Data Mining - Statistics and Dimensionality Reduction

Nhat-Quang Doan

University of Science and Technology of Hanoi

nq.doan@gmail.com

Today Objectives

- Review basic statistics
- Learn two approaches PCA and SVD for the dimensionality reduction
- Apply these techniques in real-world problems

Dataset

Data in a DM problem

- Information, data are samples consisted of attributes
- Given that a set of n samples, $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^d\}, \forall i = 1, .., n$.

$$
X = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix} = \begin{array}{c} \begin{array}{cccc} \mathbf{f}_1 & \mathbf{f}_2 & \cdots & \mathbf{f}_d \end{array} \\ \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,d} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,d} \end{pmatrix} \end{array}
$$

Dataset

Sample and Attribute

- An attribute (or variable) is a specification that defines a property of an object.
- A sample describes an object with attributes. Synonymes: point, vector (often in $\mathbb{R}^d$, $\mathbf{x}_i \in \mathbb{R}^d$ with $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,d})$ where $x_{i,j}$ is an attribute)

$$\mathbf{x}_i = \begin{pmatrix} x_{i,1} & x_{i,2} & \cdots & x_{i,d} \end{pmatrix}$$

Dataset

Questions

Learning accuracy depends on the data!

- Is data relevance?
- What is the data amount?
- What is the data quality?
  - Noise
  - Missing data
  - etc.
- Could we visualize data? any proper representation?
  $\rightarrow$ Data visualization
- How much of the data is labeled vs unlabeled?
- Is the number of features/attributes reasonable?

Dataset

Questions

Is data relevant?

- Almost all instances have the same value (no information)
- Almost all instances have unique values
- The feature is highly correlated with another feature
  $\rightarrow$ Statistical analysis

Dataset

While we are gathering data

Data availability

- More the better (in terms of number of instances, not necessarily in terms of number of dimensions/features)
- The more features you have the more data you need

$\rightarrow$ Data augmentation for creating new synthesis data

If data label is not available

- Could set up studies/experts to label data
- Use unsupervised and semi-supervised techniques

Basic Maththematics

How good is the data?

- Mean

$$\bar{\mathbf{x}} = \frac{1}{n}\sum \mathbf{x}$$

- Variance measures how far a set of (random) numbers are spread out from their means

$$var(X) = \sigma^2 = \frac{\sum(\bar{\mathbf{x}} - \mathbf{x})^2}{n}$$

- σ is the standard deviation of $\mathbf{x}$ (a vector)

Basic Maththematics

Covariance is a measure of how much two random variables change together

- Population variance:

$$cov(\mathbf{x}, \mathbf{y}) = \frac{\sum(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})}{n}$$

- Sample variance:

$$cov(\mathbf{x}, \mathbf{y}) = \frac{\sum(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})}{n - 1}$$

Basic Maththematics

Population is the whole group of data. A **sample** is a part of a population that is used to describe the characteristics (e.g. mean or standard deviation) of the whole population.

$$var(\mathbf{x}) = \sigma^2 = \frac{\sum(\bar{\mathbf{x}} - \mathbf{x})^2}{n-1}$$

$$cov(\mathbf{x}, \mathbf{y}) = \frac{\sum(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{y} - \bar{\mathbf{y}})}{n-1}$$

Basic Maththematics

- $cov(\mathbf{x}, \mathbf{x}) = var(\mathbf{x})$ and $cov(\mathbf{x}, \mathbf{y}) = cov(\mathbf{y}, \mathbf{x})$
- if $\mathbf{x}$ and $\mathbf{y}$ are independent (uncorrelated), $cov(\mathbf{x}, \mathbf{y}) = 0$
- if $\mathbf{x}$ and $\mathbf{y}$ are correlated (both dimensions increase together), $cov(\mathbf{x}, \mathbf{y}) > 0$
- if $\mathbf{x}$ and $\mathbf{y}$ are anti-correlated (one dimension increases, the other decreases), $cov(\mathbf{x}, \mathbf{y}) < 0$

Basic Maththematics

If X is a matrix of the size $N \times d$ then the covariance matrix can be computed between a pair of dimensions such as

$$C = \begin{pmatrix} cov(\mathbf{f}_1, \mathbf{f}_1) & cov(\mathbf{f}_1, \mathbf{f}_2) & \dots & cov(\mathbf{f}_1, \mathbf{f}_d) \\ cov(\mathbf{f}_2, \mathbf{f}_1) & cov(\mathbf{f}_2, \mathbf{f}_2) & \dots & cov(\mathbf{f}_2, \mathbf{f}_d) \\ \vdots & \vdots & \vdots & \vdots \\ cov(\mathbf{f}_d, \mathbf{f}_1) & cov(\mathbf{f}_n, \mathbf{f}_2) & \dots & cov(\mathbf{f}_d, \mathbf{f}_d) \end{pmatrix}$$

C is square and symmetric matrix. In order to minimize the correlation (redundancy) and maximize the variance, we would like to have a diagonal covariance matrix.

Basic Maththematics

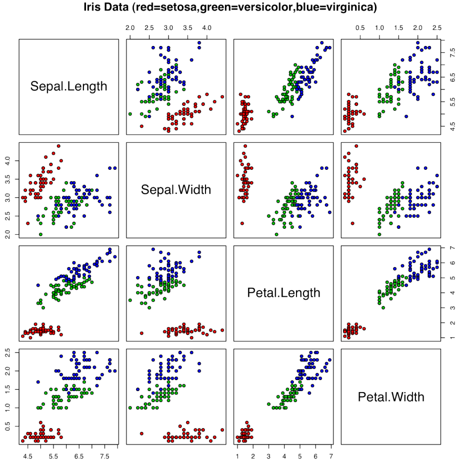
- Download any dataset from UCI, par exemple: Iris
- Calculate properties of the dataset: mean, variance, covariance
- Compute the covariance matrix

Basic Maththematics

Covariance matrix of the Iris dataset

$$\begin{pmatrix} 0.6857 & -0.0393 & 1.2737 & 0.5169 \\ -0.0393 & 0.1880 & -0.3217 & -0.1180 \\ 1.2737 & -0.3217 & 3.1132 & 1.2964 \\ 0.5169 & -0.1180 & 1.2964 & 0.5824 \end{pmatrix}$$

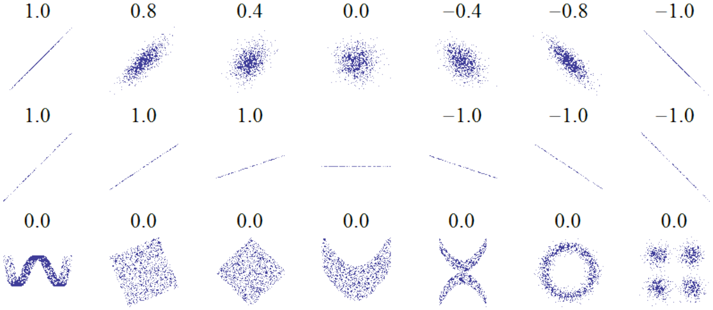
Basic Maththematics



Basic Maththematics

Correlation is a scaled version of covariance:

$$cor(\mathbf{x},\mathbf{y}) = \frac{cov(x,y)}{\sigma_x \sigma_y}$$



Basic Maththematics

Eigenvalues and eigenvectors

$$X\mathbf{v} = \lambda \mathbf{v}$$

- An **eigenvector v** is a non-zero n-by-1 vector that does not change its direction when that linear transformation is applied to it.
- An **eigenvalue** λ is a scalar.

Example:

$$\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \times \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Basic Maththematics

Eigenvalues and eigenvectors

$$X\mathbf{v} = \lambda\mathbf{v}$$

$$(X - \lambda\mathbf{I})\mathbf{v} = 0$$

where **I** is **identity matrix**, a square matrix in which all the elements of the principal diagonal are ones and all other elements are zeros.

This problem can be solve with a system of linear equations of order N (N solutions).

Basic Maththematics

Eigenvalues and eigenvectors

Find the eigenvalues and eigenvectors for the following 2×2 matrix:

$$\begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}$$

Basic Maththematics

Eigenvalues and eigenvectors

- Symmetric matrices of $n \times n$ have n eigenvectors.
- All eigenvectors are orthogonal. We say that eigenvectors are orthonormal, which means orthogonal and has length 1. Eigenvectors need to be normed as follows:

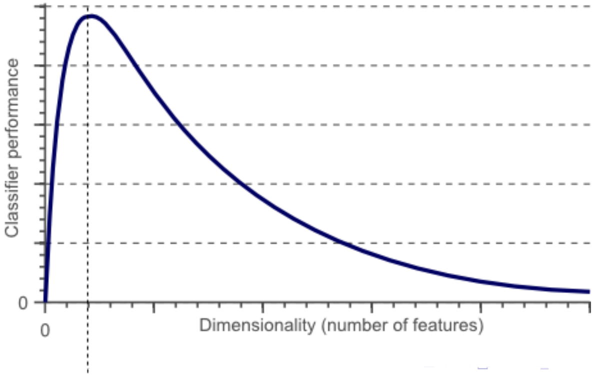
$$\mathbf{v}' = \frac{\mathbf{v}}{||\mathbf{v}||}$$

We can represent the original data in a newly defined space composed of selected eigenvectors, instead of the original space.

Introduction

Problem

In high dimensional space $X \in \mathbb{R}^d$, a large number of parameters has to be learned. Thus if the dataset is small, this will result in the curse of dimensionality and over-fitting.



Introduction

Dimensionality Reduction

The main linear technique for dimensionality reduction is to represent the data in **a lower-dimensional space** in such a way that the learning models can perform better than in the original space or even **decrease the time and memory complexity**.

- Feature Extraction: PCA - Principal Components Analysis and SVD - Singular Value Decomposition.
- Feature Selection: try to find a subset of the original variables
- and more.

To represent the data in a different space $\mathbb{R}^p$ ($p < d$) using a set of orthonormal vectors U (where $\mathbf{u}_i \in U$ is a principle component).

Introduction

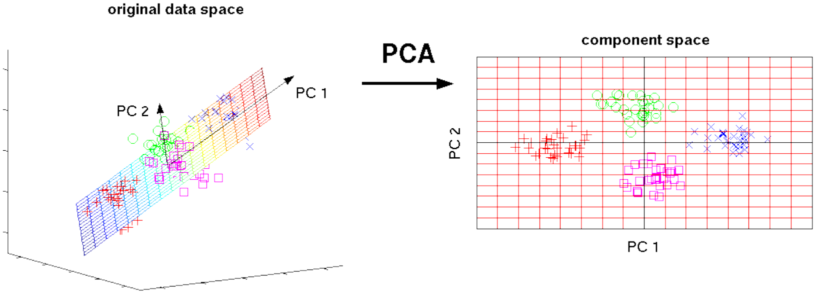
Dimensionality Reduction

Why Dimensionality Reduction?

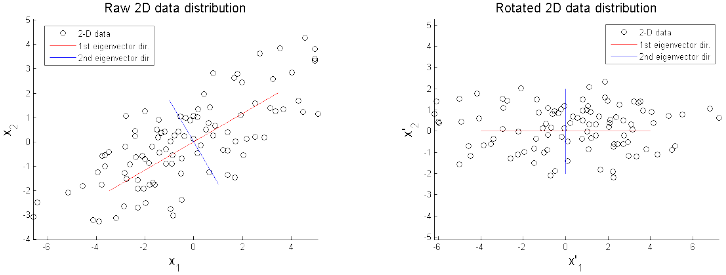
- A powerful tool for analyzing data (data visualization) and finding patterns.
- Used for compression. So you can reduce the number of dimensions without much loss of information.

Principal Components Analysis

The PCA objective is to project the data onto a lower dimensional linear space such that the variance of the projected data is maximized.

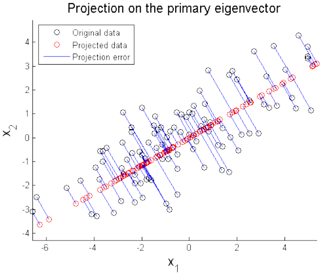


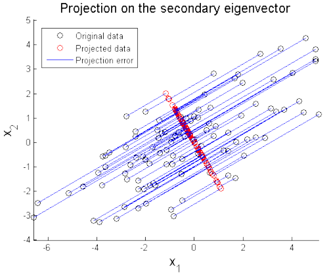
Principle Components Analysis



PCA on a synthetic dataset

Principle Components Analysis





PCA on a synthetic dataset

Principal Components Analysis

1. adjust data to the center of gravity.

$$X_{adj} = X - mean(X)$$

2. compute C, the covariance matrix of X_{adj}
3. compute $\mathbf{e}_1, ..., \mathbf{e}_d$ eigenvectors of C
4. choose p principle components, knowing that the eigenvector with the highest eigenvalue is the principle component of the data.

$$V = \begin{pmatrix} \mathbf{e}_1^T & \mathbf{e}_2^T & \dots & \mathbf{e}_p^T \end{pmatrix}$$

5. project X onto new space: $X = V^T X_{adj}^T$. X^T has the size *ntimesp*, each row represents one data object in the new space of p dimensions.

Principal Components Analysis

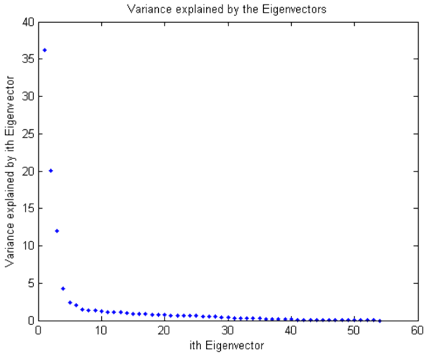
Compute the principle components for X

$$X = \begin{pmatrix} 7 & 4 & 3 \\ 4 & 1 & 8 \\ 6 & 3 & 5 \\ 8 & 6 & 1 \end{pmatrix}$$

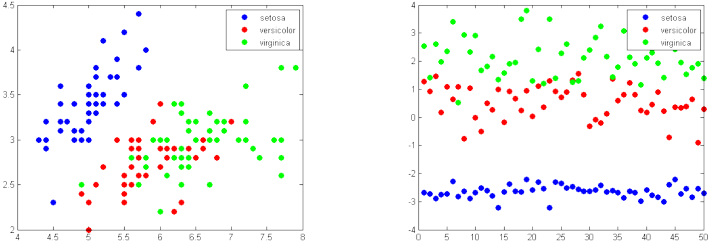
Principal Components Analysis

How many eigenvectors should we use?

- Take enough many eigen-vectors to cover 80-90% of the variance

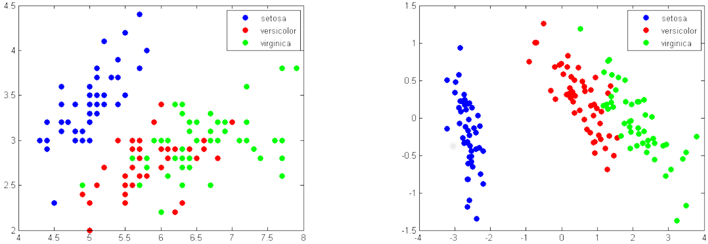


Principal Components Analysis



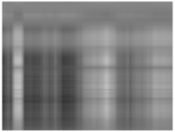
PCA on Iris dataset using one principal component

Principal Components Analysis



PCA on Iris dataset using two principal components

Principal Components Analysis



(a) 1 principal component

(b) 5 principal component

(c) 9 principal component

(d) 13 principal component

(e) 17 principal component

(f) 21 principal component

(g) 25 principal component

(h) 29 principal component

Compress the Lena's image using pprincipalcomponents

Principal Components Analysis

1. from the projected data $Y = V^T X_{adj}^T$, we have
$VY = VV^T X_{adj}^T = \mathbf{I} * X_{adj}^T$
(V is a matrix composed of principal components or unitary matrix so $VV^T = I$
2. $X_{comp} = (VY)^T + \bar{X}$

Principal Components Analysis

Limitations

What if data has a very large dimension?

- e.g: $d = 10^4$

Problem:

- covariance matrix C is size $d \times d$

Singular Value Decomposition is available to handle this issue.

Singular Value Decomposition

SVD of a matrix X is implemented to extract principal components and directions:

$$X = U\Sigma V^T$$

where $U \in \mathbb{R}^{n\times n}$ is an unitary matrix $U^T U = \mathbf{I}$
Σ is a $n \times d$ rectangular diagonal matrix with non-negative real numbers on the diagonal
$V \in \mathbb{R}^{d\times d}$ is an unitary matrix $V^T V = \mathbf{I}$

Σ is known as the singular values of X.

Singular Value Decomposition

The singular value decomposition can be computed using the following observations:

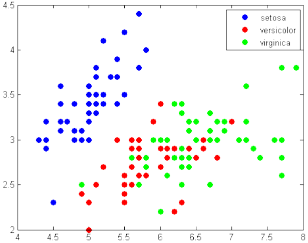
- U consists of the left-singular vectors of X, these vectors are orthogonal.

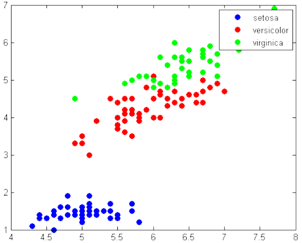
$$U = \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{pmatrix}$$

- Use p left-singular vectors in U to represent data.

$$Y = \begin{pmatrix} \mathbf{u}_1^T & \mathbf{u}_2^T & \dots & \mathbf{u}_p^T \end{pmatrix}$$

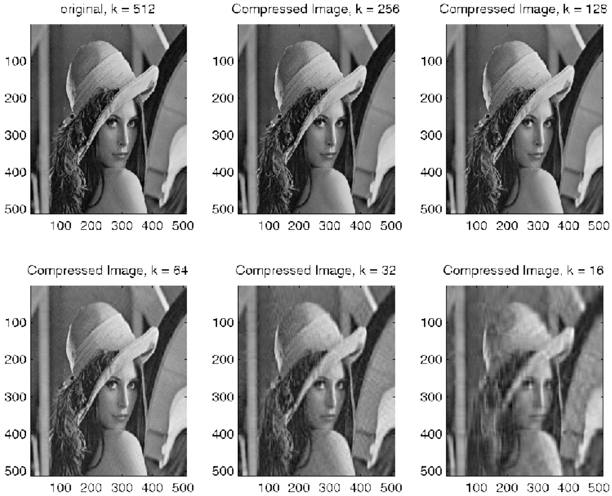
Singular Value Decomposition





SVD on Iris dataset using 1st and 3rd left-singular vectors

Singular Value Decomposition



Compress the Lena's image using SVD