Principal Component Analysis

Prerequisite-

- Machine learning fundamentals and feature space
- Linear algebra and statistics

Objectives-

- Motivation for principal component analysis(PCA).
- Understanding what is PCA.
- Steps to compute PCA and Example.

Motivation and Meaning-

Machine learning model efficiency is largely depends on the amount of data. The more and better the data we will have better chance to trained an efficient predictive model. However that large data also comes with certain limitation like curse of dimensionality. With a high dimensional data there is high possibility to have inconsistency and redundancy in data values. This will pose complexity over computation and analysis processing of data. To mitigate this limitation, a technique is proposed to reduce the dimensions while selecting most significant features of data, here principal component analysis comes into play.

principal component Analysis is a method to identify the patterns in data using their similarities and dissimilarities between the sample points. The patterns within data is hard to find especially when we cannot visualize it graphically(the data is in high dimension). The principal component analysis is a powerful tool to explore data with its hidden patterns and reduce the dimensions. The role of PCA can be understood by a toy example shown in figure 1. Let we are interested to study the motion of ideal spring. To do that we have attached a ball of mass m to a frictionless spring and release it with a small distance from the equilibrium. As result the spring start oscillating and because it is frictionless so it will oscillate indefinitely along the axis at a set frequency. Because initially we don’t know that how many dimensions are important so we assume to record the two dimensional motion of ball from three dimension by placing three camera around the system and project them. This will generate the three different distribution of oscillation and the big question is how do we get a simple equation of x from this data. [*Figure Source: A tutorial on principal component analysis- Jonathon Shlens*]

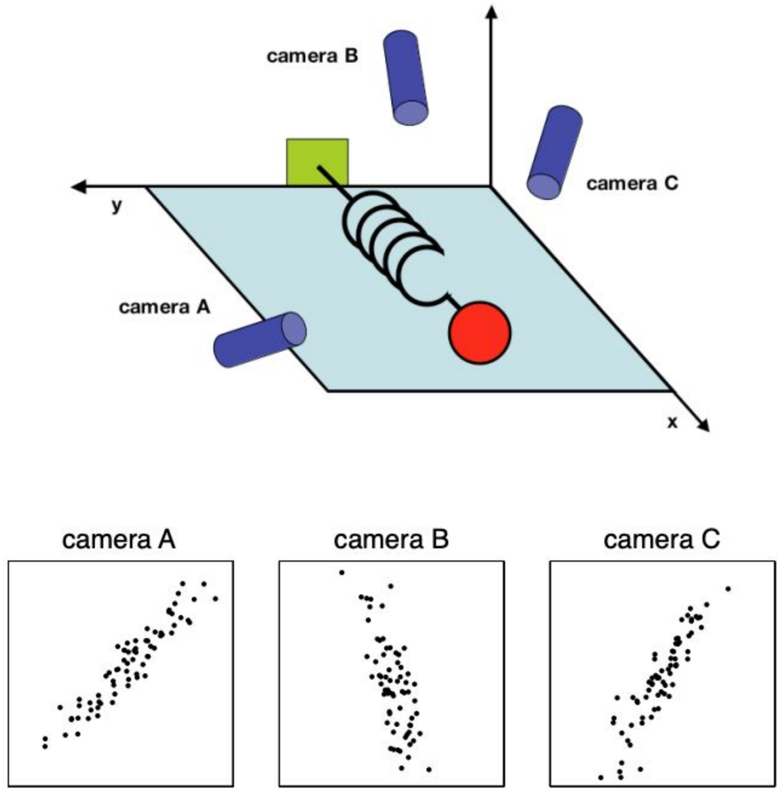

Figure 1 A toy example to understand PCA

The main goal of principal component analysis is to explore the meaningful basis to project dataset. The hope is that new basis will remove the noise and explore hidden pattern within data. The intuition to re-express the data to new basis is understood through figure 2 which shows the data from camera A point of view. It is clear from the distribution that largest direction of variation is not along the original basis of recording (x, y) but along the new line called as best fit line. The figure represents the noise and signal variance by two lines on data.

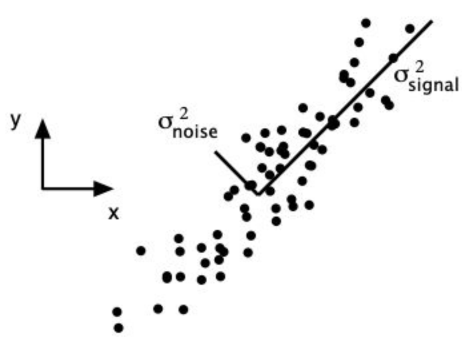

Figure 2 Change of basis (Camera A projection)

So, we can understand principal component analysis (PCA) in many ways like:

- Principal Component Analysis is a mathematical approach, which is used for better interpretation of your data.
- The main purpose of Principal Component Analysis is to reduce high dimensional data into low dimensional space.
- Principal Component Analysis is an unsupervised machine learning technique which finds insights of data without having prior knowledge. It reduces data by projecting geometrically onto lower dimensional basis know as principal components.

When to use PCA

- When you have to reduce the number of features or dimensions in the data.
- When you have to check whether the features are independent to each other or not.

Steps of principal Component Analysis

The steps of principal component analysis are summarized as shown in following chart-

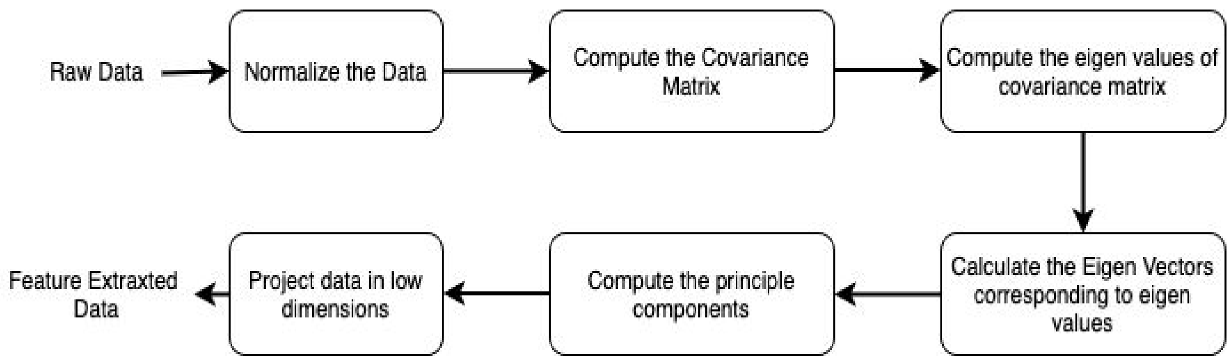

Figure 3 PCA steps

Step 1- Normalize is one of the fundamental step of data processing, aims to get an unbiased result of model. It is the process to bring all the data variables within same range of values (-1, 1). The normal value of variable is calculated using:

$$z = \frac{x-\bar{x}}{\sigma}$$

where $\bar{x}$ is the mean and σ is standard deviation of distribution.

Step 2- Second step to compute PCA is to calculate covariance matrix. A covariance matrix is describe the correlation between variables within dataset. It helps to identify heavily dependent variables. Mathematically a covariance matrix for three features data set may be defined as:

$$C = [cov(x,x)\ cov(x,y)\ cov(x,z)\ cov(y,x)\ cov(y,y)\ cov(y,z)\ cov(z,x)\ cov(z,y)\ cov(z,z)\ ]$$

Where,

$$cov\,(x,\ x) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})}{n-1}$$

and

$$cov\,(x,\ y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

- Cov(x,x) is actually variance of variable x.
- Cov(x, y) is the variance of variable x w. r. t. variable y. cov(x, y) = cov(y, x)

The covariance value within the matrix denotes the dependency of two variables with each other. A negative value denotes the two variable are indirectly proportional to each other and a positive value is for directly proportional relationship.

Step 3- The step 3 is related to find the mathematical constructs of covariance matrix to identify principal components. These mathematical constructs are eigenvalues and corresponding eigenvectors. The eigenvectors are nothing but the principal components of our covariance matrix and represent the axis of new feature space whereas eigenvalues are the magnitude of those vectors. In other words we can say that eigenvalues actually describe the contribution of each vector in terms of variance.(A high magnitude for eigen value denotes high variance along its eigen vector in feature space)

Eigenvectors and Eigenvalues- A vector whose direction remains same ever after applying linear transformation is called eigenvector. This can be expressed as:

$$A.x = \lambda.x$$

Where A is covariance square matrix with x its eigenvector and λ is a constant. The eigenvalues of the matrix A is obtained by solving the equation-

$$|A - \lambda I| = 0$$

Step 4- principal components can be computed by arranging eigenvalues with corresponding eigenvectors in descending order. The higher value eigen vectors have more significance over the data and form principal components whereas the lower value eigen vectors can be removed in order to reduce the dimensions.

Step 5- Finally the original data can be projected using principal components in reduce dimensions. This can be done by multiplying the transpose of original dataset with the transpose of computed vectors.

Example- Let we have been given with following data value in two dimensions.

X_1 : 2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2.0, 1.0, 1.5, 1.2

X_2 : 2.4, 0.7, 2.9, 2.2, 3.0, 2.7, 1.6, 1.1, 1.6, 0.9

Step 1- Standardize the values:

$$\overline{X_2} = \frac{\sum X_i}{N} = \frac{2.4+0.7+2.9+2.2+3.0+2.7+1.6+1.1+1.6+0.9}{10} = 1.91$$

$$\overline{X_2} = \frac{\sum X_i}{N} = \frac{2.4+0.7+2.9+2.2+3.0+2.7+1.6+1.1+1.6+0.9}{10} = 1.91$$

X'_1: 0.69, -1.31, 0.39, 0.09, 1.29, 0.49, 0.19, -0.81, -0.31, -0.71

X'_2: 0.49, -1.21, 0.99, 0.29, 1.09, 0.79, -0.31, -0.81, -0.31, -1.01

Step 2- Covariance Matrix-

$$C = [0.6165\ 0.6154\ 0.6154\ 0.7165\ ]$$

Step 3- Calculate eigenvalue of covariance matrix C.

$$\lambda_1\ \lambda_2\ = [0.490\ 1.284\ ]$$

and the eigenvectors are:

$$V\ = [-0.7352\ \ -0.6779\ 0.6779\ \ -0.7352\ ]$$

Step 4- Reduce the dimension and create a new feature vector (Highest eigenvalue vector is the principal component of the data). Here we choose the eigenvector with bigger eigenvalue and leave the smaller one.

$$V\ = [-0.6779\ \ -0.7352\ ]$$

Step 5- Project the data on new feature space-

$$Y\ = X\ V^T = \left[X_{1,}\ X_2\right]\left[V_1,\ V_2\right]^T$$

Y = -0.8279, 1.7775, -0.9922, -0.2742, -1.6758, -0.9129, 0.0991, 1.1445, 0.4380, 1.2238

Advantages of PCA

- Removes the correlated attributes.
- Help to reduce overfitting.

- Improves the data visualization.
- It also help to improve the performance of Algorithm,

Disadvantages of PCA

- Data normalization must be needed before applying PCA
- Some level of information loss.
- Independent variables are become less interpretable.

****