Assignment

UNIT-IV: Classification
Introduction, Decision tree induction

1) What are the different types of learnings?

Ans.) Machine learning is sub-categorized to three types:-

- Supervised learning
- Unsupervised learning
- Reinforcement learning

- **Supervised Learning:-**

Supervised Learning is the one, where you can consider the learning is guided by a teacher. We have a dataset which acts as a teacher and its role is to train the model or the machine. Once the model gets trained it can start making a prediction or decision when new data is given to it.

- **Unsupervised Learning:-**

The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes.

- **Reinforcement Learning:-**

It is the ability of an agent to interact with the environment and find out what is the best outcome. It follows the concept of hit and trial method. The agent is rewarded or penalized with a point for a correct or a wrong answer, and on the basis of the positive reward points gained the model trains itself. And again once trained it gets ready to predict the new data presented to it.

2) What is the difference between supervised and unsupervised learning?

Ans.)

Difference between supervised and unsupervised learning:-

BASIS FOR COMPARISON	SUPERVISED LEARNING	UNSUPERVISED LEARNING
Basic	Deals with labelled data.	Handles unlabeled data.
Computational complexity	High	Low
Analyzation	Offline	Real-time
Accuracy	Produces accurate results	Generates moderate results
Sub-domains	Classification and regression	Clustering and Association rule mining

3) Define classification, regression and clustering?

Ans.)

Classification:- Classification is the type of Supervised Learning in which labelled data can use, and this data is used to make predictions in a non-continuous form. The output of the information is not always continuous, and the graph is non-linear. In the classification technique, the algorithm learns from the data input given to it and then uses this learning to classify new observation. This data set may merely be bi-class, or it may be multi-class too.

Ex:- One of the examples of classification problems is to check whether the email is spam or not spam by train the algorithm for different spam words or emails.

Regression:- **Regression** is the type of Supervised Learning in which labelled data used, and this data is used to make predictions in a continuous form. The output of the input is always ongoing, and the graph is linear. Regression is a form of predictive modelling technique which investigates the relationship between a dependent variable[**Outputs**] and independent variable[**Inputs**]. This technique used for forecasting the weather, time series modelling, process optimisation.

Ex:- One of the examples of the regression technique is House Price Prediction, where the price of the house will predict from the inputs such as No of rooms, Locality, Ease of transport, Age of house, Area of a home.

Clustering:- Clustering is the type of Unsupervised Learning in which unlabeled data used, and it is the process of grouping similar entities together, and then the grouped data is used to make clusters. The goal of this unsupervised machine learning technique is to find similarities in the data point and group similar data points together and to figures out that new data should belong to which cluster.

4) What are the advantages and disadvantages of the decision tree?

Ans.)

Advantages of decision tree:-

- Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
- A decision tree does not require normalization of data.
- A decision tree does not require scaling of data as well.
- Missing values in the data also does NOT affect the process of building decision tree to any considerable extent.
- A Decision trees model is very intuitive and easy to explain to technical teams as well as stakeholders.

Disadvantages of decision tree:-

- A small change in the data can cause a large change in the structure of the decision tree causing instability.

- For a Decision tree sometimes calculation can go far more complex compared to other algorithms.
- Decision tree often involves higher time to train the model.
- Decision tree training is relatively expensive as complexity and time taken is more.
- Decision Tree algorithm is inadequate for applying regression and predicting continuous values.

5) What is entropy? What is the formula for entropy?
Ans.) Entropy is a measure of randomness. In other words, its a measure of unpredictability. Let's take an example of a coin toss. Suppose we tossed a coin 4 times, and the output of the events came as {Head, Tail, Tail, Head}. Based solely on this observation, if you have to guess what will be the output of the coin toss, what would be your guess?
Maybe.. two heads and two tails. Fifty percent probability of having head and fifty percent probability of having a tail. You can not be sure. The output is a random event between head and tail. But what if we have a biased coin, which when tossed four times, gives following output: {Tail, Tail, Tail, Head}. Here, if you have to guess the output of the coin toss, what would be your guess? Chances are you will go with Tail, and why? Because seventy-five percent chance is the output is tail based on the sample set that we have. In other words, the result is less random in case of the biased coin than what it was in case of the perfect coin.

The formal definition of the entropy H of a discrete probability distribution X is:

$$H(X) = -\Sigma ni=1 P(xi) log2 P(xi)$$

6) What is high entropy, low entropy and information gain?
Ans.)

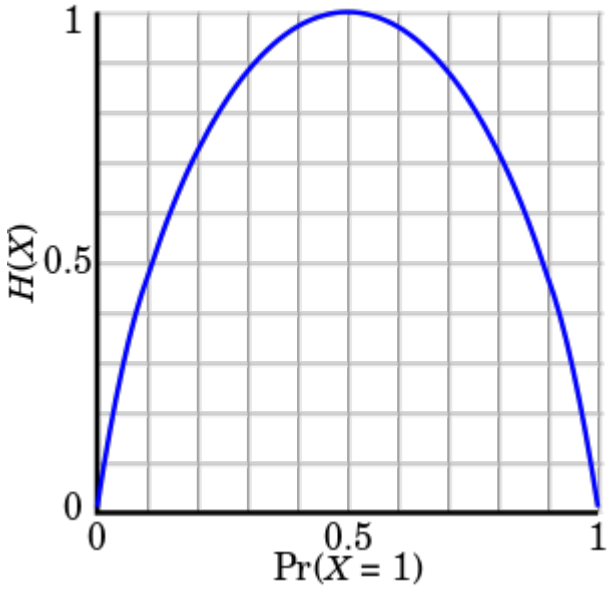
The **information entropy**, often just **entropy**, is a basic quantity in information theory associated to any random variable, which can be interpreted as the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes.
The entropy is the expected value of the self-information, a related quantity also introduced by Shannon.

The entropy can also be interpreted as the average rate at which information is produced by a stochastic source of data.

Given a random variable X, with possible outcomes x_i, each with probability $P_X(x_i)$, the entropy $H(X)$ of X is as follows:

$$H(X) = -\sum_i P_X(x_i) \log_b P_X(x_i) = \sum_i P_X(x_i) I_X(x_i) = \mathrm{E}[I_X]$$

where $I_X(x_i)$ is the self-information associated with particular outcome; I_X is the self-information of the random variable X in general, treated as a new derived random variable; and $\mathrm{E}[I_X]$ is the expected value of this new random variable, equal to the sum of the self-information of each outcome, weighted by the probability of each outcome occurring[3]; and b, the base of the logarithm, is a new parameter that can be set different ways to determine the choice of units for information entropy.



Here, the entropy is at most 1 bit, and to communicate the outcome of a coin flip (2 possible values) will require an average of at most 1 bit (exactly 1 bit for a fair coin). The result of a fair die (6 possible values) would require on average log26 bits.

Information gain = entropy (parent) – [weightes average] * entropy (children)

7) How does the decision tree work, explain with an example?

Ans.) Decision Tree Algorithm Pseudocode:-

1. Place the best attribute of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be made in such a way that each subset contains data with the same value for an attribute.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

In decision trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

We continue comparing our record's attribute values with other internal nodes of the tree until we reach a leaf node with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value. Now let's understanding how we can create the decision tree model.

Unit V: Cluster Analysis and Advance Topics

8) What is clustering and in what situations is it applied?

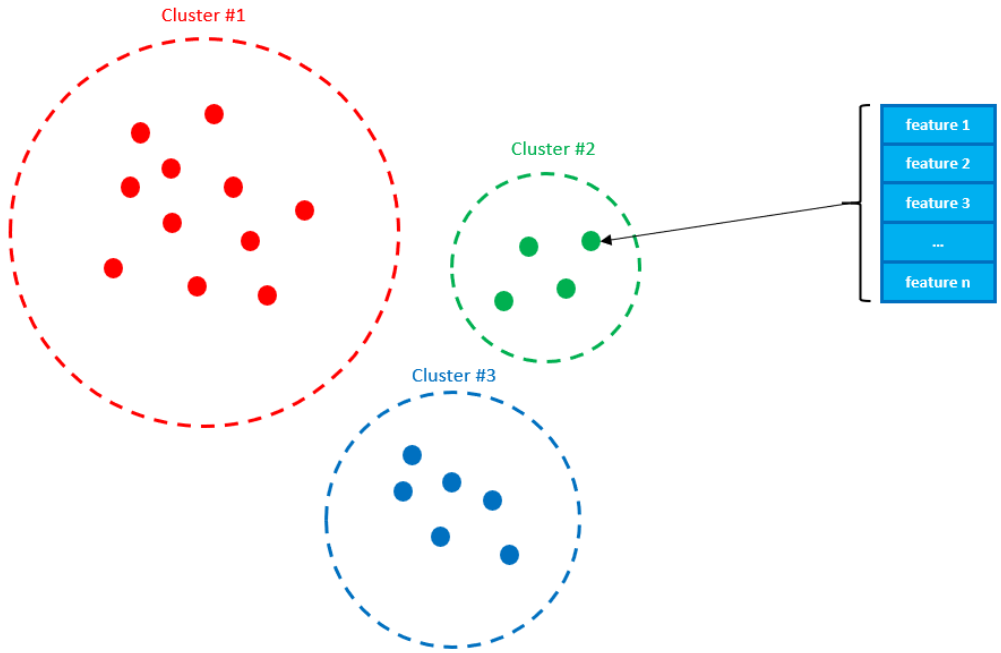
Ans.)

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Clustering algorithms are a powerful technique for machine learning on unsupervised data. The most common algorithms in machine learning are hierarchical clustering and K-Means clustering.

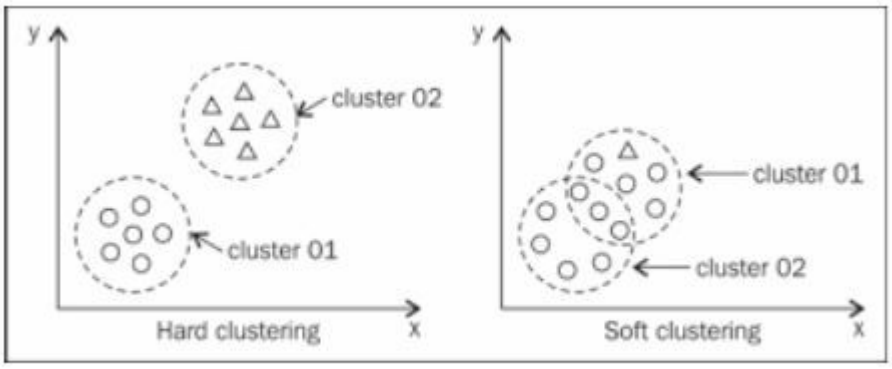
Why Clustering ?

Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for a good clustering. It depends on the user, what is the criteria they may use which satisfy their need. 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). This algorithm must make some assumptions which constitute the similarity of points and each assumption make different and equally valid clusters



Types of Clustering

Clustering can be divided into two subgroups :



- **Hard clustering:** It is about grouping the data items such that each piece is only assigned to one cluster. As an instance, we want the algorithm to read all of the tweets and determine if a tweet is a positive or negative tweet.
- **Soft Clustering**: Sometimes, we don't need a binary answer. Soft clustering is about grouping the data items such that an object can exist in multiple clusters.

Applications of Clustering

We can find clustering useful in the following areas:

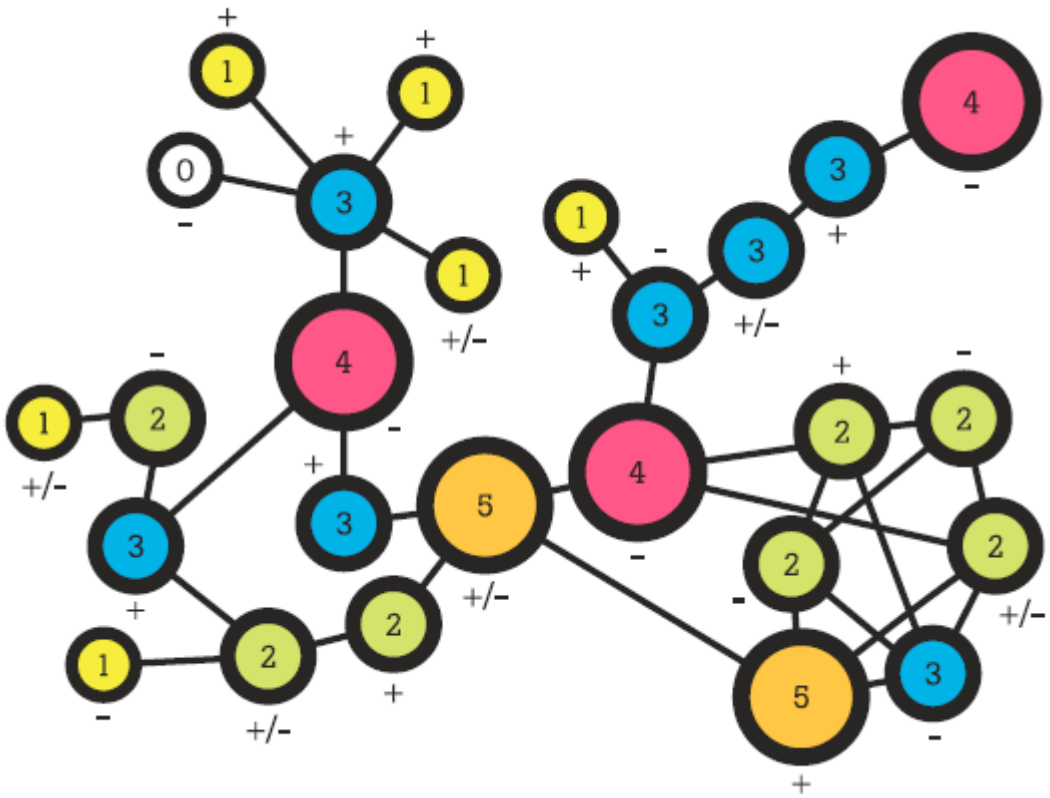
Customer Segmentation: Subdivision of customers into groups/segments such that each customer segment consists of customers with similar market characteristics — pricing, loyalty, spending behaviours etc. Some of the segmentation variables could be, e.g., the number of items bought on sale, avg transaction value, the total number of transactions.

Creating NewsFeeds: K-Means can be used to cluster articles by their similarity — it can separate documents into disjoint clusters.

Cloud Computing Environment: Clustered storage to increase performance, capacity, or reliability — clustering distributes workloads to each server, manages the transfer of workloads between servers, and provides access to all files from any server regardless of the physical location of the data.

Environmental risks: K-means can be used to analyse environmental risk in an area — environmental risk zoning of a chemical industrial area.

Pattern Recognition in images: For example, to automatically detect infected fruits or for segmentation of blood cells for leukaemia detection.



Social network analysis

Trend detection in dynamic data – Clustering can also be used for trend detection in dynamic data by making various clusters of similar trends.

Social network analysis – Clustering can be used in social network analysis. The examples are generating sequences in images, videos or audios and this approach is used in various fields.

Biological data analysis – Clustering can also be used to make clusters of images, videos; hence, it can successfully be used in biological data analysis.

9) What are the different types of clustering? Hint: You can explain any two from connectivity model, centroid model, distribution model and density model)
Ans.)

Connectivity Model -

Connectivity-based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts

of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

Connectivity-based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances), and UPGMA or WPGMA ("Unweighted or Weighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

Centroid Model -

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximate method is Lloyd's algorithm often just referred to as "k-means algorithm" (although another algorithm introduced this name). It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (k-means++) or allowing a fuzzy cluster assignment (fuzzy c-means).

10) What is the K means algorithm? Give an example and explain how does it work? (Hint: Use a pseudocode and write the algorithm)
Ans.)

k-means clusteringis a method ofvector quantization, originally fromsignal processing, that aims to*partition**n*observations into*k*clusters in which each observation belongs to theclusterwith the nearestmean(cluster centers or clustercentroid), serving as a prototype of the cluster. This results in a partitioning of the data space intoVoronoi cells. It is popular forcluster analysisindata mining.*k*-means clustering minimizes within-cluster variances (squared Euclidean distances), but not regular Euclidean distances, which would be the more difficultWeber problem: the mean optimizes

squared errors, whereas only the geometric median minimizes Euclidean distances. For instance, Better Euclidean solutions can be found usingk-mediansandk-medoids.

Algorithm:

The most common algorithm uses an iterative refinement technique. Due to its ubiquity, it is often called "the*k*-means algorithm"; it is also referred to asLloyd's algorithm, particularly in the computer science community. It is sometimes also referred to as "naive*k*-means", because there exist much faster alternatives.

Given an initial set of*k*means$m_1^{(1)},\ldots,m_k^{(1)}$(see below), the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster with the nearest mean: that with the least squaredEuclidean distance. (Mathematically, this means partitioning the observations according to theVoronoi diagramgenerated by the means.)

$$S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \leq \left\|x_p - m_j^{(t)}\right\|^2 \ \forall j, 1 \leq j \leq k\right\},$$

where each is assigned to exactly one,even if it could be assigned to two or more of them.

Update step: Recalculate means (centroids) for observations assigned to each cluster.

$$m_i^{(t+1)} = \frac{1}{\left|S_i^{(t)}\right|} \sum_{x_j \in S_i^{(t)}} x_j$$

The algorithm has converged when the assignments no longer change. The algorithm does not guarantee to find the optimum.

Pseudocode:

```
K-MEANS(P, k)
    Input: a dataset of points P = {p_1, ..., p_n}, a number of clusters k
    Output: centers {c_1, ..., c_k} implicitly dividing P into k clusters

1   choose k initial centers C = {c_1, ..., c_k}
2   while stopping criterion has not been met
3       do ▷ assignment step:
4           for i = 1, ..., N
5               do find closest center c_k ∈ C to instance p_i
6                   assign instance p_i to set C_k
7           ▷ update step:
8           for i = 1, ..., k
9               do set c_i to be the center of mass of all points in C_i
```

Example:

kmeans algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc..

Naïve Bayes Classifier

11) What is the advantages of using Naïve Bayes Classifier?
Ans.)

- Naive Bayes classifiers is a machine learning algorithm.
- Simplicity: easy to understand and implement.
- Light to train: no complicated optimisation required.
- Easily updateable if new training data is received.
- Small memory footprint.
- Although the independence assumption may seem sometimes unreasonable, its performance is usually good, even for those cases.
- It not only the prediction but also the degree of certainty, which can be very useful.
- Naive Bayes is particularly useful for Natural Language Processing.

12) What are the application of Naïve Bayes Classifier?
Ans.)

- It works well on small datasets. For most of the practical applications it hardly fits.
- NB has high bais and low variance. Hence it makes its application limited. Having said this there are no regularization or hyperparmeters tuning involved here to adjust the bias thing.
- Determining whether a given (text) document corresponds to one or more categories. In the text case, the features used might be the presence or absence of key words.
- To mark an email as **spam**, or **not spam** .
- Classify a news article about **technology**, **politics**, or **sports.**
- Check a piece of text expressing **positive** emotions, or **negative** emotions.
- Also used for face recognition softwares.