# Principal Component Analysis

Sunday, December 29, 2024 12:51 PM

PCA

Most used unsupervised algorithms

PCA is used for
- Noise Filtering
- Visualization
- Feature Extraction
- Stick Market Predictions
- Gene Data Analysis

Goal is to
- Identify patterns in Data
- Detect the correlation between variables

Reduce the dimensions of a d-dimensional dataset by projecting it onto a (k)-dimensional subspace (where k<d)

- Standardize the data.
- Obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix, or perform Singular Vector Decomposition.
- Sort eigenvalues in descending order and choose the $k$ eigenvectors that correspond to the $k$ largest eigenvalues where $k$ is the number of dimensions of the new feature subspace ($k \leq d$)/.
- Construct the projection matrix $\mathbf{W}$ from the selected $k$ eigenvectors.
- Transform the original dataset $\mathbf{X}$ via $\mathbf{W}$ to obtain a $k$-dimensional feature subspace $\mathbf{Y}$

NOT LINEAR REGRESSIOn
- PCA Learns about the relationship between X and Y values
- Find list of principal axes
- Is highly impacted by outliers in the data

# Linear Discriminant Analysis

Sunday, December 29, 2024 1:14 PM

LDA

- Used as a dimensionality reduction technique
- Used in the preprocessing step for pattern classification
- Has the goal to project a dataset onto a lower-dimensional space

LDA differs from PCA because LDA is interested in the axes that maximize the separation between multiple classes

LDA is supervised because of the relation to the dependent variable

**The goal of LDA is to project a feature space (a dataset n-dimensional samples) onto a small subspace subspace k(where $k \leq n-1$) while maintaining the class-discriminatory information.**

**Both PCA and LDA are linear transformation techniques used for dimensional reduction. PCA is described as unsupervised but LDA is supervised because of the relation to the dependent variable.**

1. Compute the $d$-dimensional mean vectors for the different classes from the dataset.
2. Compute the scatter matrices (in-between-class and within-class scatter matrix).
3. Compute the eigenvectors ($\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_d$) and corresponding eigenvalues ($\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_d$) for the scatter matrices.
4. Sort the eigenvectors by decreasing eigenvalues and choose $k$ eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\boldsymbol{W}$ (where every column represents an eigenvector).
5. Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $\boldsymbol{Y} = \boldsymbol{X} \times \boldsymbol{W}$ (where $\boldsymbol{X}$ is a $n \times d$-dimensional matrix representing the $n$ samples, and $\boldsymbol{y}$ are the transformed $n \times k$-dimensional samples in the new subspace).