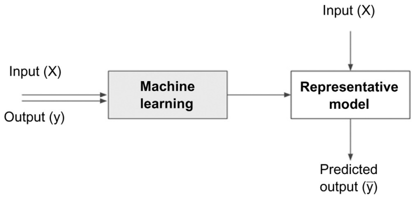
1)What is data science and why do we need data science?

Data science starts with data, in the form of numeric observations to a complex matrix ofwith millions of observations and thousands of variables. Data science utilizes certain specialized computational methods in order to discover meaningful and useful structures within a dataset.
The discipline of data science coexists and is closely associated with a number of related areas such as database systems, data engineering, visualization, data analysis, experimentation, and business intelligence.



The principal purpose of Data Science is to find patterns within data. It uses **various statistical techniques** to analyze and draw insights from the data.

2)Explain the data science classification and illustrate data science tasks.

Data science problems can be broadly categorized into supervised or unsupervised learning models.

Supervised Learning tries to use a function or relationship based on labeled training data and uses this function to map new unlabeled data. Supervised techniques predict the value of the output variables based on a set of input variables. To do this, a model is developed from a training dataset where the values of input and output are previously known. The model generalizes the relationship between the input and output variables and uses it to predict for a dataset where only input variables are known.
The output variable that is being predicted is also called a **class label or target variable.** Supervised data science needs a sufficient number of labeled records to train the model from the data. Unsupervised or undirected data science uncovers hidden patterns in unlabeled data. In unsupervised data science, there are no output variables to predict. The objective of this class of data science techniques, is to find patterns in data based on the relationship between data points themselves.
3)What is Data understanding.

Understand the attributes of the data.

Summarize the data by identifying key characteristics, such as data volume and total number of variables in the data.
Understand the problems with the data, such as missing values, inaccuracies, and outliers.
Visualize the data to validate the key characteristics of the data or unearth problems with the summary statistics.

3)Describe the various methods to understand data
Answer:
1)Data preparation

Preparing the dataset to suit a data science task is the most time-consuming part of the process. It is extremely rare that datasets are available in the form required by the data science algorithms. Most of the data science algorithms would require data to be structured.So data should be converted to the required format before it can be provide to the machine learning model.

2)Data Exploration

Data preparation starts with an in-depth exploration of the data and gaining a better understanding of the dataset. Data exploration, also known as exploratory data analysis, provides a set of simple tools to achieve basic understand ing of the data. Data exploration approaches involve computing descriptive statistics and visualization of data.

3)Data Visualization

Visualizing data is one of the most important techniques of data discovery and exploration. The visual representation of data provides easy comprehension of complex data with multiple attributes and their underlying relationships.

4)Explain the different types of data.

Data come in different formats and types.

Numeric or Continuous

Temperature expressed in Centigrade or Fahrenheit is numeric and continuous because it can be denoted by numbers and take an infinite number of values between digits.

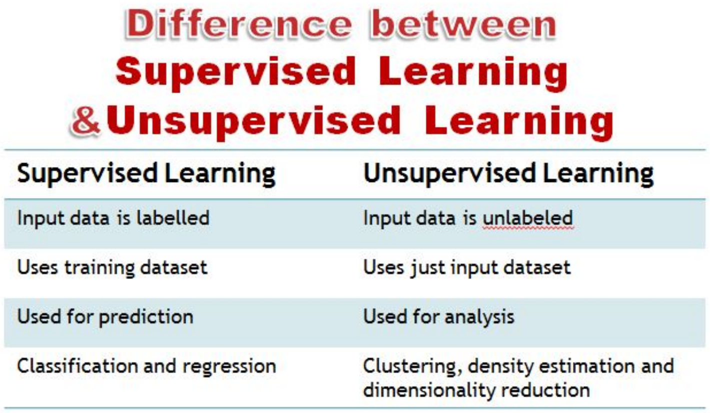
An integer is a special form of the numeric data type which does not have decimals in the value or more precisely does not have infinite values between consecutive numbers.

Categorical or Nominal

Categorical data types are attributes treated as distinct symbols or just names. The color of the iris of the human eye is a categorical data type because it takes a value like black, green, blue, gray, etc. There is no direct relationship among the data values.

5)Differentiate between supervised and unsupervised learning algorithms.

Define supervised and unsupervised learning and then write the below

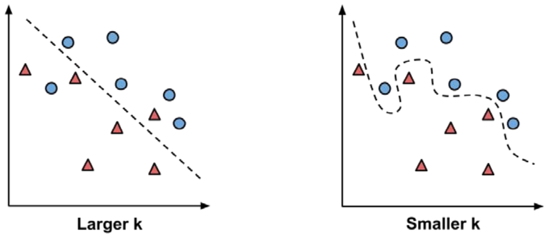


Difference between Supervised Learning &Unsupervised Learning

Supervised Learning	Unsupervised Learning
Input data is labelled	Input data is unlabeled
Uses training dataset	Uses just input dataset
Used for prediction	Used for analysis
Classification and regression	Clustering, density estimation and dimensionality reduction

6)Explain how to choose the value of k in k-NN algorithm.

Deciding how many neighbours to use for kNN determines how well the model will generalize to future data.The balance between over-fitting and under fitting the training data is a problem known as the bias-variance trade off. Choosing a large k reduces the impact or variance caused by noisy data, but can bias the learner such that it runs the risk of ignoring small, but important patterns.Typically, k is set somewhere between 3 and 10. One common practice is to set k equal to the square root of the number of training examples.



7)Explain entropy and information gain

Entropy is uncertainty/ randomness in the data, the more the randomness the higher will be the entropy. Information gain uses entropy to make decisions. If the entropy is less, information will be more.
Information gain is used in decision trees and random forest to decide the best split. Thus, the more the information gain the better the split and this also means lower the entropy.
The entropy of a dataset before and after a split is used to calculate information gain.
Entropy is the measure of uncertainty in the data. The effort is to reduce the entropy and maximize the information gain. The feature having the most information is considered important by the algorithm and is used for training the model

8)Explain the Ordinary Least Square method in regression.

Ordinary Least Squares regression (**OLS**) is a common technique for estimating coefficients of linear regression equations which describe the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).**Least square s**tand for the minimum **squares error (SSE)**. Maximum likelihood and Generalized method of moments estimator are alternative approaches to OLS.

In mathematical terms, the goal of OLS regression can be expressed as the task of

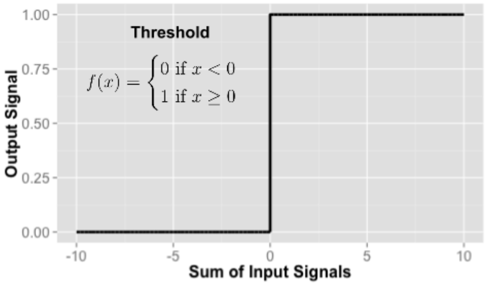
minimizing the following equation:

$$\sum (y - y_i \,\hat{}\,)^2 = \sum e_i^{\,2}$$

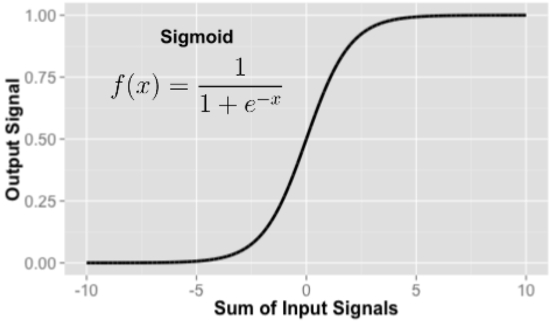
9)Define activation function. Give two examples.

An activation function is **a function used in artificial neural networks which outputs a small value for small inputs, and a larger value if its inputs exceed a threshold**. If the inputs are large enough, the activation function "fires", otherwise it does nothing.
Examples

A threshold activation function, results in an output signal only once a specified input threshold has been attained. The following figure depicts a typical threshold function; in this case, the neuron fires when the sum of input signals is at least zero. Because of its shape, it is sometimes called a unit step activation function.

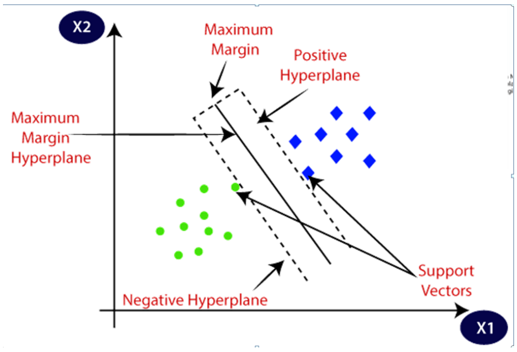


sigmoid activation function Although it shares a similar step or S shape with the threshold activation function, the output signal is no longer binary; output values can fall anywhere in the **range from 0 to 1.** Additionally, the sigmoid is differentiable, which means that it is possible to **calculate the derivative**across the entire range of inputs.



10)What is maximum margin hyper plane.

The distance between the line and the closest data points is referred to as the margin. The best or optimal line that can separate the two classes is the line that as the largest margin. This is called the Maximal-Margin hyperplane. The margin is calculated as the perpendicular distance from the line to only the closest points. Only these points are relevant in defining the line and in the construction of the classifier. These points are called the support vectors. They support or define the hyperplane. The hyperplane is learned from training data using an optimization procedure that maximizes the margin.



11)Define precision, recall and F-measure

precision (positive predictive value) in classifying the data instances. Precision is defined as follows:

$$Precision = \frac{TP}{TP + FP}$$

Precision should ideally be 1 (high) for a good classifier. *Precision* becomes 1 only when the numerator and denominator are equal i.e *TP = TP +FP*, this also means *FP* is zero.

Recall is also known as *sensitivity* or *true positive rate* and is defined as follows:

$$Recall = \frac{TP}{TP + FN}$$

Recall should ideally be 1 (high) for a good classifier. *Recall* becomes 1 only when the numerator and denominator are equal i.e *TP = TP +FN*, this also means *FN* is zero.

F1-score is a metric which takes into account both *precision* and *recall* and is defined as follows:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 Score becomes 1 only when *precision* and *recall* are both 1. *F1 score* becomes high only when both *precision* and *recall* are high. *F1 score* is the harmonic mean of *precision* and *recall* and is a better measure than *accuracy*.

12)Explain bagging and Boosting
Ensemble Learning models are based on the idea that combining multiple models can produce powerful models.
Bagging, also known as Bootstrap aggregation, is an ensemble learning method that looks for different ensemble learners by varying the training dataset. Unlike a single model trained on the entire dataset, bagging creates multiple weak learners or base models trained on a subset of the original dataset

How Does Bagging Work?

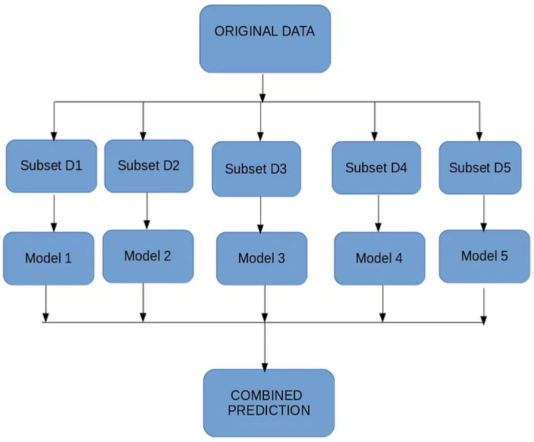
The steps involved in Bagging are:

Create multiple subsets from the training dataset by selecting observations with replacements.

Create a base model (also called a weak model).

Run base models simultaneously and independently of each other.

Combine predictions from all base models to determine the outcome.



Boosting is an ensemble technique that looks to change the training data and adjust the weight of the observations based on the previous classification. The weak learners take the results of the previous weak learner into account and adjust the weights of the data points, which converts the weak learner into a strong learner. Boosting changes the weight associated with an observation that was classified incorrectly by trying to increase the weight associated with it.

How Does Boosting work?

Following are the steps involved in the boosting technique:

1. A subset where all data points are given equal weights is created from the training dataset.
2. A based model is created for the initial dataset, and this model is used for predictions on the entire dataset.
3. Errors are calculated using predicted and actual values. The observation which was predicted incorrectly is given a higher weight.
4. The following model is created and boosting tries to correct the errors of the previous model.
5. The process is repeated for multiple models, each correcting the errors of the previous model.
6. The final model is a strong learner and is the weighted mean of all the models(weak learners.

13) Explain the various methods for visualising multivariate data.

Scatterplot

A scatterplot is one of the most powerful yet simple visual plots available. In a scatter plot, the data points are marked in Cartesian space with attributes of the dataset aligned with the coordinates. The attributes are usually of continuous data type.

Scatter Multiple

A scatter multiple is an enhanced form of a simple scatterplot where more than two dimensions can be included in the chart and studied simultaneously. The primary attribute is used for the x-axis coordinate. The secondary axis is shared with more attributes or dimensions.

Bubble Chart

A bubble chart is a variation of a simple scatterplot with the addition of one more attribute, which is used to determine the size of the data point

Density Chart

Density charts are similar to the scatterplots, with one more dimension included as a background color. The data point can also be colored to visualize one dimension, and hence, a total of four dimensions can be visualized in a density chart.

14) Explain the working of Neural Networks

Information is fed into the input layer which transfers it to the hidden layer

The interconnections between the two layers assign weights to each input randomly
A bias added to every input after weights are multiplied with them individually
The weighted sum is transferred to the activation function
The activation function determines which nodes it should fire for feature extraction
The model applies an application function to the output layer to deliver the output
Weights are adjusted, and the output is back-propagated to minimize error

15)What is Kernal trick in SVM

In many real-world applications, the relationships between variables are non-linear.As we just discovered, a SVM can still be trained on such data through the addition of a slack variable, which allows some examples to be misclassified. However, this is not the only way to approach the problem of non-linearity. A key feature of SVMs is their ability to map the problem into a higher dimension space using a process known as the kernel trick. In doing so, a non-linear relationship may suddenly appear to be quite linear.

16)Explain K means clustering algorithm

Let X = $\{x_1,x_2,x_3,……..,x_n\}$ be the set of data points and V = $\{v_1,v_2,…….,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center by taking the mean of the data point values

18)Explain cross validation and its types

Cross-validation is a technique for evaluating a machine learning model and testing its performance. CV is commonly used in applied ML tasks.

Steps

Divide the dataset into two parts: one for training, other for testing

1. Train the model on the training set
2. Validate the model on the test set
3. Repeat 1-3 steps a couple of times. This number depends on the CV method that you are using .Hold-out cross-validation

Hold-out cross-validation

It is the simplest and most common technique.

The algorithm of hold-out technique:

1. Divide the dataset into two parts: the training set and the test set. Usually, 80% of the dataset goes to the training set and 20% to the test set but you may choose any splitting that suits you better
2. Train the model on the training set
3. Validate on the test set
4. Save the result of the validation

k-Fold cross-validation

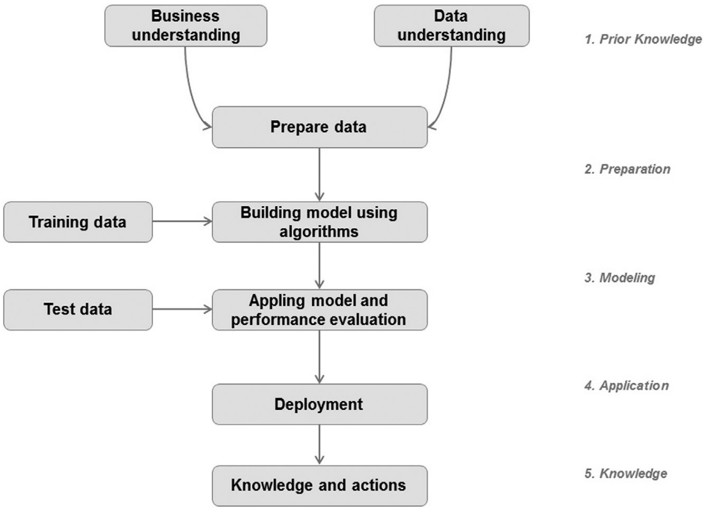
k-Fold cross-validation is a technique that minimizes the disadvantages of the hold-out method. k-Fold introduces a new way of splitting the dataset which helps to overcome the "test only once bottleneck".

The algorithm of the k-Fold technique:

1. Pick a number of folds – k. Usually, k is 5 or 10 but you can choose any number which is less than the dataset's length.
2. Split the dataset into k equal (if possible) parts (they are called folds)
3. Choose k – 1 folds as the training set. The remaining fold will be the test set
4. Train the model on the training set. On each iteration of cross-validation, you must train a new model independently of the model trained on the previous iteration
5. Validate on the test set
6. Save the result of the validation
7. Repeat steps 3 – 6 k times. Each time use the remaining fold as the test set. In the end, you should have validated the model on every fold that you have.

8. To get the final score average the results that you got on step

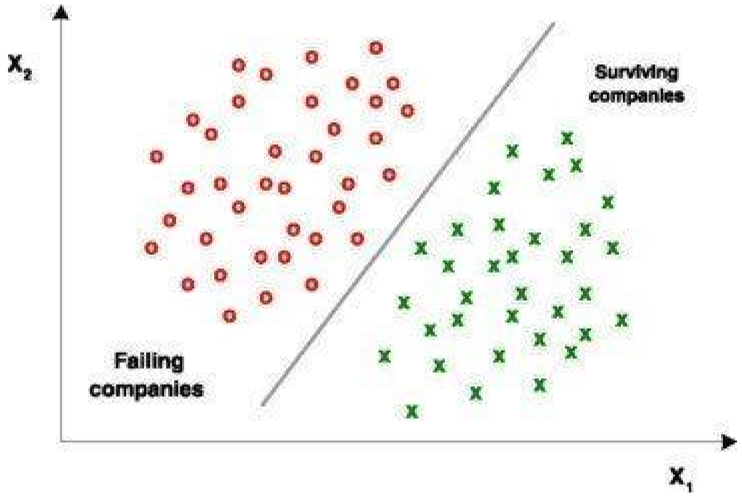
19)Explain the steps in Data science process



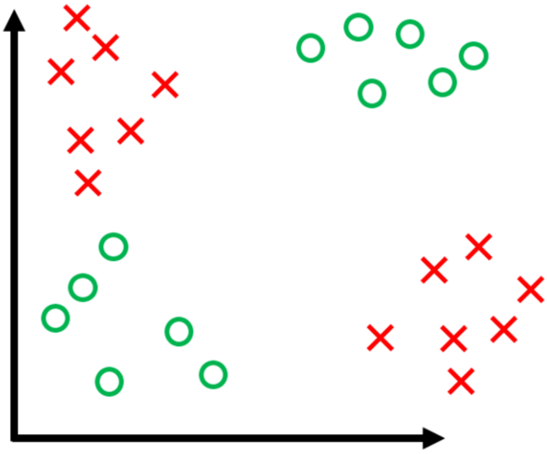
(Qno: 18 of Model Question paper)
20)Define linearly separable dataset. Give an example each of a dataset that is linearly separable and of a dataset that is not linearly separable.

Explain about the idea of linearly separable and non linearly separable problems and then add the below content

The idea of linearly separable is easiest to visualize and understand in 2 dimensions. Let the two classes be represented by colors red and green.
A dataset is said to be linearly separable if it is possible to draw a line that can separate the red and green points from each other.
Here are same examples of linearly separable data:



And here are some examples of linearly non-separable data



Example of Linearly separable dataset –Iris dataset
Example of non linearly separable dataset- make_circles dataset in sklearn