Principal Component Analysis

Learning Type: Unsupervised, Task: Dimension Reduction, Algorithm: PCA

Definition:

It is a popular technique to reduce the dimensionality/feature of the datasets. It preserves the maximum amount of information by finding new feature vectors that maximize the data spread.

Applications: It is used to visualize the data, find the patterns in high-dimension data, and image compression

Algorithm:

Step1: Standardize the features with zero mean and one standard deviation.

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \qquad \textit{where} \quad \mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)} \quad \textit{and} \quad \sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_j^{(i)} - \mu_j\right)^2} \tag{1}$$

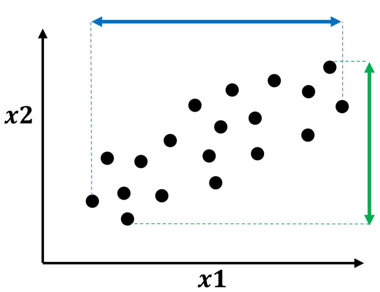
where m is the number of examples in dataset X.

Step2: Find the covariance matrix of dataset and compute eigenvalues and eigenvectors of covariance matrix using spectral decomposition, where n is the number of features.

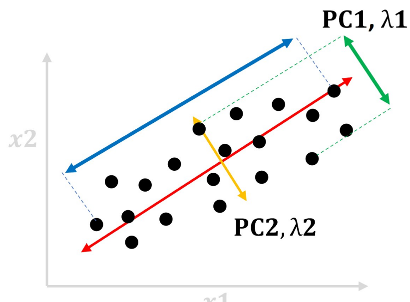
$$Cov(X) = \begin{bmatrix} \sigma_{x_1}^2 & \dots & cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ cov(x_n, x_1) & \dots & \sigma_{x_n}^2 \end{bmatrix} \qquad \textit{and} \quad Cov(X) = V\Sigma V^{-1} \tag{2}$$

Step3: Sort the eigenvalues and eigenvectors in ascending order $\Sigma_{sort} = \text{sort}(\Sigma)$ and $V_{sort} = \text{sort}\left(V, \Sigma_{\text{sort}}\right)$

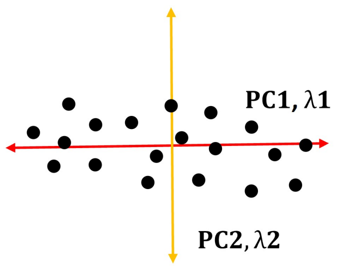
Step4: Transform the dataset X to first k eigenvectors $V_{\text{reduced}} = V[:, 0:k]$ and $X_{\text{reduced}} = X \times V_{\text{reduced}}$



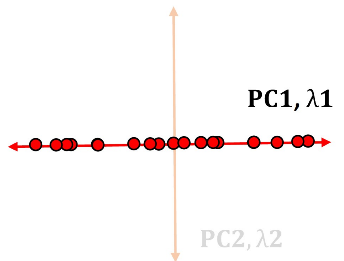
Step1: Standardize the data with mean 0 and std 1



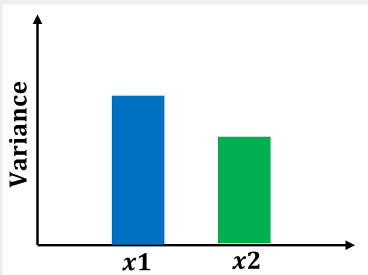
Step2: Find 2 × 2 covariance matrix then 2 eigenvalues, 2 eigenvectors

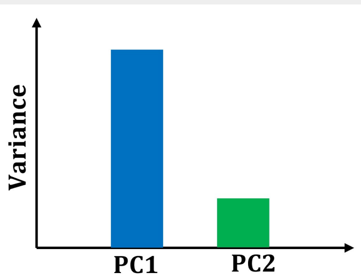


Step3: Sort PCs based on eigen value λ1>λ2 and select PC1



Step4: Map the data to PC1 vectors





1) The eigenvector with the largest eigenvalue is in the direction along which the dataset has the maximum variance which is called first principal component that is PC1.
2) Covariance matrix will be order of $n \times n$ with n eigenvalues and n eigenvectors.
3) $X_{reduced}$ will be order of $m \times k$ where X is $m \times n$ and $V_{reduced}$ is $k \times n$

For any mistake/query: alishafique3@gmail.com

For references and updates: Website, Github