Chapter 6

Dimensionality Reduction

6.1 Dimensionality Reduction

- Dimensionality reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions ensuring that it conveys similar information concisely.
- This technique is typically used while solving machine learning problems to obtain better features for classification.
- Dimensions can be reduced by:
    a. Combining features using linear or non-linear transformation
    b. Selecting subset of features

Methods for Reducing Dimensionality

1. **Missing Values Ratio**
    - Data columns with too many missing values are unlikely to carry much useful information. Thus data columns with number of missing values greater than a given threshold can be removed.
    - The higher the threshold, the more aggressive the reduction.
2. **Low Variance Filter**
    - Similarly, to the previous technique, data columns with little changes in the data carry little information. Thus all data columns with variance lower than a given threshold are removed.
    - A word of caution: variance is range dependent; therefore, normalization is required before applying this technique.
3. **High Correlation Filter**
    - Data columns with very similar trends are also likely to carry very similar information. In this case, only one of them will suffice to feed the machine learning model.
    - Here we calculate the correlation coefficient between numerical columns and between nominal columns as the Pearson’s Product Moment Coefficient and the Pearson’s chi square value respectively.
    - Pairs of columns with correlation coefficient higher than a threshold are reduced to only one.

- A word of caution: correlation is scale sensitive; therefore, column normalization is required for a meaningful correlation comparison.

4. **Random Forests / Ensemble Trees**
    - Decision Tree Ensembles, also referred to as random forests, are useful for feature selection in addition to being effective classifiers.
    - One approach to dimensionality reduction is to generate a large and carefully constructed set of trees against a target attribute and then use each attribute's usage statistics to find the most informative subset of features.
    - Specifically, we can generate a large set (2000) of very shallow trees (2 levels), with each tree being trained on a small fraction (3) of the total number of attributes.
    - If an attribute is often selected as best split, it is most likely an informative feature to retain.
    - A score calculated on the attribute usage statistics in the random forest tells us – relative to the other attributes – which are the most predictive attributes.
5. **Principal Component Analysis (PCA)**
    - Principal Component Analysis (PCA) is a statistical procedure that orthogonally transforms the original 'n' coordinates of a data set into a new set of 'n' coordinates called **principal components**.
    - As a result of the transformation, the first principal component has the largest possible variance; each succeeding component has the highest possible variance under the constraint that it is orthogonal to (i.e., uncorrelated with) the preceding components.
    - Keeping only the first $m < n$ components reduces the data dimensionality while retaining most of the data information, i.e. the variation in the data. Notice that the PCA transformation is sensitive to the relative scaling of the original variables.
    - Data column ranges need to be normalized before applying PCA. Also notice that the new coordinates (PCs) are not real system-produced variables anymore.
    - Applying PCA to given data set, loses its interpretability. If interpretability of the results is important for our analysis, PCA is not the transformation for the project.
6. **Backward Feature Elimination**

- In this technique, at a given iteration, the selected classification algorithm is trained on n input features.
- Then we remove one input feature at a time and train the same model on n-1 input features n times.
- The input feature whose removal has produced the smallest increase in the error rate is removed, leaving us with n-1 input features.
- The classification is then repeated using n-2 features, and so on. Each iteration k produces a model trained on n-k features and an error rate e(k).
- Selecting the maximum tolerable error rate, we define the smallest number of features necessary to reach that classification performance with the selected machine learning algorithm.

7. **Forward Feature Construction**
    - This is the inverse process to the Backward Feature Elimination.
    - We start with 1 feature only, progressively adding 1 feature at a time, i.e. the feature that produces the highest increase in performance.
    - Both algorithms, Backward Feature Elimination and Forward Feature Construction, are quite time and computationally expensive.
    - They are practically only applicable to a data set with an already relatively low number of input columns.

Advantages of Dimensionality Reduction

- Space required to store the data is reduced as the number of dimensions comes down
- Less dimensions lead to less computation/training time
- Some algorithms do not perform well when we have a large dimensions. So reducing these dimensions needs to happen for the algorithm to be useful
- It takes care of multi-collinearity by removing redundant features. For example, you have two variables – ‘time spent on treadmill in minutes’ and ‘calories burnt’. These variables are highly correlated as the more time you spend running on a treadmill, the more calories you will burn. Hence, there is no point in storing both as just one of them does what you require

- It helps in visualizing data. As discussed earlier, it is very difficult to visualize data in higher dimensions so reducing our space to 2D or 3D may allow us to plot and observe patterns more clearly

6.2 Principal Component Analysis (PCA)

1. The main idea of PCA is to reduce dimensionality from a given data set.
2. It basically solves the problem of overfitting.
3. Given data set consists of many variables, correlated to each other either heavily or lightly.
4. Reducing is done while retaining the variation present in data set up to a maximum extend.
5. The same is done by transforming variables to a new set of variables which are known as principal components.
6. Principal Components are orthogonal, ordered such that retention of variation present in original components decreases as we move down in order.
7. So, in this way, the first principal component will have maximum variation that was present in orthogonal component.
8. Principal Components are Eigen Vectors of Covariance matrix and hence they are called as orthogonal.
9. The data set on which PCA is to be applied must be scaled.
10. The result of PCA is sensitive to relative scaling.
11. Properties of Principal Components:
    - Principal components are linear combination of original variables.
    - Principal components are orthogonal.
    - Variation present in principal components decreases as we move from first principal component to last principal component.

Implementation of PCA

1. **Normalize the data**
   - Data as input to PCA process must be normalized to work PCA properly.
   - This can be done by subtracting the respective means from numbers in respective columns.
   - If we have two dimensions X and Y, then for all X becomes $(X - \bar{X})$ and Y becomes $(Y - \bar{Y})$.

- The result gives us a dataset whose mean is zero.

2. **Calculate the covariance matrix**
   - Since we have taken 2 dimensional data set, the covariance matrix will be

$$Covariance\ Matrix = \begin{bmatrix} Var(X1) & Cov(X1, X2) \\ Cov(X2, X1) & Var(X2) \end{bmatrix}$$

3. **Finding Eigen Values and Eigen Vectors**
   - In this step we have to find Eigen Values and Eigen Vectors for covariance matrix.
   - It is possible because it is a square matrix.
   - The λ will Eigen value of matrix A.
   - If it satisfies following condition: |A - λI| = 0, then we can find Eigen vector for each Eigen value λ by calculating (A – λI)V = 0.
4. **Choosing components and forming feature vectors**
   - We order the Eigen values from highest to lowest.
   - So we get components in order.
   - If a data set has n variables, then we will have n Eigen Values and Eigen Vectors.
   - It turns out that Eigen Vector with highest Eigen Value is principal component of dataset.
   - Now we have to decide how many Eigen Values for further processing.
   - We choose first p Eigen Values and discard others. We do lose out some information in this process. But if Eigen Values are small, we do not lose much.
   - Now we form a feature vector. Since we are working on 2D data, we can choose either greater Eigen Values or simply take both.
   - Feature Vector = (Eigen_Value1, Eigen_Value2)
5. **Forming Principal Components**
   - In this step, we develop Principal Components based on data from previous steps.
   - We take transpose of feature vector and transpose of scaled dataset and multiply it to get Principal Component.
   - New Data = Feature Vector^T x Scaled Dataset^T

     New Data = Matrix of Principal Components

Q. Use Principal Component Analysis (PCA) to arrive at the transformed matrix for the given matrix A.

$$A^T = \begin{bmatrix} 2 & 1 & 0 & -1 \\ 4 & 3 & 1 & 0.5 \end{bmatrix}$$

→

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\times(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
2	4	1.5	1.875	2.8185	2.25	3.5156
1	3	0.5	0.875	0.4375	0.25	0.7656
0	1	-0.5	-1.125	0.5625	0.25	1.2656
-1	0.5	-1.5	-1.625	2.4375	2.25	2.6406

$\therefore \bar{x} = \dfrac{2+1+0-1}{4} = 0.5$ $\qquad \bar{y} = \dfrac{4+3+1+0.5}{4} = 2.125$

Now, find the covariance values.

$$cov(x,y) = \sum_{i=1}^{n} \frac{(x-\bar{x})\times(y-\bar{y})}{n-1} = \frac{6.25}{3} = 2.08$$

Similarly, $cov(y,x) = 2.08$

$$cov(x,x) = \sum_{i=1}^{n} \frac{(x-\bar{x})\times(x-\bar{x})}{n-1} = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1} = \frac{5}{3} = 1.67$$

$$cov(y,y) = \sum_{i=1}^{n} \frac{(y-\bar{y})\times(y-\bar{y})}{n-1} = \sum_{i=1}^{n} \frac{(y-\bar{y})^2}{n-1} = \frac{8.1875}{3} = 2.73$$

$\therefore$ Covariance matrix becomes

$$C = \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix} = \begin{bmatrix} 1.67 & 2.08 \\ 2.08 & 2.73 \end{bmatrix}$$

As given data is in two dimensional form, there will be two principal components.

Now, find the eigen values using the characteristic equation of the form

$$\left| C - \lambda I \right| = 0$$

where C is the covariance matrix

λ is the set of eigen values

I is the identity matrix with size same as C.

$$\therefore \left| \begin{bmatrix} 1.67 & 2.08 \\ 2.08 & 2.73 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\therefore \begin{vmatrix} 1.67-\lambda & 2.08 \\ 2.08 & 2.73-\lambda \end{vmatrix} = 0 \quad \text{——— (1)}$$

i.e. $(1.67-\lambda)(2.73-\lambda) - (2.08)(2.08) = 0$

i.e. $4.56 - 1.67\lambda - 2.73\lambda + \lambda^2 - 4.33 = 0$

$\therefore \lambda^2 - 4.4\lambda + 0.23 = 0$

Solving above equation, we get

$\lambda_1 = 4.35$ and $\lambda_2 = 0.053$

Next, find eigen vectors for principal components by substituting λ_1 and λ_2 in equation (1).

i.e. $$\begin{bmatrix} 1.67-\lambda & 2.08 \\ 2.08 & 2.73-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

where x and y are eigen vectors.

Take $\lambda_1 = 4.35$

$$\therefore \begin{bmatrix} 1.67-4.35 & 2.08 \\ 2.08 & 2.73-4.35 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

i.e. $-2.68x_1 + 2.08y_1 = 0$ ——— (2)

$2.08x_1 - 1.62y_1 = 0$ ——— (3)

Note:- If we solve above equations, we will get eigen vectors as 0.

To avoid this, we use equation of orthogonal transformation relation, which is $x_1^2 + y_1^2 = 1$ —— (4)

From equation (2), we get

$$y_1 = \frac{2.68}{2.08} x_1 \quad \text{i.e. } y_1 = 1.2885\, x_1 \quad \text{—— (5)}$$

Substitute this value of y_1 in equation (4)

$$\therefore \; x_1^2 + (1.2885\, x_1)^2 = 1$$

$$\therefore \; 2.66\, x_1^2 = 1$$

$$\text{i.e. } x_1^2 = \frac{1}{2.66} = 0.3759$$

$$\therefore \; \boxed{x_1 = 0.613}$$

Substitute value of x_1 in equation (5)

$$\therefore \; y_1 = 1.2885 \times 0.6131$$

$$\therefore \; \boxed{y_1 = 0.789}$$

Now take $\lambda_2 = 0.053$

$$\therefore \begin{bmatrix} 1.67 - 0.053 & 2.08 \\ 2.08 & 2.73 - 0.053 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 0$$

$$\text{i.e. } 1.617\, x_2 + 2.08\, y_2 = 0 \quad \text{—— (6)}$$

$$2.08\, x_2 + 2.677\, y_2 = 0 \quad \text{—— (7)}$$

Note :- If we solve above equations, we will get eigen vectors as 0.
To avoid this, we use equation of orthogonal transformation relation, which is $x_2^2 + y_2^2 = 1$ —— (8)

From equation (6), we get

$$y_2 = \frac{-1.617}{2.08} x_2 \quad \text{i.e. } y_2 = -0.7774\, x_2 \quad \text{—— (9)}$$

Substitute this value of y_2 in equation (8)

$$\therefore \; x_2^2 + (-0.7774\, x_2)^2 = 1$$

$$\therefore \; 1.6044\, x_2^2 = 1$$

$$\text{i.e. } x_2^2 = \frac{1}{1.6044} = 0.6233$$

$$\therefore \boxed{x_2 = 0.789}$$

Substitute value of x_2 in equation 9

$$\therefore y_2 = -0.7774 \times 0.789$$

$$\therefore \boxed{y_2 = -0.613}$$

Now we represent these eigen vectors in matrix form as

$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}^T = \begin{bmatrix} 0.613 & 0.789 \\ 0.789 & -0.613 \end{bmatrix}$$

Now we need to find Principal components of the given dataset. For this we take into consideration the maximum eigen value and correspondingly its eigen vectors will be the principal components.

We have $\lambda_1 = 4.35$ as maximum eigen value.

$\therefore$ The principal components are $[x_1 \quad y_1]^T = [0.613 \quad 0.789]$

Q. Use Principal Component Analysis (PCA) to arrive at the transformed matrix for the given matrix A.

$$A^T = \begin{bmatrix} 2.5 & 0.5 & 2.2 & 1.9 & 3.1 & 2.3 & 2 & 1 & 1.5 & 1.1 \\ 2.4 & 0.7 & 2.9 & 2.2 & 3.0 & 2.7 & 1.6 & 1.1 & 1.6 & 0.9 \end{bmatrix}$$

→

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})\times(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
2.5	2.4	0.69	0.49	0.3381	0.4761	0.2401
0.5	0.7	-1.31	-1.21	1.5851	1.7161	1.4641
2.2	2.9	0.39	0.99	0.3861	0.1521	0.9801
1.9	2.2	0.09	0.29	0.0261	0.0081	0.0841
3.1	3.0	1.29	1.09	1.4061	1.6641	1.1881
2.3	2.7	0.49	0.79	0.3871	0.2401	0.6241
2	1.6	0.19	-0.31	-0.0589	0.0361	0.0961
1	1.1	-0.81	-0.81	0.6561	0.6561	0.6561
1.5	1.6	-0.31	-0.31	0.0961	0.0961	0.0961
1.1	0.9	-0.71	-1.01	0.7171	0.5041	1.0201

We get $\bar{x} = 1.81$ and $\bar{y} = 1.91$

Now, find the covariance values.

$$\text{cov}(x,y) = \text{cov}(y,x) = \sum_{i=1}^{n} \frac{(x-\bar{x})(y-\bar{y})}{n-1} = \frac{5.539}{9} = 0.6154$$

$$\text{cov}(x,x) = \sum_{i=1}^{n} \frac{(x-\bar{x})^2}{n-1} = \frac{5.549}{9} = 0.6165$$

$$\text{cov}(y,y) = \sum_{i=1}^{n} \frac{(y-\bar{y})^2}{n-1} = \frac{6.449}{9} = 0.7165$$

∴ Covariance matrix becomes

$$C = \begin{bmatrix} \text{cov}(x,x) & \text{cov}(x,y) \\ \text{cov}(y,x) & \text{cov}(y,y) \end{bmatrix} = \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix}$$

As given data is in two dimensional form, there will be two principal components.

Now, to find the eigen values we use characteristic equation of the form

$$|C - \lambda I| = 0$$

where C is the covariance matrix
λ is the set of eigen values
I is the identity matrix with size same as C.

$$\therefore \left| \begin{bmatrix} 0.6165 & 0.6154 \\ 0.6154 & 0.7165 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

$$\therefore \begin{vmatrix} 0.6165-\lambda & 0.6154 \\ 0.6154 & 0.7165-\lambda \end{vmatrix} = 0 \quad \text{——— (1)}$$

i.e. $(0.6165-\lambda)(0.7165-\lambda) - (0.6154)(0.6154) = 0$

i.e. $0.4417 - 0.6165\lambda - 0.7165\lambda + \lambda^2 - 0.3787 = 0$

$\therefore \lambda^2 - 1.333\lambda + 0.063 = 0$

Solving the above equation, we get
$\lambda_1 = 1.2839$ and $\lambda_2 = 0.0491$

Next, find eigen vectors for principal components by substituting λ_1 and λ_2 in equation (1)

i.e. $$\begin{bmatrix} 0.6165-\lambda & 0.6154 \\ 0.6154 & 0.7165-\lambda \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 0$$

where x and y are eigen vectors.

Take $\lambda_1 = 1.2839$

$$\therefore \begin{bmatrix} 0.6165 - 1.2839 & 0.6154 \\ 0.6154 & 0.7165 - 1.2839 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

i.e. $-0.6674x_1 + 0.6154y_1 = 0$ ——— (2)

$0.6154x_1 - 0.5674y_1 = 0$ ——— (3)

Note :- If we solve above equations, we will get eigen vectors as 0. To avoid this, we use equation of orthogonal transformation relation, which is $x_1^2 + y_1^2 = 1$ —— (4)

From equation (2), we get

$$y_1 = \frac{0.6674}{0.6154} x_1 \quad \text{i.e. } y_1 = 1.0845\, x_1 \quad \text{—— (5)}$$

Substitute this value of y_1 in equation (4)

$$\therefore\ x_1^2 + (1.0845\, x_1)^2 = 1$$

$$\therefore\ 2.1761\, x_1^2 = 1$$

$$\text{i.e. } x_1^2 = \frac{1}{2.1761} = 0.4595$$

$$\therefore\ \boxed{x_1 = 0.6778}$$

Substitute value of x_1 in equation (5)

$$\therefore\ y_1 = 1.0845 \times 0.6778$$

$$\therefore\ \boxed{y_1 = 0.7351}$$

Now take $\lambda_2 = 0.0491$

$$\therefore\ \begin{bmatrix} 0.6165 - 0.0491 & 0.6154 \\ 0.6154 & 0.7165 - 0.0491 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 0$$

$$\text{i.e. } 0.5674\, x_2 + 0.6154\, y_2 = 0 \quad \text{—— (6)}$$

$$0.6154\, x_2 + 0.6674\, y_2 = 0 \quad \text{—— (7)}$$

Note:- If we solve above equations, we will get eigen vectors as 0. To avoid this, we use equation of orthogonal transformation relation, which is $x_2^2 + y_2^2 = 1$ —— (8)

From equation (6), we get

$$y_2 = -\frac{0.5674}{0.6154} x_2 \quad \text{i.e. } y_2 = -0.9220\, x_2 \quad \text{—— (9)}$$

Substitute this value of y_2 in equation (8)

$$\therefore\ x_2^2 + (-0.9220\, x_2)^2 = 1$$

$$\therefore\ 1.8501\, x_2^2 = 1$$

i.e. $x_2^2 = \dfrac{1}{1.8501} = 0.5405$

$\therefore \boxed{x_2 = 0.7351}$

Substitute value of x_2 in equation (9)

$\therefore y_2 = -0.9220 \times 0.7351$

$\therefore \boxed{y_2 = -0.6778}$

Now we represent these eigen vectors in matrix form as

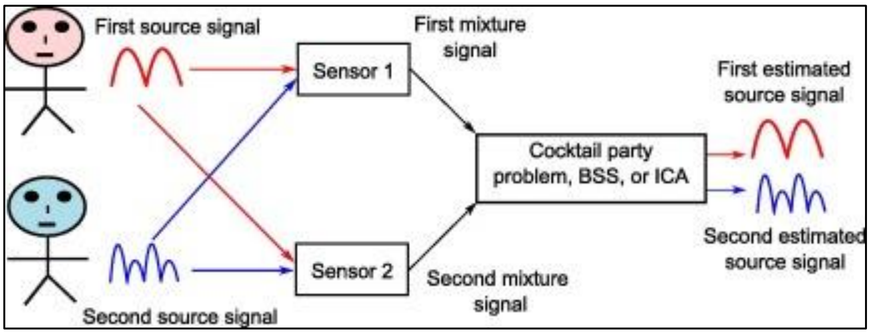
$$\begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \end{bmatrix}^T = \begin{bmatrix} 0.6778 & 0.7351 \\ 0.7351 & -0.6778 \end{bmatrix}$$

Now we need to find Principal components of the given dataset. For this we take into consideration the maximum eigen value and correspondingly its eigen vectors will be the principal components. We have $\lambda_1 = 1.2839$ as maximum eigen value.

$\therefore$ The principal components are $\begin{bmatrix} x_1 & y_1 \end{bmatrix}^T = \begin{bmatrix} 0.6778 & 0.7351 \end{bmatrix}$

6.3 Independent Component Analysis (ICA)

- **Independent component analysis** (ICA) is a statistical and computational technique for revealing hidden factors that underlie sets of random variables, measurements, or signals.
- ICA defines a generative model for the observed multivariate data, which is typically given as a large database of samples.
- In the model, the data variables are assumed to be linear mixtures of some unknown latent variables, and the mixing system is also unknown.
- The latent variables are assumed non-gaussian and mutually independent, and they are called the independent components of the observed data.
- These independent components, also called sources or factors, can be found by ICA.
- ICA is superficially related to principal component analysis and factor analysis.
- ICA is a much more powerful technique, however, capable of finding the underlying factors or sources when these classic methods fail completely.
- The data analyzed by ICA could originate from many different kinds of application fields, including digital images, document databases, economic indicators and psychometric measurements.
- In many cases, the measurements are given as a set of parallel signals or time series; the term blind source separation is used to characterize this problem.
- Typical examples are mixtures of simultaneous speech signals that have been picked up by several microphones, brain waves recorded by multiple sensors, interfering radio signals arriving at a mobile phone, or parallel time series obtained from some industrial process.
- ICA is used for solving the Blind Source Separation (BSS) problem.
- We are given two linear mixtures of two source signals which we know to be independent of each other, i.e. observing the value of one signal does not give any information about the value of the other.
- The BSS problem is then to determine the source signals given only the mixtures.
- Putting this into mathematical notation, we model the problem by $\mathbf{x} = \mathbf{A}\mathbf{s}$
  where $\mathbf{s}$ is a two-dimensional random vector containing the independent source signals, $\mathbf{A}$ is the two-by-two mixing matrix, and $\mathbf{x}$ contains the observed (mixed) signals.

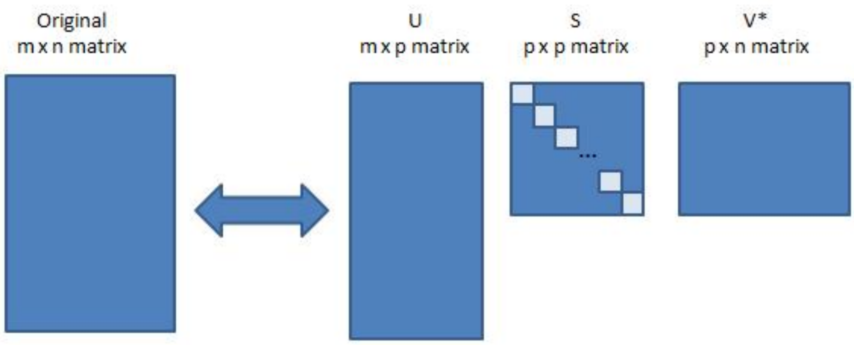


Difference between PCA and ICA

PRINCIPAL COMPONENT ANALYSIS	INDEPENDENT COMPONENT ANALYSIS
It is a statistical transformation that takes information from second order statistics.	It is a statistical transformation that takes information from high order statistics.
It reduces the dimensions to avoid the problem of overfitting.	It decomposes the mixed signal into its independent sources' signals.
It deals with the Principal Components.	It deals with the Independent Components.
PCA removes correlations, but not higher order dependence.	ICA removes correlations and higher order dependence.
It focuses on maximizing the variance.	It doesn't focus on the issue of variance among the data points.
Some components are more important than others (recall eigenvalues).	All components are equally important.
It focuses on the mutual orthogonality property of the principal components.	It doesn't focus on the mutual orthogonality of the components.
It doesn't focus on the mutual independence of the components.	It focuses on the mutual independence of the components.

6.4 Singular Value Decomposition (SVD)

- SVD, or Singular Value Decomposition, is one of several techniques that can be used to reduce the dimensionality, i.e., the number of columns, of a data set.
- Why would we want to reduce the number of dimensions? In predictive analytics, more columns normally means more time required to build models and score data. If some columns have no predictive value, this means wasted time, or worse, those columns contribute noise to the model and reduce model quality or predictive accuracy.
- Dimensionality reduction can be achieved by simply dropping columns, for example, those that may show up as collinear with others or identified as not being particularly predictive of the target as determined by an attribute importance ranking technique.
- But it can also be achieved by deriving new columns based on linear combinations of the original columns.
- In both cases, the resulting transformed data set can be provided to machine learning algorithms to yield faster model build times, faster scoring times, and more accurate models.
- While SVD can be used for dimensionality reduction, it is often used in digital signal processing for noise reduction, image compression, and other areas.
- SVD is an algorithm that factors an m x n matrix, M, of real or complex values into three component matrices, where the factorization has the form USV*.
- U is an m x p matrix.
- S is a p x p diagonal matrix.
- V is an n x p matrix, with V* being the transpose of V, a p x n matrix, or the conjugate transpose if M contains complex values.
- The value p is called the rank.
- The diagonal entries of S are referred to as the singular values of M.
- The columns of U are typically called the left-singular vectors of M, and the columns of V are called the right-singular vectors of M.
- Consider the following visual representation of these matrices:



- One of the features of SVD is that given the decomposition of *M* into *U*, *S*, and *V*, one can reconstruct the original matrix *M*, or an approximation of it.
- The singular values in the diagonal matrix *S* can be used to understand the amount of variance explained by each of the singular vectors.

Applications of Singular Value Decomposition (SVD)

- Image Compression
- Image Recovery
- Eigenfaces
- Spectral Clustering
- Background Removal from Videos

/

Q.1. Find the singular value decomposition of $A = \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix}$

→ Any real $m \times n$ matrix can be factored as

$$A = U \Sigma V^T$$

where U is an $m \times m$ orthogonal matrix whose columns are the eigen vectors of AA^T

V is an $n \times n$ orthogonal matrix whose columns are the eigen vectors of $A^T A$

Σ is an $m \times n$ diagonal matrix of the form

$$\Sigma = \begin{bmatrix} \sigma_1 & & & 0 \\ & \ddots & & \\ & & \sigma_r & \\ & & & 0 \\ 0 & & & \ddots \ 0 \end{bmatrix}$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and $r = \text{rank}(A)$.

Here, $\sigma_1, \cdots, \sigma_r$ are the square roots of eigen values of $A^T A$. They are called Singular values of A.

Compute $A^T A$:-

$$A^T A = \begin{bmatrix} 2 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$$

Compute eigen values such that $|A^T A - \lambda I| = 0$

$$\therefore \left| \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

i.e. $\begin{vmatrix} 5-\lambda & 3 \\ 3 & 5-\lambda \end{vmatrix} = 0$

$\therefore (5-\lambda)^2 - 3^2 = 0$

i.e. $25 - 10\lambda + \lambda^2 - 9 = 0$

i.e. $\lambda^2 - 10\lambda + 16 = 0$

i.e. $(\lambda - 8)(\lambda - 2) = 0$

$\therefore \lambda_1 = 8$ and $\lambda_2 = 2$

Now compute eigen vectors such that $\left[A^TA - \lambda_i I\right]\begin{bmatrix} x_i \\ y_i \end{bmatrix} = 0$

i.e. $\begin{bmatrix} 5-\lambda_i & 3 \\ 3 & 5-\lambda_i \end{bmatrix}\begin{bmatrix} x_i \\ y_i \end{bmatrix} = 0$

Use $\lambda_1 = 8$ $\quad \therefore \begin{bmatrix} 5-8 & 3 \\ 3 & 5-8 \end{bmatrix}\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$

$\therefore -3x_1 + 3y_1 = 0$ —— ①

$3x_1 - 3y_1 = 0$ —— ②

Note:- If we solve above equations, we will get eigen vectors as 0. To avoid this, we use equation of orthogonal transformation relation, which is $x_1^2 + y_1^2 = 1$ —— ③

From equation ①, $y_1 = \frac{3x_1}{3}$ i.e. $y_1 = x_1$ —— ④

Substitute this value of y_1 in equation ③

$\therefore x_1^2 + x_1^2 = 1$

i.e. $2x_1^2 = 1$

$\therefore \boxed{x_1 = 1/\sqrt{2}}$

From equation ④, $\boxed{y_1 = 1/\sqrt{2}}$

Now we take $\lambda_2 = 2$

$\therefore \begin{bmatrix} 5-2 & 3 \\ 3 & 5-2 \end{bmatrix}\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 0$

$\therefore 3x_2 + 3y_2 = 0$ —— ⑤

$3x_2 + 3y_2 = 0$ —— ⑥

Note:- If we solve above equations, we will get eigen vectors as 0. To avoid this, we use equation of orthogonal transformation relation, which is $x_2^2 + y_2^2 = 1$ —— ⑦

From equation ⑤, $y_2 = -\frac{3x_2}{3}$ i.e. $y_2 = -x_2$ —— ⑧

Substitute this value of y_2 in equation (7)

$$\therefore \ x_2^2 + (-x_2)^2 = 1$$

$$\text{i.e.} \quad 2x_2^2 = 1$$

$$\therefore \ \boxed{x_2 = 1/\sqrt{2}}$$

From equation (8),

$$\boxed{y_2 = -1/\sqrt{2}}$$

$$\therefore \text{ Eigen vectors } \quad v_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad v_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Rank of matrix $A = 2$

$$\therefore \ \sigma_1 = \sqrt{\lambda_1} = \sqrt{8} = 2\sqrt{2}$$

$$\sigma_2 = \sqrt{\lambda_2} = \sqrt{2}$$

We have $\quad Av_1 = \sigma_1 u_1$

$$\therefore \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = 2\sqrt{2} \ u_1$$

$$\text{i.e.} \quad u_1 = \frac{1}{2\sqrt{2}} \begin{bmatrix} 2\sqrt{2} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Similarly, $\quad Av_2 = \sigma_2 u_2$

$$\therefore \begin{bmatrix} 2 & 2 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \sqrt{2} \ u_2$$

$$\text{i.e.} \quad u_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 0 \\ -\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$\therefore \; A = U \Sigma V^T$

i.e. $A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 2\sqrt{2} & 0 \\ 0 & \sqrt{2} \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$

Q.2. Find the singular value decomposition of $A = \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix}$

$\longrightarrow$

Compute A^TA :-

$$A^TA = \begin{bmatrix} 3 & 4 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} = \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix}$$

Compute eigen values such that $|A^TA - \lambda I| = 0$

$$\therefore \left| \begin{bmatrix} 25 & 20 \\ 20 & 25 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = 0$$

i.e. $\begin{vmatrix} 25-\lambda & 20 \\ 20 & 25-\lambda \end{vmatrix} = 0$

$\therefore \; (25-\lambda)^2 - (20)^2 = 0$

i.e. $625 - 50\lambda + \lambda^2 - 400 = 0$

i.e. $\lambda^2 - 50\lambda + 225 = 0$

i.e. $(\lambda - 45)(\lambda - 5) = 0$

$\therefore \; \lambda_1 = 45$ and $\lambda_2 = 5$

Now compute eigen vectors such that $[A^TA - \lambda_i I] \begin{bmatrix} x_i \\ y_i \end{bmatrix} = 0$

i.e. $\begin{bmatrix} 25-\lambda_i & 20 \\ 20 & 25-\lambda_i \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} = 0$

Use $\lambda_1 = 45$

$$\therefore \begin{bmatrix} 25-45 & 20 \\ 20 & 25-45 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$\therefore -20x_1 + 20y_1 = 0 \quad \text{——— (1)}$$
$$20x_1 - 20y_1 = 0 \quad \text{——— (2)}$$

Note :- If we solve above equations, we will get eigen vectors as 0. To avoid this, we use equation of orthogonal transformation relation, which is $x_1^2 + y_1^2 = 1$ ——— (3)

From equation (1), $y_1 = \frac{20x_1}{20}$ i.e. $y_1 = x_1$ ——— (4)

Substitute this value of y_1 in equation (3)

$$\therefore x_1^2 + x_1^2 = 1$$
$$\text{i.e. } 2x_1^2 = 1$$
$$\therefore \boxed{x_1 = 1/\sqrt{2}}$$

From equation (4),

$$\boxed{y_1 = 1/\sqrt{2}}$$

Now we take $\lambda_2 = 5$

$$\therefore \begin{bmatrix} 25-5 & 20 \\ 20 & 25-5 \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = 0$$

$$\therefore 20x_2 + 20y_2 = 0 \quad \text{——— (5)}$$
$$20x_2 + 20y_2 = 0 \quad \text{——— (6)}$$

Note :- If we solve above equations, we will get eigen vectors as 0. To avoid this, we use equation of orthogonal transformation relation, which is $x_2^2 + y_2^2 = 1$ ——— (7)

From equation (5), $y_2 = -\frac{20x_2}{20}$ i.e. $y_2 = -x_2$ ——— (8)

Substitute this value of y_2 in equation 7

$$\therefore \; x_2^2 + (-x_2)^2 = 1$$

$$\text{i.e. } 2x_2^2 = 1$$

$$\therefore \boxed{x_2 = 1/\sqrt{2}}$$

From equation 8,

$$\boxed{y_2 = -1/\sqrt{2}}$$

$\therefore$ Eigen vectors $v_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$ and $v_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$

Rank of matrix $A = 2$

$$\therefore \sigma_1 = \sqrt{\lambda_1} = \sqrt{45}$$

$$\sigma_2 = \sqrt{\lambda_2} = \sqrt{5}$$

We have, $A v_1 = \sigma_1 u_1$

$$\therefore \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \sqrt{45}\, u_1$$

$$\therefore u_1 = \frac{1}{\sqrt{45}} \begin{bmatrix} 3/\sqrt{2} \\ 9/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 1/\sqrt{10} \\ 3/\sqrt{10} \end{bmatrix} = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Similarly, $A v_2 = \sigma_2 u_2$

$$\therefore \begin{bmatrix} 3 & 0 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \sqrt{5}\, u_2$$

$$\therefore u_2 = \frac{1}{\sqrt{5}} \begin{bmatrix} 3/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{10} \\ -1/\sqrt{10} \end{bmatrix} = \frac{1}{\sqrt{10}} \begin{bmatrix} 3 \\ -1 \end{bmatrix}$$

We have, $A = U \Sigma V^T$

$$\therefore A = \frac{1}{\sqrt{10}} \begin{bmatrix} 1 & 3 \\ 3 & -1 \end{bmatrix} \begin{bmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{bmatrix} \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$