Principal Component Analysis (PCA)

Goal

PCA finds a new set of dimensions such that all the dimensions are orthogonal (and hence linearly independent) and ranked according to the variance of data along them.

Find a transformation such that
- The transformed features are linearly independent
- Dimensionality can be reduced by taking only the dimensions with the highest importance
- Those newly found dimensions should minimize the projection error
- The projected points should have maximum spread, i.e. maximum variance

Variance

How much variation or spread the data has.

$$Var(X) = \frac{1}{n}\sum(X_i - \bar{X})^2$$

Covariance Matrix

Indicates the level to which two variables vary together.

$$Cov(X,Y) = \frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})^T$$

$$Cov(X,X) = \frac{1}{n}\sum(X_i - \bar{X})(X_i - \bar{X})^T$$

Eigenvector, Eigenvalues

The eigenvectors point in the direction of the maximum variance, & the corresponding eigenvalues indicates the importance of its corresponding eigen vector.

$$A\vec{v} = \lambda\vec{v}$$

Approach

- Substract the mean from X
- Calculate cov(X,X)
- Calculate eigenvector & eigenvalues of covariance matrix
- Sort the eigenvectors according to their eigenvalues in decreasing order
- Choose first k eigenvectors & that will be the new k dimensions
- Transform the original n dimensional data points into k dimensions (= Projections with dot product)