Principal Component Analysis (PCA)

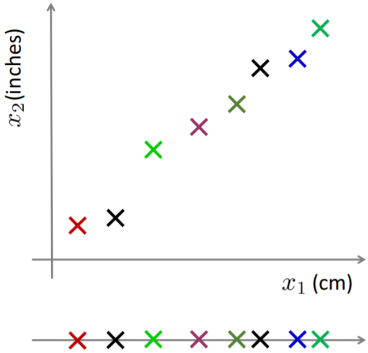
Tuan Nguyen

Ngày 20 tháng 1 năm 2021

Overview

Dimensional reduction

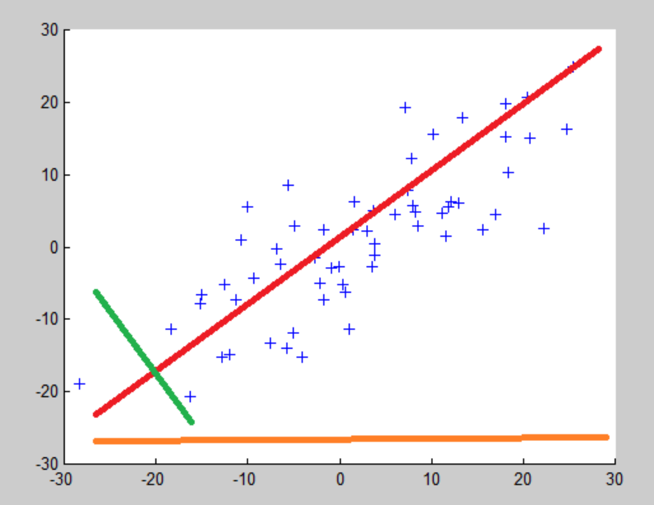
Given the vector $x \in R^n$, we want to find $\tilde{x} \in R^m$ (normally $m < n$) that are as similar to the original data points as possible.



Hình 1: Reduce 2D to 1D

Why dimentional reduction?

Dimensional reduction (cont.)



Hình 2: Which line to project?

Dimensional reduction (cont.)

For n samples $x_1, x_2, ..., x_n$.
The mean is the average of the numbers.

$$\mu = \frac{1}{N} \sum_{i=1}^{n} x_i$$

The variance is the average of the squared differences from the mean or how the samples spread from mean.

$$\text{var} = \sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2$$

When we project the data, we want to have large or small variance?

Problem setting

we consider an independent and identically distributed (i.i.d.) dataset X $= \{x_1, x_2, ..., x_n\}$ with mean 0, $x_n \in \mathrm{R}^D$

we assume there exists a low-dimensional compressed representation

$$z_n = B^T x_n \in \mathrm{R}^M$$

of x_n where projection matrix

$$B := [b_1, b_2, ..., b_M] \in \mathrm{R}^{D \times M}$$

we assum that the columns of B are orthonormal so that $b_i^T b_j = 0$ if i $\neq$ j and $b_j^T b_j = 1$.

$\mathbb{E}_z[z] = \mathbb{E}_x[B^T x] = B^T \mathbb{E}_x[x] = 0 \Rightarrow$ mean z is 0.

Maximum Variance

We need to find a matrix B that retains as much information as possible when compressing data by projecting it onto the subspace spanned by the columns $b_1, b_2, ..., b_M$ of B.

Retaining most information after data compression is equivalent to capturing the largest amount of variance in the low-dimensional code (source).

We maximize the variance of the low-dimensional code using a sequential approach. We start by seeking a single vector $b_1 \in \mathbb{R}^D$ that maximizes the variance of the projected data, i.e., we aim to maximize the variance of the first coordinate z_1 of $z \in \mathrm{R}^M$ so that

$$V_1 = \frac{1}{N} \sum_{n=1}^{N} z_{1n}^2$$

is maximized, where

$$z_{1n} = b_1^T x_n$$

Maximum Variance (cont.)

$$V_1 = \frac{1}{N}\sum_{n=1}^{N}(b_1^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} b_1^T x_n x_n^T b_1$$

$$= b_1^T(\frac{1}{N}\sum_{n=1}^{N} x_n x_n^T)b_1 = b_1^T S b_1$$

where S is called the data covariance matrix.

Remark: increasing the magnitude of the vector b1 increases V. Therefore, we restrict all solutions to $\|b\|^2 = 1$, which results in a constrained optimization problem in which we seek the direction along which the data varies most.

Maximum Variance (cont.)

Constrained optimization problem

$$\max_{b_1} b_1^T S b_1$$
$$\text{subject to } \|b_1\|^2 = 1$$

The Lagrangian:

$$\mathcal{L}(b_1, \lambda_1) = b_1^T S b_1 + \lambda_1 (1 - b_1^T b_1)$$
$$\frac{\partial \mathcal{L}}{\partial b_1} = 2 b_1^T S - 2\lambda_1 b_1^T, \frac{\partial \mathcal{L}}{\partial \lambda_1} = 1 - b_1^T b_1$$

Setting these partial derivaties to 0 gives us

$$b_1^T b_1 = 1$$
$$S b_1 = \lambda_1 b_1$$

We can see that b_1, λ_1 is an eigenvector and an eigenvalue of S respectively.

Maximum Variance (cont.)

$$V = b_1^T S b_1 = b_1^T \lambda_1 b_1 = \lambda_1$$

The variance of the data projected onto a one-dimensional subspace equals the eigenvalue that is associated with the basis vector b_1 that spans this subspace.

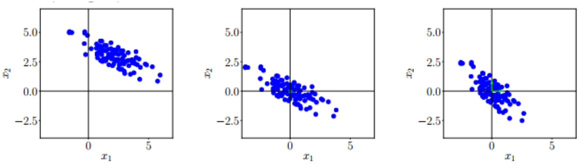
Therefore, to maximize the variance of the low-dimensional code, we choose the basis vector associated with the largest eigenvalue principal component of the data covariance matrix.

This eigenvector is called the first principal component. The second component is the projection of data onto the eigenvector corresponding with the second largest eigenvalue.

Steps of PCA in Practice

1. Mean subtraction and standardization, get the data to the normal distribution.
2. Eigendecomposition of the covariance matrix: Compute the data covariance matrix and its eigenvalues and corresponding eigenvectors. Because covariance matrix is symmetric eigenvectors are orthogonal.
3. Projection: Project the data onto the eigenvectors.

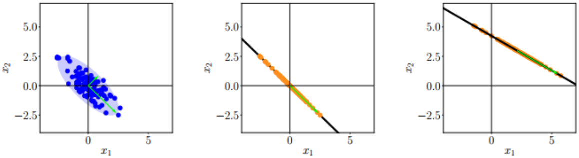
Steps of PCA in Practice (cont.)



(a) Original dataset.

(b) Step 1: Centering by subtracting the mean from each data point.

(c) Step 2: Dividing by the standard deviation to make the data unit free. Data has variance 1 along each axis.

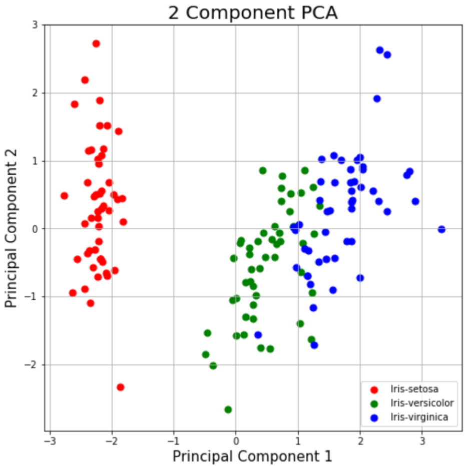


(d) Step 3: Compute eigenvalues and eigenvectors (arrows) of the data covariance matrix (ellipse).

(e) Step 4: Project data onto the principal subspace.

(f) Undo the standardization and move projected data back into the original data space from (a).

Hình 3: PCA steps



Hình 4: PCA steps