# Singular Value Decomposition

Singular Value Decomposition (SVD) is an important theorem in Linear Algebra which describes how any matrix can be decomposed in sub-components. The theorem states that for any m x n matrix A with rank r, there exists a decomposition consisting of matrices U, S and V such that:

(1)

U is m x r matrix

S is r x r diagonal matrix

V is r x n matrix

This is also called economic decomposition of matrix A. Please note that columns of U are orthonormal eigen vectors of AAT while columns of V are orthonormal eigenvectors of ATA. The diagonal elements of S are square root of non-zero eigen values of AAT or ATA. Although number of eigen values of AAT and ATA may differ, but non-zero eigen values are the same for both matrices. For the ith eigen vector, the following equation holds:

This equation directly follows from equation (1) when matrix V gets multiplied to both the sides. Please note that VTV = I since V is orthonormal. This equation is the eigen value equation equivalent for a rectangular matrix.

**Lemma**: The non-zero eigenvalues of AAT are same as ATA.

**Proof:** We have AAT v = λ v, where λ is a non-zero eigenvalue and v is the corresponding eigenvector.

Multiply with AT on both the sides, we get (ATA) (AT v) = λ (AT v) which shows that λ is also an eigen value corresponding to eigen vector AT v.

# Principal Component Analysis

Principal Component Analysis (PCA) is used in for reducing number of features in machine learning. Let us assume we have several colors in our color palette and some colors are sparingly used. Thus, the color variations in a picture can be explained in terms of few principal colors. Further only a handful or principal colors may be needed as other colors can be derived from these principal colors. PCA is the systematic approach to get these principal colors. Please note that these principal colors can be a combination of the existing colors in our color palette.

The following table shows matrix used typically in a machine learning model. There are n input features and m distinct observations of these features. The problem at hand is how to compress these features into only a handful since there can be interdependence among features. Furthermore, a feature may not vary much in our observations and hence may not have much prediction power.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Observation/Features | Feature 1 |  | Feature 2 |  |  |  | Feature n |
| Observation 1 |  |  |  |  |  |  |  |
| Observation 2 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Observation m |  |  |  |  |  |  |  |
| Stats | Mean = µ1  Stddev = σ1 |  |  |  |  |  |  |

Let us call observation-feature matrix as D. We first center the observations such that each feature has mean of 0. For each observation of feature k, normalize observations as follows:

Please note that . Hence mean of our scaled observations is 0. Let us call our centered matrix as A.

Next, we find the eigen vectors of matrix ATA which is also called covariance matrix. The covariance matrix shows the dependence of one feature on another. The eigen-vectors of the covariance matrix are orthogonal to each other and first k eigen vectors with non-zero eigen value, form the basis of row space of matrix A. Please note that row space of A has dimension k and k eigen vectors of ATA are linearly independent since they are orthogonal to each other. Let us now find the new co-ordinates in the new basis. Please note that the new basis represents the new feature space and only a max of k features are needed since dimension of the new feature space is k. Thus, we may automatically drop some features as a result of transformation to new feature space.

If we denote the row vector of centered matrix A as R, then the ***dot*** product of R and vi is the projection of vector R on vi. This projection or shadow is the new co-ordinate for basis vector vi.

New co-ordinates in the basis [v1 v2 ... vk] are = [R. v1 R. v2 ... R. vk].

(x, y)

x

The projection of vector [x, y] on the x axis is the dot product with standard basis vector e1 [1 0]T which is equal to x.

In the preceding discussion we covered only one row and thus one observation of matrix A. If we project all m observations on vector v1 , the resulting co-ordinates for basis vector v­1 will be Av1. It is easy to check that such a vector will have mean 0 since it is a linear combination of columns with mean 0 since matrix A is centered.

**Lemma**: The variance of this projected coordinate across observations is a function of eigen value corresponding to the v1.

**Proof:** Let Y = Av1

Variance = YT Y = v1T ATA v1 = v1T λ1 v1 , since v1 is the eigen vector for matrix ATA

After simplification, we get variance = λ1 since vector v1 has norm 1.

In other words, variance of across observations is highest for the eigen vector with the largest eigen value. Thus, we come to the main idea behind PCA. If we take first few eigen vectors corresponding to the largest eigen values, we can sense a large variation in feature values. The features corresponding to low eigen values will have low variance and will contribute little to the predictive power in a machine learning model. Hence, we take only a handful of eigen vectors which we call principal components to transform our original features by projecting the original observations into the feature space defined by the principal components.