UNIT - I

1 Introduction of Pattern Recognition

1.1 : What is Pattern Recognition ?

Q.1 What is pattern ?

Ans. : Pattern is defined as composite of features that are characteristic of an individual.

Q.2 What do you mean by pattern recognition ? Explain.

Ans. : Pattern recognition can be defined as the categorization of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant detail.

Q.3 What is basic function of pattern recognition system ?

Ans. : The basic functions of a pattern recognition system are to detect and extract common features from the patterns describing the objects that belong to the same pattern class and to recognize this pattern in any new environment and classify it as a member of one of the pattern classes under consideration.

Q.4 Define pattern and pattern class.

Ans. : • **Pattern :** A pattern is the description of an object.

• **Pattern Class :** It is a category determined by some given common attributes.

Q.5 List the classification of Pattern Recognition System.

Ans. : Classification of Pattern Recognition System are Rule based system, classical fuzzy system, Bayesian system,
Neural networks system and Fuzzy neural networks systems.

1.2 : Data Sets for Pattern Recognition

Q.6 Discuss the design process of the pattern recognition system with suitable block diagram.

Ans. : • Pattern recognition is the science of making inferences from perceptual data, using tools from statistics, probability, computational geometry, machine learning, signal processing, and algorithm design.

- Pattern recognition can be defined as the categorization of input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant detail.
- The identification of implicit objects, types or relationships in raw data by an animal or machine.
- A pattern is an object, process or event that can be given a name.
- A pattern class (or category) is a set of patterns sharing common attributes and usually originating from the same source.

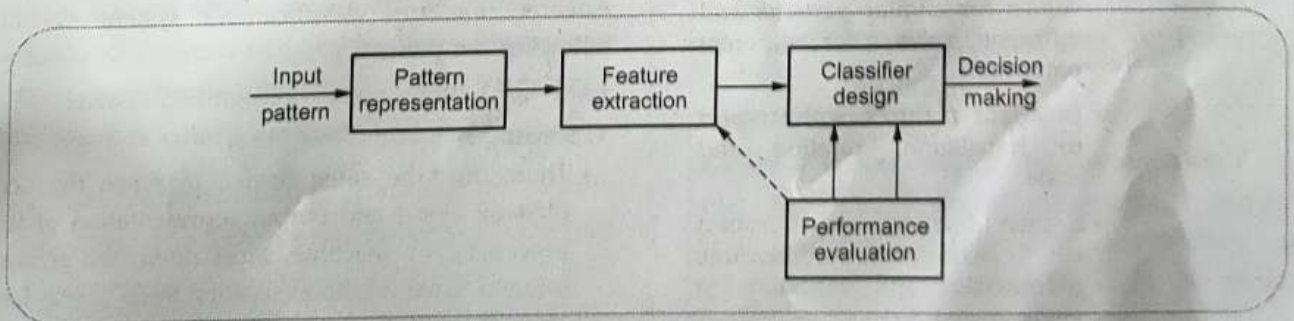


Fig. Q.6.1

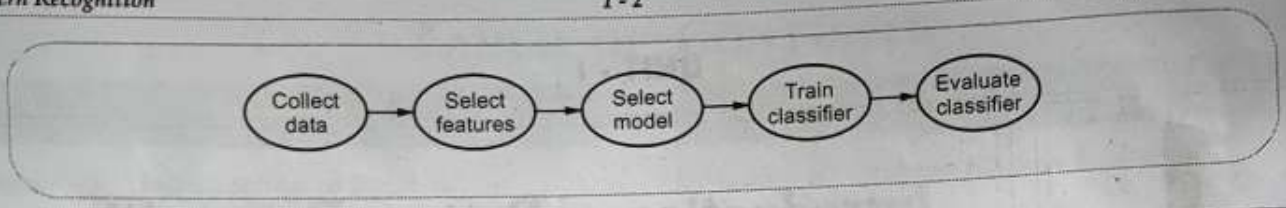


Fig. Q.6.2

- During recognition (or classification) given objects are assigned to prescribed classes. A classifier is a machine which performs classification.
- The purpose of a pattern recognition program is to analyze a scene in the real world and to arrive at a description of the scene which is useful for the accomplishment of some task.
- The real world observations are gathered through sensors and pattern recognition system classifies or describes these observations.
- A feature extraction mechanism computes numeric or symbolic information from these observations. These extracted features are then classified or described using a classifier.
- The process used for pattern recognition consists of many procedures that ensure efficient description of the patterns.

Design Principle of Pattern Recognition :

- Fig. Q.6.2 shows design cycle of pattern recognition.

Steps	Description
Data Collection	• Collecting training and testing data. • It is difficult to identify an adequately large and representative set of samples.
Feature Selection	• Domain dependence and prior information. • Computational cost and feasibility. • Discriminative features : Similar values for similar patterns and different values for different patterns. • Invariant features with respect to translation, rotation and scale. • Robust features with respect to occlusion, distortion, deformation and variations in environment.
Model Selection	• Definition of design criteria. • Parametric vs. non-parametric models. • Handling of missing features. • Computational complexity. • Types of models : templates, decision-theoretic or statistical, syntactic or structural, neural and hybrid.
Training	• How can we learn the rule from data ? • Supervised learning : A teacher provides a category label or cost for each pattern in the training set. • Unsupervised learning : The system forms clusters or natural groupings of the input patterns. • Reinforcement learning : No desired category is given but the teacher provides feedback to the system such as the decision is right or wrong.
Evaluation	• How can we estimate the performance with training samples ? • How can we predict the performance with future data ?

Q.7 Explain components of pattern recognition system.

Ans. : • Pattern classification system contains following components :

- Fig. Q.7.1 shows pattern recognition system.

1. Sensing :

- There must be some device to sense the actual physical object and output representation of it for processing by machine. Most often, the sensor is selected from existing sensors built for a larger class of problems.

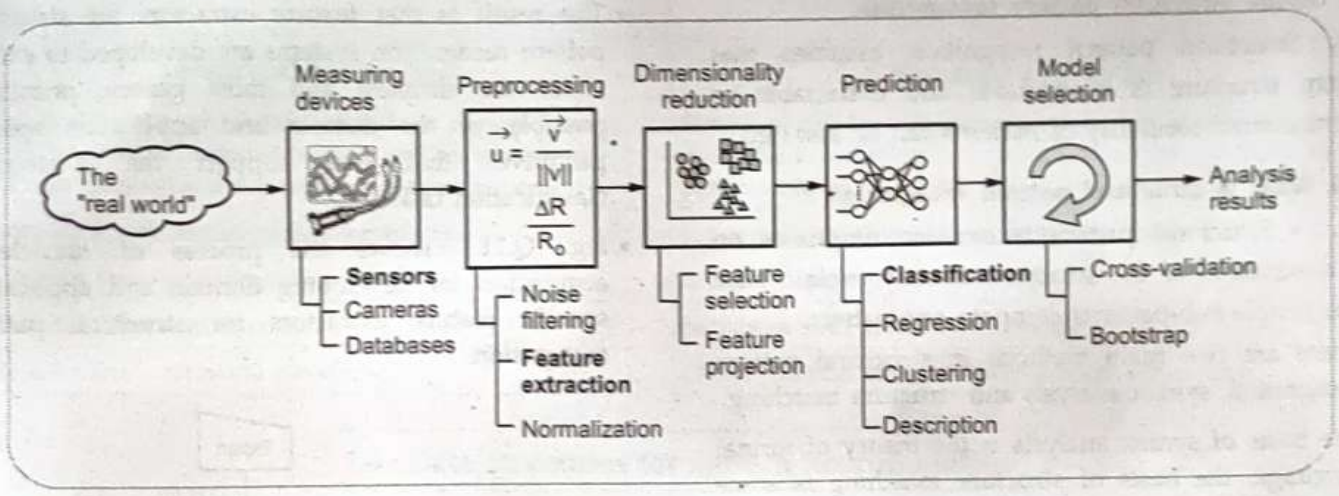


Fig. Q.7.1

- Device used for this purpose of transducer, such as a camera or a microphone array.
- Pattern reorganization system depends on the bandwidth, the resolution, sensitivity distortion of the transducer.

2. **Segmentation and grouping :**
   - Break up the image into meaningful or perceptually similar regions. Compact representation for image data in terms of a set of components.
   - Components share "common" visual properties. Properties can be defined at different level of abstractions. Patterns should be well separated and should not overlap.

3. **Feature extraction :**
   - A set of characteristic measurements (numerical or non-numerical) and their relations are extracted to represent patterns for further process.
   - The traditional goal of the feature extractor is to characterize an object to be recognized by measurements whose values are very similar for objects in the same category and very different for objects in different categories.

- **Two types of criteria are commonly used :**

a) **Signal representation** : The goal of feature selection is to accurately represent the samples accurately in a lower-dimensional space.

b) **Classification :** The goal of feature selection is to enhance the class-discriminatory information in the lower-dimensional space.

4. **Classification :**
   - The process or events with same similar properties are grouped into a class. The number of classes is task-dependent. The task of a classifier is to partition feature space into class-labeled decision regions.
   - The borders between decision regions are called decision boundaries. The classification of feature vector x consists of determining which decision region it belongs to and assigns x to this class.
   - A classifier can be represented as a set of Discriminant functions.

5. **Post processing :**
   - Considering the effects of context and the cost of errors.
   - The post-processor uses the output of the classifier to decide on the recommended action.

1.3 : Different Paradigms for Pattern Recognition

Q.8 What is statistical pattern recognition ?

Ans. : Statistical pattern recognition attempts to classify patterns based on a set of extracted features and an underlying statistical model for the generation of these patterns.

Q.9 Define structural pattern recognition.

Ans. : Structural pattern recognition assumes that pattern structure is quantifiable and extractable so that structural similarity of patterns can be assessed.

Q.10 What is structural pattern recognition ?

Ans. : • Structural pattern recognition emphases on the description of the structure, namely explain how some simple sub-patterns compose one pattern.

- There are two main methods in structural pattern recognition, syntax analysis and structure matching.
- The basis of syntax analysis is the theory of formal language, the basis of structure matching is some special technique of mathematics based on sub-patterns.
- When consider the relation among each part of the object, the structural pattern recognition is best.
- Different from other methods, structural pattern recognition handle with symbol information, and this method can be used in applications with higher level, such as image interpretation.
- Structural pattern recognition always associates with statistic classification or neural networks through which we can deal with more complex problem of pattern recognition, such as recognition of multidimensional objects.
- Structural pattern recognition, sometimes referred to as syntactic pattern recognition due to its origins in formal language theory, relies on syntactic grammars to discriminate among data from different groups based upon the morphological interrelationships present within the data.

Q.11 Explain the process of knowledge acquisition for developing domain and application specific feature extractors for structural pattern recognition.

Ans. : • The description task of a structural pattern recognition system is difficult to implement because there is no general solution for extracting structural features, commonly called primitives, from data.

- The lack of a general approach for extracting primitives puts designers of structural pattern recognition systems in an awkward position : feature extractors are necessary to identify primitives in the data, and yet there is no established methodology for deciding which primitives to extract.
- The result is that feature extractors for structural pattern recognition systems are developed to extract either the simplest and most generic primitives possible, or the domain and application specific primitives that best support the subsequent classification task.
- Fig. Q.11.1 shows the process of knowledge acquisition for developing domain and application specific feature extractors for structural pattern recognition.

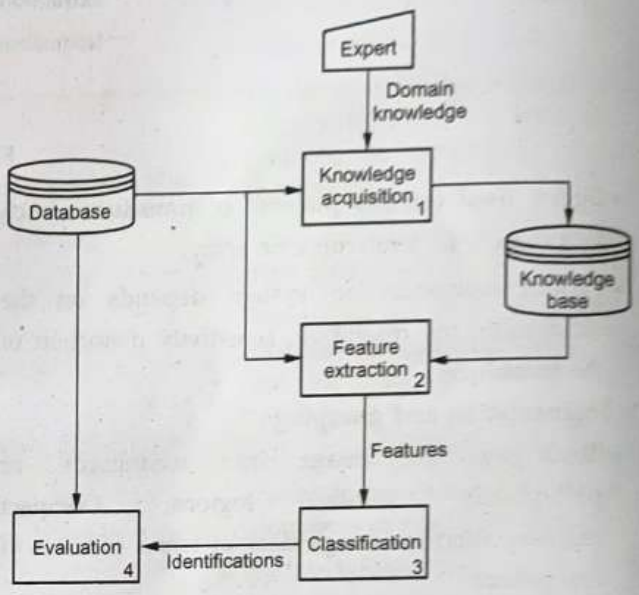


Fig. Q.11.1

- Some structural pattern recognition systems justify the use of a particular set of feature extractors by claiming that the same set had been used successfully by a previous system developed for a similar application within the same domain; such claims simply shift the burden of feature extractor development onto previously implemented systems.
- Simplistic primitives are domain independent, but capture a minimum of structural information and postpone deeper interpretation until the classification step.
- At the other extreme, domain and application specific primitives can be developed with the assistance of a domain expert, but obtaining and formalizing knowledge from a domain expert, called knowledge acquisition, can be problematic.

Q.12 Explain the difference between statistical and structural approaches of pattern recognition.

Ans. :

Statistical approaches	Structural approaches
Statistical pattern recognition attempts to classify patterns based on a set of extracted features and an underlying statistical model for the generation of these patterns.	Structural pattern recognition assumes that pattern structure is quantifiable and extractable so that structural similarity of patterns can be assessed.
It depends upon statistical decision theory.	It depends upon human perception and cognition.
It ignores feature relationships.	It captures primitive relationship.
Semantics from feature position.	Semantics from primitive encoding.
Classification is statistical classifiers.	Classification is semantics from primitive encoding.

1.4 : Data Structures for Pattern Representation

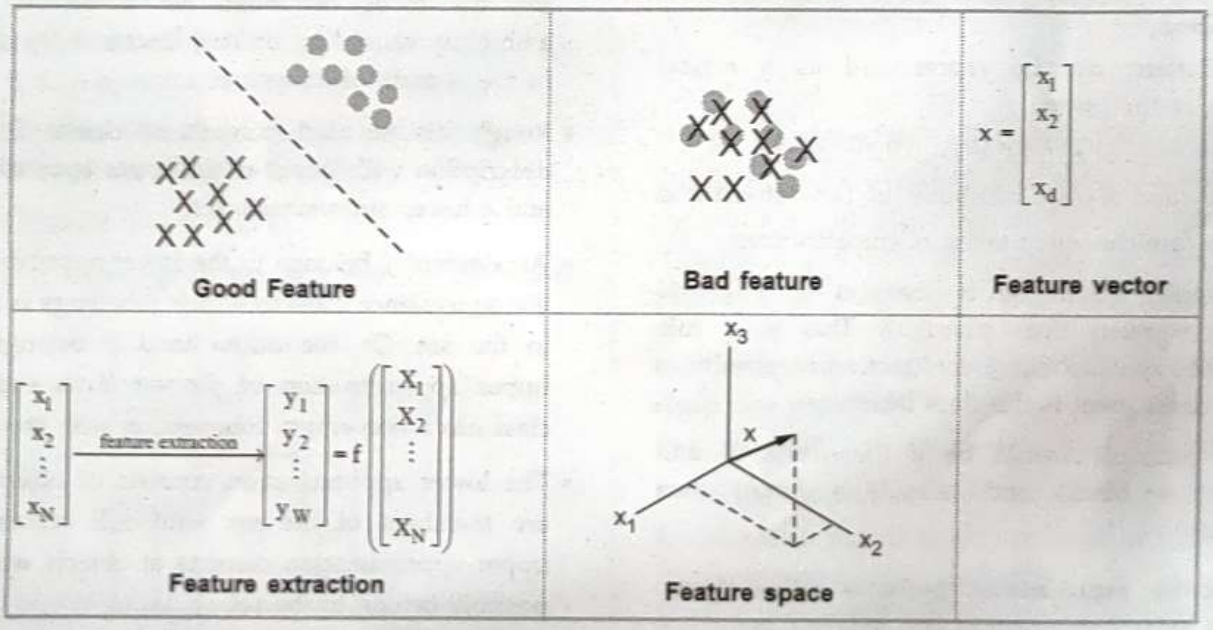
Q.13 List the various data structure used for representing pattern.

Ans. : Pattern is represented by vector, string, logical descriptions, fuzzy and rough pattern set.

Q.14 Define pattern, feature, feature vector, feature space and feature extraction.

Ans. :

- **Pattern :** Pattern is defined as composite of features that are characteristic of an individual.
- **Feature :** Feature is any distinctive aspect, quality or characteristic. It may be symbolic (i.e., color) or numeric (i.e., height).
- **Feature vector :** The feature vector is a single row column matrix of N elements and its computation involves time and memory.
- **Feature space :** A feature space is a collection of features related to some properties of the object or event under study.
- **Feature extraction :** Transforming the existing features into a lower dimensional space. Feature extraction is the process of transforming the input data into a set of features which can very well represent the input data. It is a special form of dimensionality reduction.



Q.15 How pattern is represented as vector ?

Ans. : • Pattern is represented as a vector. Each elements of the vector can represent one attribute of the pattern.

- Here, the training dataset may be represented as a matrix of size (n x d), where each row corresponds to a pattern and each column represents a feature.
- Each attribute/feature/variable is associated with a domain. A domain is a set of numbers, each number pertains to a value of an attribute for that particular pattern.
- The class label is a dependent attribute which depends on the 'd' in-dependent attributes.
- Each element of the vector can represent one attribute of the pattern. The first element of the vector will contain the value of the first attribute for the pattern being considered.
- While representing spherical objects, (25, 1) may be represented as an spherical object with 25 units of weight and 1 unit diameter. The class label can form a part of the vector.
- If spherical objects belong to class 1, the vector would be (25, 1, 1), where the first element represents the weight of the object, the second element, the diameter of the object and the third element represents the class of the object.

Q.16 Explain logical description of pattern representation.

Ans. : • Pattern can be represented as a logical description of the form :

$$X_1 = (a_1,\ldots,a_2) \wedge X_2 = (b_1,\ldots,b_2) \wedge \ldots..$$

- Where X_1 and X_2 are attributes of the pattern and a_i and b_i are the value taken of the attributes.
- An example would be if (beak(x) = red) and (colour(x) = green) then parrot(x). This is a rule where the antecedent is a conjunction of primitives and the consequent is the class label.
- Another example would be if (has-trunk(x)) and (colour(x) = black) and (size(x) = large) then elephant(x).

Q.17 Describe representing patterns using fuzzy and rough sets.

Ans. :

- The features in a fuzzy pattern may consist of linguistic values, fuzzy numbers and intervals.
- For example, linguistic values can be like tall, medium, short for height which is very subjective and can be modelled by fuzzy membership values.
- A feature in the pattern maybe represented by an interval instead of a single number. This would give a range in which that feature falls. An example of this would be the pattern (3, small, 6.5, [1, 10]).
- The above example gives a pattern with 4 features. The 4^{th} feature is in the form of an interval.
- In this case the feature falls within the range 1 to 10. This is also used when there are missing values. When a particular feature of a pattern is missing, looking at other patterns, we can find a range of values which this feature can take. This can be represented as an interval.
- The example pattern given above has the second feature as a linguistic value. The first feature is an integer and the third feature is a real value.
- Rough set theory was developed by Pawlak for classificatory analysis of data tables. The main goal of rough set theoretic analysis is to synthesise approximation (upper and lower) of concepts from the acquired data.
- While fuzzy set theory assigns to each object a grade of belongingness to represent an imprecise set, the focus of rough set theory is on the ambiguity caused by limited discernibility of objects in the domain of discourse.
- Rough sets are used to represent classes. So, a class description will consist of an upper approximate set and a lower approximate set.
- An element y belongs to the lower approximation if the equivalence class to which y belongs is included in the set. On the other hand y belongs to the upper approximation of the set if its equivalence class has a non-empty intersection with the set.
- The lower approximation consists of objects which are members of the set with full certainty. The upper approximation consists of objects which may possibly belong to the set.

Q.18 Which method are used for represented dataset by using tree ?

Ans. : Representing a dataset as a tree are minimum Spanning Tree, Frequent Pattern Trees (FP trees) and k-dimensional trees.

Q.19 Explain k-dimensional tree.

Ans. : • k-dimensional tree is also called as K-d tree.

- It is a data structure for storing patterns in k-dimensional space by partitioning the space.
- The k-d tree is a binary tree structure for storing and performing operations on data containing k-dimensional keys.
- The K-d tree is used to represent a set of patterns which have been partitioned. Each region of the K-d tree represents a subset of the patterns.
- Consider the set of patterns represented by Fig. Q.19.1.

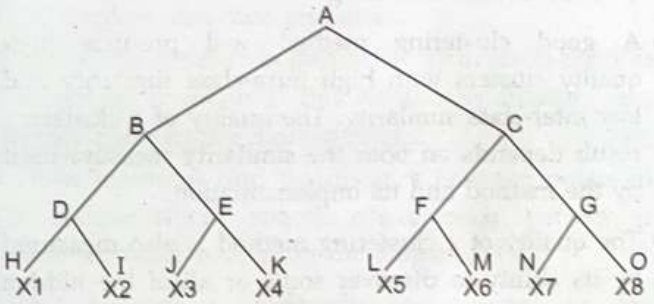


Fig. Q.19.1

- A represents the complete set of patterns. At each point, the set of patterns is partitioned into two blocks giving rise to a binary tree.
- A is partitioned into B and C. B is partitioned into D and E. C is partitioned into F and G.
- D is partitioned into H and I. E is partitioned into J and K. F is partitioned into L and M.
- G is partitioned into N and O.
- Fig. Q.19.2 shows gives the K-d tree for the above data.

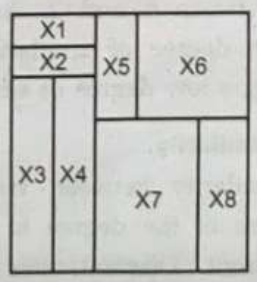


Fig. Q.19.2

Q.20 Describe Frequent Pattern Trees (FP trees).

Ans. : • FP tree is used to represent the patterns in a transaction database. An example of a transaction database is the collection of transactions at a supermarket.

- In this database, it is necessary to find the association between items in the database.
- It is necessary to first find the frequent items in the database. Then the FP-tree can be constructed to help in efficiently mine the frequent item-sets in the database.
- The FP tree is a compressed tree structure where only items which occur more frequently are considered. The number of times an item occurs in the transaction database is the support of that item.
- When we say that only frequently occurring items are included, it means that items which have a support below a threshold are left out.
- The items in the transactions are then reordered so that more frequent items occur earlier and the less frequent items occur later in the transaction.
- The first transaction is drawn with the items as nodes and links connecting the nodes. Each item has a count of one.
- The second transaction is then drawn. If it has a overlap with the existing link then it is not redrawn but the count is increased of those items.

1.5 : Representations of Clusters

Q.21 Explain clustering.

Ans. : Clustering is a process of partitioning a set of data in a set of meaningful subclasses. Every data in the subclass shares a common trait. It helps a user to understand the natural grouping or structure in a data set.

Q.22 What are the desirable properties of a clustering algorithm ?

Ans. : 1. Scalability (in terms of both time and space)

2. Ability to deal with different data types
3. Minimal requirements for domain knowledge to determine input parameters
4. Interpretability and usability

Q.23 What is the goal of clustering ?

Ans. : • The goal of clustering is to identify distinct groups in a dataset.

- But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering.
- Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding "natural clusters" and describe their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection).

Q.24 What do you mean clustering ?

Ans. : Clustering : • Clustering of data is a method by which large sets of data are grouped into clusters of smaller sets of similar data. Clustering can be considered the most important unsupervised learning problem.

- A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. Fig. Q.24.1 shows cluster.

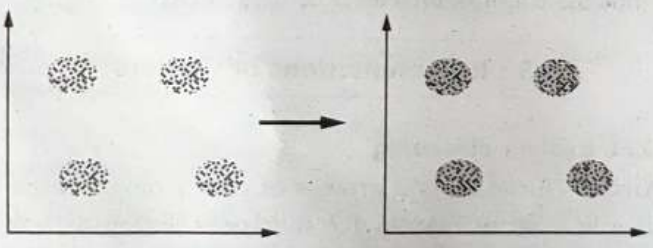


Fig. Q.24.1 Cluster

- Clustering means grouping of data or dividing a large data set into smaller data sets of some similarity.
- A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets.
- The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering ? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering.
- Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
- Clustering analysis helps construct meaningful partitioning of a large set of objects. Cluster analysis has been widely used in numerous applications, including pattern recognition, data analysis, image processing, etc.
- Clustering algorithms may be classified as listed below :
  1. Exclusive clustering
  2. Overlapping clustering
  3. Hierarchical clustering
  4. Probabilistic clustering
- A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation.
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

1.6 : Proximity Measures

Q.25 List the various proximity measures.

Ans. : Various proximity measures are distance measures, weighted distance measures, edit distance, non-metric similarity function and kernel functions.

Q.26 Define similarity.

Ans. : The similarity between two objects is a numerical measure of the degree to which the two objects are alike. Similarities are usually non-negative and are often between 0 and 1. A small distance indicating a high degree of similarity and a large distance indicating a low degree of similarity.

Q.27 Define dissimilarity.

Ans. : The dissimilarity between two objects is a numerical measure of the degree to which the two objects are different. Dissimilarities are lower for more similar pairs of objects.

Q.28 Mention any three measures of similarity.

Ans. : Similarity measures are cosine similarity, Euclidean distance and Manhattan distance.

Q.29 What is Euclidean distance ?

Ans. : The Euclidean distance is the most common distance metric used in low dimensional data sets. It is also known as the L2 norm. The Euclidean distance is the usual manner in which distance is measured in real world.

Q.30 Define Mahalanobis distance.

Ans. : Mahalanobis distance is also called quadratic distance. Mahalanobis distance is a distance measure between two points in the space defined by two or more correlated variables. Mahalanobis distance takes the correlations within a data set between the variable into considerations.

Q.31 Explain distance measures.

Ans. :

- The distance between two patterns is used as a proximity measure. If this distance is smaller, then the two patterns are more similar.
- These measures find the distance between points in a d-dimensional space, where each pattern is represented as a point in the d-space.
- The distance is inversely proportional to the similarity. If d(X,Y) gives the distance between X and Y, and s(X,Y) gives the similarity between X and Y, then,

$$d(X,Y) \propto \frac{1}{s(X,Y)}$$

- The Euclidian distance is the most popular distance measure. If we have two patterns X and Y, then the Euclidian distance will be

$$d_2(X,Y) = \sqrt{\sum_{k=1}^{d}(x_k - y_k)^2}$$

- The generalization of this equation gives the Minkowski distance which is

$$d_m(X,Y) = \left(\sum_{k=1}^{d}|x_k - y_k|^m\right)^{\frac{1}{m}}$$

Where, m = 1,2..., ∞

Depending on the value of m, we get different distance measures.

Q.32 Write short note on Euclidean distance.

Ans. : The Euclidean distance

- The Euclidean distance is the most common distance metric used in low dimensional data sets. It is also known as the L_2 norm. The Euclidean distance is the usual manner in which distance is measured in real world.

  $$d_{euclidean}(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

  Where x and y are m-dimensional vectors and denoted by

  $x = (x_1, x_2, x_3, \ldots x_m)$ and

  $y = (y_1, y_2, y_3, \ldots y_m)$ represent the m attribute values of two records.

- While Euclidean metric is useful in low dimensions, it does not work well in high dimensions and for categorical variables. The drawback of Euclidean distance is that is ignores the similarity between attributes. Each attribute is treated as totally different from all of the attributes.

Q.33 How the Manhattan distance between the two objects is calculated ?

Ans. : The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components. The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components.

The formula for this distance between a point.

$X = (X_1, X_2, \text{etc.})$ and a point $Y = (Y_1, Y_2, \text{etc.})$ is :

$$d = \sum_{i=1}^{n}|x_i - y_i|$$

Where n is the number of variables, and X_i and Y_i are the values of the i^{th} variable, at points X and Y respectively.

The distance between two points measured along axes at right angles. In a plane with p_1 at (x_1, y_1) and p_2 at (x_2, y_2), it is $|x_1 - x_2| + |y_1 - y_2|$.

Let the reference point for calculating cartesian coordinates be node 1.

In that case, the coordinates of node 1 is (0, 0)
the coordinates of node 2 is (1, 0)
the coordinates of node 10 is (0, 1)
the coordinates of node 11 is (1, 1)
the coordinates of node 21 is (2, 2)
the coordinates of node 30 is (2, 3)
the coordinates of node 41 is (4, 4)
the coordinates of node 81 is (8, 8)

Now use the formula to calculate the Manhattan distance.

The Manhattan distance between nodes 1 and 2 is $1 = |0 - 1| + |0 - 0|$.

The Manhattan distance between nodes 1 and 10 is $1 = |0 - 0| + |0 - 1|$.

The Manhattan distance between nodes 1 and 11 is $2 = |0 - 1| + |0 - 1|$.

The Manhattan distance between nodes 1 and 3 is 2.

The Manhattan distance between nodes 1 and 12 is 3.

Q.34 What is weighted distance measure ?

Ans. : • When some features are more significant than others, the weighted distance measure is used. The weighted distance is as follows :

$$d(X,Y) = \left(\sum_{k=1}^{d} w_k (x_k - y_k)^m\right)^{\frac{1}{m}}$$

- The weight for each feature depends on its importance.
  - The greater the significance of the feature, the larger the weight.
  - The weighted distance measure using Euclidean distance (m = 2) would be,

$$d_2(X,Y) = w_1 * (x_1 - y_1)^2 + w_2 * (x_2 - y_2)^2 + \ldots + w_d * (x_d - y_d)^2$$

- A generalized version of the weighted distance is the squared Mahalanobis distance (MD).
- Let a class be represented by N (μ, Σ), where μ is the mean and Σ is the covariance matrix, then the squared Mahalanobis distance between a pattern X and the class is

$$(X-\mu)^t \sum^{-1} (X-\mu)$$

If $\Sigma = I$ (Identity Matrix), them MD = ED

Q.35 What is k-Median Distance ?

Ans. : • The k-Median distance is a non-metric distance function. This finds the distance between two vectors.

- The difference between the values for each element of the vector is found. Putting this together, we get the difference vector.
- If the values in the difference vector are put in non-decreasing order, the kTH value gives the k-Median distance.

1.7 : Abstractions of the Data Set

Q.36 What is data ?

Ans. : Data is collection of data objects and their attributes.

Q.37 What is data abstraction ?

Ans. : Data abstraction is the process of extracting a simple and compact representation of a data set. Here simplicity is either from the perspective of automatic analysis, so that a machine can perform further processing efficiently or it is human-oriented , so that the representation obtained is easy to comprehend and intuitively appealing. In the clustering context a typical data abstraction is a compact description of each cluster usually in terms of cluster prototypes or representative patterns such as the centroid.

Q.38 Define various data types.

Ans. : Types of data are record data, data matrix, document data, transaction data, graph data and ordered data.

Q.39 List the types of data sets.

Ans. : Types of data sets are as follows :

1. Record : Data Matrix, Document Data and Transaction Data
2. Graph : World Wide Web and Molecular Structures
3. Ordered : Spatial Data - Temporal Data - Sequential Data - Genetic Sequence Data

1.8 : Feature Extraction and Feature Selection

Q.40 What is feature reduction ?

Ans. : Feature reduction refers to the mapping of the original high-dimensional data onto a lower-dimensional space.

Q.41 What is difference between feature selection vs dimensionality reduction ?

Ans. : Dimensionality Reduction :

- When classifying novel patterns, all features need to be computed.
- New features are combinations of the original features.

Feature Selection :

- When classifying novel patterns, only a small number of features need to be computed.
- New features is just a subset of the original features.

Q.42 Define feature extraction.

Ans. : Feature extraction is the process of converting raw data to numerical information which a computer can recognize. The computer cannot recognize the raw text information directly.

Q.43 What is feature selection ?

Ans. : Feature selection is the process of removing redundant and irrelevant features, that is, to select the most effective feature from the raw features. It can reduce the dimensionality of the datasets and improve the performance of the algorithm.

Q.44 What is dimension reduction ? List the components of dimension reduction.

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.

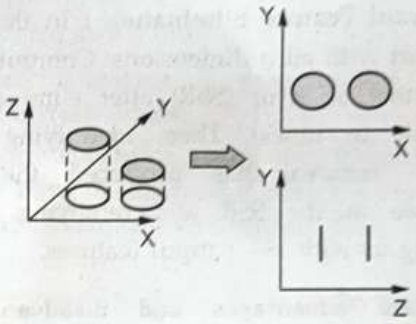


Fig. Q.44.1

- There are two components of dimensionality reduction :
- Feature selection : In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways : Filter, Wrapper and Embedded.
- Feature extraction : This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Q.45 What are the common methods to perform dimension reduction ?

Ans. : There are many methods to perform Dimension reduction.

1. **Missing Values :** While exploring data, if we encounter missing values, what we do ? Our first step should be to identify the reason then impute missing values / drop variables using appropriate methods. But, what if we have too many missing values ? Should we impute missing values or drop the variables ?
2. **Low Variance :** Let's think of a scenario where we have a constant variable in our data set.
3. **Decision Trees :** It can be used as a ultimate solution to tackle multiple challenges like missing values, outliers and identifying significant variables.
4. **Random Forest :** Similar to decision tree is Random Forest.

5. **High Correlation :** Dimensions exhibiting higher correlation can lower down the performance of model. Moreover, it is not good to have multiple variables of similar information or variation also known as "Multicollinearity".
6. **Backward Feature Elimination :** In this method, we start with all n dimensions. Compute the sum of square of error (SSR) after eliminating each variable (n times). Then, identifying variables whose removal has produced the smallest increase in the SSR and removing it finally, leaving us with n – 1 input features.

Q.46 Explain advantages and disadvantages of dimensionality reduction.

Ans. : Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.
- We may not know how many principal components to keep- in practice, some thumb rules are applied.

Q.47 What is principal component analysis ?

Ans. : Principal Component Analysis (PCA) is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retains most of the sample's information and useful for the compression and classification of data.

Q.48 What are the different feature extraction techniques ?

Ans. : • Feature extraction involves reducing the number of resources required to describe a large set of data.

- Different techniques are as follows :

1. Independent component analysis.
2. Kernel PCA.
3. Latent semantic analysis.
4. Partial least squares.
5. Principal component analysis.
6. Multifactor dimensionality reduction.
7. Nonlinear dimensionality reduction.

Q.49 Write short note on PCA.

Ans. : • This method was introduced by Karl Pearson. It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

- Principal Component Analysis (PCA) is to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retains most of the sample's information and useful for the compression and classification of data.

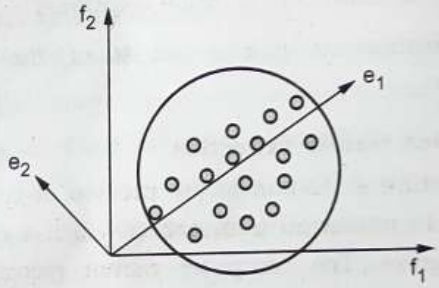


Fig. Q.49.1 PCA

- In PCA, it is assumed that the information is carried in the variance of the features, that is, the higher the variation in a feature, the more information that feature carries.
- Hence, PCA employs a linear transformation that is based on preserving the most variance in the data using the least number of dimensions.
- It involves the following steps :

1. Construct the covariance matrix of the data.
2. Compute the eigenvectors of this matrix.
3. Eigenvectors corresponding to the largest eigen values are used to reconstruct a large fraction of variance of the original data.

- The data instances are projected onto a lower dimensional space where the new features best represent the entire data in the least squares sense.

- It can be shown that the optimal approximation, in the least square error sense, of a d-dimensional random vector $x_2 < d$ by a linear combination of independent vectors is obtained by projecting the vector x onto the eigenvectors ei corresponding to the largest eigen values li of the covariance matrix (or the scatter matrix) of the data from which x is drawn.
- The eigenvectors of the covariance matrix of the data are referred to as principal axes of the data, and the projection of the data instances on to these principal axes are called the principal components. Dimensionality reduction is then obtained by only retaining those axes (dimensions) that account for most of the variance, and discarding all others.
- In the figure below, Principal axes are along the eigenvectors of the covariance matrix of the data. There are two principal axes shown in the figure, first one is closed to origin, the other is far from origin.
- If $X = X_1, X_2, \ldots\ldots, X_N$ is the set of n patterns of dimension d, the sample mean of the data set is

$$m = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The sample covariance matrix is

$$C = (X-m)(X-m)^T$$

- C is a symmetric matrix. The orthogonal basis can be calculated by finding the eigenvalues and eigenvectors.

The eigenvectors g_i and the corresponding eigenvalues λ_i are solutions of the equation

$$C*g_i = \lambda_i * g_i \quad i = 1, \ldots\ldots, d$$

The eigenvector corresponding to the largest eigenvalue gives the direction of the largest variance of the data. By ordering the eigenvectors according to the eigenvalues, the directions along which there is maximum variance can be found.

If E is the matrix consisting of eigenvectors as row vectors, we can transform the data X to get Y .

$$Y = E(X - m)$$

- The original data X can be got from Y as follows :

$$X = E^t Y + m$$

- Instead of using all d eigenvectors, the data can be represented by using the first k eigenvectors where $k < d$.
- If only the first k eigenvectors are used represented by E_K , then

$$Y = E_K (X - m) \text{ and } X' = E_k^t Y + m$$

Q.50 Explain feature construction and transformation techniques.

Ans. : • Feature construction involves transforming a given set of input features to generate a new set of more powerful features which can then used for prediction.

- Feature construction methods may be applied to pursue two distinct goals : reducing data dimensionality and improving prediction performance.
- **Steps :**
  1. Start with an initial feature space F_0
  2. Transform F_0 to construct a new feature space F_N
  3. Select a subset of features F_i from F_N
  4. If some terminating criteria is achieved : Go back to step 3 otherwise set $F_T = F_i$
  5. F_T is the newly constructed feature space.
- The initial feature space F_0 consists of manually constructed features that often encode some basic domain knowledge.
- The task of constructing appropriate features is often highly application specific and labour intensive. Thus building auto-mated feature construction methods that require minimal user effort is challenging. In particular we want methods that :
  1. Generate a set of features that help improve prediction accuracy.
  2. Are computationally efficient.
  3. Are generalizable to different classifiers.
  4. Allow for easy addition of domain knowledge.
- Genetic programming is an evolutionary algorithm based technique that starts with a population of individuals, evaluates them based on some fitness function and constructs a new population by applying a set of mutation and crossover operators on high scoring individuals and eliminating the low scoring ones.

- In the feature construction paradigm, genetic programming is used to derive a new feature set from the original one. Individuals are often tree like representations of features, the fitness function is usually based on the prediction performance of the classifier trained on these features while the operators can be applications specific.
- The method essentially performs a search in the new feature space and helps generate a high performing subset of features. The newly generated features may often be more comprehensible and intuitive than the original feature set, which makes GP-related methods well-suited for such tasks.
- In decision trees, the model explicitly selects features that are highly correlated with the label. In particular, by limiting the depth of the decision tree, one can at least hope that the model will be able to throw away irrelevant features.
- In the case of K-nearest neighbors, the situation is perhaps more terrible. Since KNN weighs each feature just as much as another feature, the introduction of irrelevant features can completely mess up KNN prediction.
- **Feature extraction** is a process that extracts a set of new features from the original features through some functional mapping.
- Transformation studies ways of mapping original attributes to new features. Different mappings can be employed to extract features. In general the mappings can be categorized into linear or nonlinear transformations. One could categorize transformations along two dimensions linear and labeled, linear and non labeled, nonlinear and labeled, nonlinear and non labeled.
- Many data mining techniques can be used in transformation such as EM k - Means k - Medoidsand multi layer perceptron etc

Q.51 What is feature selection ? Explain the role of feature selection in machine learning. Explain feature selection algorithm.

Ans. : • Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.

- Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often.
- Perhaps the word "groovy" appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature? It's a dangerous endeavor because it's hard to tell with just one training example if it is really correlated with the positive class, or is it just noise.
- You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.
- There are three general classes of feature selection algorithms : filter methods, wrapper methods and embedded methods.
- The role of feature selection in machine learning is,
  1. To reduce the dimensionality of feature space
  2. To speed up a learning algorithm
  3. To improve the predictive accuracy of a classification algorithm
  4. To improve the comprehensibility of the learning results.
- Features Selection Algorithms are as follows :
  1. **Instance based approaches :** There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.
  2. **Nondeterministic approaches :** Genetic algorithms and simulated annealing are also used in feature selection.
  3. **Exhaustive complete approaches :** Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

Q.52 What is subset selection ? Explain its types.

Ans. : Subset Selection

- Finding the best subset of the set of features is main aim of subset selection. The best subset contains the least number of dimensions that most contribute to accuracy.

Ans. : • As an example, let us consider two patterns of class 1. The patterns are (1, 1) and (1, 2). Let us consider two patterns of class 2. The patterns are (4, 4) and (5, 4). The covariance matrix is

$$C_x = \begin{bmatrix} 4.24 & 2.92 \\ 2.92 & 2.25 \end{bmatrix}$$

The eigen value of C_x are

$$\lambda = \begin{bmatrix} 6.366 \\ 0.1635 \end{bmatrix}$$

The first eigenvalue is very much larger than the second eigenvalue. The eigenvector corresponding to this is

$$\text{eigen}_1 = \begin{bmatrix} 0.814 \\ 0.581 \end{bmatrix}$$

The pattern (1, 1) gets transformed to

$$[0.814 \quad 0.581] \times \begin{bmatrix} -1.75 \\ -1.75 \end{bmatrix} = -2.44$$

- Similarly, the patterns (1, 2), (4, 4) and (5, 4) get transformed to – 1.86, 1.74 and 2.56.
- When we try to get the original data from the transformed, some information gets lost. After transformation, pattern (1, 1) becomes,

$$[0.814 \quad 0.581]^*(-2.44) + \begin{bmatrix} 2.75 \\ 2.75 \end{bmatrix} = \begin{bmatrix} 0.76 \\ 1.332 \end{bmatrix}$$

1.9 : Evaluation of Classifier, Evaluation of Clustering

Q.60 What is classification ?

Ans. : Classification is a technique used to predict group membership for data instances. For example, you may wish to use classification to predict whether the weather on a particular day will be "sunny", "rainy" or "cloudy".

Q.61 What is rule classification ?

Ans. : The term rule-based classification can be used to refer to any classification scheme that make use of IF-THEN rules for class prediction. They are also used in the class prediction algorithm to give a ranking to the rules which will be then be utilized to predict the class of new cases.

Q.62 Mention types of classifier techniques.

Ans. : Types of classifier techniques are back-propagation, support vector machines, and k-nearest-neighbor classifiers.

Q.63 What is association based classification ?

Ans. : • Association-based classification, which classifies documents based on a set of associated, frequently occurring text patterns.

- Notice that very frequent terms are likely poor discriminators. Thus only those terms that are not very frequent and that have good discriminative power will be used in document classification.
- Such an association-based classification method proceeds as follows :

1. Keywords and terms can be extracted by information retrieval and simple association analysis techniques.
2. Concept hierarchies of keywords and terms can be obtained using available term classes, such as WordNet, or relying on expert knowledge, or some keyword classification systems.

Q.64 What is classification ? Explain with diagram.

Ans. : • Classification predicts categorical labels (classes), prediction models continuous-valued functions. Classification is considered to be supervised learning.

- Classifies data based on the training set and the values in a classifying attribute and uses it in classifying new data. **Prediction** means models continuous-valued functions, i.e., predicts unknown or missing values.
- Preprocessing of the data in preparation for classification and prediction can involve data cleaning to reduce noise or handle missing values, relevance analysis to remove irrelevant or redundant attributes, and data transformation, such as generalizing the data to higher level concepts or normalizing data.
- Fig. Q.64.1 shows the classification.

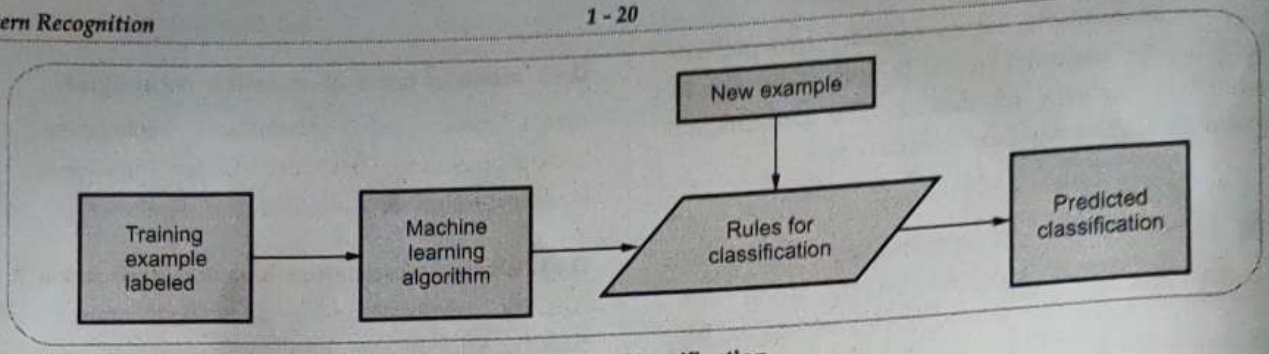


Fig. Q.64.1 Classification

Aim : To predict categorical class labels for new samples

Input : Training set of samples, each with a class label

Output : classifier is based on the training set and the class labels

- **Prediction** is similar to classification. It constructs a model and uses the model to predict unknown or missing value.
- Classification is the process of finding a model that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data.
- Classification and prediction may need to be preceded by relevance analysis, which attempts to identify attributes that do not contribute to the classification or prediction process.
- Numeric prediction is the task of predicting continuous values for given input. For example, we may wish to predict the salary of college employee with 15 years of work experience, or the potential sales of a new product given its price.
- Some of the classification methods like back-propagation, support vector machines, and k-nearest-neighbor classifiers can be used for prediction.

Fill in the Blanks for Mid Term Exam

Q.1 Data _____ is the process of extracting a simple and compact representation of a data set.

Q.2 The _____ distance is the most common distance metric used in low dimensional data sets.

Q.3 Feature _____ refers to the process of identifying and combining certain features of pattern.

Q.4 The k-Median distance is a ____ distance function.

Q.5 The distance between two patterns is used as a _____. If this distance is smaller, then the two patterns are more similar.

Q.6 The minimum spanning tree is the spanning tree which gives the total _____ cost.

Q.7 A _____ is a tree which consists of a subset of all the edges in the graph which spans all the nodes of the graph.

Q.8 A pattern represents a physical ____ or an abstract notion.

Multiple Choice Questions for Mid Term Exam

Q.1 ____ distance is also called quadratic distance.

a Euclidean b Mahalanobis
c k-Median d None of these

Q.2 Types of classifier techniques are _____.

a back-propagation
b support vector machines
c k-nearest-neighbor classifiers
d all of these

Q.3 _____ pattern recognition attempts to classify patterns based on a set of extracted features and an underlying statistical model for the generation of these patterns.

a Statistical b Structural
c Syntactically d None