2] CSDN, "Lasso regression."

[3] Osborne, Michael R., Brett Presnell, and Berwin A. Turlach. "On the lasso and its dual." *Journal of Computational and Graphical statistics* 9.2 (2000): 319-337.

[4] Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996): 267-288.