UNIT – 5

Pattern Recognition:

- Introduction of Design principles of pattern recognition system
- Statistical Pattern recognition
- Parameter estimation methods
- Principle Component Analysis (PCA)
- Linear Discriminant Analysis (LDA).

Classification Techniques

- Nearest Neighbor (NN) Rule
- Bayes Classifier
- Support Vector Machine (SVM)
- K – means clustering.

ANUJ KHANNA
ASSISTANT PROFESOR(KIOT, Kanpur)

Short Questions & Answers

Ques1. What is pattern recognition?

Ans. Pattern recognition is a branch of ***machine learning*** that focuses on the recognition of patterns and regularities in ***data***,It is the study of how machines can observe the environment intelligently , learn to distinguish patterns of interest from their backgrounds and make reasonable & correct decisions about the different classes of objects. Patterns may be a finger print image, handwritten cursive word, a human face, iris of human eye ora speech signal. These examples are called input stimuli. Recognition establishes a close match between some new stimulus and previously stored stimulus patterns. Pattern recognition systems are in many cases trained from labeled "training" data ***(supervised learning),*** but when no labeled data are available other algorithms can be used to discover previously unknown patterns ***(unsupervised learning).At the most abstract level patterns can also be some ideas , concepts , thoughts , procedures Activated in human brain and body. This is known as the study of human psychology (Cognitive Science)***

Example: In automatic sorting of integrated circuit amplifier packages, there can be three possible types : metal –cane , dual –in-line and flat pack. The unknown object should be classified as being one of these types.

Ques 2. Define Measurement space and Feature space in classification process for objects.

Ans: Measurement space: This is the set of all pattern attributes which are stored in a vector form.

It is a range of characteristic attribute values. In vector form measurement space is also called observation space /data space. E.g : W = [W_1 , W_2 ,……,Wn-1, W_n] for n pattern classes.

W is a pattern vector. Let X = $\begin{bmatrix} x1 \\ x2 \end{bmatrix}$, X is a pattern vector for flower , x1 is petal length and x2 is petal width.

Feature Space: The range of subset of attribute values is called Feature Space F. This subset represents a reduction in attribute space and pattern classes are divided into sub classes. Feature space signifies the most important attributes of a pattern class observed in measurement space.

Ques 3. What is dimensionality reduction problem?

Ans. In machine learning classification problems, there are often too many factors on the basis of which the final classification is done. These factors are basically variables called features. ***The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant.*** This is where dimensionality reduction algorithms come into play. ***Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables***. It can be divided into feature selection and feature extraction. The various methods used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

Ques 4. State some advantages and disadvantage with application of LDA.

Ans. Advantages of Linear Discriminant Analysis

- Suitable for larger data set.
- Calculations of scatter matrix in LDA is much easy as compared to covariance matrix

Disadvantages of Linear Discriminant Analysis

- \More redundancy in data.
- Memory requirement is high.
- More Noisy.
-

Applications Of Linear Discriminant Analysis

- Face Recognition.
- Earth Sciences.
- Speech Classification.

Ques 5. Write some disadvantages of K Nearest Neighbor.

Ans. Disadvantages of Using K-NN

(a).Expensive. (b) High Space Complexity (c)High Time Complexity.
(d)Data Storage Required . (e) High-Dimensionality of Data

Ques 7. How K-Mean is different by KNN.

Ans.

K- Means Clustering	K- Nearest Neighbor Classification
1. This is an unsupervised learning technique	Supervised Learning Technique
2. All the variables are independent	All the variables are dependent
3. Splits data point into K number of clusters	Determines classification of a point .
4. The points in each cluster tend to be near each other.	Combines the classification of the K nearest points

Ques 8. What is clustering?

Ans. **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (clusters). It is a main task of exploratory ***data mining***, and a common technique for statistical data analysis, used in many fields, including ***machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.***

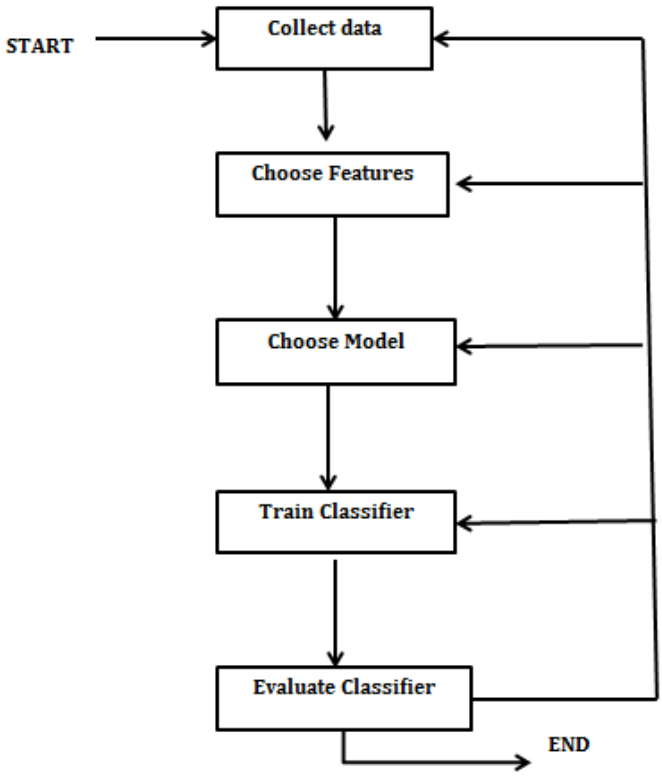
Ques 9. What is partitioning clustering?

Ans. Partitioning algorithms are clustering techniques that subdivide the data sets into a set of k groups, where k is the number of groups pre-specified by the analyst. There are different types of partitioning clustering methods. ***The most popular is the K-means clustering*** , in which, each cluster is represented by the center or means of the data points belonging to the cluster. The K-means method is sensitive to outliers.

Long Questions & Answers

Ques 10. Explain Design Cycle of a Pattern Recognition System.

Ans : Design Cycle of a Pattern Recognition System



Pattern classification involves finding three major attribute spaces:

(a) Measurement space (b) Feature space (c) decision space.

After this appropriate neural network set up is trained with these attribute sets to make system learn for unknown set of patterns and objects. Steps of classification process are as follows:

Step 1. Stimuli produced by the objects are perceived by sensory devices. Important attributes like (shape , size , color , texture) produce the strongest inputs. ***Data collection involves identification of attributes of objects and creating Measurement space.***

Measurement space: This is the set of all pattern attributes which are stored in a vector form. It is a range of characteristic attribute values. In vector form measurement space is also called ***observation space /data space***. E.g : W = [W_1 , W_2 ,……,Wn-1, W_n] for n pattern classes. W is a pattern vector. Let X = $\begin{bmatrix} x1 \\ x2 \end{bmatrix}$, X is a pattern vector for flower , x1 is petal length and x2 is petal width. Pattern classes can be W_1= Lilly , W_2= Rose , W_3 = Sunflower.

Step 2.After this features are selected and feature space vector is designed. The range of subset of attribute values is called **Feature Space F**. This subset represents a reduction in attribute space and pattern classes are divided into sub classes. Feature space signifies the most important attributes of a pattern class observed in measurement space. Feature space is shown in smaller size than M- space.

Step 3.AI models based on probability theory, E.g : Bayesian Model and Hidden Markov Models are Used for grouping or clustering the objects. Attributes selected are those which provide ***High Inter Class and Low Inter Class groupings.***

Step 4. Using ***Unsupervised(for feature extraction)*** or ***Supervised Learning techniques(classification)*** training of classifiers is performed. When we present a pattern recognition with a set of classified patterns so that it can learn the characteristics of the set, we call it ***training.***

Step 5.In ***evaluation of classifier testing*** is performed. In this an unknown pattern is given to the PR System for identifying its correct class .Using the selected attribute values, object/class characterization models are learned by forming generalized prototype descriptors, Classification rules or Decision Functions. ***The range of decision function values is known as Decision space D of r – dimensions.*** We also evaluate performance, efficiency of the classifier for further improvement.

$$D = \begin{bmatrix} d1 \\ d2 \\ d3 \\ . \\ . \\ . \\ dn \end{bmatrix}$$

Recognition of familiar objects is achieved through the application of the rules learned in step 4, by comparing and matching of objects feature with stored models.

Ques 11. What are the design principles of a Pattern Recognition System ? What are major steps involved in this process?

Ans : Design principles of a Pattern Recognition System are as mentioned below **:**

i. Designing of a pattern recognition system is based on the construction of following AI techniques :
- Multi layer perceptron in Artificial Neural Network.
- Decision tree implementation.
- Nearest neighbor classification.
- Segmentation of large objects.

ii. Designing of a robust PR system against the variation in illumination and brightness in environment.
iii. Designing parameters based on translation, scaling and rotation.
iv. Color and texture representation by histograms.
v. Designing brightness based and feature based PR systems.

. ***This system comprises of mainly five components namely sensing, segmentation, feature extraction, classification and post processing***. All of these together generates a System and works as follows:

1. Sensing and Data Acquisition: It includes, various properties that describes the object, such as its entities and attributes which are captured using sensing device.

2. Segmentation: Data objects are segmented into smaller segments in this step.

3. Post Processing & Decision: Certain refinements and adjustments are done as per the changes in features of the data objects which are in the process of recognition. Thus, decision making can be done once, post processing is completed.

Need of Pattern Recognition System

Pattern Recognition System is responsible for generating patterns and similarities among given problem/data space, that can further be used to generate solutions to complex problems effectively and efficiently. Certain problems that can be solved by humans, can also be made to be solved by machine by using this process. Affective computing which gives a computer the ability to recognize and express emotions, to respond intelligently to human emotions that contribute to rational decision making.

Ques12. Discuss about the four best approaches for a Pattern Recognition system. Also Discuss some of the main application area with example of PR system.

Ans : Approaches of PR system are as mentioned below :

1). Template Matching 2). Statistical Approach 3). Syntactic Approach 4). ANN Approach.

TEMPLATE MATCHING: This approach of pattern recognition is based on finding the similarity between two entities (points , curves / shapes) of same type. A 2-D shape or a prototype of a pattern to be recognized is available. Template is a d x d mask or window. Pattern to be recognized is matched against stored template in a knowledge base.

STATISTICAL APPROACH: Each pattern is represented in terms of d- features in d- dimension space. Goal is to select those features that allow pattern vectors belonging to different categories to occupy compact and disjoint regions. Separation of pattern classes is determined. ***Decision surfaces and lines are drawn which are determined by probability distribution of random variables w.r.t each pattern class.***

SYNTACTIC APPROACH: This approach solves complex pattern classification problems. A hierarchal rules are defined. E.g: Grammar rules for natural language, syntax tree structure. ***These are used to decompose complex patterns into simpler sub patterns.*** Patterns can be viewed as sentences where sentences are decomposed into words and further words are sub divided into letters.

NEURAL NETWORKS APPROACH: Artificial neural networks are massively parallel computing systems consisting of extremely large number of simple processors with many interconnections.

Network Models attempt to use some principles like Learning , Generalization , Adaptivity, Fault Tolerance , Distributed representation & computation. Learning process involves updating network architecture and connection mapping and weights so that network may perform better clustering.

Applications of PR System with Examples

Problem Domain	Application	Input pattern	Pattern Classes
Bioinformatics	Sequence analysis	DNA / Protein sequence	Known types of genes patterns
Data Mining	Searching for meaningful data	Points in multidimensional space	Compact & well separated clusters.
Document classification	Internet Searching	Text document	Semantic categories(sports, movies, business, science)
Document image analysis	Reading machines for blinds	Document image	Alphanumeric characters, words
Industrial automation	Printed circuit board inspection	Intensity or range image	Defective / non defective nature of product
Biometrics	Personal identification	Face, iris , finger prints	Authorized users for access control
Speech recognition	Searching content on google voice assistance.	Speech waveforms	Spoken words.

Ques 13. Write short notes on :

(A) Decision theoretic classification (B) Optimum Statistical Classifier

Ans : (A) Decision theoretic classification: This is a statistical pattern recognition technique. Which is based on the use of decision functions to classify the objects. A decision function maps pattern vectors X into decision regions of D **i.e**. f : X → D. These functions are also termed as ***Discriminant Functions***.

- Given a set of Objects O = { O_1 , O_2 ….O_n }. Let each D_i have K- observable attributes (Measurement space and relations are V = { V_1 , V_2 , …..V_k}).
- Determine the following parameters :
  a) A subset of m ≤ *k* of V_i , X = [X_1 , X_2 ,…..X_m] whose values uniquely characterize O_i.
  b) C ≥ 2 grouping of Oi which exhibits High Inter Class and Low Inter Class similarities, such that a decision function d (X) can be found which partition D into C disjoint regions. These regions are used to classify each object Oi for some class.
- For W pattern Classes , W = [W_1 , W_2 ,……,Wn-1, W_n]to find W decision functions **$d_1(x)$, d_2 (x) ,……d_w (x).** with property that if a pattern X belongs to class **W_i , then $d_i(X) > d_j$ (x) , for j = 1, 2, ….,w ; j ≠ i.**
- Linear decision function can be in the form of line equation as : **d (X) = W_1 X_1 + W_1 X_1** for a 2-D pattern vector.

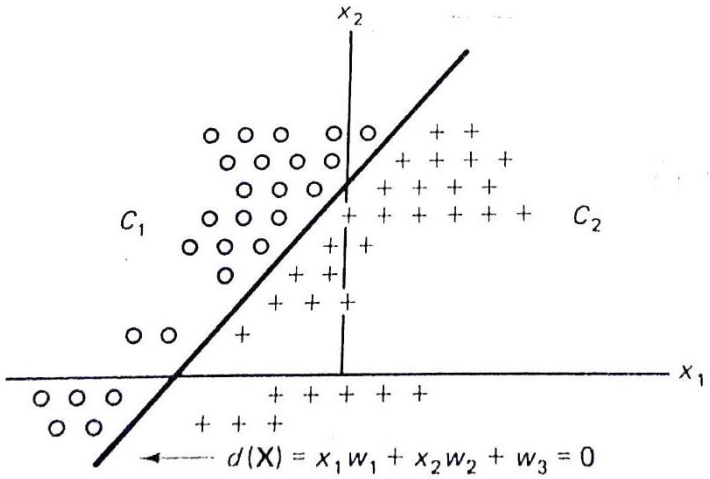


Fig : (a)

An object belongs to class W1 or C1 if d(x) < 0.else for d(x) > 0 it belongs to class W2 or C2.

If d (x) =0 then it is indeterminate.

Fig(a) is linearly separable Class.

Decision Boundary: d_i (x) – d_j (x) = 0 . Aim is to identify decision boundary between two classes by a Single function d_{ij} (x) = d_i(x) – d_j (x) =0.

When a line can be found that separates classes into two or more clusters we say classes are ***Linearly Separable*** else they are called ***Non Linear Separable Classes.***

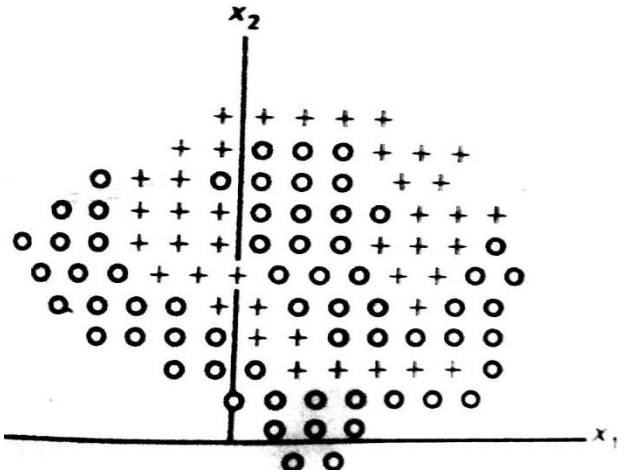
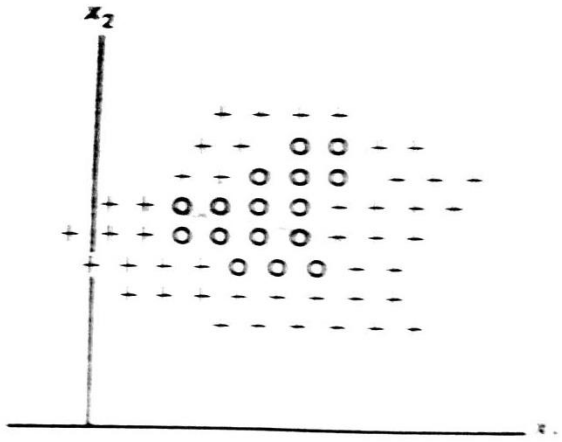


Fig : (b)



Fig(c)

Fig(b) and Fig (c) are for Non linear separable classes.

(B) Optimum Statistical Classifier: This is a pattern classification approach developed on the basis of probabilistic technique because of randomness under which pattern classes are normally generated.

It is based on ***Bayesian theory and conditional probabilities***." Probability that a particular pattern x is from class W_i is denoted as **P (Wi | x)**". If a pattern classifier decides that x came from W_j when it actually came from W_i it incurs a ***loss*** L_{ij}.

Average Loss incurred in assigning x to class W_j is given by following equation:

$$r_j(x) = \sum_{k=1}^{W} L_{kj}\, P(W_k \mid x) \quad \text{..... Eq (1)} \qquad \text{[W are total no. of classes]}$$

This is called ***Conditional average Risk / Loss.***

By Baye's Theorem: $P(A \mid B) = \dfrac{P(B \mid A).P(A)}{P(B)}$, So Eq (1) can be modified as :

$$\therefore r_j(x) = \sum_{k=1}^{W} \frac{L_{kj}\, P(x \mid W_k).P(W_k)}{P(x)} \quad \text{..... Eq (2)}$$

$P(x \mid W_k)$ ***is Probability density function*** of the patterns from class W_k and P (W_k) is the probability of occurrence of class W_k.P(x) is priori probability and independent of k (Has same value for all the classes), Hence equation can be again rewritten as below :

$$r_j(x) = \frac{1}{P(x)} \sum_{k=1}^{W} L_{kj}\, P(x \mid W_k).\, P(W_k) \text{.....Eq (3)}$$

Ques 14. Explain Maximum Likelihood technique under Parameter Estimation of Classification.

Ans : Estimation model consists of a number of parameters. So, in order to calculate or estimate the parameters of the model, the concept of Maximum Likelihood is used.

- Whenever the probability density functions of a sample are unknown, they can be calculated by taking the parameters inside sample as quantities having unknown but fixed values.
- Consider we want to calculate the height of a number of boys in a school. But, it will be a time consuming process to measure the height of all the boys. So, the unknown mean and unknown variance of the heights being distributed normally, by maximum likelihood estimation we can calculate the mean and variance by only measuring the height of a small group of boys from the total sample.

Let we separate a collection of samples as per the class, having **C data sets , D_1 , D_2 ,….D_c** with samples in Dj drawn accurately to probability p (x | W_j). Let this has a known parametric form and is determined by value θ_j. E.g : **p (x | W_j) ~ N (μ_i , Σ_j) , θ_j consists of these parameter**.

To show dependence we have :

p (x | W_j, θ_j) . Objective is to use information provided by training samples to achieve good estimates for unknown parameter vectors $\boldsymbol{\theta_1}$, $\boldsymbol{\theta_2}$, $\boldsymbol{\theta_3}$ …. . $\boldsymbol{\theta_{c-1}}$, $\boldsymbol{\theta_c}$ associated with each category. Assume samples in Di give no information about θ_j, if i ≠ j i.e Parameters of Different Classes are functionally independent. Let set D has n samples [X_1 , X_2 ,…..X_n],

$$\therefore \mathbf{p\ (\ D\ |\ \theta\ )} = \prod_{k=1}^{n} P(\ X_k\ |\theta\ ).$$

p (D | θ) is likelihood of θ w.r.t set of samples." Maximum likelihood estimate of $\boldsymbol{\theta}$ **is by definition value** $\widehat{\boldsymbol{\theta}}$ that maximizes p (D | θ).

Logarithmic Form : Since Log makes the expressions simpler in the form of addition , θ that maximizes log likelihood also maximizes likelihood. If number of parameters to be estimated is p, we let $\boldsymbol{\theta}$ **denote p – component vector i.e** $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \boldsymbol{\theta_3} \ldots . \boldsymbol{\theta_{p-1}}, \boldsymbol{\theta_p})^t$.

Let gradient operator $\nabla_\theta = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ \cdots \\ \frac{\partial}{\partial\theta_p} \end{bmatrix}$, Define l ($\theta$) : log likelihood function.

Therefore, l ($\boldsymbol{\theta}$) = ln p ($D \mid \theta$) $\Rightarrow \hat{\theta}$ = arg max_θ l(θ)

l ($\boldsymbol{\theta}$) = $\sum_{k=1}^{n}$ **ln** $\boldsymbol{p}$ $(X_{k \mid \theta})$ **and** $\nabla_\theta\, l = \sum_{k=1}^{n} \nabla_\theta\, ln\, p\, (X_{k \mid \theta})$

For maximum likelihood $\nabla_\theta\, l = 0$

Ques 15. (A) Write down the steps for K nearest Neighbor estimation.
(B) Mention some of the advantages and disadvantages of KNN technique.

Ans. (A) K –Nearest Neighbor Estimation:

1. Calculate "$d(x,x_i)$" i =1, 2,….., **n**; where **d** denotes the Euclidean distance between the points.
2. Arrange the calculated **n** Euclidean distances in non-decreasing order.
3. Let **k** be a +ve integer, take the first **k** distances from this sorted list.
4. Find those **k**-points corresponding to these **k**-distances.
5. Let $\mathbf{k}_i$ denotes the number of points belonging to the i^{th} class among **k** points i.e. $k \geq 0$
6. If $k_i > k_j\ \forall\ i \neq j$ then put x in class i.

(B) Advantages of KNN :

1. Easy to understand
2. No assumptions about data
3. Can be applied to both classification and regression
4. Works easily on multi-class problems

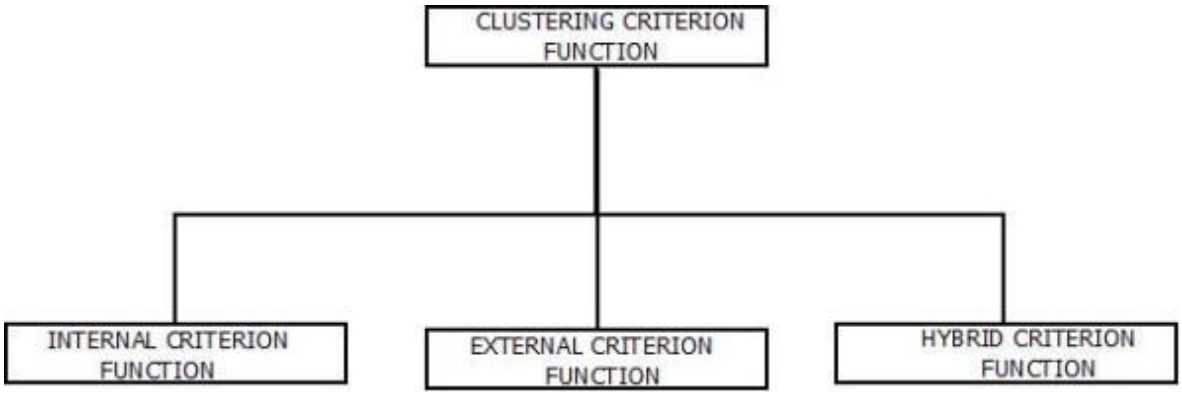
Disadvantages are:

1. Memory Intensive / Computationally expensive
2. Sensitive to scale of data
3. Not work well on rare event (skewed) target variable
4. Struggle when high number of independent variables

Ques 16.. Explain the function of clustering.

Ans. To measure the quality of clustering ability of any partitioned data set, criterion function is used.

1. Consider a set , B = { $x_1,x_2,x_3…x_n$} containing "n" samples, that is partitioned exactly into "t" disjoint subsets i.e. B_1, B_2,…..,B_t.
2. The main highlight of these subsets is, every individual subset represents a cluster.
3. Sample inside the cluster will be similar to each other and dissimilar to samples in other clusters.
4. To make this possible, criterion functions are used according the occurred situations.



Criterion Function For Clustering

1. **Internal Criterion Function**

a) This class of clustering is an intra-cluster view.

b) Internal criterion function optimizes a function and measures the quality of clustering ability various clusters which are different from each other.

2. **External Criterion Function**

a) This class of clustering criterion is an inter-class view.

b) External Criterion Function optimizes a function and measures the quality of clustering ability of various clusters which are different from each other.

3. **Hybrid Criterion Function**

a) This function is used as it has the ability to simultaneously optimize multiple individual Criterion Functions unlike as Internal Criterion Function and External Criterion Function

Ques17. Solve it with the help of K-mean clustering.

Subject	A	B
1	**1.0**	**1.0**
2	**1.5**	**2.0**
3	**3.0**	**4.0**
4	**5.0**	**7.0**
5	**3.5**	**5.0**
6	**4.5**	**5.0**
7	**3.5**	**4.5**

Centre points are : (1,1) and (5,7)

Ans.

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

This data set is to be grouped into two clusters. As a first step in finding a sensible initial partition, let the A & B values of the two individuals furthest apart (using the Euclidean distance measure), define the initial cluster means, giving:

	Individual	Mean Vector (centroid)
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

The remaining individuals are now examined in sequence and allocated to the cluster to which they are closest, in terms of Euclidean distance to the cluster mean. The mean vector is recalculated each time a new member is added. This leads to the following series of steps:

	Cluster 1		Cluster 2	
Step	Individual	Mean Vector (centroid)	Individual	Mean Vector (centroid)
1	1	(1.0, 1.0)	4	(5.0, 7.0)
2	1, 2	(1.2, 1.5)	4	(5.0, 7.0)
3	1, 2, 3	(1.8, 2.3)	4	(5.0, 7.0)
4	1, 2, 3	(1.8, 2.3)	4, 5	(4.2, 6.0)
5	1, 2, 3	(1.8, 2.3)	4, 5, 6	(4.3, 5.7)
6	1, 2, 3	(1.8, 2.3)	4, 5, 6, 7	(4.1, 5.4)

Now the initial partition has changed, and the two clusters at this stage having the following characteristics:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2, 3	(1.8, 2.3)
Cluster 2	4, 5, 6, 7	(4.1, 5.4)

But we cannot yet be sure that each individual has been assigned to the right cluster. So, we compare each individual's distance to its own cluster mean and to that of the opposite cluster. And we find:

Individual	Distance to mean (centroid) of Cluster 1	Distance to mean (centroid) of Cluster 2
1	1.5	5.4
2	0.4	4.3
3	2.1	1.8
4	5.7	1.8
5	3.2	0.7
6	3.8	0.6
7	2.8	1.1

Only individual 3 is nearer to the mean of the opposite cluster (Cluster 2) than its own (Cluster 1). In other words, each individual's distance to its own cluster mean should be smaller that the distance to the other cluster's mean (which is not the case with individual 3). Thus, individual 3 is relocated to Cluster 2 resulting in the new partition:

	Individual	Mean Vector (centroid)
Cluster 1	1, 2	(1.3, 1.5)
Cluster 2	3, 4, 5, 6, 7	(3.9, 5.1)

Ques : 18 What is Dimensionality reduction in pattern classification methods? Explain Principal Component analysis with its application in AI.

Ans : A very common problem in statistical pattern recognition is of Feature Selection i.e. a process of transforming Measurement Space to Feature Space (Set of data which are of interest).

Transformation reduces the dimensionality of data features . Let we have a m- dimensional vector , X = $[X_1, X_2, \ldots . X_m]$ and we want to convert it in l-dimensions (where l << m) .

$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \dots \\ \dots \\ X_m \end{bmatrix}$ **, after reducing the dimensions vector X =** $\begin{bmatrix} X_1 \\ X_2 \\ \dots . \\ \dots . \\ X_L \end{bmatrix}$

This reduction causes mean square error . So we need to find that does there exist an invertible transform T, such that truncation of T_x is optimal in terms of Mean Square Error. So T must have some components of low variance(σ^2) where $\boldsymbol{\sigma^2}$ **= E [** $(\boldsymbol{x} - \boldsymbol{\mu})^{\boldsymbol{2}}$ **]** , E is expectation function, x is random variable , and μ is mean value**.** $\boldsymbol{\mu = \frac{1}{x} \sum_{k=1}^{m} X_k}$

Definition of PCA: This is a mathematical procedure that uses Orthogonal transforms to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variables known as Principal Components. ***So here we preserve the most variance with reduced dimensions and minimum mean square error.***

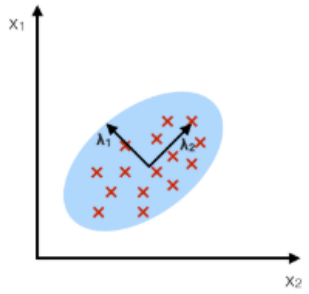
- Number of principal components are less than or equal to number of original variables.
- First Principal component has largest variance. Successively it decreases.
- These are defined by Best Eigen Vectors of Covariance Matrix of vector X.

Geometrical analysis of PCA:

i. PCA projects data along directions where σ^2 is maximum.
ii. These directions are determined by eigen vectors of covariance matrix corresponding to largest eigen values.
iii. Magnitude of variance is variance of data along the directions of eigen values.
iv. Eigen Values are characteristic values given as AX = λX , A is m x n matrix , λ *is eigen values*.

PCA:

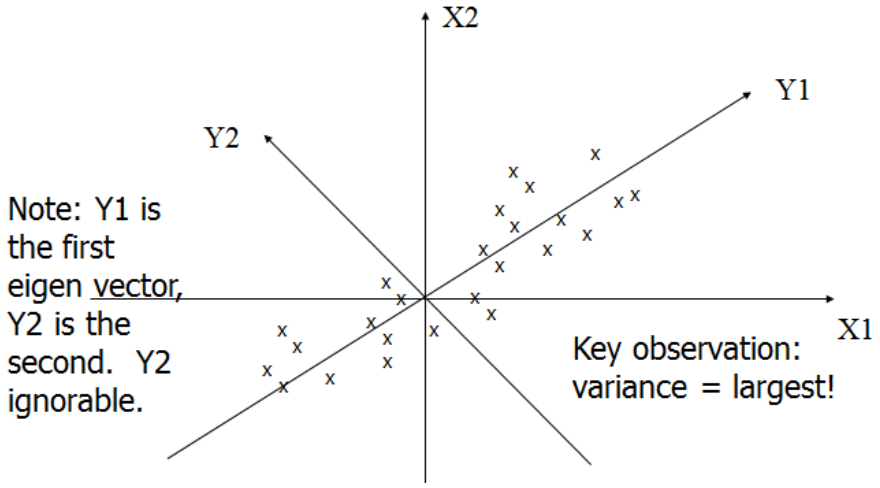
component axes that maximize the variance



- Degree to which the variables are linearly correlated is represented by their covariance.
- PCA uses Euclidean Distance calculated from the *p* variables as the measure of dissimilarity among the n objectsThe eigenvalues (latent roots) of S are solutions (λ) to the characteristic equation
- $|\mathbf{S} - \lambda\mathbf{I}| = 0$

- The eigenvalues, λ_1, λ_2, ... λ_p are the variances of the coordinates on each principal component axis. Coordinates of eac…h object *i* on the k^{th} principal axis, known as the scores on PC *k*, are computed as mentioned below:

$$z_{ki} = u_{1k}x_{1i} + u_{2k}x_{2i} + \cdots + u_{pk}x_{pi}$$



Steps of PCA

- Let μ be the mean vector (taking the mean of all rows)
- Adjust the original data by the mean $\varphi = X_k - \mu$
- Compute the covariance matrix C of adjusted X
- Find the eigenvectors and eigenvalues of C.
- For matrix *C*, vectors **e** (=column vector) having same direction as *C***e** :
- *eigenvectors* of *C* is **e** such that *C***e**=λ**e**,
- λ is called an *eigenvalue* of *C*. *C***e**=λ**e** ⇔ (*C*-λI)**e**=0

Applications of PCA in AI:

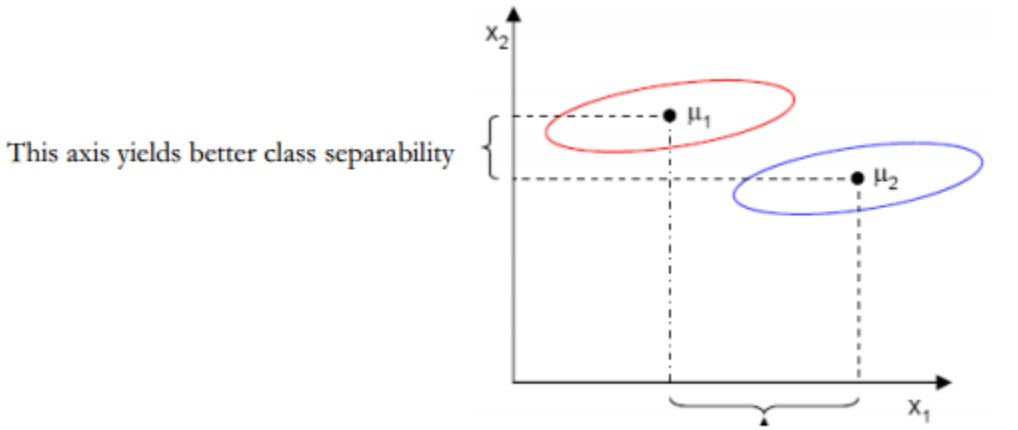
- Face recognition
- Image compression
- Gene expression analysis

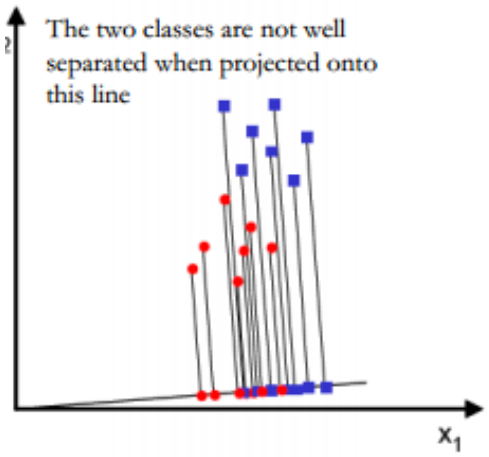
Ques 19. Explain Linear Discriminant Analysis with its derivation.

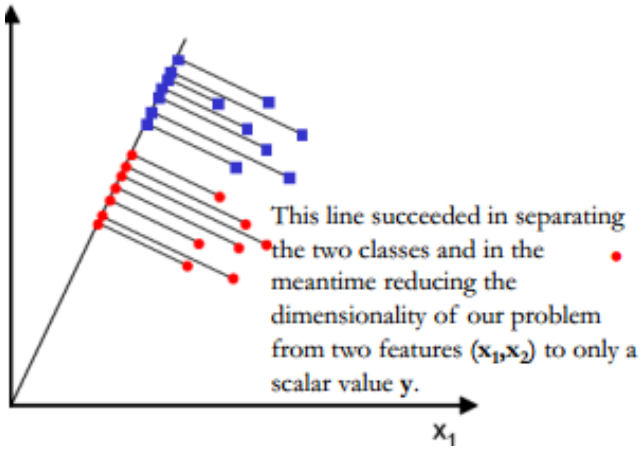
Ans : PCA finds components that are useful for data representation , but drawback is that PCA can not discriminate components /data between different classes. If we group all the samples , then those directions that are discarded by PCA might be exactly the directions needed for distinguishing between classes.

- ***PCA is based on representation for efficient direction***
- ***LDA is based on discrimination for efficient direction.***

Objective of LDA is to perform dimensionality reduction while preserving as much of the class discrimination information as possible. Here in LDA data is projected from d – dimensions onto a line. If the samples formed well separated compact clusters in d- space then projection onto an arbitrary line will usually produce poor recognition performance. By rotating the line we can find an orientation for which projected samples are well separated.







- Assume we have *m*-dimensional samples $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, N_1 of which belong to ω_1 and N_2 belong to ω_2.
- We seek to obtain a scalar **y** by projecting the samples **x** onto a line (C-1 space, C = 2).

$$y = w^T x \quad where \quad x = \begin{bmatrix} x_1 \\ . \\ . \\ x_m \end{bmatrix} \quad and \quad w = \begin{bmatrix} w_1 \\ . \\ . \\ w_m \end{bmatrix}$$

  - where **w** is the projection vectors used to project **x** to **y**.
- **Of all the possible lines we would like to select the one that maximizes the separability of the scalars.**

In order to find a good projection vector, we need to define a measure of separation between the projections.

- The mean vector of each class in **x** and **y** feature space is:

$$\mu_i = \frac{1}{N_i}\sum_{x\in\omega_i} x \quad and \quad \tilde{\mu}_i = \frac{1}{N_i}\sum_{y\in\omega_i} y = \frac{1}{N_i}\sum_{x\in\omega_i} w^T x$$

$$= w^T \frac{1}{N_i}\sum_{x\in\omega_i} x = w^T \mu_i$$

  - i.e. projecting **x** to **y** will lead to projecting the mean of **x** to the mean of **y**.
- We could then choose the distance between the projected means as our objective function

$$J(w) = \left|\tilde{\mu}_1 - \tilde{\mu}_2\right| = \left|w^T\mu_1 - w^T\mu_2\right| = \left|w^T(\mu_1 - \mu_2)\right|$$

The solution proposed by ***Fisher is to maximize*** a function that represents the difference between the means, normalized by a measure of the within-class variability, or the ***so-called scatter***. • For each class we define the scatter, an equivalent of the variance, as; ***(sum of square differences between the projected samples and their class mean).***

$$\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$$

The Fisher linear discriminant is defined as the linear function w Tx that maximizes the criterion function: (the distance between the projected means normalized by ***the within class scatter*** of the projected samples.

$$J(w) = \frac{\left|\tilde{\mu}_1 - \tilde{\mu}_2\right|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

In order to find the optimum projection w*, we need to express J(w) as an explicit function of w.. We will define a measure of the ***scatter in multivariate feature space x*** which are denoted as scatter matrices.

$$S_i = \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^T$$

$$S_w = S_1 + S_2$$

Where Si is the covariance matrix of class ωi, and Sw is called the within-class scatter matrix. Similarly, the difference between the projected means (in y-space) can be expressed in terms of the means in the original feature space (x-space)

$$(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^T \mu_1 - w^T \mu_2)^2$$

$$= w^T \underbrace{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{S_B} w$$

$$= w^T S_B w = \tilde{S}_B$$

The matrix S_B is called the between-class scatter of the original samples/feature vectors, while is the between-class scatter of the projected samples y.

Ques 20 : Explain Support Vector Machines in detail. What are advantages and disadvantages of SVM ?

Ans : Support Vector Machines: This is a linear machine with a case of separable patterns that may arise in the context of pattern classification . Idea is to construct **a HYPERPLANE** as a direction surface in such a way that the margin of separation between positive and negative examples is maximized.

A good example of such a system is ***classifying a set of new* documents *into positive or negative sentiment groups,*** based on other documents which have already been classified as positive or negative. Similarly, we could ***classify new emails into spam or non-spam, based on a large corpus of documents*** that have already been marked as spam or non-spam by humans. SVMs are highly applicable to such situations.

- SVM is an approximate implementation of ***Structural Risk Minimization.***
- Error rate of a machine on test data is bounded by the sum of training error rate and term that depends on ***Vapnik Chervonenki's dimension.***
- SVM sets first term to zero and minimizes second term. We use SVM learning algorithm to construct following three types of learning machines :

(i) Polynomial learning machine

(ii) Two layer perceptrons

(iii) Radial basis function N/W

Condition is as : Test error rate ≤ Train error rate + f (N , h , p).

Where N: Size of training set , h : Measure of model complexity

P: Probability that this bound fails.

If we consider an element of our p-dimensional feature space, i.e. $\rightarrow x = (x1,...,xp) \in Rp$, then we can ***mathematically define an affine Hyper plane by the following equation:*** $b_0+b_1x_1+...+b_px_p=0$, $b0 \neq 0$ gives us an affine plane (i.e. it does not pass through the origin). We can use a more succinct notation for this equation by introducing the summation sign: $b_0+p\sum j=1 b_j x_j=0$. The line that maximizes the minimum margin is better. Maximum margin separator is determined by a subset of data points. Data points in the subset are called Support Vectors. Support vectors are used to decide which side of separator a test case is ON.

Consider a training set **{ (** $X_{i,}$ $d_{i,}$ **) } for i= 1 to n ,** where X_i is input pattern for i^{th} example.

And d_i is the desired response (Target output). ***Let*** $d_{i,} = +1$ ***and*** $d_{i,} = -1$ ***Pattern classes for positive and negative examples are linearly separable.*** Hyper Plane decision surface is given as below equation:

W^T X + b = 0 , then d_i =0(when data point is on the line)

where W : adjustable weight factor and b is Bias .

Therefore, $W^T X_i + b_i \geq 0$ for $d_{i,} = +1$ and $W^T X_i + b_i < 0$ for $d_{i,} = -1$.

Closest data point is called ***Margin of Separation***. Denoted by ρ. Objective of SVM is to maximize ρ for **Optimal Hyper plane.**

Ques 21: What is Nearest Neighbor rule of classification? Mention some of the metrics used in method.

Ans : Nearest neighbor algorithm assigns to a test pattern the class label of its closest neighbor.

Let n training patterns ($X_{1,} \theta_{1,}$) , ($X_{2,} \theta_{2,}$) ……., (X_n, θ_n) where X_i is of dimension d and $\theta_{i,}$ *is ith pattern.* If P is the test pattern then if d (P , X_k) = min { d (P , X_i) }, i = 1 to n.

Error: In NN classifier error is at most twice the Bayes Error , when the number of training samples tends to infinity.

$$\mathbf{E}(\alpha_{bayes}) \leq \mathbf{E}(\alpha_{nn}) \leq E(\alpha_{bayes}) \left[2 - \frac{C\,E\,(\alpha_{bayes})}{C-1}\right]$$

Distance metrics Used in Nearest Neighbor Classification:

(1). Euclidian Distance: L (a , b) = $\|a-b\| = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}$

Properties are as following:

I. L (a , b) ≥ 0

II. L (a , b) = 0 iff a = b

III. L (a , b) = L (b ,a)

IV. L (a , b) + L (b , c) ≥ $L(a, c)$

(3). Mahalanobis Distance:

L (a , b) = $\sqrt{(a-b)^T C^{-1} (a-b)}$

C : Positive definite matrix called Covariance Matrix.

(2). Minkowiski Metric: $\mathbf{L_k(a, b)} = \left[\sum_{i=1}^{d} |ai - bi|^k\right]^{1/k}$, L_1 is also called Manhattan or City Block Distance. **L_k : are Norms**

[END OF 5th UNIT]