PAGE NO .....

Unit -3

Introduction to Statistical Learning Theory

Features are the variables that can be defined and observed. Feature extraction is the process of identifying and selecting the most important information or characteristics from a data-set without losing vital information.

Feature extraction essentially is the process of converting raw data into numerical features that can be processed while preserving the information in the original data-set. Feature extraction is helpful when training a machine learning model. It leads to:

→ An improvement in model accuracy
→ A Boost in Training speed
→ A reduction in risk of overfitting
→ Better data visualization

Feature extraction techniques:

- Principle Components Analysis (PCA)
- Independant Component Analysis (ICA)
- Linear Discriminant Analysis (LDA)
- Locally Linear Embedding (LLE)
  etc.

Challenges in Feature Extraction:

↳ Choosing the right method
↳ Loss of information
↳ Computational complexity

Dimensionality reduction :- In machine learning model, the higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, many of these features are correlated or redundant. This is where dimensionality reduction algorithms come into play.

Dimensionality reduction can be done in 2 ways:

→ Feature extraction
→ Feature selection

The feature extraction technique gives us new features which are a linear combination of the existing features. The new set of features will have different values as compared to the original feature value. The main aim is that fewer features will be required to capture the same info.

PAGE NO. .....

Principal Component Analysis :-

PCA used in data analysis, particularly for reducing the dimensionality of datasets while preserving crucial information. It does; by transforming the original variables into a set of new, uncorrelated variables called principal components.

PCA key aspects :-

↳ Dimensionality Reduction
↳ Linear transformation
↳ feature selection
↳ Data Compression
↳ Matrix Requirement PCA work with symmetric correlation or covariance matrices and requires numeric, standardized data.
↳ Eigen values and Eigen vectors Eigen values represent variance magnitude, and Eigen vectors indicate variance direction.

⇒ Limitation : It may loss information.

How PCA work?

1. Standardize the data
   If the features of dataset are on different scales, it's essential to standardize them (subtract the mean and divide by the standard deviation).
2. Compute the covariance matrix
   Calculate the covariance matrix for the standardized dataset.
3. Compute Eigen vectors and Eigenvalues of covariance matrix.
4. Sort Eigenvectors by Eigenvalues in descending order
5. Choose principal components
   Select the top K eigenvectors (principal components) where K is the desired dimensionality of the reduced dataset.
6. Transform the Data
   Multiply the original standardized data by the selected principal components to obtain the new, lower-dimensional representation of the data.

PAGE NO. .....

Singular Value Decomposition :-

SVD is primarily used for dimensionality reduction, information extraction, and noise reduction.

Importance in Data Science :

→ Dimensionality Reduction
→ Noise Reduction
→ Recommendation System

SVD is a method for decomposing a matrix :

For a given matrix A, it can be decomposed into three matrices U, Σ and V^T such that

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

Given a matrix A of dimension m x n, the SVD decomposes it into three matrices :

1. U : An m x m orthogonal matrix, called the left singular matrix.
2. Σ : An m x n diagonal matrix, with non-negative real numbers on the diagonal. The values on the diagonal are the "singular values" and are usually ordered in decreasing order.

3. V^T: An $n \times n$ orthogonal matrix, where V^T is the transpose of V, called the right singular matrix.

So, the decomposition can be represented as:

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

The columns of U are called the left singular vectors, the columns of V are called the right singular vectors, and the non-zero values in Σ are the singular values of A.

Eg:-

Compute the SVD for small matrix

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$$

(using Numpy we get the values)

$$\rightarrow U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

A is already a diagonal matrix, So SVD just gives us back the matrix A itself in Σ and U, V^T are identity matrix.

for Non-diagonal matrices, SVD decomposition will complex but method will same.

→ Decompose any matrix into three other matrices representing orthogonal, transformation and Scaling.

PAGE NO. .....

The diagonal element of Σ are known as singular values and are non-negative. The columns of U and V^T are ~~ortho~~ orthonormal eigenvectors of AA^T and A^TA, respectively.

$$M = [U_1\ U_2 \ldots U_n] \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix} [U_1\ U_2 \ldots U_n]^T$$

eigenvectors — eigenvalues

$U.U^T = I$
$V.V^T = I$

for eg:

$$\begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$A^T.A = (V\Sigma^T U^T).(U\Sigma V^T)$$

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

$$A.A^T = (U\Sigma V^T).(V\Sigma^T U^T)$$

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

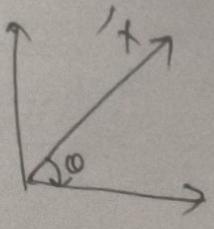
rotation . Stretch . rotation

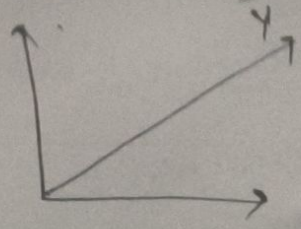
$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}$ ← unitary transformation

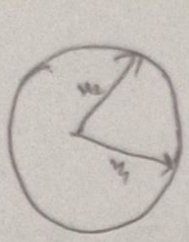
Stretch

$\begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}$

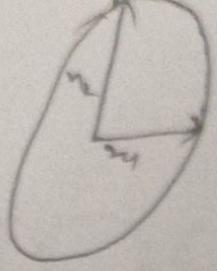
Rotation







after streching



Eg:- $A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$

we compute ev, $e\vec{v}$

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

from def^n

$A^T \cdot A = V$

~~[illegible]~~

$A \cdot A^T = U$

$A \cdot A^T = \begin{bmatrix} 11 & 1 \\ 1 & 11 \end{bmatrix}$

Characteristic eq.

$(A - \lambda I) = 0$

$\begin{bmatrix} (11-\lambda) & 1 \\ 1 & (11-\lambda) \end{bmatrix} = 0$

$(11-\lambda)^2 - 1^2 = 0$

$(11-\lambda+1)(11-\lambda-1) = 0$

$11 - \lambda + 1 = 0 \Rightarrow \boxed{12 = \lambda_1}$ ev

$11 - \lambda - 1 = 0 \Rightarrow \boxed{10 = \lambda_2}$

Put the λ values in eq.

~~[illegible]~~ $\lambda_2 = 12$ and $\lambda_1 = 10$

$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \sqrt{1^2 + 1^2}$

$= \sqrt{1^2 + (-1)^2}$

Orthogonal done with the help of Gram-Schmidt orthogonalization process.

POORNIMA
COLLEGE OF ENGINEERING
DETAILED LECTURE NOTES

PAGE NO. .....

$\sqrt{1^2+1^2}$ $\sqrt{1+(-1)^2}$

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} = U$$

$$A = \begin{bmatrix} 3 & 1 & 1 \\ -1 & 3 & 1 \end{bmatrix}$$

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

$(A - \lambda I)X = 0$

$$A^T \cdot A = \begin{bmatrix} 10 & 0 & 2 \\ 0 & 10 & 4 \\ 2 & 4 & 2 \end{bmatrix}$$

minor of diagonal

$4 + 16 + 100 = 120$

$\begin{bmatrix} 10 & 4 \\ 4 & 2 \end{bmatrix} \begin{bmatrix} 10 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$

$$\begin{bmatrix} 10-\lambda & 0 & 2 \\ 0 & (10-\lambda) & 4 \\ 2 & 4 & (2-\lambda) \end{bmatrix}$$

$$\lambda^3 - S_1\lambda^2 + S_2\lambda - S_3$$

trace (A) — minor of diagonal — determinant of (A)

8-

$$\begin{matrix} -2 & 0 & 2 \\ 0 & -2 & 4 \\ 2 & 4 & -10 \end{matrix}$$

trace $= 10 + 10 + 2 = 22$

$\det(A) = 10(4) + 2(-20)$

$= 40 - 40 = 0$

we get

$\lambda^3 - 22\lambda^2 + 120\lambda = 0$

$\lambda(\lambda^2 - 22\lambda + 120) = 0$

$\lambda_1 = 0$

$\lambda^2 - 22\lambda + 120 = 0$

$\lambda_2 = 12$

$\lambda_3 = 10$

we get λ values

$$\begin{bmatrix} -2 & 0 & 2 \\ 0 & -2 & 4 \\ 2 & 4 & -10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

$(A - \lambda I)X = 0$

By Cramers rule

$$\frac{x_1}{4} = \frac{-x_2}{-8} = \frac{x_3}{4}$$

Divide by 4

$$\frac{1}{1} = \frac{2}{2} = \frac{1}{1}$$

$$V = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 2 \\ 1 & 0 & -5 \end{bmatrix}$$

$$V^T = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 0 \\ 1 & 2 & -5 \end{bmatrix} \begin{matrix} \sqrt{6} \\ \sqrt{5} \\ \sqrt{30} \end{matrix}$$

$$V^T = \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{5}} & \frac{0}{\sqrt{5}} \\ \frac{1}{\sqrt{30}} & \frac{2}{\sqrt{30}} & \frac{-5}{\sqrt{30}} \end{bmatrix}$$

$\sqrt{8}$ $\sqrt{}$

20 + 16

$20 - 4$

42 16

$\frac{4+16}{2}$ (4)

$16 + 4$

$\sqrt{20}$

$\sqrt{8}$ $\sqrt{8}$

$\sqrt{20}$ $16+4$

$\sqrt{4+16+100}$

$\sqrt{120}$

$\sqrt{8}$ $\sqrt{20}$ $\sqrt{20}$

$2\sqrt{2}$ $2\sqrt{5}$ 4×2 4×5 $2\sqrt{50}$

$\sqrt{2}$ $\sqrt{50}$

$\sqrt{2}$ $\sqrt{5}$ $\sqrt{50}$

$-2x_4 + 2x_4 = 0$

POORNIMA
COLLEGE OF ENGINEERING
An autonomous institution approved by RTU, AICTE & UGC • NAAC A+ Accredited

DETAILED LECTURE NOTES

PAGE NO.....

Principal Component Analysis:-

For the given dataset in table, reduce the dimension from 2 to 1 using PCA.

Feature	Eg 1	Eg 2	Eg 3	Eg 4
x	4	8	13	7
y	11	4	5	14

No. of features = 2
No. of samples = 4

Step 1: Calculate mean of each feature

$\bar{x} = (4+8+13+7)/4 = 8$

$\bar{y} = 8.5$

Step 2: computation of co-variance matrix,

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

$$cov(x,x) = \frac{1}{N-1} \sum_{i=1}^{N} (4-8)^2 + (8-8)^2 + (13-8)^2 + (7-8)^2$$

$$= \frac{1}{3}(16 + 0 + 25 + 1) = 14$$

$$cov(x,y) = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

$$cov(x,y) = \frac{1}{N-1} \sum_{i=1}^{N} (4-8.5)(11-8.5) + (8-8)(4-8.5) + (13-8)(5-8.5) + (7-8)(14-8.5)$$

$$= -11$$

$cov(y,y) = 23$

$cov(y,x) = -11$

$$C = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}$$

Step 3: Compute the eigen vector and eigenvalues of the covariance matrix.

i) Eigen value:

$$\det(C - \lambda I) = 0$$

$$\begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = 0$$

$$\frac{1}{29}\sqrt{b^2 - 4ac}$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} = 0$$

$$\lambda^2 - 37\lambda + 201 = 0$$

$$\Rightarrow \lambda_1 = 30.38, \quad \lambda_2 = 6.61$$

} Arrange in descending order

Step 4: Computation of eigen vectors:

eigenvector corresponding to λ_1 is a

$$u_1 = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$(C - \lambda_1 I)\, u_1 = 0$$

$$\begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = 0 \Rightarrow \begin{matrix} (14-\lambda)u_1 + (-11)u_2 = 0 \quad (1) \\ -11u_1 + (23-\lambda)u_2 = 0 \quad (2) \end{matrix}$$

from ①

$$\frac{u_1}{11} = \frac{u_2}{14-\lambda} = t \quad ; \quad U_1 = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix}$$

$u_1 = 11t$

$u_2 = (14-\lambda)t$

[we consider largest value of λ]

To find a unit eigenvector, we compute the length $= U_1$

$$\|U_1\| = \sqrt{11^2 + (14-\lambda)^2}$$

$$\|U_1\| = 19.7348$$

Unit eigenvector

$$e_1 = \begin{bmatrix} 11/\|U_1\| \\ (14-\lambda_1)/\|U_1\| \end{bmatrix}$$

$$= \begin{bmatrix} 11/19.73 \\ 14-30.38/19.73 \end{bmatrix}$$

$$e_1 = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

for $\lambda = 5$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 5 : Computation of first Principal Component

$$e_1^T \begin{bmatrix} x_{1k} - \bar{x}_1 \\ x_{2k} - \bar{x}_2 \end{bmatrix}$$

$$= [0.5574 \quad -0.8303] \begin{bmatrix} x_{11} - \bar{x}_1 \\ x_{21} - \bar{x}_2 \end{bmatrix}$$

$$= 0.5574 (x_{11} - \bar{x}_1) - .8303 (x_{21} - \bar{x}_2)$$

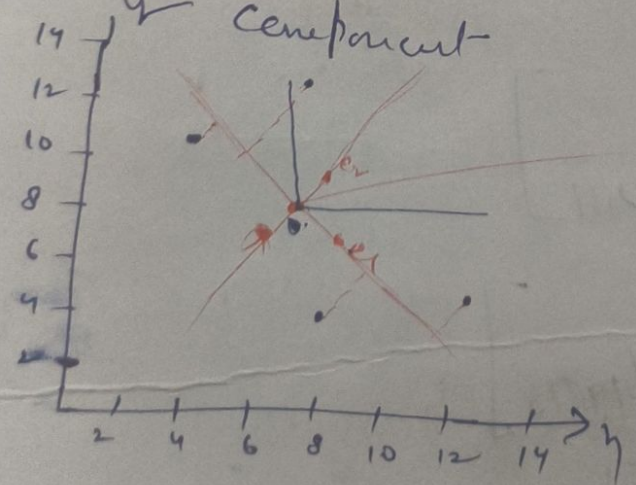
$$= .55 (4-8) - .83 (11-8.5)$$

$$= -4.3053$$

So The result of calculation is summarized in table:

x	4	8	13	7
y	11	4	5	14
PC	-4.3053	3.7361	5.6728	-5.1230

Step 6 : Geometric meaning of first principal component



PCA in machine learning:

--PCA stands for Principal Component Analysis.

--PCA, or Principal Component Analysis, is a common technique used in machine learning for dimensionality reduction.

--Dimensionality reduction refers to the process of reducing the number of features or variables in a dataset while preserving the most important information. This is important because datasets with a large number of features can be difficult to work with and can lead to overfitting and decreased performance of machine learning models.

--remove the problem of curse of dimensionality. The curse of dimensionality refers to the difficulties that arise when working with high-dimensional data in machine learning. As the number of

dimensions (or features) of a dataset increases, the amount of data required to obtain statistically significant results increases exponentially. This can lead to problems such as overfitting, increased computational complexity, and reduced model performance. This makes it difficult to find meaningful patterns or relationships within the data, and it becomes more challenging to distinguish between relevant and irrelevant features.

--PCA works by identifying the directions in the data that contain the most variance, and then projecting the data onto a new coordinate system that is aligned with these directions. The new coordinate system consists of a set of orthogonal axes (principal components) that capture the largest possible variance in the data. The first principal component captures the most variance, followed by the second, and so on.

Advantages of PCA Algorithm: PCA (Principal Component Analysis) is a widely used technique in machine learning for dimensionality reduction. Here are some advantages of using PCA:

1. Reduces the number of features: PCA can help reduce the number of features in a dataset while retaining most of the information. This can be particularly useful when dealing with high-dimensional data, where the number of features is much larger than the number of samples.

2 Improves model performance: By reducing the number of features, PCA can help improve the performance of machine learning models. This is because it reduces the risk of overfitting, which can occur when there are too many features relative to the number of samples.

3. Speeds up training: With fewer features, training machine learning models can be faster and more efficient. This is because there are fewer calculations to perform, and the resulting models are simpler and easier to interpret.

4. Helps with visualization: PCA can be used to visualize high-dimensional data in two or three dimensions. This can help identify patterns and relationships in the data that may not be apparent in higher dimensions.

5. Removes correlated features: PCA can help remove correlated features, which can cause problems for some machine learning algorithms. By removing these features, PCA can help improve the stability and accuracy of the resulting models.

data. This can lead to suboptimal results if the dataset contains outliers.

to understand the underlying patterns in the data or explain the results of a machine learning model to others.

3. Assumes linearity: PCA assumes that the underlying relationships between the features are linear. However, in some cases, the relationships may be non-linear, which can lead to suboptimal results.

4. Computationally expensive: PCA involves computing the eigenvectors and eigenvalues of the covariance matrix of the data, which can be computationally expensive for large datasets. This can make PCA impractical for some applications.

5. Sensitivity to outliers: PCA is sensitive to outliers, as outliers can have a large influence on the covariance matrix of the

Disadvantages of PCA Algorithm: While PCA (Principal Component Analysis) is a powerful technique for dimensionality reduction in machine learning, it also has some disadvantages. Here are some of the main disadvantages of PCA:

1. **Information loss:** PCA works by finding a new set of features that capture most of the variance in the data. However, this can result in some information loss, as the new features may not capture all of the information in the original features. This can be particularly problematic if the discarded information is important for the task at hand.

2. **Interpretability:** The new features generated by PCA are linear combinations of the original features, which can make them difficult to interpret. This can be a problem if we want