Principal Component Analysis(PCA)

1. Introduction

We represent data as a d-dimensional vector space.

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_d \end{bmatrix}$$

The expectation value is given as:

$$\mathbf{E}[\mathbf{X}] = \begin{bmatrix} \mathbf{E}[\mathbf{x}_1] \\ \mathbf{E}[\mathbf{x}_2] \\ \vdots \\ \mathbf{E}[\mathbf{x}_d] \end{bmatrix}$$

2. Covariance Matrix

Covariance matrix is a dxd matrix which signifies the covariance between each pair of variables. In other words, it is a measure of how much two random variables change together. The covariance matrix is given by:

$$\mathbf{Cov}[\mathbf{X}] = \begin{bmatrix} \mathbf{Cov}[\mathbf{x}_1, \mathbf{x}_1] & \mathbf{Cov}[\mathbf{x}_1, \mathbf{x}_2] & \cdots & \mathbf{Cov}[\mathbf{x}_1, \mathbf{x}_d] \\ \mathbf{Cov}[\mathbf{x}_2, \mathbf{x}_1] & \mathbf{Cov}[\mathbf{x}_2, \mathbf{x}_2] & \cdots & \mathbf{Cov}[\mathbf{x}_2, \mathbf{x}_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Cov}[\mathbf{x}_d, \mathbf{x}_1] & \mathbf{Cov}[\mathbf{x}_d, \mathbf{x}_2] & \cdots & \mathbf{Cov}[\mathbf{x}_d, \mathbf{x}_d] \end{bmatrix}$$

where,

$$\begin{aligned} \mathbf{Cov}[\mathbf{x}_i, \mathbf{x}_j] &= \mathbf{E}[(\mathbf{x}_i - \mathbf{E}[\mathbf{x}_i])(\mathbf{x}_j - \mathbf{E}[\mathbf{x}_j])] \\ &= \mathbf{E}[\mathbf{x}_i \mathbf{x}_j] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{x}_j] \\ &= \mathbf{E}[\mathbf{x}_i \mathbf{x}_j] - \mathbf{E}[\mathbf{x}_i]\mathbf{E}[\mathbf{x}_j] \end{aligned}$$

Covariance matrix can also be represented as:

$$\sigma = \mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])(\mathbf{X} - \mathbf{E}[\mathbf{X}])^T]$$

where, $\mathbf{E}[\mathbf{X}]$ is a d-dimensional vector and $\mathbf{E}[(\mathbf{X} - \mathbf{E}[\mathbf{X}])(\mathbf{X} - \mathbf{E}[\mathbf{X}])^T]$ is a dxd matrix.

Dimensionality Reduction

PCA is a data reduction technique. It is done with two perspectives in mind: Lower dimension and orthogonality of new dimensions.
PCA is an unsupervised learning algorithm.
We find new axes in such a way that the variance of the data along the new axes is maximum, i.e. the loss of information is minimum.
The new axes are called principal components.
To do this, we take help of the covairance matrix. It is a matrix that helps us find the correlation

between the variables.
A covariance matrix is:

- Symmetric
- Positive semi-definite(all eigenvalues are non-negative)
- Eigenvectors are orthogonal

We will use a two dimensional space for demonstration.

Let the covariance matrix be:

$$C_x = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

In general, C_x is a dxd matrix which looks like:

$$C_x = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$$

In order to reduce dimensionality, following are the steps:

1. Make the data uncorrelated:
   Let C_y be the covariance matrix of the transformed data.
   Then C_y is diagonal matrix with the eigenvalues of C_x on the diagonal. C_y looks as follows:

$$C_y = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix}$$

In two dimensional case,

$$C_y = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

In order to do this diagonalization, we need to find the eigenvectors of C_x.

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

where, $\mathbf{v}$ is the eigenvector and λ is the eigenvalue.
Let $\mathbf{v}_1$ and $\mathbf{v}_2$ be the eigenvectors of C_x.
Then we can say that:

$$C_x\mathbf{v}_1 = \lambda_1\mathbf{v}_1$$

and

$$C_x\mathbf{v}_2 = \lambda_2\mathbf{v}_2$$

Let

$$U = [\mathbf{v}_1 \quad \mathbf{v}_2]$$

Then,

$$C_xU = U\begin{bmatrix}\lambda_1 & 0 \\ 0 & \lambda_2\end{bmatrix}$$

Let the eigenvalue matrix be represented as Λ.
Then,

$$C_xU = U\Lambda$$

Hence,

$$U^TC_xU = \Lambda$$

This is owing to the fact that U is an orthogonal matrix.
Hence, the required covariance matrix is

$$C_y = U^TC_xU = \Lambda$$

Now that we have the diagonalized covariance matrix, we need to find the new feature vectors.

$$C_y = U^TE((x - x_{mean})(x - x_{mean})^T)U$$

Let $(x - x_{mean})$ be represented as $\mathbf{z}$.
Then,

$$C_y = U^TE(\mathbf{zz}^T)U$$
$$C_y = E(U^T\mathbf{zz}^TU)$$

So,

$$E(YY^T) = E(U^T\mathbf{zz}^TU)$$
$$E(YY^T) = E(U^TZ(U^TZ)^T)$$

So we get,

$$Y = U^TZ$$
$$Y = U^T(x - x_{mean})$$

So, we first standardize the dataset to make the mean zero and then multiply it with the eigenvectors of the covariance matrix to get the new feature vectors.

2. Choosing the maximum variance axes:
   Let e_1 be the vector on which X is projected.
   Then, projected vector will be

$$\mathbf{y}_1 = \mathbf{e}_1^T\mathbf{x}$$

Variance of the projection is given by:

$$\sigma_1^2 = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{e}_1^T\mathbf{x}_i - e^T\mathbf{x}_{mean})^2$$

where, $\mathbf{x}_{mean}$ is the mean of the dataset.

After some algebraic manipulation, we get:

$$\sigma_1^2 = e_1^TC_xe_1$$

where, C_x is the covariance matrix of the dataset.
We have maximize this variance subject to the constraint that $e_1^T e_1 = 1$.

$$L = \sigma_1^2 - \lambda(e_1^T e_1 - 1)$$

$$\frac{\partial L}{\partial e_1} = 0$$

After taking partial derivative we get an equation of the form

$$Ae_1 = \lambda e_1$$

where, A is the covariance matrix of the dataset.
Hence we conclude that the eigenvector of the covariance matrix with the largest eigenvalue is the principal component.

3. Main Objective is to minimize the loss of information:
   To do this, order the eigenvalues in decreasing order and remake the C_y matrix.

$$C_y = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{bmatrix}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$.

$$C_y = U^T C_x U = \Lambda$$

4. Dimensionality Reduction:
   Let k be the number of dimensions we want to reduce to. (k < d)
   So, we can say that,

$$C_y = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix}$$

So,

$$C_y = U^T C_x U = \Lambda$$

except this time, the dimension of Λ is $k \times k$, and the dimension of U is $k \times d$.

So, the transformation will be

$$Y = U^T(x - x_{mean})$$

where, U is the matrix of eigenvectors of C_x of dimension $k \times d$.

The original dataset can be reconstructed from the new feature vectors as follows:

$$X = UY + x_{mean}$$