UNIT- 5

Lecture: 35

Introduction to Pattern Recognition

→ Patterns are recognised by the help of algorithm used in Machine learning. Recognizing patterns is the process of classifying the data based on the model that is created by training data, which then detects patterns & characteristics from the patterns.

→ Pattern Recognition is the process which can detect different categories and get information about particular data.

→ Features of Pattern Recognition :-

1) Pattern Recognition learns from the data.
2) Automatically recognizes the patterns even when partially visible.
3) Should be able to recognise pattern which are familiar.
4) The pattern should be recognized from different angles & shapes.

→ Applications of Pattern Recognition :-

1) Computer vision
2) Engineering
3) Speech Recognition
4) Image Recognition
5) Biometrics.
6) Geology.
7) Intrusion Detection.
8) Civil Administration

Lecture : 36.

* Design Principles of Pattern Recognition :

In pattern Recognition systems, for recognising patterns or structures 2 basic approaches are used which can be implemented in different techniques. These are

1) Statistical Approach
2) Structural Approach

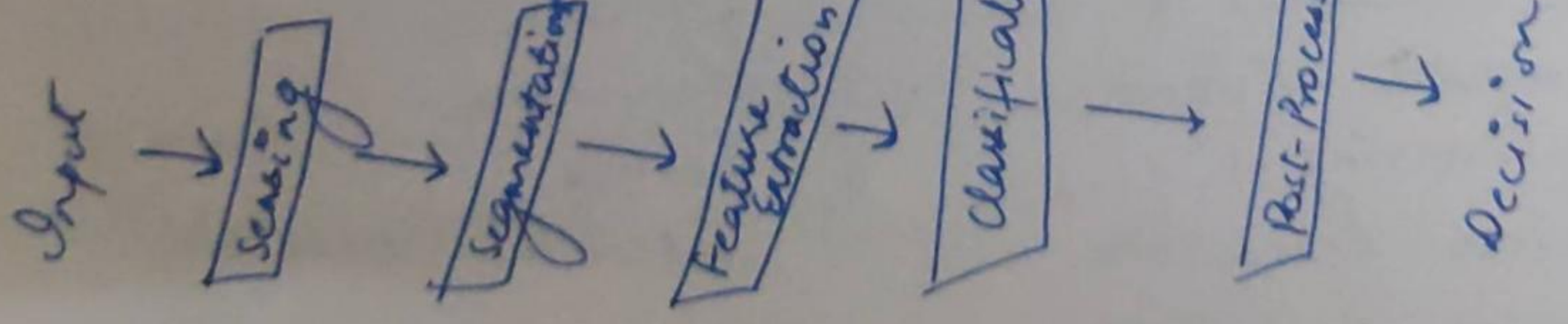
1) Statistical Methods are mathematical formulas, models & techniques that are used in statistical analysis of raw research data. The application of statistical method extract information from research data and provides different ways to assess the robustness of research output.

The 2 main statistical methods are:

a) Descriptive - summarizes data from a sample using indexes.
b) Inferential - Draws conclusion from data that are subject to random variation.

2) Structural Approach is a technique wherein the learner masters the pattern of sentence. structures are different arrangements of words in one accepted style or others. Some of them are sentence or Phrase Pattern, Formulas and Idioms.

* Phases in Pattern Recognition :



Lecture : 37

Statistical Pattern Recognition

→ It refers to the use of statistics to learn from exmpl
It means to collect observation, study & digest them
in order to infer general rules / concepts that can be
applied to the new, unseen observation.

→ It relates to the use of statistical techniques for
analysing data measurement in order to extract info.
and justified decision.

→ Statistical approach in PR is represented in
D features in d dimensional space as a point.
→ objective to establish decision boundaries in feature
space which separate pattern of different classes.
→ Discriminate Analysis based approach for classificati
and uses mean squared error criteria.
→ Construct the decision boundaries of the specified
form.

→ Statistical Approach follows up:
1) Linear Discriminant Analysis (LDA)
2) Principal Component Analysis (PCA)
3) Posterior Probability Estimation
4) Window - Density based classifier.
5) Edited K-NN Rule.

Lecture: 38.

(PEM)
Parameter Estimation Method.

① Principal Component Analysis } → It is a process of computing the principal components & using them to perform a change of basis on the data.

↓

There are collection of points in a real p-space are a sequence of p-direction vectors.

→ used in exploratory data analysis
→ used in making predictive models
→ commonly used for dimensionality reduction.

↓

By projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible.

→ PCA is defined as orthogonal linear transformation that transform the data to a new co-ordinate system such that the greatest variances by some scalar projection of the data comes to lie on the first co-ordinate; similarly the second variance on the second co-ordinate.

Numerical Example: PCA

* Consider the dataset with 5 observation of Math, English & Arts marks; arranged in matrix A

$$A = \begin{bmatrix} 90 & 60 & 90 \\ 90 & 90 & 30 \\ 60 & 60 & 60 \\ 60 & 60 & 90 \\ 30 & 30 & 30 \end{bmatrix}$$

Matrix A

$$\bar{A} = [66 \quad 60 \quad 60]$$

Mean of matrix A

- Compute the covariance matrix of the whole dataset

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

find the covariance matrix of A, the result would be square matrix of d x d dimension.

→ Covariance Matrix —

	M	E	A
M	504	360	180
E	360	360	0
A	180	0	720

least variance — more variance

→ Compute Eigen values & Eigen vectors.

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

$$\det\left(\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\right) = \det\begin{pmatrix} 504 & 360 & 180 \\ 360 & 360 & 0 \\ 180 & 0 & 720 \end{pmatrix} - \begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix}$$

$$= \det\begin{pmatrix} 504-\lambda & 360 & 180 \\ 360 & 360-\lambda & 0 \\ 180 & 0 & 720-\lambda \end{pmatrix}$$

Determinant : $-\lambda^3 + 1584\lambda^2 - 641520\lambda + 25660800 = 0$

on solving we get -

Eigen values $\lambda \simeq 44.82 ; 629.11 ; 910.069$ (appx.)

- Solving eigenvector we get -

$$\begin{pmatrix} -3.75100 \\ 4.28441 \\ 1 \end{pmatrix} \begin{pmatrix} -0.50494 \\ -0.67548 \\ 1 \end{pmatrix} \begin{pmatrix} 1.05594 \\ 0.69108 \\ 1 \end{pmatrix}$$

- Sort the eigenvectors by decreasing eigenvalues & choose k eigenvectors with the largest eigenvalue to form a $d \times k$ dimensional matrix W

- After sorting the eigenvalues in decreasing order,

$$\begin{pmatrix} 910.06945 \\ 629.11039 \\ 44.81966 \end{pmatrix}$$

- Eigen vectors corresponding to 2 max. eigenvalues are :

$$W = \begin{bmatrix} 1.05594 & -0.50494 \\ 0.69108 & -0.67548 \\ 1 & 1 \end{bmatrix}$$

- Transform the samples onto the new subspace.

$$y = W' \times x$$

↓

transpose of the matrix W

Lecture: 39

PEM ② Linear Discriminant Analysis

→ The objective of LDA is to perform dimensionality reduction while preserving as much of the class discriminatory info. as possible.

→ most famous example of dimensionality reduction is - "PCA"; but question is how to utilize the label info. in finding info. projections? Thus, for that purpose - fisher-LDA consider maximizing the following obj.

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

S_B → B/w classes scatter matrix.

S_W → within classes scatter matrix.

→ LDA approximates the Bayes classifier one by modeling conditional class densities as multivariate normals. For classes $C \in 1 \ldots K$ & a feature vector $X \in \mathbb{R}^p$. This can be expressed:

$$P(X = x \mid C = j) = N(\mu_j, \Sigma)$$

, each class j has its own mean $\mu_j \in \mathbb{R}^p$ but classes together share a covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$

→ Various steps of LDA process are:

1. Centre Data
2. Calculate Estimators
3. Sphere variables
4. Project using P_m — (covariance matrix Δ)

Example: Fisher Linear Discriminant Analysis.

Data: 1) Class 1 has 5 samples $c_1 = [(1,2), (2,3), (3,3), (4,5), (5,5)]$
2) " 2 has 6 " $c_2 = [(1,0), (2,1), (3,1), (3,2), (5,3), (6,5)]$

Step 1: Arrange data in 2 separate matrices:

$$C_1 = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 3 \\ 4 & 5 \\ 5 & 5 \end{bmatrix} \qquad C_2 = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & 1 \\ 3 & 2 \\ 5 & 3 \\ 6 & 5 \end{bmatrix}$$

* PCA performs poorly on this due to direction of largest variance is not helpful for classification.

Step 2: Compute the mean for each class

$$\mu_1 = \text{mean}(C_1) = [3 \quad 3.6]$$
$$\mu_2 = \text{mean}(C_2) = [3.3 \quad 2]$$

Step 3: Compute the scatter matrices S_1 & S_2 for each class

$$S_1 = 4 * \text{cov}(C_1) = \begin{bmatrix} 10 & 8.0 \\ 8.0 & 7.2 \end{bmatrix}$$

$$S_2 = 5 * \text{cov}(C_2) = \begin{bmatrix} 17.3 & 16 \\ 16 & 16 \end{bmatrix}$$

Step 4: Within the class scatter,

$$S_w = S_1 + S_2 = \begin{bmatrix} 27.3 & 24 \\ 24 & 23.2 \end{bmatrix}$$

No need to solve for eigenvalues, as it has full rank

Step 5: Inverse $(S_w^{-1}) = \begin{bmatrix} 0.39 & -0.41 \\ -0.41 & 0.47 \end{bmatrix}$

$$v = S_w^{-1}(\mu_1 - \mu_2) = \begin{bmatrix} -0.79 \\ 0.89 \end{bmatrix}$$

→ As long as the line has the right direction its exact position doesn't matter

Step 6: Compute the actual 1D vector y for each class.

$$y_1 = v^t C_1^t = [-0.65 \quad 0.73] \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 3 & 5 & 5 \end{bmatrix} = [0.81 \quad 0.4]$$

$$y_2 = v^t C_2^t = [-0.65 \quad 0.73] \begin{bmatrix} 1 & 2 & 3 & 3 & 5 & 6 \\ 0 & 1 & 1 & 2 & 3 & 5 \end{bmatrix} = [-0.65 \quad -0.25]$$

— X —

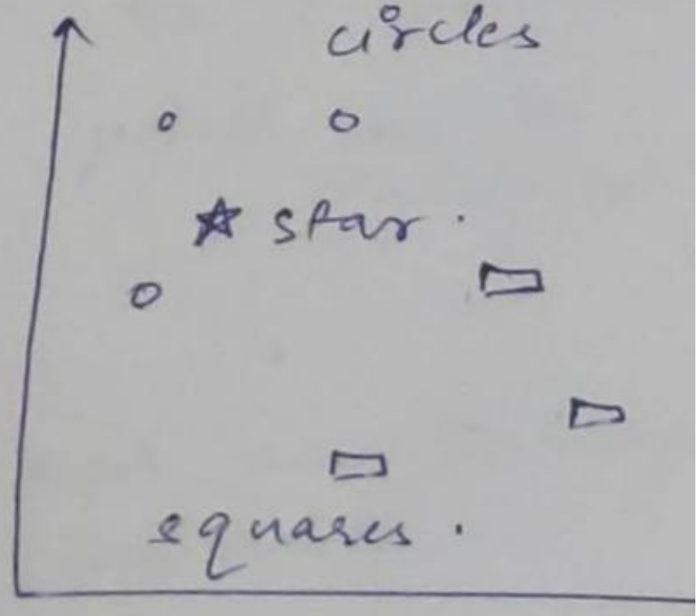
Lecture : 40.

Classification Technique : Nearest Neighbour Rule

→ It is the simplest decision procedures — NN Rule.
① It classifies a sample based on the category of its nearest neighbour.
② These classifier essentially involve finding the similarity b/w the test pattern & every pattern in training set.
③ Ease to get and interpret output.
↓
Q calculation time is less
↳ ⑤ More predictive power

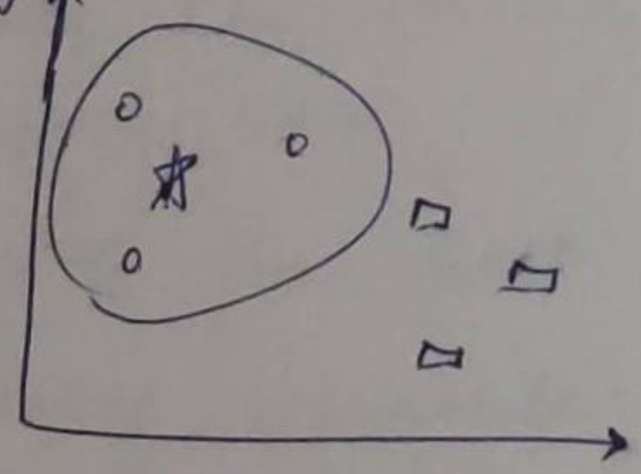
→ It can be used for both the classification & regression predictive problems.

→ for eg: Find the class of star.



can be circle square

say k = 3 for first time.



① • As all the circles are nearest to the star. So, the confidence level is high.

• So, the choice becomes very obvious.

• As the observations become constant, we can make boundaries of each class.

• Gradually after each iteration the boundaries gets specific and smoother and clean; sorted form.
• Training depends on no of observation.
• Get the predicted class at end.

Lecture: 41

Bayes classifier

→ Important for statistical learning.

→ A classifier is a machine learning model that is used to discriminate different object based on certain features.

→ A Naive Bayes classifier is a probabilistic M/c learning model that's used for classification task.

→ Base is on Bayes theorem:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

→ Using this Bayes Theorem, we can find the prob. of A happening given that B has occurred. Here B is the evidence & A is the hypothesis.

→ Here, the predictors are independent. It means presence of one feature doesn't affect the other.

Lecture: 42

Support Vector Machine

→ It is a supervised learning technique that can be used for both classification & regression problem.

→ Basic Concepts :

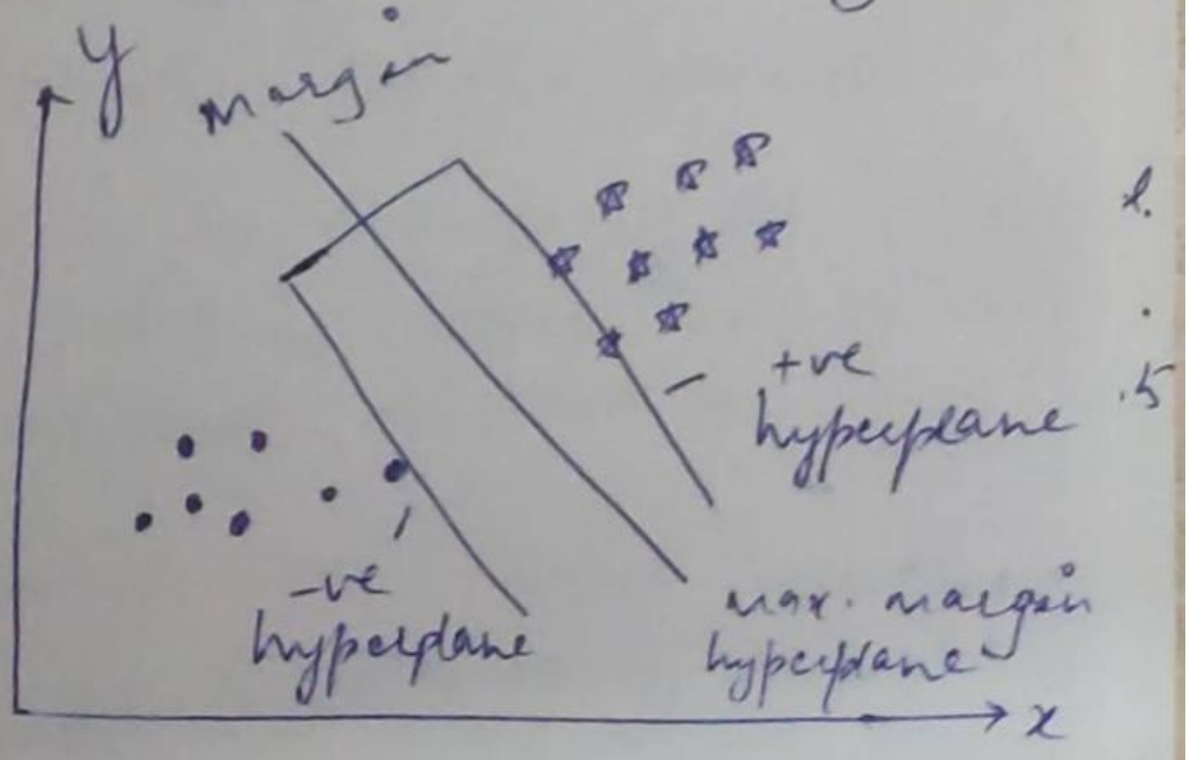
a) Support vectors - These are simply the co-ordinates of individual observation. These are near to hyperplane and influence the position & orientation of hyperplane.

b) Hyperplane - These are the decision boundaries that help classify the data points

c) Margin - This is one ⊥lar from both hyperplane & more the margin more better machine. As it provides some re-inforcement so that future data points can be classified.

→ Working of SVM -

1) Take the linearly separable data and non-linearly separable.
2) Apply SVM on both
3) With the help of SVM kernel tricks to make it linearly separable. and then solve.
4) It take data points in consideration. & gives out a hyperplane, which divides the both classes.
5) Many hyperplanes can be generated but the best is the one which divided the classes from maximum distance.



Lecture: 43

K- Means clustering

→ It is an iterative algorithm that tries to partition the dataset into 'k'- pre-defined distinct non-overlapping sub-groups where each data points belong to only one group.

→ It tries to make the intra-cluster data points as similar as possible while keeping the cluster as different as possible.

→ It assign data points to a cluster such that the sum of squared distance b/w the data points and cluster's centroid is at the minimum.

→ The less variation we've within cluster the more homogeneous the data points are within the same cluster.

→ Working of K- means:

1) Specify the no of clusters k.
2) Initialize centroids by 1st shuffling the dataset & then randomly selecting k data points for the centroids.
3) Keep iterating until there is no change to the centroid. i.e the assignment of data point to cluster isn't changing.
   - Compute the sum of squared distance b/w data points & centroid.
   - Assign each data point to the closest cluster.
   - Compute the centroid for the cluster by taking the avg. of the all data points that belong to each cluster.