# Top 100+ Google Data Science Interview Questions

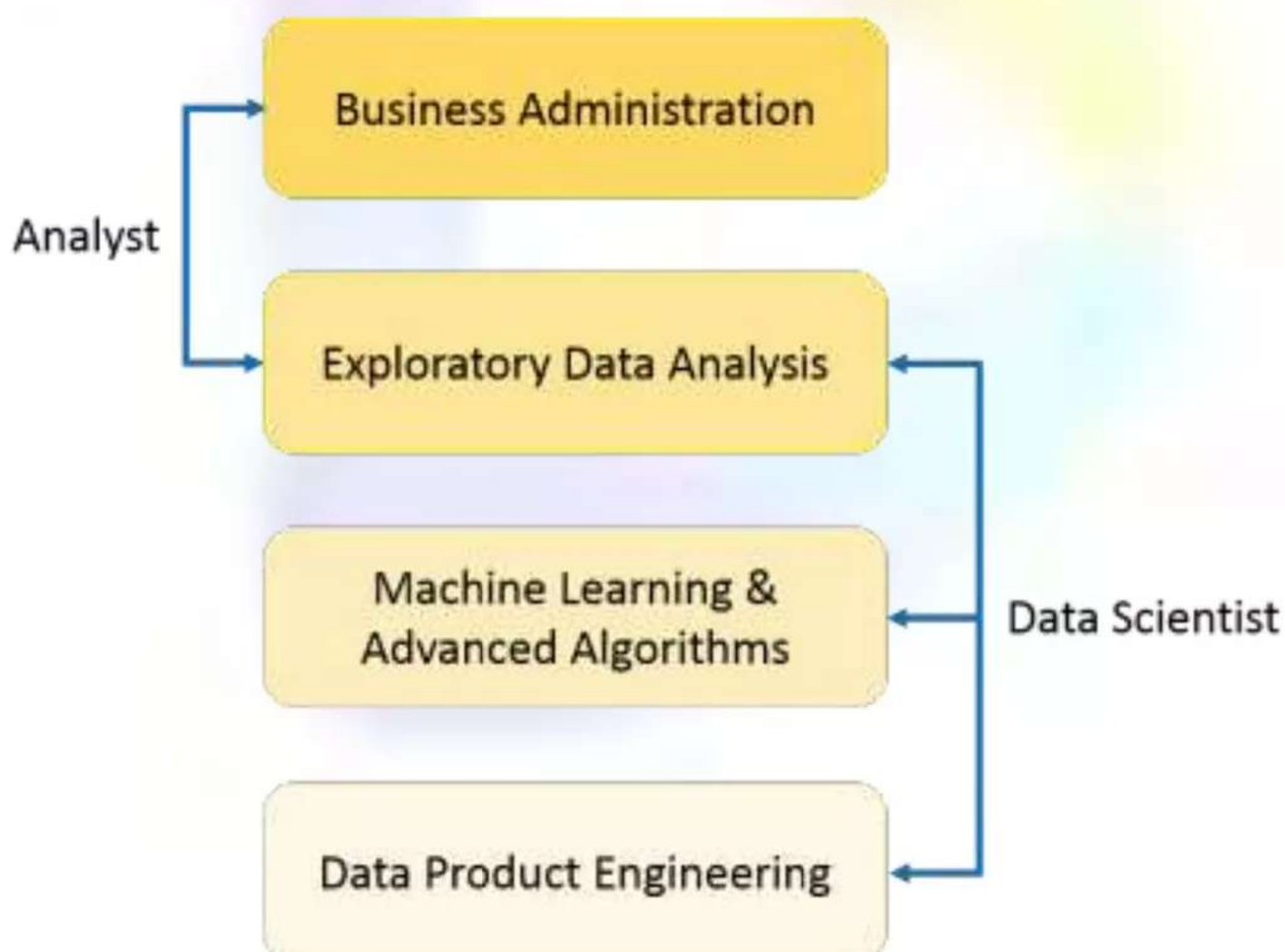
## 1. Why R is used in Data Visualization?

R is used in data visualization as it has many inbuilt functions and libraries that help in data visualizations. These libraries include ggplot2, leaflet, lattice, etc.

R helps in exploratory data analysis as well as feature engineering. Using R, almost any type of graph can be created. Customizing graphics is easier in R than using python.

## 2. What is Data Science? List the differences between supervised and unsupervised learning.

Data Science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns in the raw data. How is this different from what statisticians have been doing for years?
The answer lies in the difference between explaining and predicting.




The differences between supervised and unsupervised learning are as follows;

| Supervised Learning | Unsupervised Learning |
|---|---|
| Input data is labeled. | Input data is unlabelled. |
| Uses a training data set. | Uses the input data set. |
| Used for prediction. | Used for analysis. |
| Enables classification and | Enables Classification, Density Estimation, & Dimension Reduction |

| | |
|---|---|
| regression. | |

# 3. What is Selection Bias?

Selection bias is a kind of error that occurs when the researcher decides who is going to be studied. It is usually associated with research where the selection of participants isn't random. It is sometimes referred to as the selection effect. It is the distortion of statistical analysis, resulting from the method of collecting samples. If the selection bias is not taken into account, then some conclusions of the study may not be accurate.

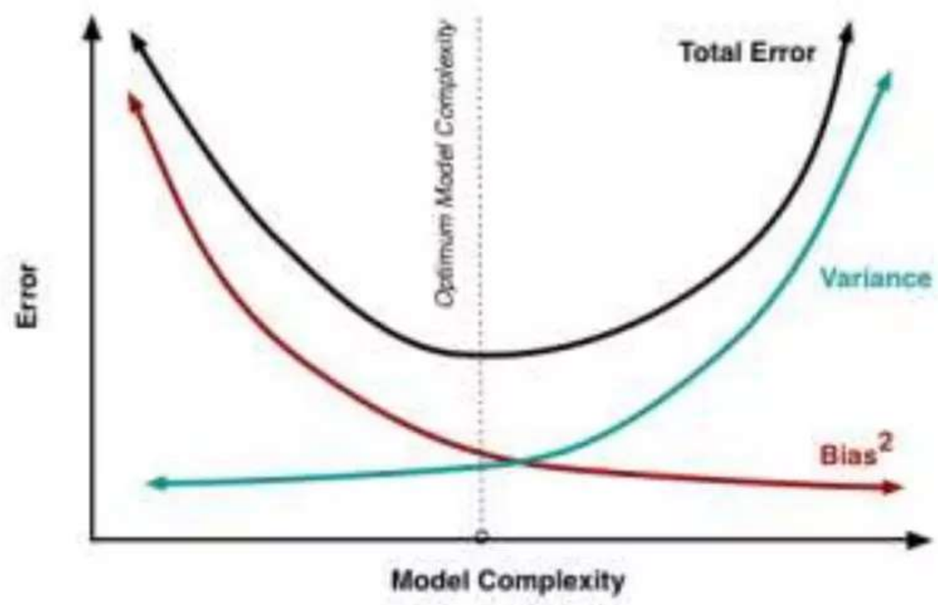
The types of selection bias include:

1. Sampling bias: It is a systematic error due to a non-random sample of a population causing some members of the population to be less likely to be included than others resulting in a biased sample.
2. Time interval: A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean.
3. Data: When specific subsets of data are chosen to support a conclusion or rejection of bad data on arbitrary grounds, instead of according to previously stated or generally agreed criteria.
4. Attrition: Attrition bias is a kind of selection bias caused by attrition (loss of participants) discounting trial subjects/tests that did not run to completion.

# 4. What is a bias-variance trade-off?

Bias: Bias is an error introduced in your model due to the oversimplification of the machine learning algorithm. It can lead to underfitting. When you train your model at that time model makes simplified assumptions to make the target function easier to understand.
Low bias machine learning algorithms — Decision Trees, k-NN, and SVM High bias machine learning algorithms — Linear Regression, Logistic Regression

Variance: Variance is the error introduced in your model due to a complex machine learning algorithm, your model learns noise also from the training data set and performs badly on the test data set. It can lead to high sensitivity and overfitting.
Normally, as you increase the complexity of your model, you will see a reduction in error due to lower bias in the model. However, this only happens until a particular point. As you continue to make your model more complex, you end up over-fitting your model and hence your model will start suffering from high variance.




Bias-Variance trade-off: The goal of any supervised machine learning algorithm is to have low bias and low variance to achieve good prediction performance.

1. The k-nearest neighbor algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute to the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

There is no escaping the relationship between bias and variance in machine learning. Increasing the bias will decrease the variance. Increasing the variance will decrease bias.

# 5. In your choice of language, write a program that prints the numbers ranging from one to 50

The python code to print numbers ranging from 1 to 50 is as follows-

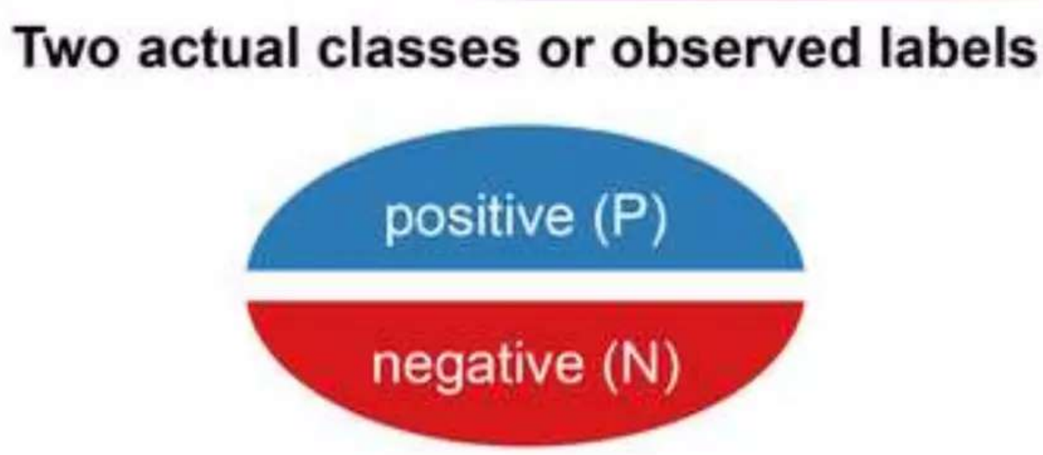
| | |
|---|---|
| 123 | `For i in` |
| 2 | range(1,51): |
| 3 | print(i) |

# 6. What is a confusion matrix?

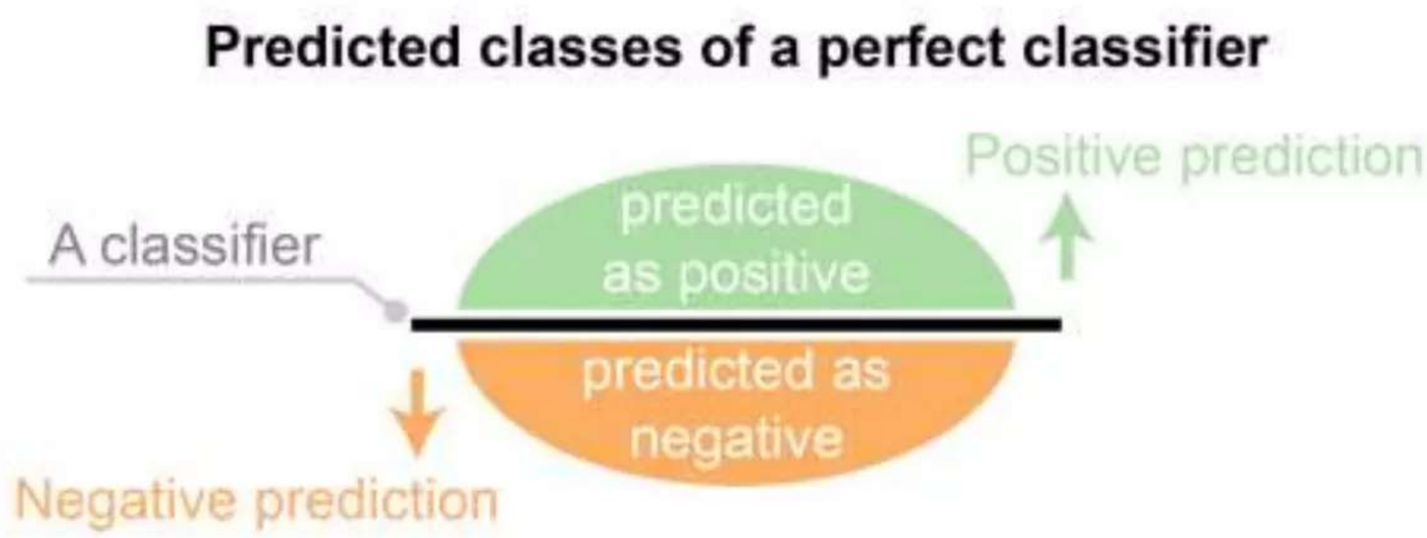
The confusion matrix is a 2X2 table that contains 4 outputs provided by the binary classifier. Various measures, such as error rate, accuracy, specificity, sensitivity, precision, and recall are derived from it. *Confusion Matrix*

| | | Predicted class | |
|---|---|---|---|
| | | $P$ | $N$ |
| Actual Class | $P$ | True Positives (TP) | False Negatives (FN) |
| | $N$ | False Positives (FP) | True Negatives (TN) |

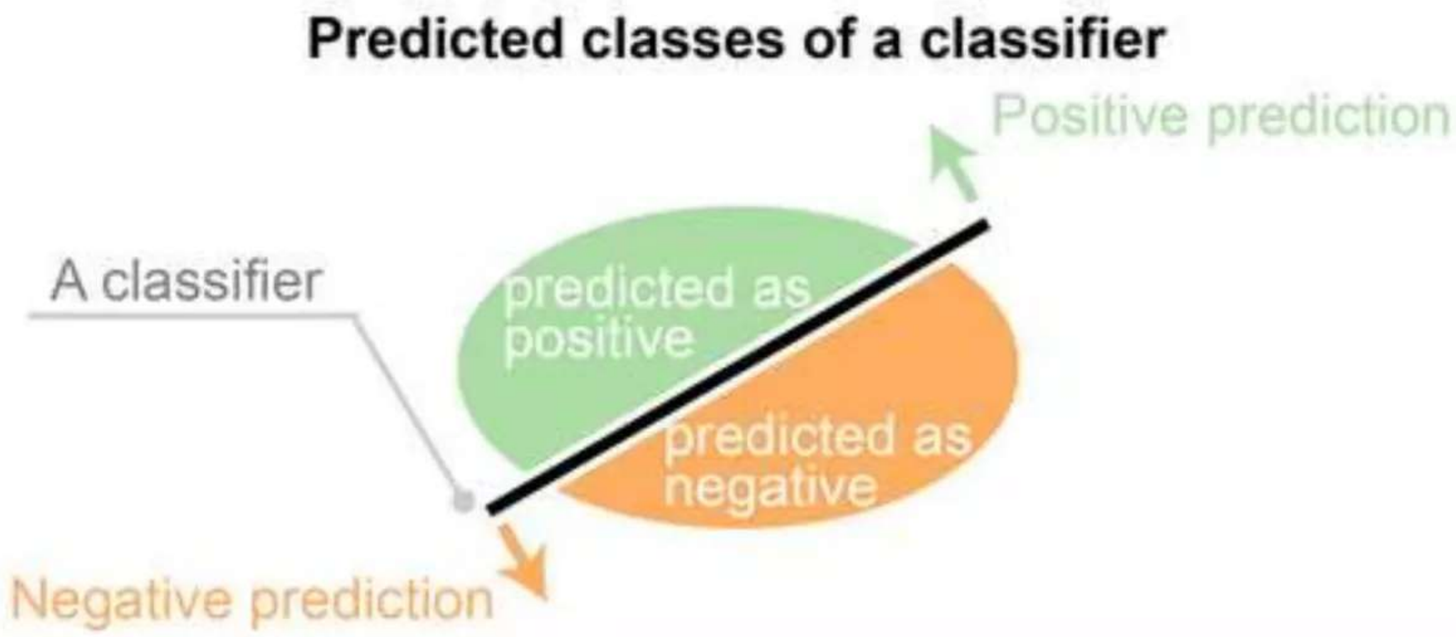
A data set used for performance evaluation is called a test data set. It should contain the correct labels and predicted labels.




The predicted labels will exactly the same if the performance of a binary classifier is perfect.

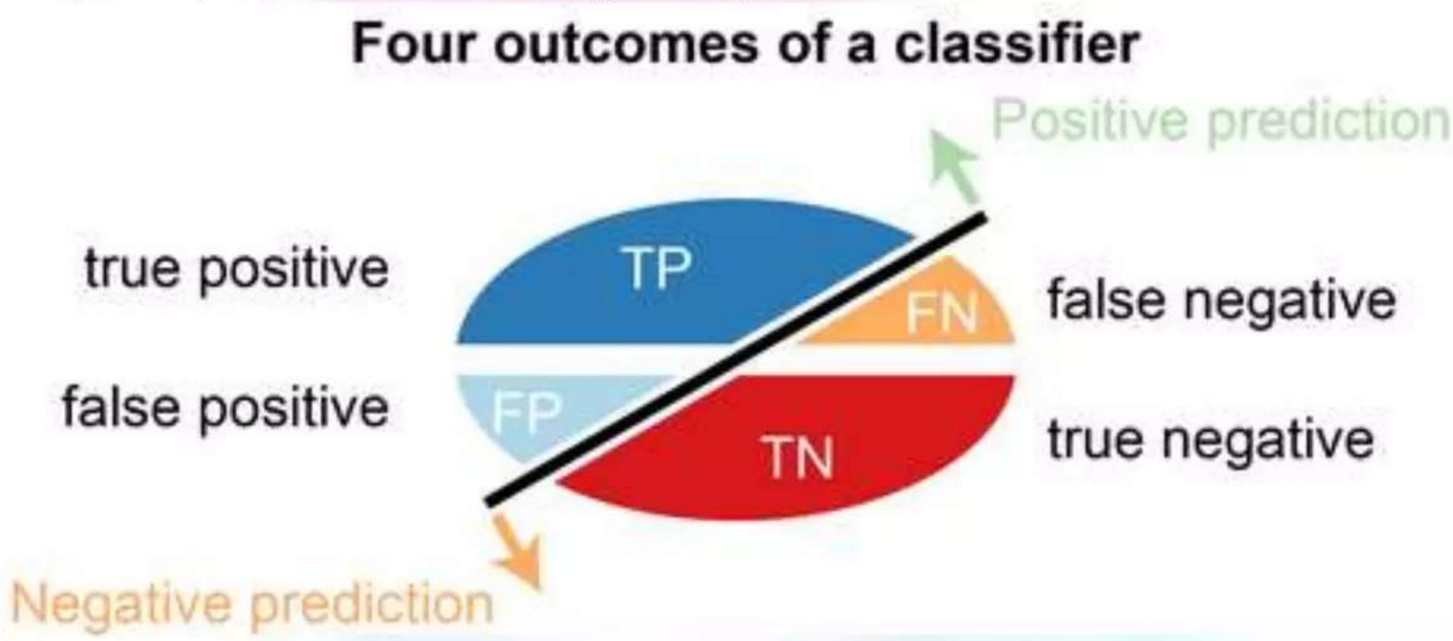



The predicted labels usually match with part of the observed labels in real-world scenarios.




A binary classifier predicts all data instances of a test data set as either positive or negative. This produces four outcomes-

1. True-positive(TP) — Correct positive prediction
2. False-positive(FP) — Incorrect positive prediction
3. True-negative(TN) — Correct negative prediction
4. False-negative(FN) — Incorrect negative prediction

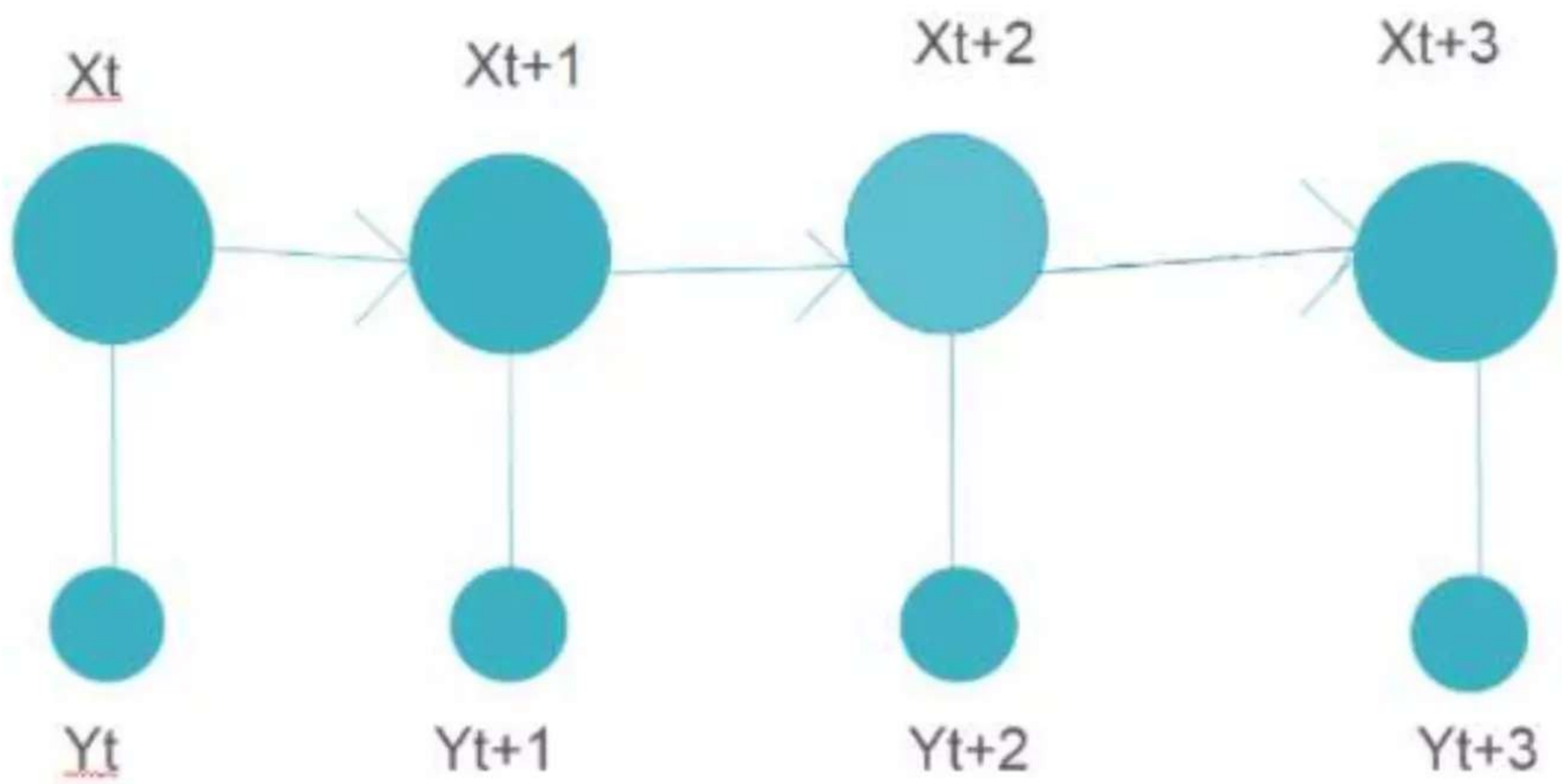



Basic measures derived from the confusion matrix-

1. Error Rate = (FP+FN)/(P+N)
2. Accuracy = (TP+TN)/(P+N)
3. Sensitivity(Recall or True positive rate) = TP/P
4. Specificity(True negative rate) = TN/N
5. Precision(Positive predicted value) = TP/(TP+FP)
6. F-Score(Harmonic mean of precision and recall) = (1+b)(PREC.REC)/($b^2$PREC+REC) where b is commonly 0.5, 1, 2.

# 7. Describe Markov chains.

Markov chain is a type of stochastic process. In Markov chains, the future probability of any state depends only on the current state.




The above figure represents a Markov chain model where each step has an output that depends on the current state only.

An example can be word recommendation. When we type a paragraph, the next word is suggested by the model which depends only on the previous word and not on anything before it. The Markov chain model is trained previously on a similar paragraph where the next word to a given word is stored for all the words in the training data. Based on this training data output, the next words are suggested.

## 8. What do you understand by true positive rate and false-positive rate?

True Positive rate(TPR) is the ratio of True Positives to True Positives and False Negatives. It is the probability that an actual positive will test as positive.

TPR=TP/TP+FN

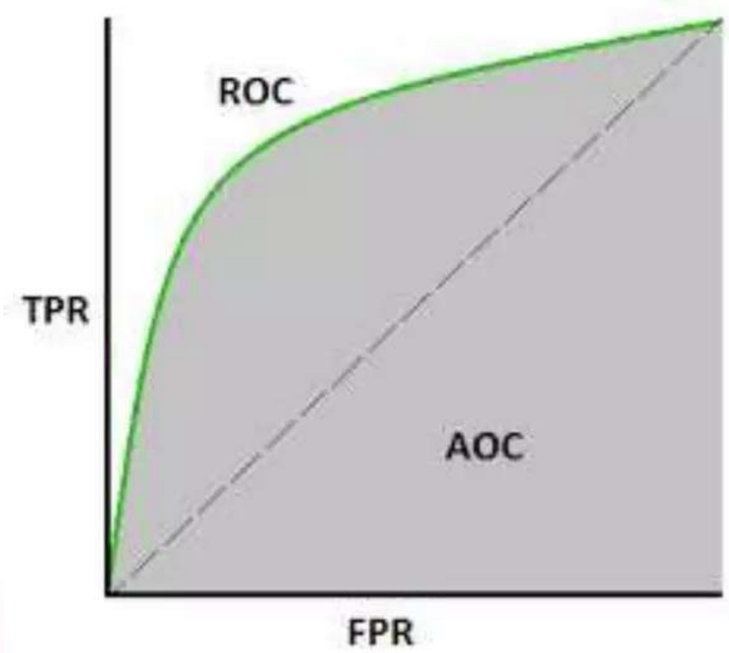
The False Positive Rate(FPR) is the ratio of the False Positives to all the positives(True positives and false positives). It is the probability of a false alarm, i.e., a positive result will be given when it is actually negative.

FPR=FP/TP+FP

## 9. What is the ROC curve?

The ROC curve is a graph between the False positive rate on the x-axis and the True positive rate on the y-axis. The true positive rate is the ratio of True positives to the total number of positive samples. The false positive rate is the ratio of False positives to the total number of negative samples.

The FPR and TPR are plotted on several threshold values to construct the ROC curve. The area under the ROC curve ranges from 0 to 1. A completely random model has a ROC of 0.5, which is represented by a straight line. The more the ROC curve deviates from this straight line, the better the model is. ROC curves are used for binary classification. The below image shows an example of a ROC curve.




# 10. What are dimensionality reduction and its benefits?

Reducing the number of features for a given dataset is known as dimensionality reduction. There are many techniques used to reduce dimensionality such as-

## Feature Selection Methods

## Matrix Factorization

## Manifold Learning

## Autoencoder Methods

### *Linear Discriminant Analysis (LDA)*

### *Principal component analysis (PCA)*

One of the main reasons for dimensionality reduction is the curse of dimensionality. When the number of features increases, the model becomes more complex. But if the number of data points is less, the model will start learning or overfitting the data. The model will not generalize the data. This is known as the curse of dimensionality.

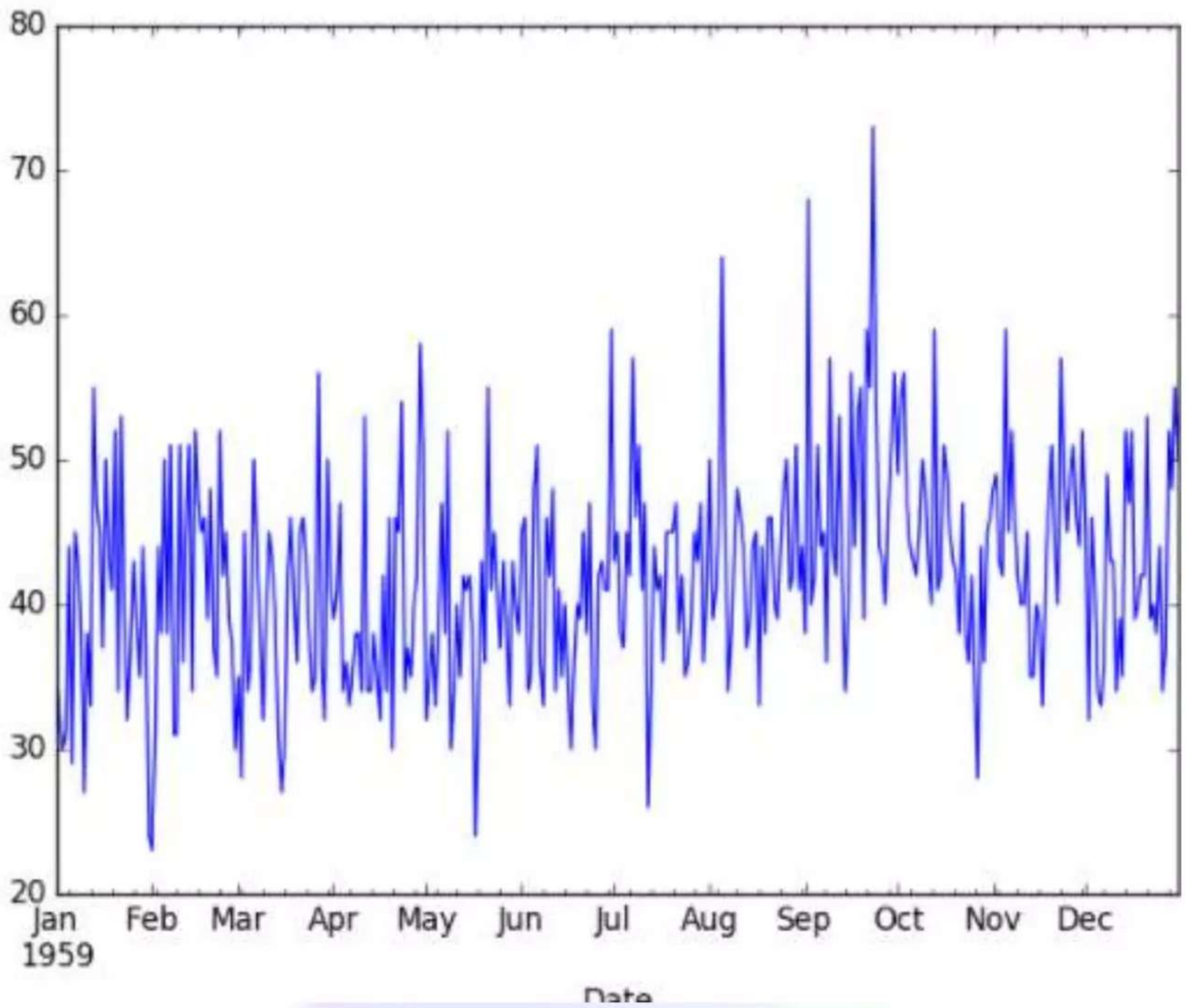
Other benefits of dimensionality reduction include-

- The time and storage space is reduced.

- It becomes easier to visualize and visually represent the data in 2D or 3D.
- Space complexity is reduced.

# 11. How do you find RMSE and MSE in a linear regression model?

Root Mean Squared Error(RMSE) is used to test the performance of the linear regression model. It evaluates how much the data is spread around the line of best fit. Its formula is-




Where

y_hat is the predicted value

y_i is the actual value of the output variable.

N is the number of data points

Mean Squared Error(MSE) tells how close the line is to the actual data. The difference of the line from the data points is taken and squared. The MSE value should be low for a good model. This means that the error between actual and predicted output values should be low. It is calculated as-

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where $N$ is the number of data points, $f_i$ the value returned by the model and $y_i$ the actual value for data point $i$.

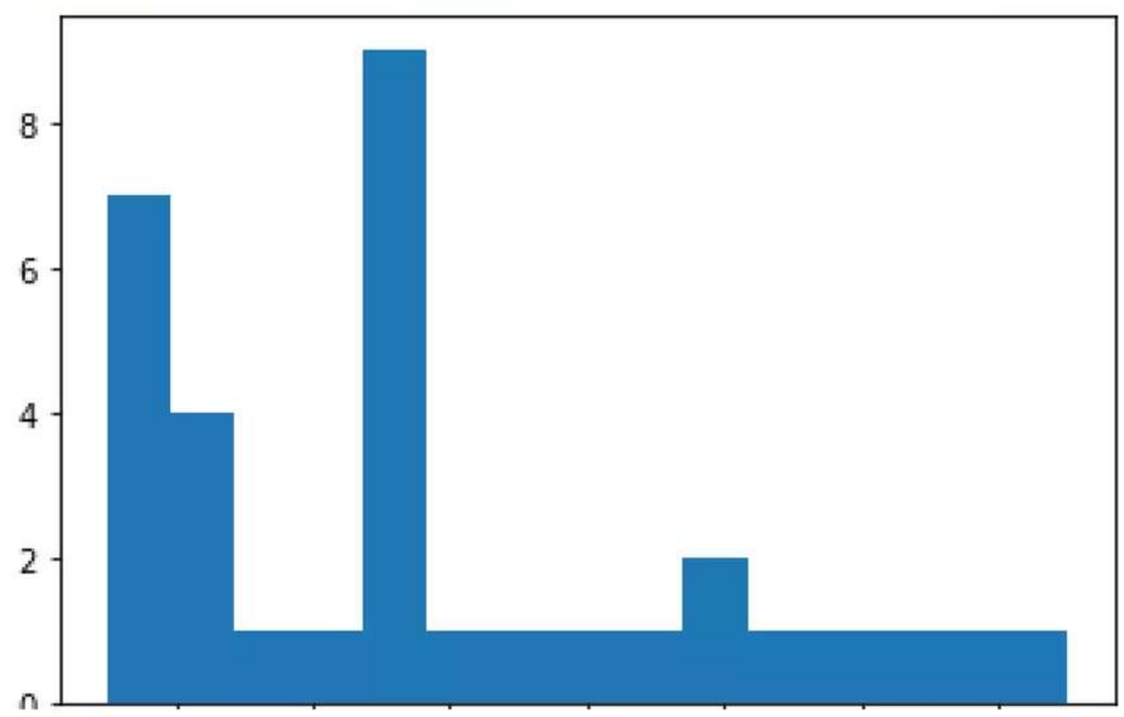
## 12. How to deal with unbalanced binary classification?

While doing binary classification, if the data set is imbalanced, the accuracy of the model can't be predicted correctly using only the R2 score. For example, if the data belonging to one of the two classes is very less in quantity as compared to the other class, the traditional accuracy will take a very small percentage of the smaller class. If only 5% of the examples are belonging to the smaller class, and the model classifies all outputs belonging to the other class, the accuracy would still be around 95%. But this will be wrong. To deal with this, we can do the following-

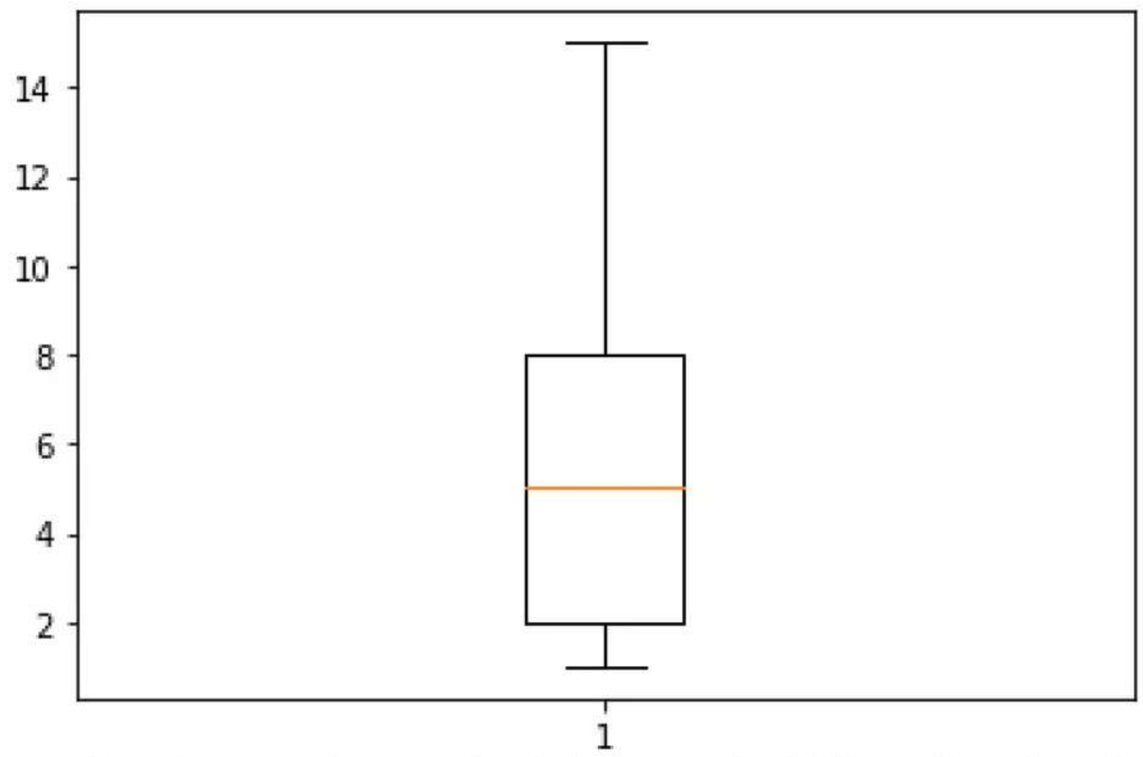
1. Use other methods for calculating the model performance like precision/recall, F1 score, etc.
2. Resample the data with techniques like undersampling(reducing the sample size of the larger class), and oversampling(increasing the sample size of the smaller class using repetition, SMOTE, and other such techniques.
3. Using K-fold cross-validation
4. Using ensemble learning such that each decision tree considers the entire sample of the smaller class and only a subset of the larger class.

## 13. What is the difference between a box plot and a histogram?

Both histograms and box plots are used to visually represent the frequency of values of a certain feature. The figure below shows a histogram.




While the figure below shows a boxplot for the same data.




The histogram is used to know the underlying probability distribution of data. While boxplots are used more to compare several datasets. Boxplots have fewer details and take up less space than histograms.

# 14. What does NLP stand for?

NLP stands for Natural language processing. It is the study of programming computers to learn large amounts of textual data. Examples of NLP include tokenization, stop words removal, stemming, sentiment analysis, etc.

# 15. Walkthrough of the probability fundamentals

The possibility of the occurrence of an event, among all the possible outcomes, is known as its probability. The probability of an event always lies between(including) 0 and 1.

$$P(A) = \frac{(Number\ of\ times\ event\ A\ occurred)}{(Number\ of\ times\ experiment\ performed)}$$

Factorial -it is used to find the total number of ways n number of things can be arranged in n places without repetition. Its value is n multiplied by all natural numbers till n-1.eg. 5!=5X4X3X2X1=120

Permutation– It is used when replacement is not allowed, and the order of items is important. Its formula is-

$${}_nP_r = \frac{n!}{(n-r)!}$$

Where,

n is the total number of items

R is the number of ways items are being selected

Combination– It is used when replacement is not allowed, and the order of items is not important. Its formula is-

$${}_nC_r = \frac{n!}{(n-r)!r!}$$

Some rules for probability are-

Addition Rule
P(A or B)= P(A) + P(B) – P(A and B)

Conditional probability
It is the probability of event B occurring, assuming that event A has already occurred.

P(A and B)= P(A). P(B|A)

Central Limit theorem
It states that when we draw random samples from a large population and take the mean of these samples, they form a normal distribution.

## 16. Describe different regularization methods, such as L1 and L2 regularization

There are 3 important regularization methods as follows-

L2 regularization-(Ridge regression)– In L2 regularization, we add the sum of the squares of all the weights, multiplied by a value lambda, to the loss function. The formula for Ridge regression is as follows-

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} Wi^2$$

Top 100+ Google Data Science Interview Questions : All You Need to know to Crack it 60

As you can see, if the value of the weights is multiplied by the data value for a particular data point and the feature becomes very large, the original loss will become small. But the added value of lambda multiplied with the sum of squares of weights will become large as well. Similarly, if the original loss value becomes very large, the added value will become small. Thus it will control the final value from becoming too large or too small.

L1 Regularization-(Lasso regression)– In L1 regularization, we add the sum of the absolute values of all the weights, multiplied by a value lambda, to the loss function. The formula for Lasso regression is as follows-

$$\text{Loss} = \sum_{j=1}^{m}\left(Yi - Wo - \sum_{i=1}^{n} Wi\,Xji\right)^2 + \lambda \sum_{i=1}^{n} |Wi|$$

Top 100+ Google Data Science Interview Questions : All You Need to know to Crack it 61

The loss function along with the optimization algorithm brings parameters near to zero but not actually zero, while lasso eliminates less important features and sets respective weight values to zero.

### *Dropout*

This is used for regularization in neural networks. Fully connected layers are more prone to overfitting. Dropout leaves out some neurons with a 1-p probability in neural networks. Dropout reduces overfitting, improves training speed, and makes the model more robust.

## 17. How should you maintain a deployed model?

After a model has been deployed, it needs to be maintained. The data being fed may change over time. For example, in the case of a model predicting house prices, the prices of houses may rise over time or fluctuate due to some other factor. The accuracy of the model on new data can be recorded. Some common ways to ensure accuracy include-

1. The model should be frequently checked by feeding negative test data. If the model gives low accuracy with negative test data, it is fine.
2. An Auto Encoder should be built that Uses anomaly detection techniques, the AE model will calculate the Reconstruction error value. If the Reconstruction error value is high, it means the new data does not follow the old pattern learned by the model.

If the model shows good prediction accuracy with new data, it means that the new data follows the pattern or the generalization learned by the model on old data. So, the model can be retrained on the new data. If the accuracy of new data is not that good, the model can be retrained on the new data with feature engineering on the data features along with the old data.

If the accuracy is not good, the model may need to be trained from scratch.

## 18. Write the equation and calculate the precision and recall rate?

Precision quantifies the number of correct positive predictions made. Precision is calculated as the number of true positives divided by the total number of true positives and false positives.

Precision = True Positives / (True Positives + False Positives)
Precision is defined as the number of correct positive predictions made out of all positive predictions that could have been made. The recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

Recall = True Positives / (True Positives + False Negatives)

## 19. Why do we use the summary function?

Summary functions are used to give the summary of all the numeric values in a data frame. Eg. The describe() function can be used to provide the summary of all the data values given to it.

column_name.describe() will give the following values of all the numeric data in the column-

1. Count
2. Mean
3. Std-Standard deviation
4. Min-Minimum
5. 25%
6. 50%
7. 75%
8. max-Maximum

## 20. How will you measure the Euclidean distance between the two arrays in NumPy?

The Euclidean distance between 2 arrays A[1,2,3,] and [8,9,10] can be calculated by taking the Euclidean distance of each point respectively. The built-in function NumPy.linalg.norm()can be used as follows-

```
import numpy as np
```

```
[10] A=np.array([1,2,3])
     B=np.array([8,9,10])
```

```
[11] numpy.linalg.norm(A-B)
```

```
12.12435565298214
```



## 21. What is the difference between an error and a residual error?

An error refers to the difference between the predicted value and the actual value. The most popular means for calculating errors in data science are Mean Absolute Error(MAE), Mean Squared Error(MSE), and Root Mean Squared Error(RMSE). While residual is the difference between a group of values observed and their arithmetical mean. An error is generally unobservable while a residual error can be visualized on a graph. Error represents how observed data differs from the actual population. While a residual represents the way observed data differs from the sample population data.

## 22. Difference between Normalisation and Standardization?

Normalization, also known as min-max scaling, is a technique where all the data values are converted such that they lie between 0 and 1.

The formula for Normalization is-

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$



Where,

X_max is the maximum value of the feature

X_min is the minimum value of the feature

Standardization refers to converting our data such that the data has a normal distribution with its mean as 0 and standard deviation as 1.

The formula for Standardization is-

$$X' = \frac{X - \mu}{\sigma}$$



So, while normalization rescales the data into the range from 0 to 1 only, standardization ensures data follows the standard normal distribution.

# STATISTICS INTERVIEW QUESTIONS

## 23. What is the difference between "long" and "wide" format data?

In the wide format, a subject's repeated responses will be in a single row, and each response is in a separate column. In the long format, each row is a one-time point per subject. You can recognize data in wide format by the fact that columns generally represent groups.

| Name | Height | Weight |
|---|---|---|
| John | 160 | 67 |
| Christopher | 182 | 78 |

***Figure:*** *Wide Format*

| Name | Attribute | Value |
|---|---|---|
| John | Height | 160 |
| John | Weight | 67 |
| Christopher | Height | 182 |
| Christopher | Weight | 78 |

***Figure:*** *Long Format*

# 24. What do you understand by the term Normal Distribution?

Data is usually distributed in different ways with a bias to the left or to the right or it can all be jumbled up.

However, there are chances that data is distributed around a central value without any bias to the left or right and reaches normal distribution in the form of a bell-shaped curve.

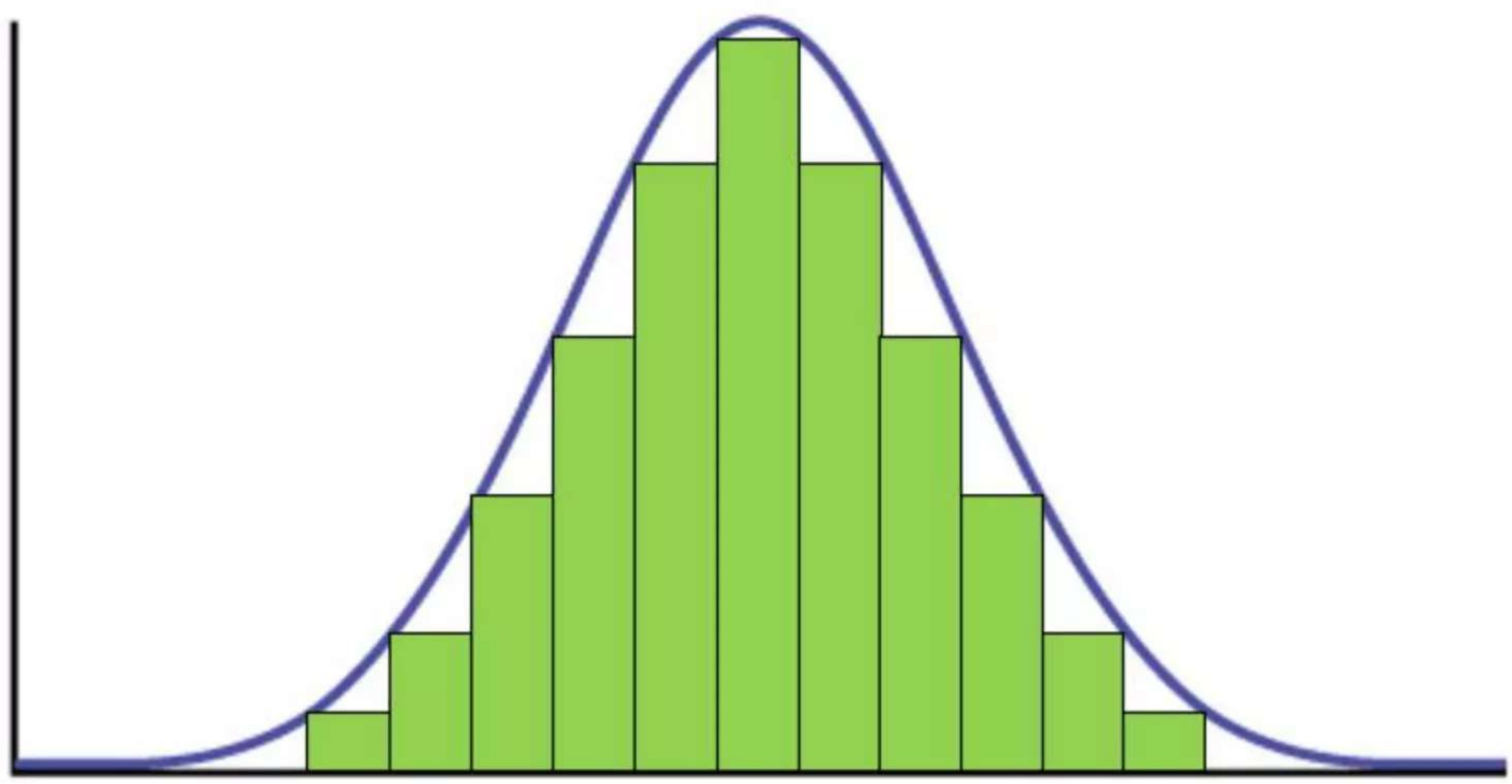


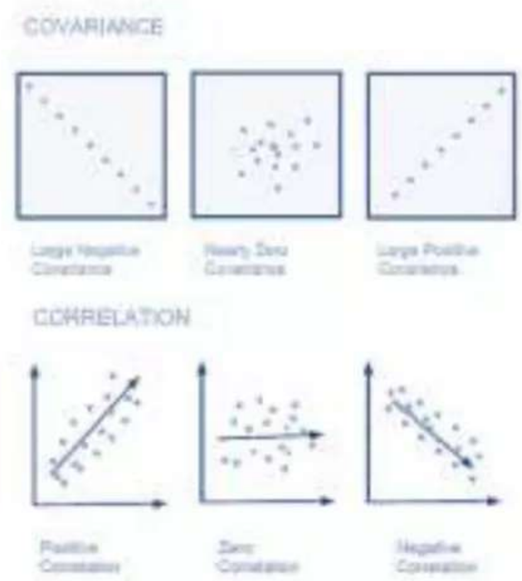
Figure:*Normal distribution in a bell curve*
The random variables are distributed in the form of a symmetrical, bell-shaped curve.

The properties of Normal Distribution are as follows;

1. Unimodal -one mode
2. Symmetrical -left and right halves are mirror images
3. Bell-shaped -maximum height (mode) at the mean
4. Mean, Mode, and Median are all located in the center
5. Asymptotic

# 25. What are correlation and covariance in statistics?

Covariance and Correlation are two mathematical concepts; these two approaches are widely used in statistics. Both Correlation and Covariance establish the relationship and also measure the dependency between two random variables. Though the work is similar between these two in mathematical terms, they are different from each other.




Correlation: Correlation is considered or described as the best technique for measuring and also for estimating the quantitative relationship between two variables. Correlation measures how strongly two variables are related.
Covariance: In covariance two items vary together and it's a measure that indicates the extent to which two random variables change in a cycle. It is a statistical term; it explains the systematic relation between a pair of random variables, wherein changes in one variable are reciprocal by a corresponding change in another variable.

# 26. What is the difference between Point Estimates and Confidence Intervals?

Point Estimation gives us a particular value as an estimate of a population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

A confidence interval gives us a range of values that is likely to contain the population parameter. The confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called the Confidence Level or Confidence coefficient and is represented by 1 — alpha, where alpha is the level of significance.

# 27. What is the goal of A/B Testing?

It is a hypothesis testing for a randomized experiment with two variables A and B.

The goal of A/B Testing is to identify any changes to the web page to maximize or increase the outcome of interest. A/B testing is a fantastic method for figuring out the best online promotional and marketing strategies for your business. It can be used to test everything from website copy to sales emails to search ads

An example of this could be identifying the click-through rate for a banner ad.

## 28. What is the p-value?

When you perform a hypothesis test in statistics, a p-value can help you determine the strength of your results. a p-value is a number between 0 and 1. Based on the value will denote the strength of the results. The claim which is on trial is called the Null Hypothesis.

A low p-value (≤ 0.05) indicates strength against the null hypothesis which means we can reject the null Hypothesis. A high p-value (≥ 0.05) indicates strength for the null hypothesis which means we can accept the null Hypothesis p-value of 0.05 indicates the Hypothesis could go either way. To put it in another way,

High P values: your data are likely with a true null. Low P values: your data are unlikely with a true null.

## 29. In any 15-minute interval, there is a 20% probability that you will see at least one shooting star. What is the probability that you see at least one shooting star in the period of an hour?

The probability of not seeing any shooting star in 15 minutes is

= 1 – P( Seeing one shooting star )

= 1 – 0.2 = 0.8

Probability of not seeing any shooting star in the period of one hour

= (0.8) ^ 4 = 0.4096

Probability of seeing at least one shooting star in the one hour

= 1 – P( Not seeing any star )

= 1 – 0.4096 = 0.5904

## 30. How can you generate a random number between 1 – 7 with only a die?

- Any die has six sides from 1-6. There is no way to get seven equal outcomes from a single rolling of a die. If we roll the die twice and consider the event of two rolls, we now have 36 different outcomes.

- To get our 7 equal outcomes we have to reduce this 36 to a number divisible by 7. We can thus consider only 35 outcomes and exclude the other one.
- A simple scenario can be to exclude the combination (6,6), i.e., to roll the die again if 6 appears twice.
- All the remaining combinations from (1,1) till (6,5) can be divided into 7 parts of 5 each. This way all seven sets of outcomes are equally likely.

## 31. A certain couple tells you that they have two children, at least one of which is a girl. What is the probability that they have two girls?

In the case of two children, there are 4 equally likely possibilities

BB, BG, GB, and GG;
where B = Boy and G = Girl and the first letter denotes the first child.
From the question, we can exclude the first case of BB. Thus from the remaining 3 possibilities of BG, GB & BB, we have to find the probability of the case with two girls.
Thus, P(Having two girls given one girl) = 1 / 3

## 32. A jar has 1000 coins, of which 999 are fair and 1 is double-headed. Pick a coin at random, and toss it 10 times. Given that you see 10 heads, what is the probability that the next toss of that coin is also a head?

There are two ways of choosing the coin. One is to pick a fair coin and the other is to pick the one with two heads.

Probability of selecting fair coin = 999/1000 = 0.999
Probability of selecting unfair coin = 1/1000 = 0.001
Selecting 10 heads in a row = Selecting a fair coin * Getting 10 heads + Selecting an unfair coin

P (A) = 0.999 * (1/2)^5 = 0.999 * (1/1024) = 0.000976
P (B) = 0.001 * 1 = 0.001
P( A / A + B ) = 0.000976 / (0.000976 + 0.001) = 0.4939
P( B / A + B ) = 0.001 / 0.001976 = 0.5061
Probability of selecting another head = P(A/A+B) 0.5 + P(B/A+B) 1 = 0.4939 * 0.5 + 0.5061 = 0.7531

## 33. What do you understand by the statistical power of sensitivity and how do you calculate it?

Sensitivity is commonly used to validate the accuracy of a classifier (Logistic, SVM, Random Forest, etc.).

Sensitivity is nothing but “Predicted True events/ Total events”. True events here are the events that were true and the model also predicted them as true.

The calculation of seasonality is pretty straightforward.

Seasonality = ( True Positives ) / ( Positives in Actual Dependent Variable )

# 34. Why Is Re-sampling Done?

Resampling is done in any of these cases:

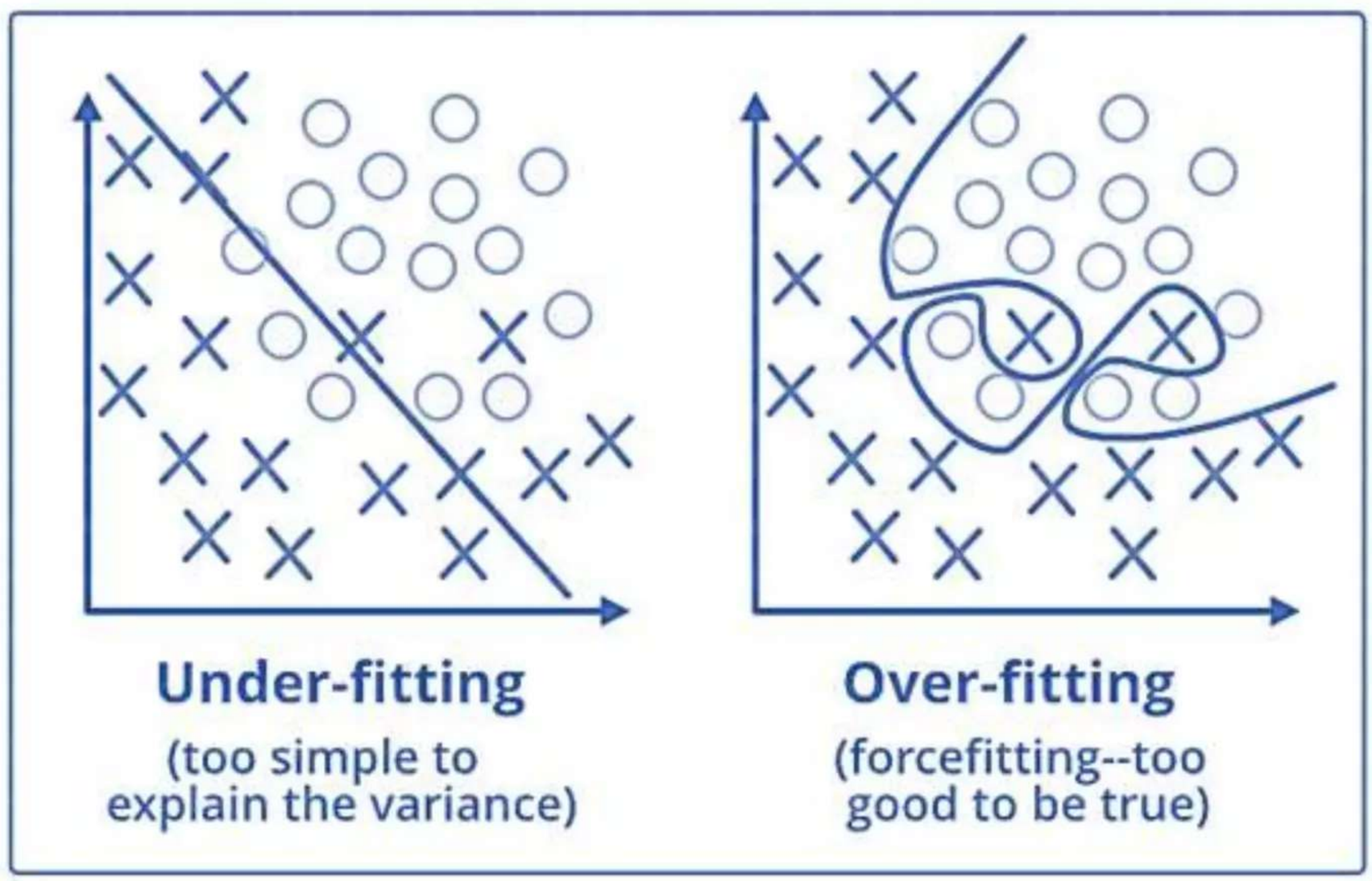
- Estimating the accuracy of sample statistics by using subsets of accessible data or drawing randomly with replacement from a set of data points
- Substituting labels on data points when performing significance tests
- Validating models by using random subsets (bootstrapping, cross-validation)

# 35. How to combat Overfitting and Underfitting?

To combat overfitting and underfitting, you can resample the data to estimate the model accuracy (k-fold cross-validation) and by having a validation dataset to evaluate the model.

# 36. What are the differences between over-fitting and under-fitting?

In statistics and machine learning, one of the most common tasks is to fit a *model* to a set of training data, so as to be able to make reliable predictions on general untrained data.




In overfitting, a statistical model describes random error or noise instead of the underlying relationship. Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. A model that has been overfitted, has poor predictive performance, as it overreacts to minor fluctuations in the training data.

Underfitting occurs when a statistical model or machine learning algorithm cannot capture the underlying trend of the data. Underfitting would occur, for example, when fitting a linear model to non-linear data. Such a model too would have poor predictive performance.

## 37. What is regularisation? Why is it useful?

Regularisation is the process of adding tuning parameters to a model to induce smoothness in order to prevent overfitting. This is most often done by adding a constant multiple to an existing weight vector. This constant is often the L1(Lasso) or L2(ridge). The model predictions should then minimize the loss function calculated on the regularized training set.

## 38. What Is the Law of Large Numbers?

It is a theorem that describes the result of performing the same experiment a large number of times. This theorem forms the basis of frequency-style thinking. It says that the sample means, the sample variance, and the sample standard deviation converge to what they are trying to estimate.

## 39. What Are Confounding Variables?

In statistics, a confounder is a variable that influences both the dependent variable and the independent variable.

For example, if you are researching whether a lack of exercise leads to weight gain,

lack of exercise = independent variable
weight gain = dependent variable.
A confounding variable here would be any other variable that affects both of these variables, such as the age of the subject.

## 40. What Are the Types of Biases That Can Occur During Sampling?

- Selection bias
- Under coverage bias
- Survivorship bias

## 41. What is Survivorship Bias?

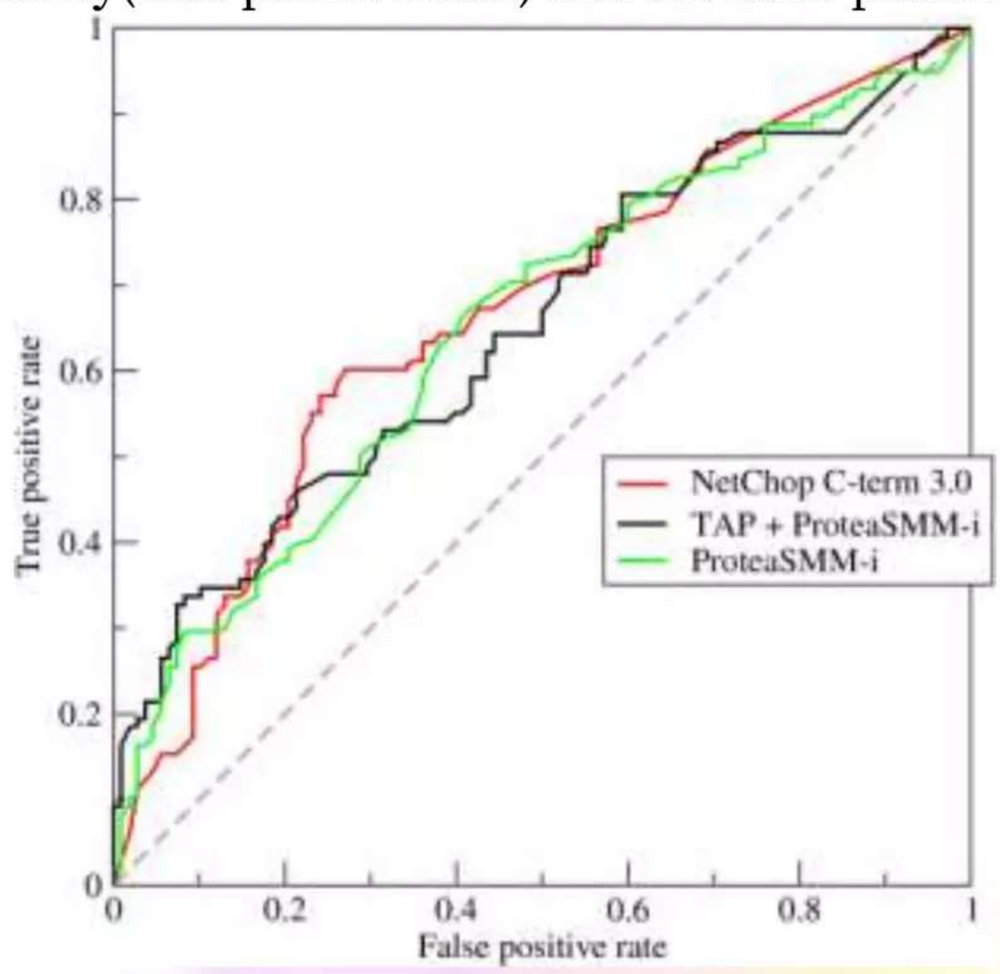
It is the logical error of focusing on aspects that support surviving some process and casually overlooking those that did not work because of their lack of prominence. This can lead to wrong conclusions in numerous different means.

## 42. What is selection Bias?

Selection bias occurs when the sample obtained is not representative of the population intended to be analyzed.

## 43. Explain how a ROC curve works.?

The ROC curve is a graphical representation of the contrast between true-positive rates and false-positive rates at various thresholds. It is often used as a proxy for the trade-off between the sensitivity(true positive rate) and the false-positive rate.




## 44. What is TF/IDF vectorization?

TF–IDF is short for term frequency-inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining.

The TF–IDF value increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

## 45. Why do we generally use Softmax non-linearity function as test operation in-network?

It is because it takes in a vector of real numbers and returns a probability distribution. Its definition is as follows. Let x be a vector of real numbers (positive, negative, whatever, there are no constraints).

Then the i'th component of Softmax(*x*) *is* —

$$P(y=j \mid \Theta^{(i)}) = \frac{e^{\Theta^{(i)}}}{\sum_{j=0}^{k} e^{\Theta_k^{(i)}}}$$

Softmax function

where $\Theta = w_0x_0 + w_1x_1 + \ldots + w_kx_k = \sum_{i=0}^{k} w_ix_i = w^Tx$

Top 100+ Google Data Science Interview Questions : All You Need to know to Crack it 70

It should be clear that the output is a probability distribution: each element is non-negative and the sum over all components is 1.

# DATA ANALYSIS INTERVIEW QUESTIONS

## 46. Python or R – Which one would you prefer for text analytics?

We will prefer Python because of the following reasons:

- Python would be the best option because it has the Pandas library that provides easy-to-use data structures and high-performance data analysis tools.
- R is more suitable for machine learning than just text analysis.
- Python performs faster for all types of text analytics.

## 47. How does data cleaning plays a vital role in the analysis?

Data cleaning can help in analysis because:

- Cleaning data from multiple sources helps to transform it into a format that data analysts or data scientists can work with.
- Data Cleaning helps to increase the accuracy of the model in machine learning.
- It is a cumbersome process because as the number of data sources increases, the time taken to clean the data increases exponentially due to the number of sources and the volume of data generated by these sources.
- It might take up to 80% of the time just clean data making it a critical part of the analysis task.

## 48. Differentiate between univariate, bivariate, and multivariate analysis.

*Univariate analyses* are descriptive statistical analysis techniques that can be differentiated based on the number of variables involved at a given point in time. For example, the pie charts of sales based on territory involve only one variable, and can the analysis can be referred to as univariate analysis.

*The bivariate analysis* attempts to understand the difference between two variables at a time as in a scatterplot. For example, analyzing the volume of sales and spending can be considered as an example of bivariate analysis.

*The multivariate analysis* deals with the study of more than two variables to understand the effect of variables on the responses.

## 49. What is Cluster Sampling?

Cluster sampling is a technique used when it becomes difficult to study the target population spread across a wide area and simple random sampling cannot be applied. A cluster Sample is a probability sample where each sampling unit is a collection or cluster of elements.

For eg., A researcher wants to survey the academic performance of high school students in Japan. He can divide the entire population of Japan into different clusters (cities). Then the researcher selects a number of clusters depending on his research through simple or systematic random sampling.

Let's continue our Data Science Interview Questions blog with some more statistics questions

## 50. Explain Star Schema?

It is a traditional database schema with a central table. Satellite tables map IDs to physical names or descriptions and can be connected to the central fact table using the ID fields; these tables are known as lookup tables and are principally useful in real-time applications, as they save a lot of memory. Sometimes star schemas involve several layers of summarization to recover information faster.

## 51. What is Systematic Sampling?

*Systematic sampling* is a statistical technique where elements are selected from an ordered sampling frame. In systematic sampling, the list is progressed in a circular manner so once you reach the end of the list, it is progressed from the top again. The best example of systematic sampling is the equal probability method.

## 52. What are Eigenvectors and Eigenvalues?

*Eigenvectors* are used for understanding linear transformations. In data analysis, we usually calculate the eigenvectors for a correlation or covariance matrix. Eigenvectors are the directions along which a particular linear transformation acts by flipping, compressing, or stretching.

*Eigenvalue* can be referred to as the strength of the transformation in the direction of the eigenvector or the factor by which the compression occurs.

## 53. Can you cite some examples where a false positive is more important than a false negative?

Let us first understand what false positives and false negatives are.

- False Positives are the cases where you wrongly classified a non-event as an event a.k.a Type I error.
- False Negatives are the cases where you wrongly classify events as non-events, a.k.a Type II error.

*Example 1:* In the medical field, assume you have to give chemotherapy to patients. Assume a patient comes to that hospital and he is tested positive for cancer, based on the lab prediction but he actually doesn't have cancer. This is a case of false positives. Here it is of the utmost danger to start chemotherapy on this patient when he actually does not have cancer. In the absence of cancerous cells, chemotherapy will do certain damage to normal healthy cells and might lead to severe diseases, even cancer.

*Example 2:* Let's say an e-commerce company decided to give a $1000 Gift voucher to the customers whom they assume to purchase at least $10,000 worth of items. They send free voucher mail directly to 100 customers without any minimum purchase condition because they assume to make at least 20% profit on sold items above $10,000. Now the issue is if we send the $1000 gift vouchers to customers who have not actually purchased anything but are marked as having made $10,000 worth of purchases.

## 54. Can you cite some examples where a false negative important than a false positive?

*Example 1*: Assume there is an airport 'A' which has received high-security threats and based on certain characteristics they identify whether a particular passenger can be a threat or not. Due to a shortage of staff, they decide to scan passengers being predicted as risk positives by their predictive model. What will happen if a true threat customer is flagged as non-threat by the airport model?

*Example 2*: What if the Jury or judge decides to make a criminal go free?

*Example 3*: What if you rejected marrying a very good person based on your predictive model and you happen to meet him/her after a few years and realize that you had a false negative?

## 55. Can you cite some examples where both false positives and false negatives are equally important?

In the Banking industry giving loans is the primary source of making money but at the same time if your repayment rate is not good you will not make any profit, rather you will risk huge losses.

Banks don't want to lose good customers and at the same point in time, they don't want to acquire bad customers. In this scenario, both the false positives and false negatives become very important to measure.

## 56. Can you explain the difference between a Validation Set and a Test Set?

A *Validationset* can be considered as a part of the training set as it is used for parameter selection and to avoid overfitting the model being built.
On the other hand, a *Test Set* is used for testing or evaluating the performance of a trained machine-learning model.
In simple terms, the differences can be summarized as; the training set is to fit the parameters i.e. weights, and the test set is to assess the performance of the model i.e. evaluating the predictive power and generalization.

## 57. Explain cross-validation.

Cross-validation is a model validation technique for evaluating how the outcomes of statistical analysis will generalize to an independent dataset. Mainly used in backgrounds where the objective is forecast and one wants to estimate how accurately a model will accomplish in practice.
The goal of cross-validation is to term a data set to test the model in the training phase (i.e. validation data set) in order to limit problems like overfitting and get an insight into how the model will generalize to an independent data set.

# MACHINE LEARNING INTERVIEW QUESTIONS

## 58. What is Machine Learning?

Machine Learning explores the study and construction of algorithms that can learn from and make predictions on data. Closely related to computational statistics. Used to devise complex models and algorithms that lend themselves to a prediction which in commercial use is known as predictive analytics. Given below, is an image representing the various domains Machine Learning lends itself to.

## 59. What is Supervised Learning?

Supervised learning is the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples.

Algorithms: Support Vector Machines, Regression, Naive Bayes, Decision Trees, K-nearest Neighbor Algorithm, and Neural Networks

*E.g. If you built a fruit classifier, the labels will be "this is an orange, this is an apple and this is a banana", based on showing the classifier examples of apples, oranges, and bananas.*

## 60. What is Unsupervised learning?

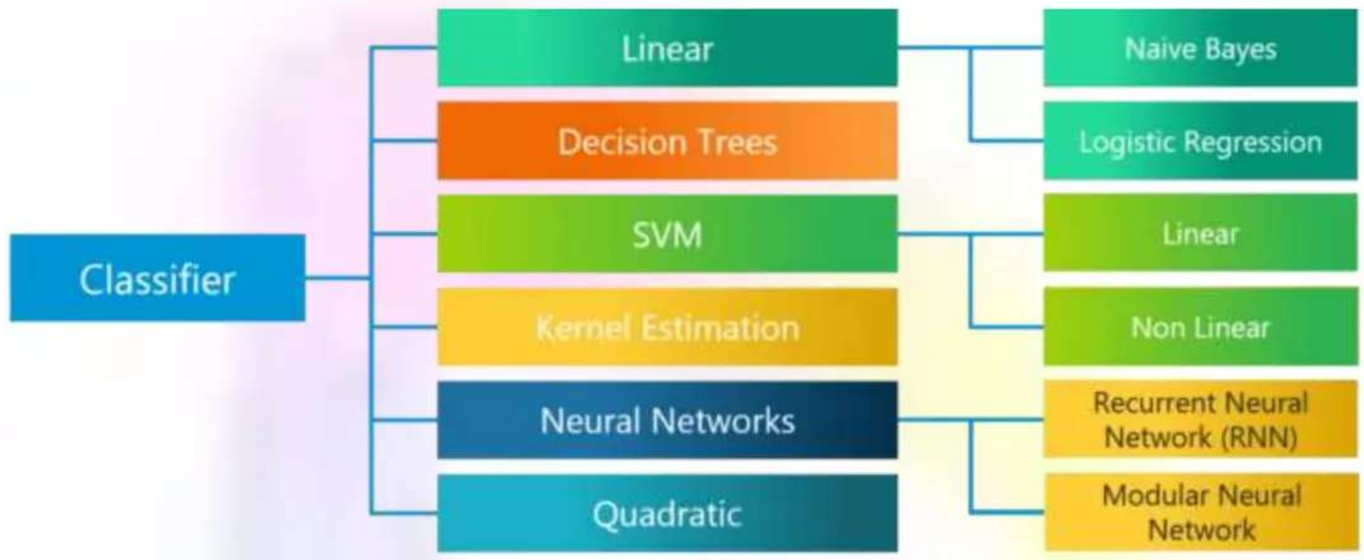
Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

Algorithms: Clustering, Anomaly Detection, Neural Networks, and Latent Variable Models

*E.g. In the same example, a fruit clustering will categorize as "fruits with soft skin and lots of dimples", "fruits with shiny hard skin" and "elongated yellow fruits".*

## 61. What are the various classification algorithms?

The diagram lists the most important classification algorithms



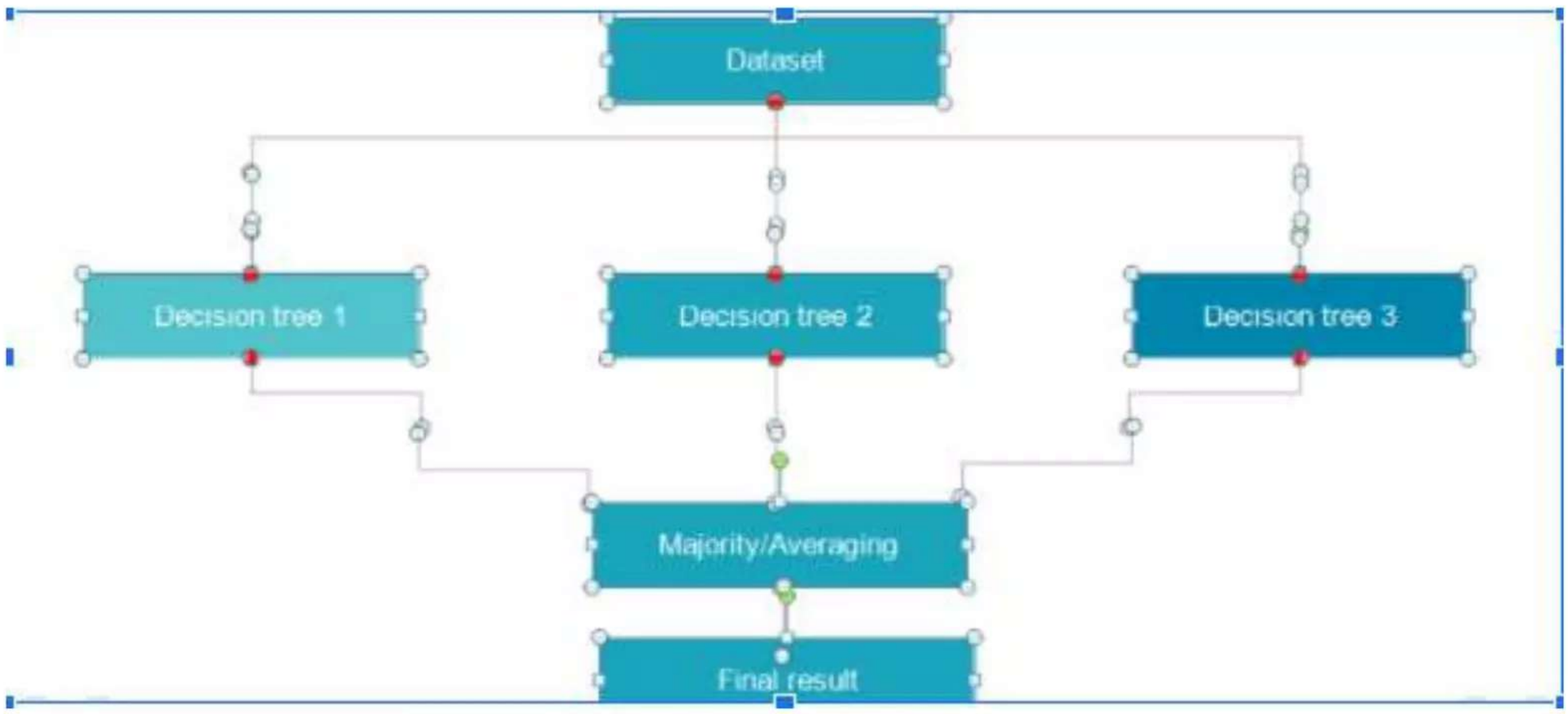

## 62. What is 'Naive' in a Naive Bayes?

The Naive Bayes Algorithm is based on the Bayes Theorem. Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event.

The Algorithm is 'naive' because it makes assumptions that may or may not turn out to be correct.

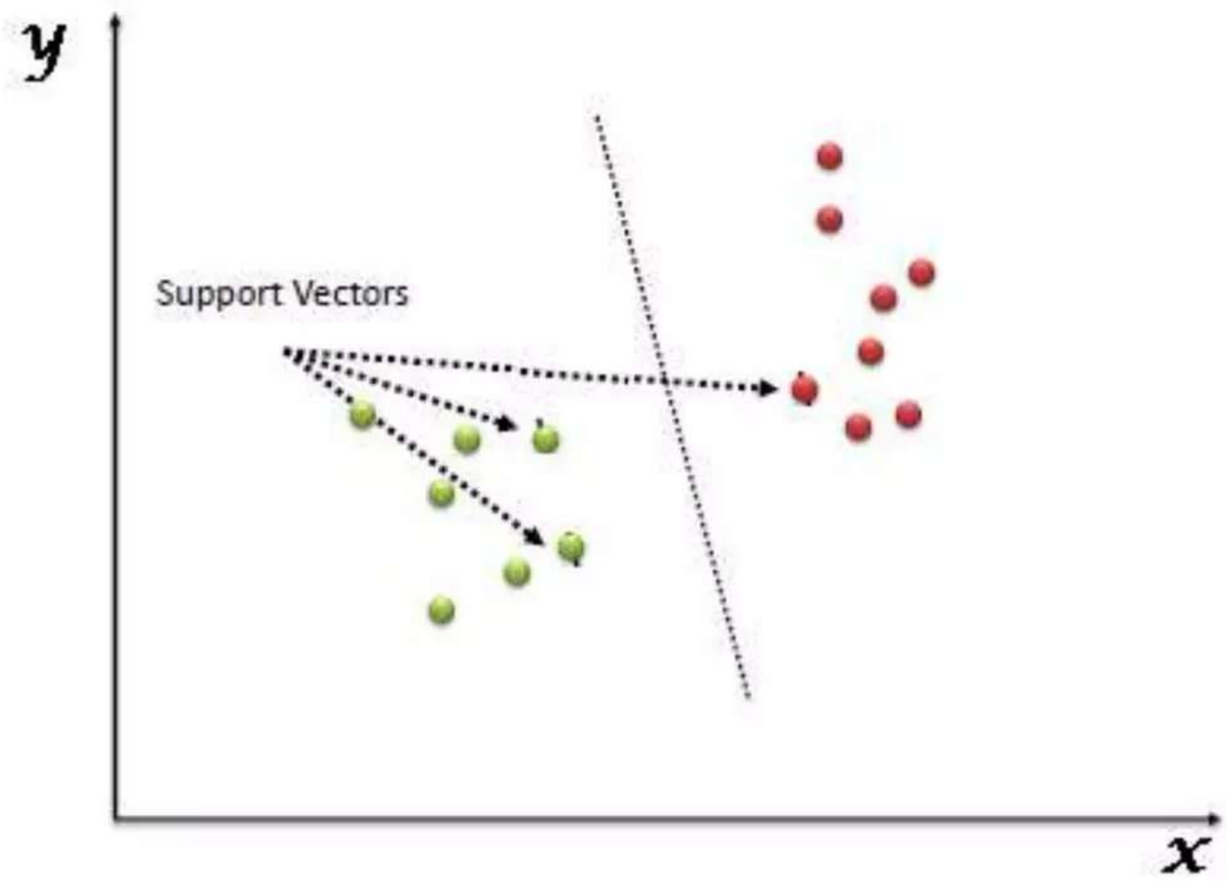
## 63. How do you build a random forest model?

A random forest model combines many decision tree models together. The decision trees chosen have high bias and low variance. These decision trees are placed parallelly. Each decision tree takes a sample subset of rows and columns with replacement. The result of each decision tree is noted and the majority, which is the mode in case of a classification problem, or mean and median in case of a regression problem is taken as the answer.

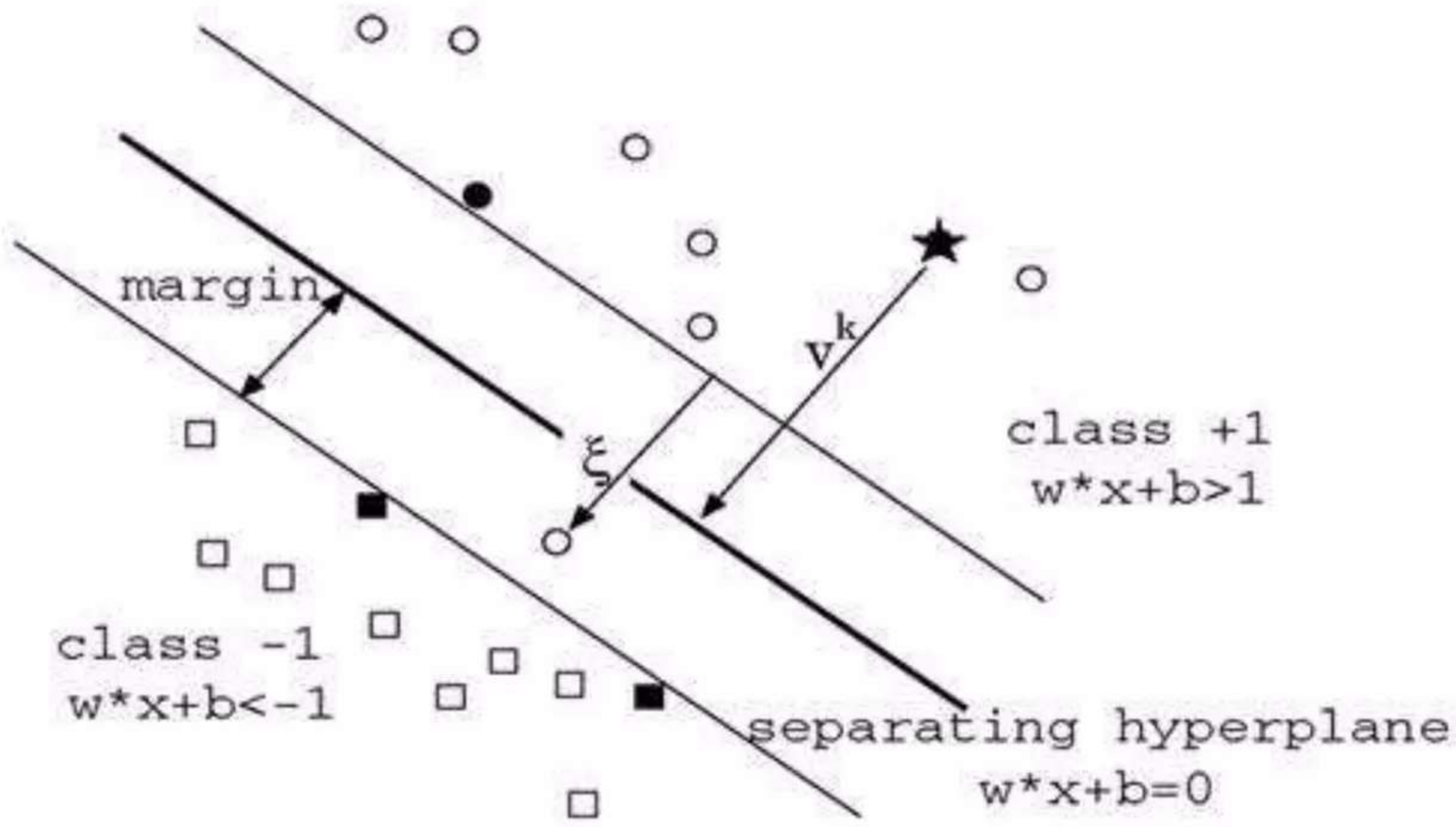



# 64. Explain the SVM algorithm in detail.

SVM stands for support vector machine, it is a supervised machine learning algorithm that can be used for both *Regression* and Classification. If you have n features in your training data set, SVM tries to plot it in n-dimensional space with the value of each feature being the value of a particular coordinate. SVM uses hyperplanes to separate out different classes based on the provided kernel function.



## 65. What are the support vectors in SVM?




In the diagram, we see that the thinner lines mark the distance from the classifier to the closest data points called the support vectors (darkened data points). The distance between the two thin lines is called the margin.

## 66. What are the different kernels in SVM?

There are four types of kernels in SVM.

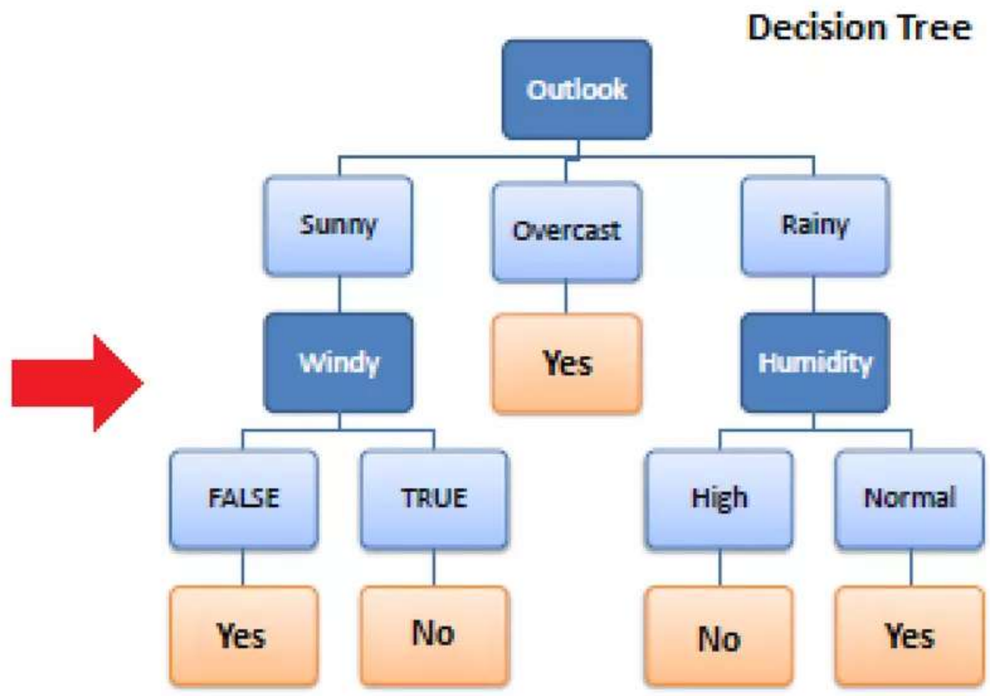
1. Linear Kernel
2. Polynomial kernel
3. Radial basis kernel
4. Sigmoid kernel

## 67. Explain the Decision Tree algorithm in detail.

A decision tree is a supervised machine-learning algorithm mainly used for Regression and Classification. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision tree can handle both categorical and numerical data.

Predictors | Target

| Outlook | Temp. | Humidity | Windy | Play Golf |
|---|---|---|---|---|
| Rainy | Hot | High | False | No |
| Rainy | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | Yes |
| Sunny | Cool | Normal | False | Yes |
| Sunny | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Rainy | Mild | High | False | No |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Mild | Normal | False | Yes |
| Rainy | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Sunny | Mild | High | True | No |

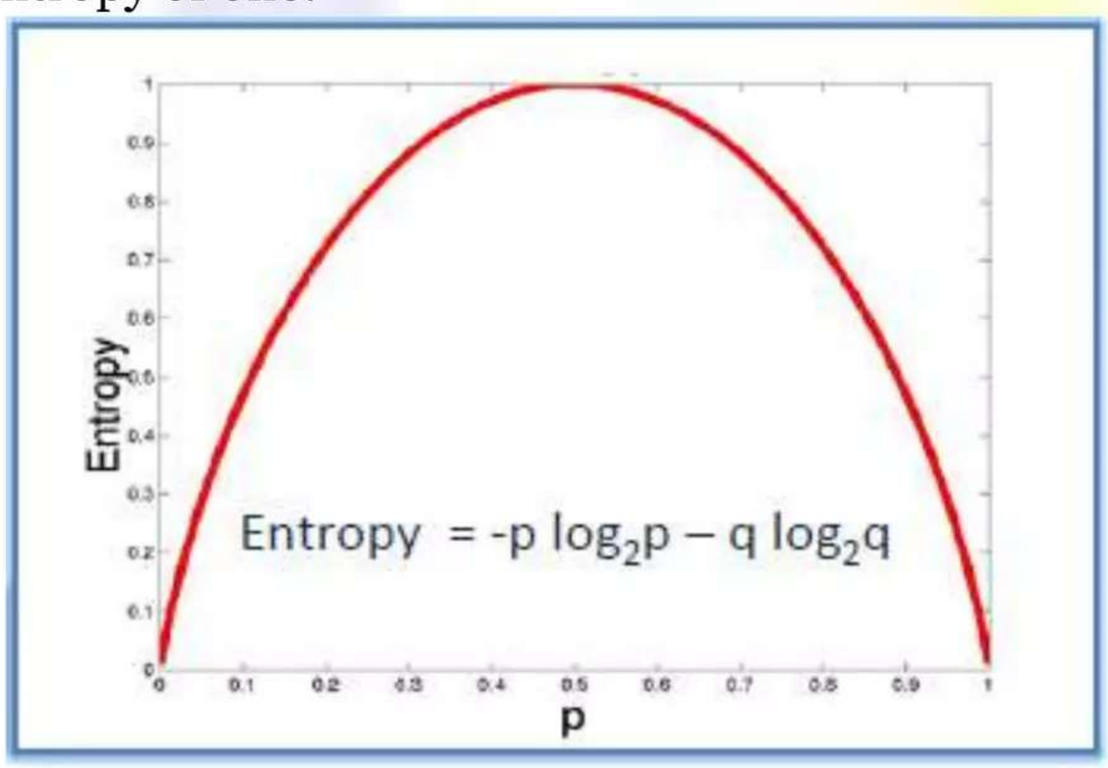



# 68. What are Entropy and Information gained in the Decision tree algorithm?

The core algorithm for building a decision tree is
called ID3. ID3 uses Entropy and Information Gain to construct a decision tree.

## *Entropy*

A decision tree is built top-down from a root node and involves the partitioning of data into homogeneous subsets. ID3 uses entropy to check the homogeneity of a sample. If the sample is completely homogeneous then the entropy is zero and if the sample is equally divided it has an entropy of one.




$$\text{Entropy} = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

## *Information Gain*

The Information Gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attributes that return the highest information gain.

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Outlook | Sunny | 3 | 2 |
| | Overcast | 4 | 0 |
| | Rainy | 2 | 3 |
| Gain = 0.247 | | | |

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Temp. | Hot | 2 | 2 |
| | Mild | 4 | 2 |
| | Cool | 3 | 1 |
| Gain = 0.029 | | | |

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Humidity | High | 3 | 4 |
| | Normal | 6 | 1 |
| Gain = 0.152 | | | |

| | | Play Golf | |
|---|---|---|---|
| | | Yes | No |
| Windy | False | 6 | 2 |
| | True | 3 | 3 |
| Gain = 0.048 | | | |

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

**G**(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)

= 0.940 – 0.693 = 0.247

## 69. What is pruning in a Decision Tree?

Pruning is a technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. So, when we remove sub-nodes of a decision node, this process is called *pruning* or the opposite process of splitting.

## 70. What is logistic regression? State an example of when you have used logistic regression recently.

Logistic Regression often referred to as the logit model is a technique to predict the binary outcome from a linear combination of predictor variables.

For example, if you want to predict whether a particular political leader will win the election or not. In this case, the outcome of the prediction is binary i.e. 0 or 1 (Win/Lose). The predictor variables here would be the amount of money spent for election campaigning of a particular candidate, the amount of time spent in campaigning, etc.

## 71. What is Linear Regression?

Linear regression is a statistical technique where the score of a variable Y is predicted from the score of a second variable X. X is referred to as the predictor variable and Y is the criterion variable.

## 72. What Are the Drawbacks of the Linear Model?

Some drawbacks of the linear model are:

- The assumption of linearity of the errors.
- It can't be used for count outcomes or binary outcomes
- There are overfitting problems that it can't solve

# 73. What is the difference between Regression and classification ML techniques?

Both Regression and classification machine learning techniques come under Supervised machine learning algorithms. In a Supervised machine learning algorithm, we have to train the model using the labeled data sets, While training we have to explicitly provide the correct labels and the algorithm tries to learn the pattern from input to output. If our labels are discrete values then it will a classification problem, e.g A, B, etc. but if our labels are continuous values then it will be a regression problem, e.g 1.23, 1.333, etc.

# 74. What are Recommender Systems?

Recommender Systems are a subclass of information filtering systems that are meant to predict the preferences or ratings that a user would give to a product. Recommender systems are widely used in movies, news, research articles, products, social tags, music, etc.

*Examples include movie recommenders in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations, and game recommendations on Xbox.*

# 75. What is Collaborative filtering?

The process of filtering is used by most recommender systems to find patterns or information by collaborating viewpoints, various data sources, and multiple agents.

| Movie | Alice | Bob | Carol | Dave |
|---|---|---|---|---|
| Shutter Island | 4 | 3 | 5 | 1 |
| Fight Club | 5 | 4 | 4 | 2 |
| Dark Knight | 5 | 3 | 4 | ? |
| 21 | 4 | 3 | ? | 5 |
| Home Alone | 4 | 4 | 5 | 5 |

**Figure**: *Predicting the rating of Dave for Dark Knight and Carol for 21 using Collaborative Filtering*

Top 100+ Google Data Science Interview Questions : All You Need to know to Crack it 80
An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others' ratings for all movies. This concept is widely used in recommending movies on IMDB, Netflix & BookMyShow, product recommenders on e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations, and game recommendations on Xbox.

# 76. How can outlier values be treated?

Outlier values can be identified by using univariate or any other graphical analysis method. If the number of outlier values is few then they can be assessed individually but for a large number of outliers, the values can be substituted with either the 99th or the 1st percentile values.

All extreme values are not outliers values. The most common ways to treat outlier values

1. To change the value and bring it within a range.
2. To just remove the value.

## 77. What are the various steps involved in an analytics project?

The following are the various steps involved in an analytics project:

1. Understand the Business problem
2. Explore the data and become familiar with it.
3. Prepare the data for modeling by detecting outliers, treating missing values, transforming variables, etc.
4. After data preparation, start running the model, analyze the result and tweak the approach. This is an iterative step until the best possible outcome is achieved.
5. Validate the model using a new data set.
6. Start implementing the model and track the result to analyze the performance of the model over a period of time.

## 78. During analysis, how do you treat missing values?

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst has to concentrate on them as it could lead to interesting and meaningful business insights.

If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored. Assigning a default value which can be the mean, minimum, or maximum value. Getting into the data is important.

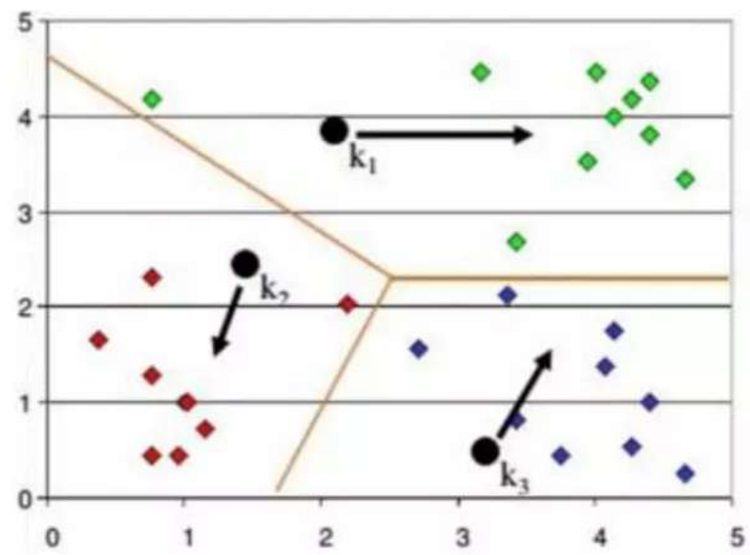
If it is a categorical variable, the default value is assigned. The missing value is assigned a default value. If you have a distribution of data coming, for normal distribution give the mean value.

If 80% of the values for a variable are missing then you can answer that you would be dropping the variable instead of treating the missing values.

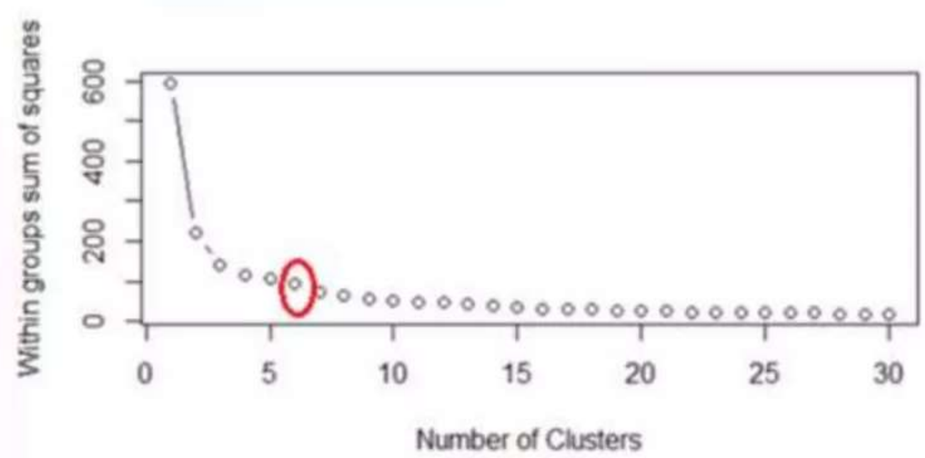
## 79. How will you define the number of clusters in a clustering algorithm?

Though the Clustering Algorithm is not specified, this question is mostly in reference to K-Means clustering where “K” defines the number of clusters. The objective of clustering is to group similar entities in a way that the entities within a group are similar to each other but the groups are different from each other.

For example, the following image shows three different groups.




Within Sum of squares is generally used to explain the homogeneity within a cluster. If you plot WSS for a range of a number of clusters, you will get the plot shown below.




The Graph is generally known as *Elbow Curve*.

- Red circled a point in the above graph i.e. *Number of Cluster* =6 is the point after which you don't see any decrement in WSS.
- This point is known as the bending point and is taken as K in K – Means.

This is the widely used approach but few data scientists also use Hierarchical clustering first to create dendrograms and identify the distinct groups from there.

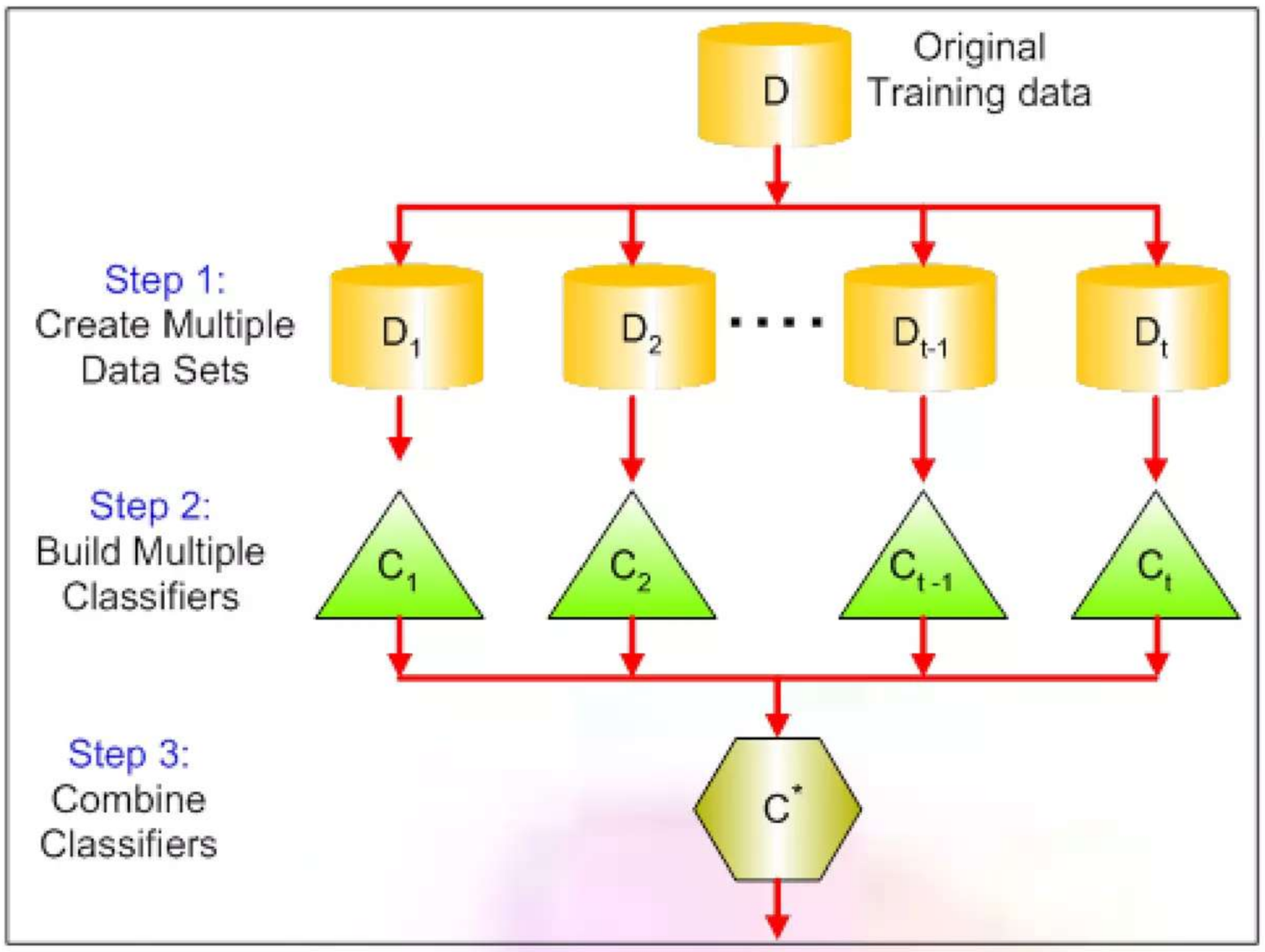
# 80. What is Ensemble Learning?

Ensemble Learning is basically combining a diverse set of learners(Individual models) together to improvise on the stability and predictive power of the model.

# 81. Describe in brief any type of Ensemble Learning.

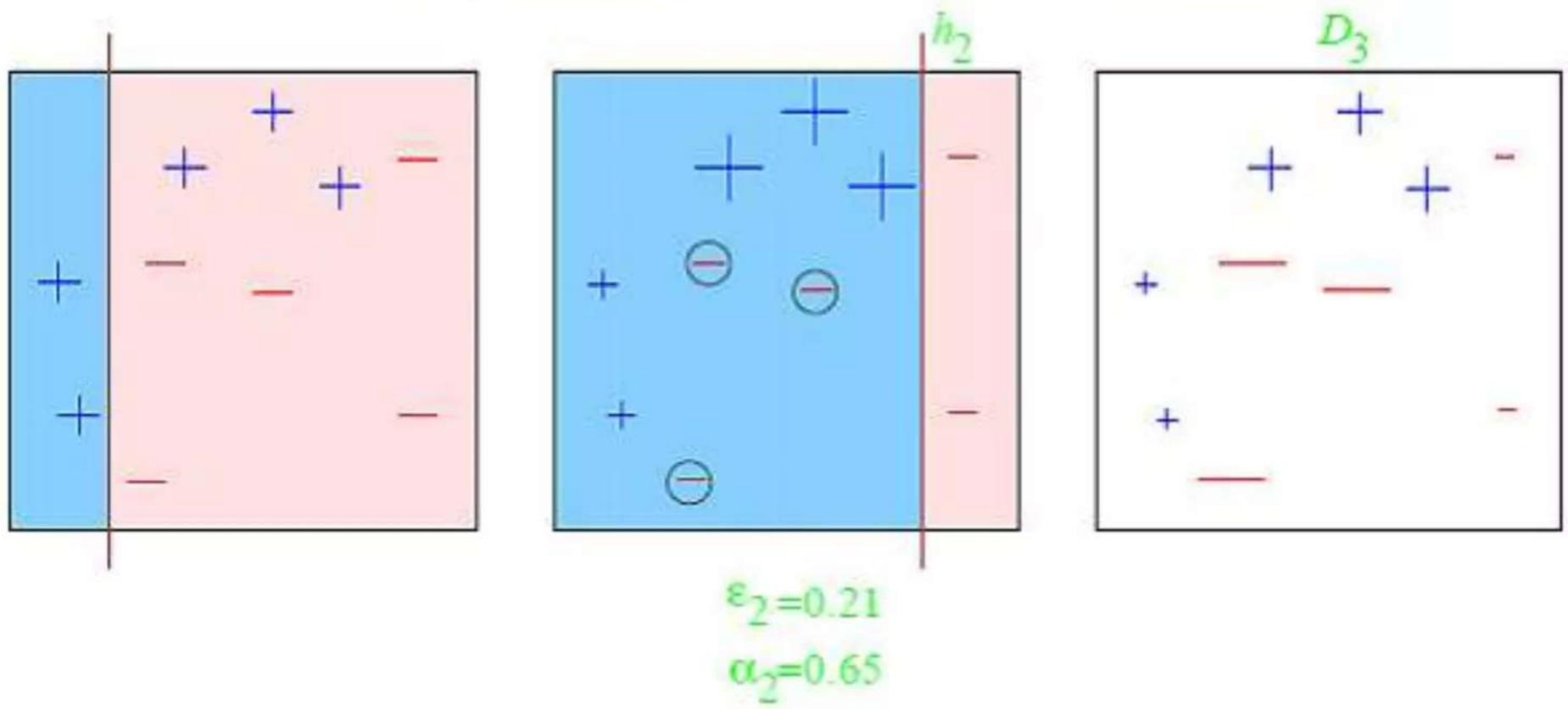
Ensemble learning has many types but two more popular ensemble learning techniques are mentioned below.

Bagging
Bagging tries to implement similar learners on small sample populations and then takes a mean of all the predictions. In generalized bagging, you can use different learners in different populations. As you expect this helps us to reduce the variance error.



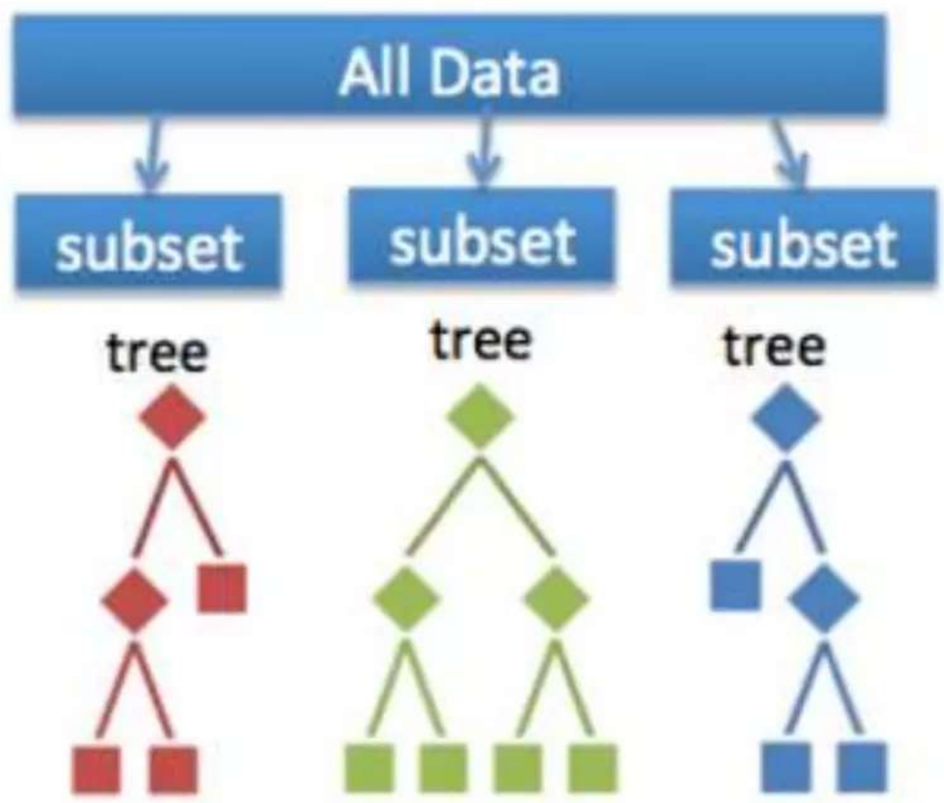

Boosting

Boosting is an iterative technique that adjusts the weight of an observation based on the last classification. If an observation was classified incorrectly, it tries to increase the weight of this observation and vice versa. Boosting in general decreases the bias error and builds strong predictive models. However, they may overfit the training data.




## 82. What is a Random Forest? How does it work?

Random forest is a versatile machine-learning method capable of performing both regression and classification tasks. It is also used for dimensionality reduction, and treats missing values, and outlier values. It is a type of ensemble learning method, where a group of weak models combines to form a powerful model.




In Random Forest, we grow multiple trees as opposed to a single tree. To classify a new object based on attributes, each tree gives a classification. The forest chooses the classification having the most votes(Overall the trees in the forest) and in the case of regression, it takes the average of outputs by different trees.

## 83. How Do You Work Towards a Random Forest?

The underlying principle of this technique is that several weak learners combined to provide a keen learner. The steps involved are

- Build several decision trees on bootstrapped training samples of data
- On each tree, each time a split is considered, a random sample of mm predictors is chosen as split candidates, out of all pp predictors
- Rule of thumb: At each split m=p√m=p
- Predictions: At the majority rule

## 84. What cross-validation technique would you use on a time series data set?

Instead of using k-fold cross-validation, you should be aware of the fact that a time series is not randomly distributed data — It is inherently ordered by chronological order.

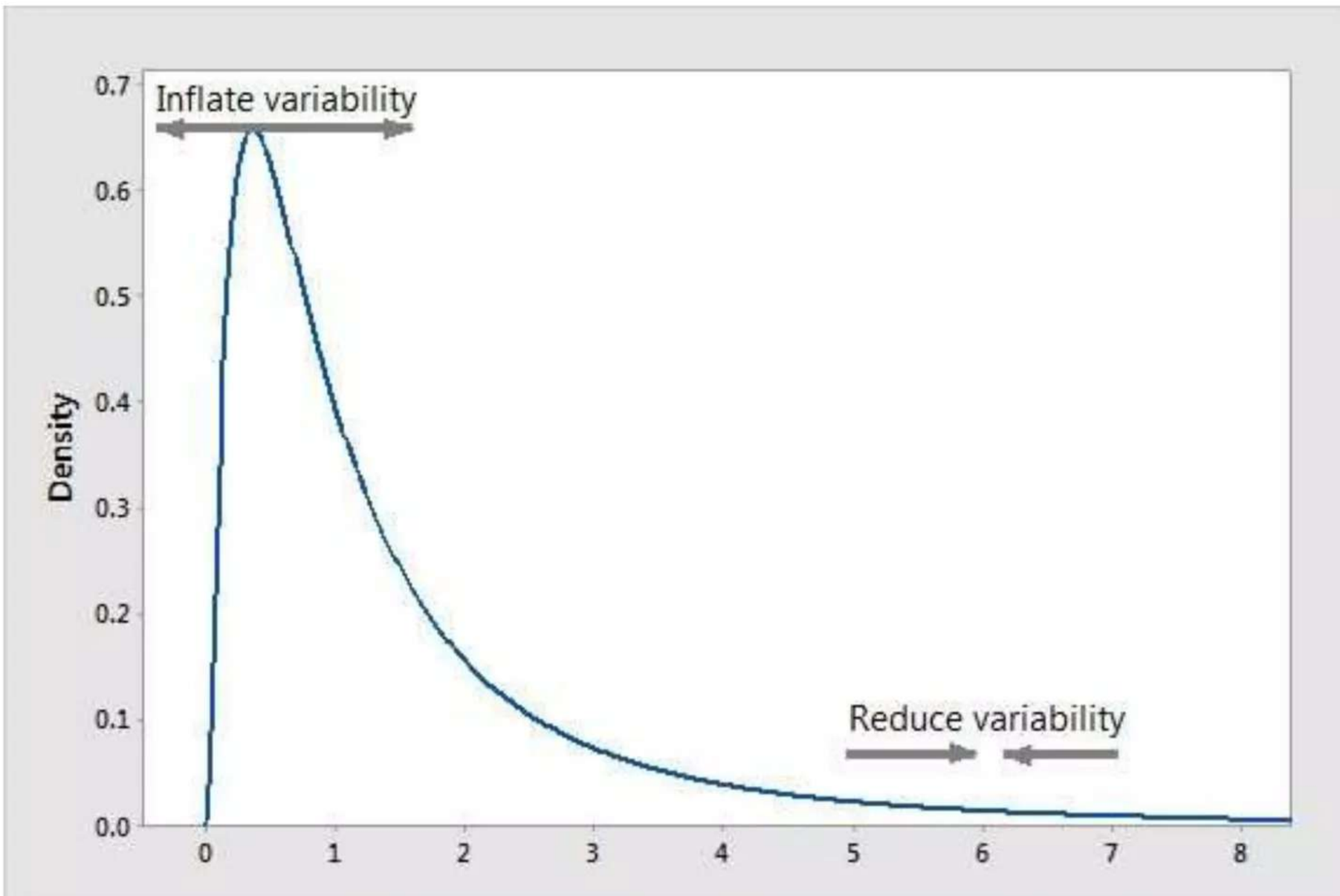
In the case of time series data, you should use techniques like forward=chaining — Where you will be model past data and then look at forward-facing data.

fold 1: training[1], test[2]
fold 1: training[1 2], test[3]
fold 1: training[1 2 3], test[4]
fold 1: training[1 2 3 4], test[5]

## 85. What is a Box-Cox Transformation?

The dependent variable for a regression analysis might not satisfy one or more assumptions of an ordinary least squares regression. The residuals could either curve as the prediction

increases or follow the skewed distribution. In such scenarios, it is necessary to transform the response variable so that the data meets the required assumptions. A Box cox transformation is a statistical technique to transform non-normal dependent variables into a normal shape. If the given data is not normal then most of the statistical techniques assume normality. Applying a box cox transformation means that you can run a broader number of tests.




A Box-Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques, if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. The Box-Cox transformation is named after statisticians *George Box* and *Sir David Roxbee Cox* who collaborated on a 1964 paper and developed the technique.

# 86. How Regularly Must an Algorithm be Updated?

You will want to update an algorithm when:

- You want the model to evolve as data streams through infrastructure
- The underlying data source is changing
- There is a case of non-stationarity
- The algorithm underperforms/ results in a lack of accuracy

## 87. If you are having 4GB RAM in your machine and you want to train your model on 10GB data set. How would you go about this problem? Have you ever faced this kind of problem in your machine learning/data science experience so far?

First of all, you have to ask which ML model you want to train.

For Neural networks: Batch size with Numpy array will work.
Steps:

1. Load the whole data in the Numpy array. A NumPy array has a property to create a mapping of the complete data set, it doesn't load the complete data set in memory.
2. You can pass an index to the Numpy array to get the required data.
3. Use this data to pass to the Neural network.
4. Have a small batch size.

For SVM: Partial fit will work
Steps:

1. Divide one big data set into small-size data sets.
2. Using a partial fit method of SVM requires a subset of the complete data set.
3. Repeat step 2 for other subsets.

# DEEP LEARNING INTERVIEW QUESTIONS

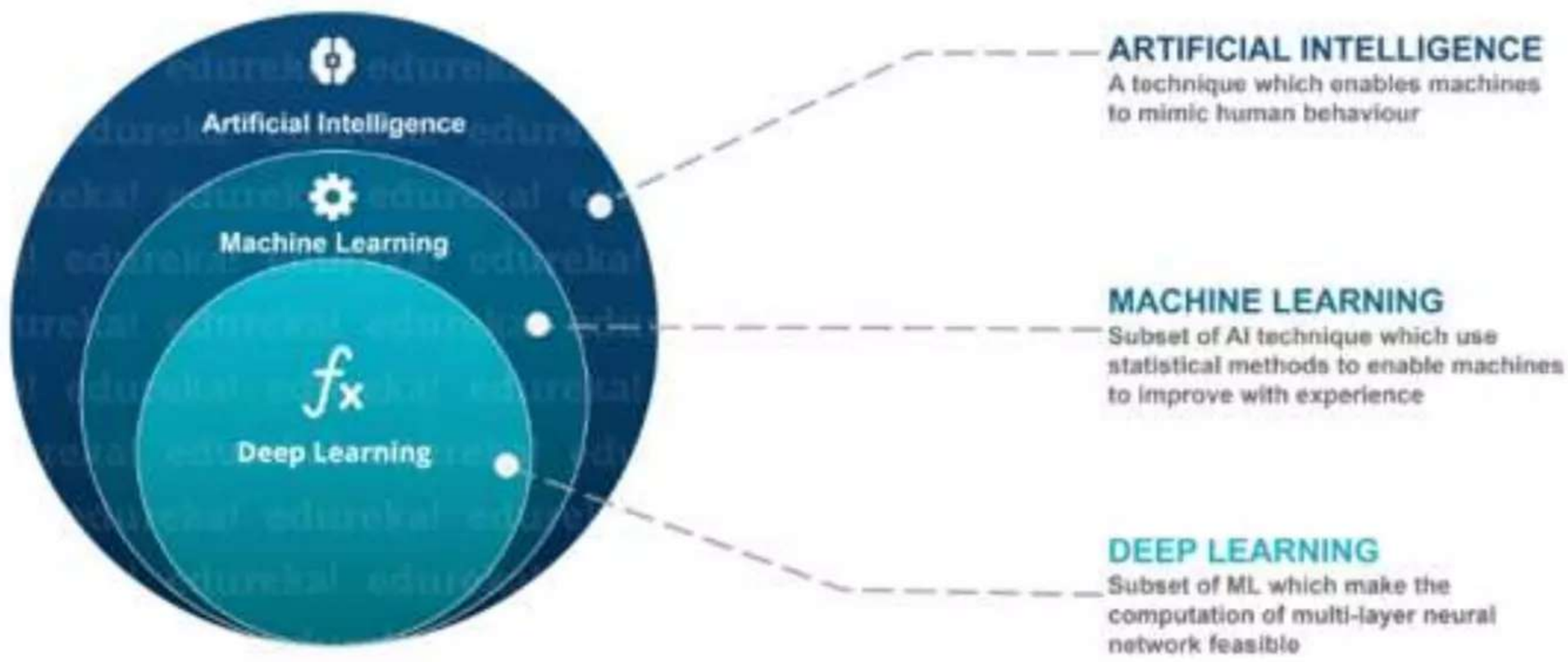
## 88. What do you mean by Deep Learning?

Deep Learning is nothing but a paradigm of machine learning which has shown incredible promise in recent years. This is because of the fact that Deep Learning shows a great analogy with the functioning of the human brain.

## 89. What is the difference between machine learning and deep learning?

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. Machine learning can be categorized into the following three categories.

1. Supervised machine learning,
2. Unsupervised machine learning,
3. Reinforcement learning





Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.

## 90. What, in your opinion, is the reason for the popularity of Deep Learning in recent times?

Now although Deep Learning has been around for many years, the major breakthroughs from these techniques came just in recent years. This is because of two main reasons:

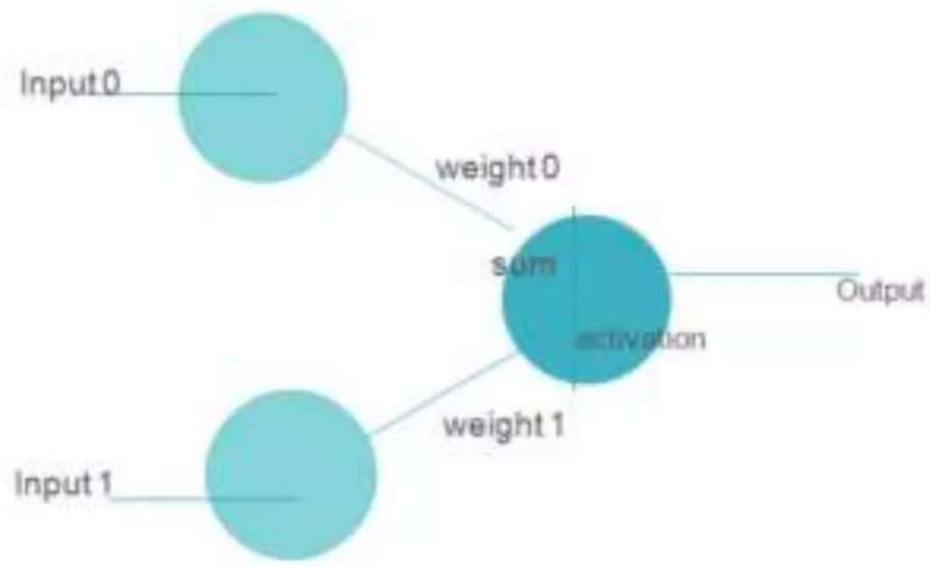
- The increase in the amount of data generated through various sources
- The growth in hardware resources required to run these models

GPUs are multiple times faster and they help us build bigger and deeper deep-learning models in comparatively less time than we required previously.

## 91. Explain Neural Network Fundamentals

A neural network in data science aims to imitate a human brain neuron, where different neurons combine together and perform a task. It learns the generalizations or patterns from data and uses this knowledge to predict output for new data, without any human intervention.

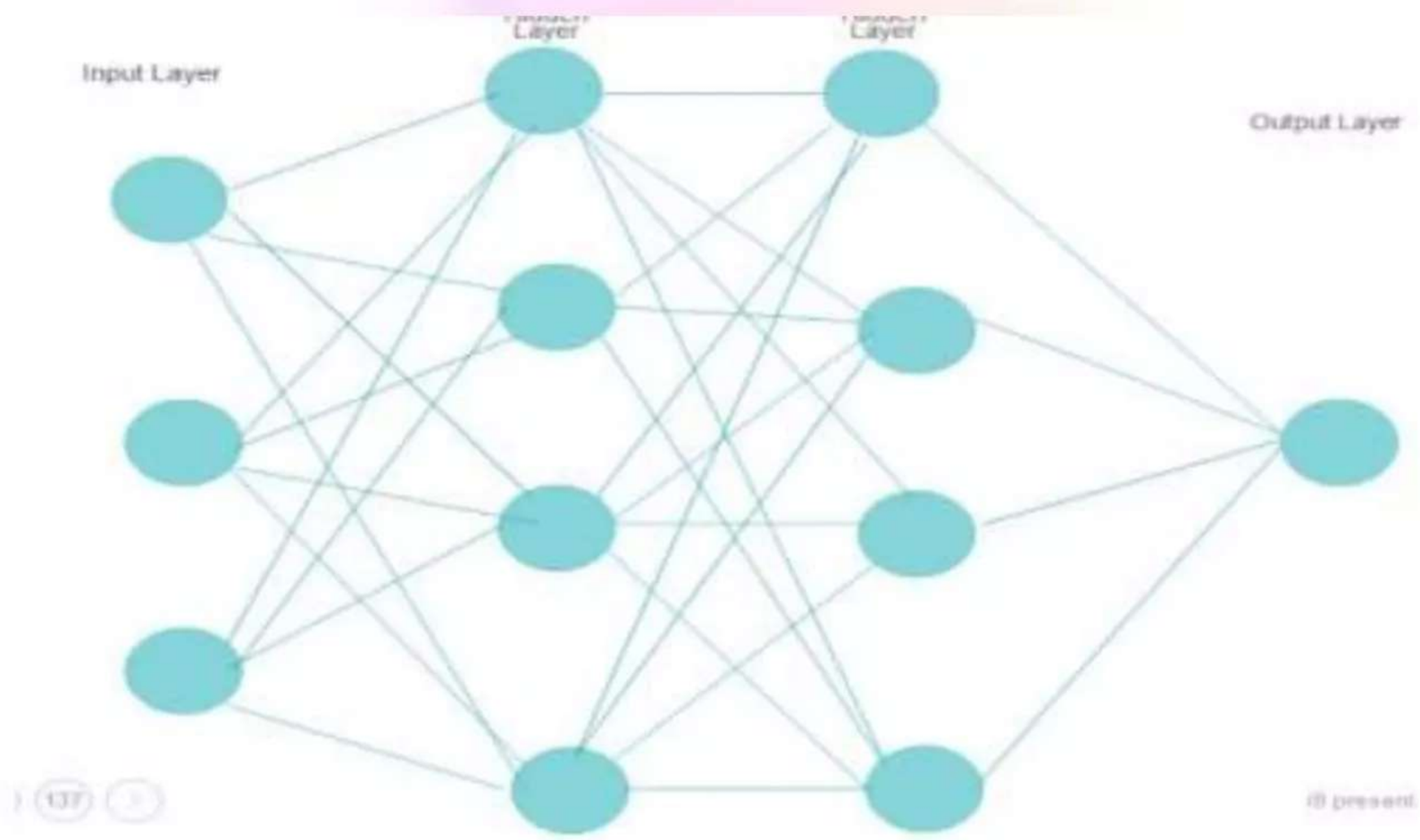
The simplest neural network can be a perceptron. It contains a single neuron, which performs the 2 operations, a weighted sum of all the inputs, and an activation function.



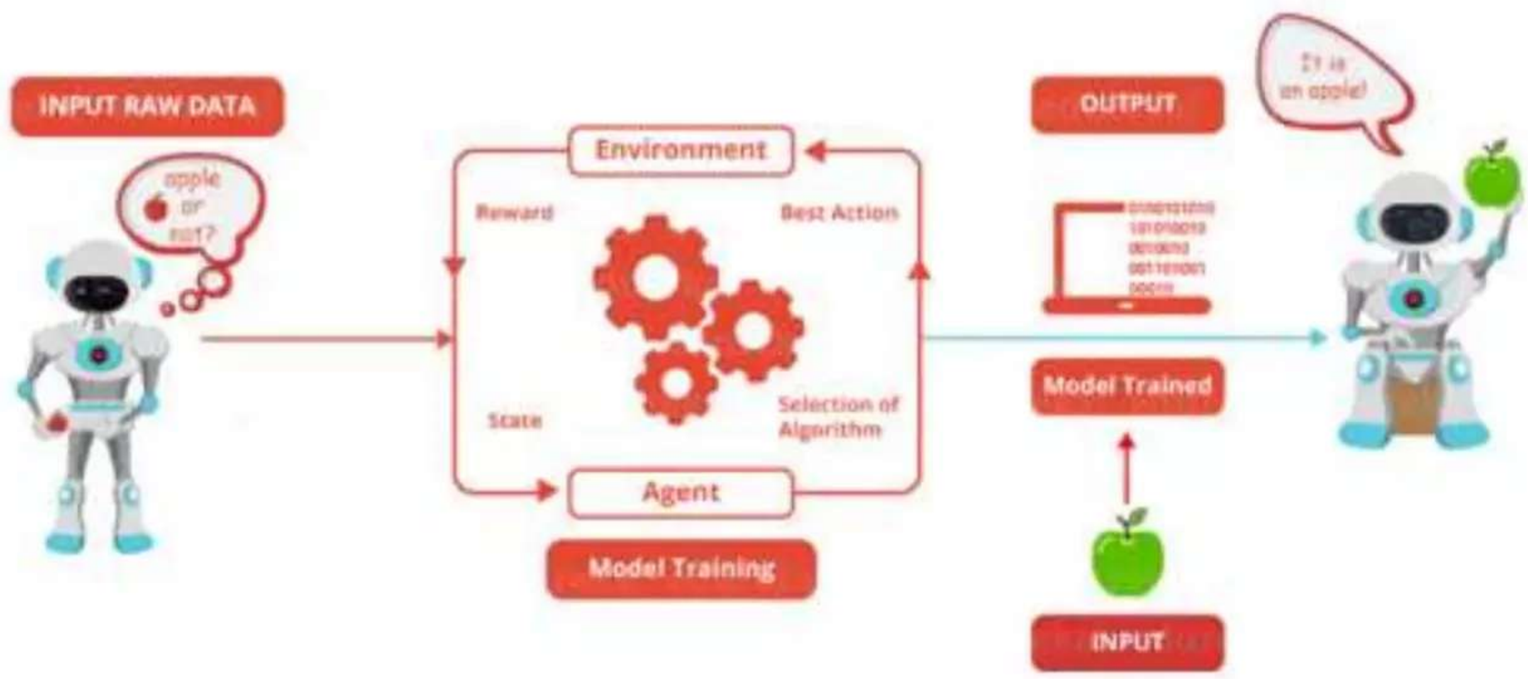

More complicated neural networks consist of the following 3 layers-

1. Input Layer– It receives the input
2. Hidden Layer-These is the layers between the input and output layers. The initially hidden layers generally help detect low-level patterns, while the further layers combine output from previous layers to find more patterns.
3. Output Layer– the output layer is the final layer that outputs the prediction.

The figure below shows a neural network-



# 92. What is reinforcement learning?



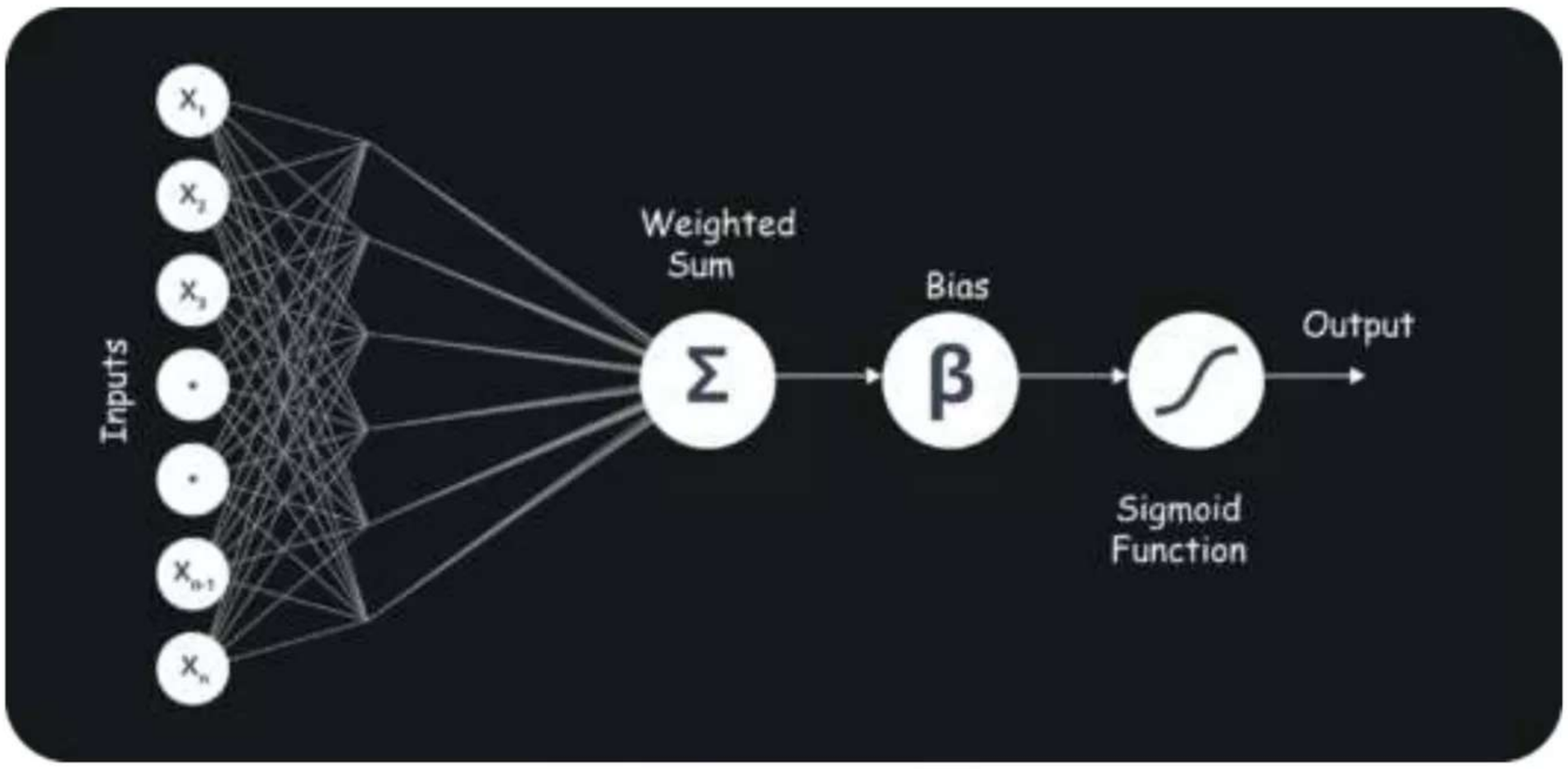

Reinforcement Learning is learning what to do and how to map situations to actions. The end result is to maximize the numerical reward signal. The learner is not told which action to take but instead must discover which action will yield the maximum reward. Reinforcement learning is inspired by the learning of human beings, it is based on the reward/penalty mechanism.

# 93. What are Artificial Neural Networks?

Artificial Neural networks are a specific set of algorithms that have revolutionized machine learning. They are inspired by biological neural networks. Neural Networks can adapt to changing the input so the network generates the best possible result without needing to redesign the output criteria.

# 94. Describe the structure of Artificial Neural Networks.

Artificial Neural Networks works on the same principle as biological Neural Network. It consists of inputs that get processed with weighted sums and Bias, with the help of Activation Functions.




## 95. How Are Weights Initialized in a Network?

There are two methods here: we can either initialize the weights to zero or assign them randomly.

Initializing all weights to 0: This makes your model similar to a linear model. All the neurons and every layer perform the same operation, giving the same output and making the deep net useless.

Initializing all weights randomly: Here, the weights are assigned randomly by initializing them very close to 0. It gives better accuracy to the model since every neuron performs different computations. This is the most commonly used method.

## 96. What Is the Cost Function?

Also referred to as "loss" or "error," cost function is a measure to evaluate how good your model's performance is. It's used to compute the error of the output layer during backpropagation. We push that error backward through the neural network and use that during the different training functions.

## 97. What Are Hyperparameters?

With neural networks, you're usually working with *hyperparameters* once the data is formatted correctly. A hyperparameter is a parameter whose value is set before the learning process begins. It determines how a network is trained and the structure of the network (such as the number of hidden units, the learning rate, epochs, etc.)

## 98. What Will Happen If the Learning Rate Is Set inaccurately (Too Low or Too High)?

When your learning rate is too low, training of the model will progress very slowly as we are making minimal updates to the weights. It will take many updates before reaching the minimum point.

If the learning rate is set too high, this causes undesirable divergent behavior to the loss function due to drastic updates in weights. It may fail to converge (the model can give a good output) or even diverge (data is too chaotic for the network to train).

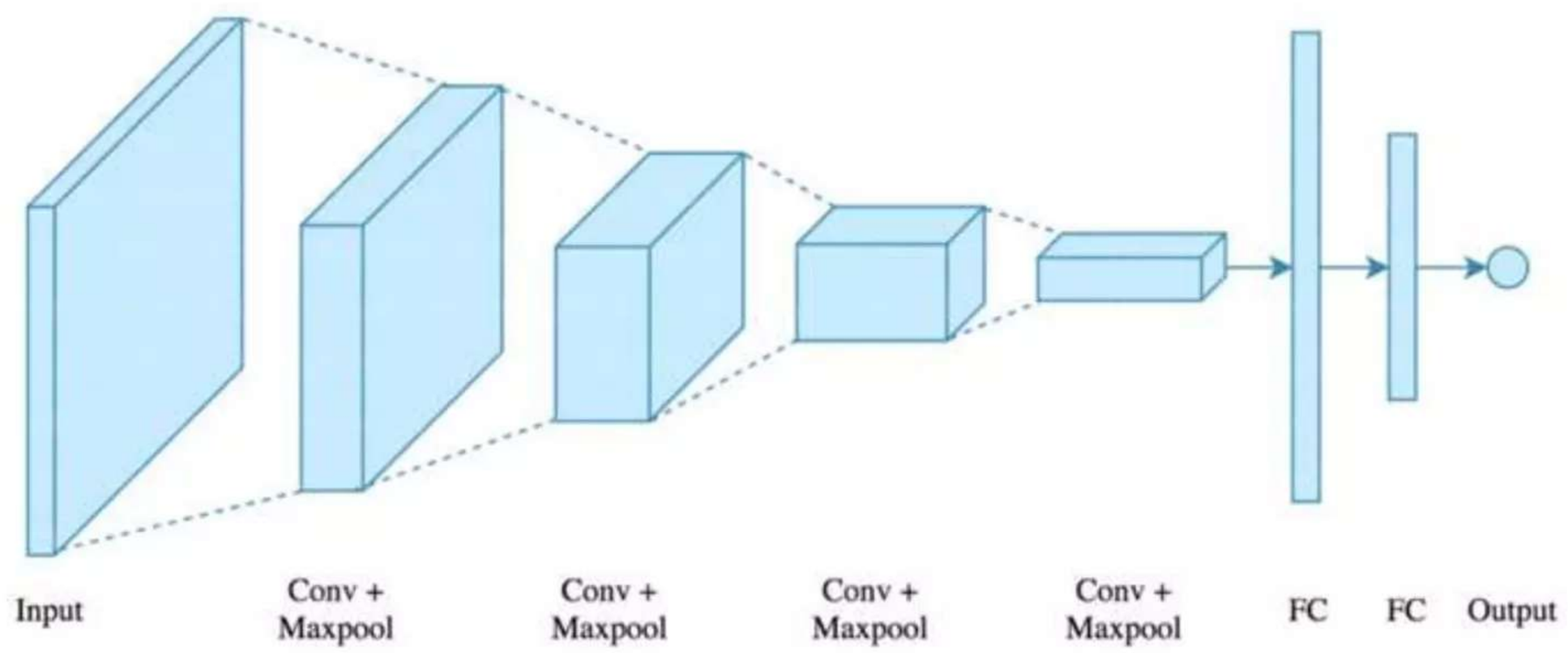
## 99. What Is the Difference Between Epoch, Batch, and Iteration in Deep Learning?

- Epoch – Represents one iteration over the entire dataset (everything put into the training model).
- Batch – Refers to when we cannot pass the entire dataset into the neural network at once, so we divide the dataset into several batches.
- Iteration – if we have 10,000 images as data and a batch size of 200. then an epoch should run 50 iterations (10,000 divided by 50).

## 100. What Are the Different Layers on CNN?

There are four layers in CNN:

1. Convolutional Layer – the layer that performs a convolutional operation, creating several smaller picture windows to go over the data.
2. ReLU Layer – it brings non-linearity to the network and converts all the negative pixels to zero. The output is a rectified feature map.
3. Pooling Layer – pooling is a down-sampling operation that reduces the dimensionality of the feature map.
4. Fully Connected Layer – this layer recognizes and classifies the objects in the image.

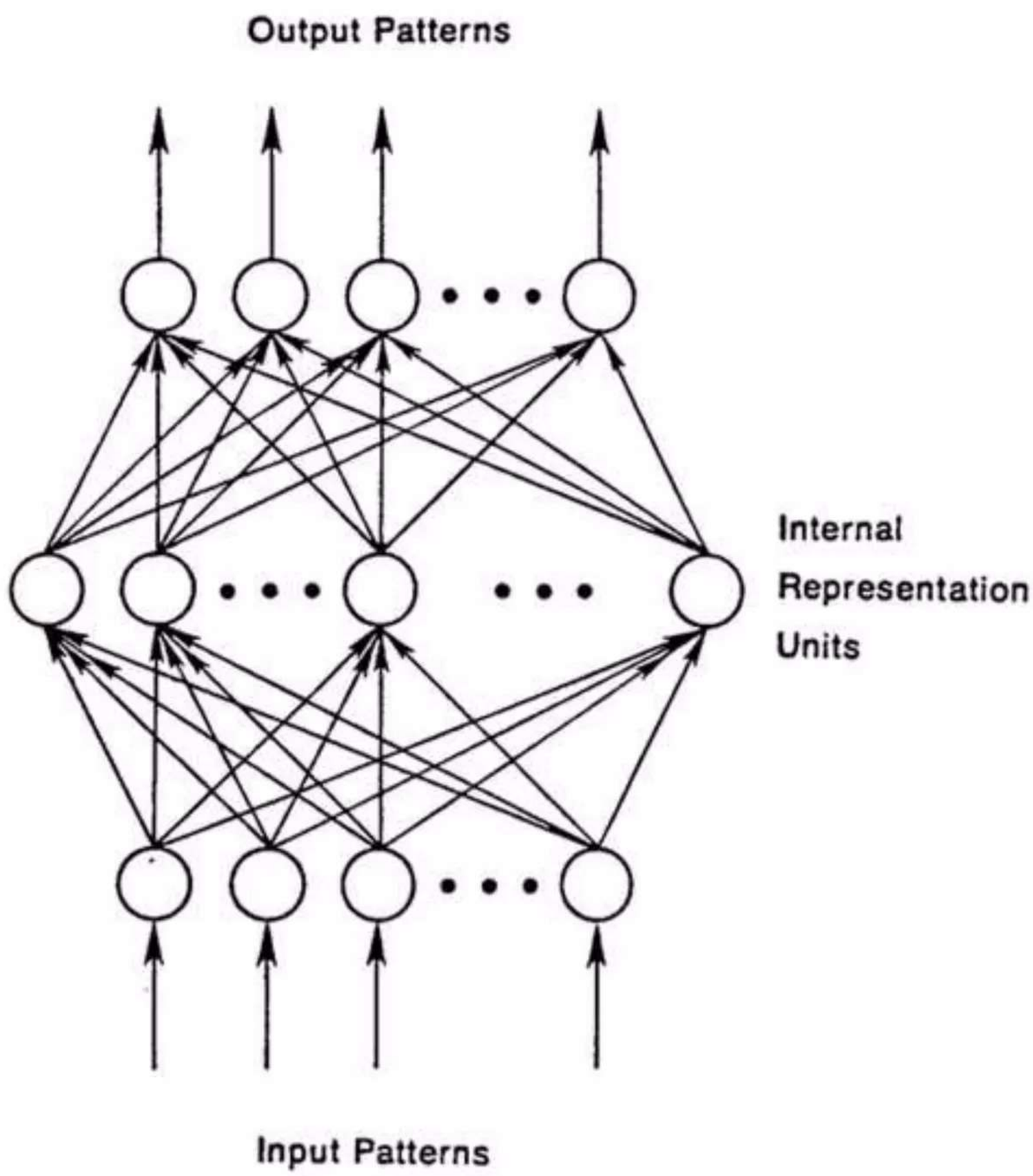



## 101. What Is Pooling on CNN, and How Does It Work?

Pooling is used to reduce the spatial dimensions of a CNN. It performs down-sampling operations to reduce the dimensionality and creates a pooled feature map by sliding a filter matrix over the input matrix.

# 102. What are Recurrent Neural Networks(RNNs)?

*RNNs* are a type of artificial neural network designed to recognize the pattern from the sequence of data such as Time series, stock market, government agencies, etc. To understand recurrent nets, first, you have to understand the basics of feedforward nets. Both these networks RNN and feed-forward named after the way they channel information through a series of mathematical orations performed at the nodes of the network. One feeds information through straight(never touching the same node twice), while the other cycles it through a loop, and the latter are called recurrent.




Recurrent networks, on the other hand, take as their input, not just the current input example they see, but also what they have perceived previously in time.

The decision a recurrent neural network reached at time t-1 affects the decision that it will reach one moment later at time t. So recurrent networks have two sources of input, the present and the recent past, which combine to determine how they respond to new data, much as we do in life.

The error they generate will return via backpropagation and be used to adjust their weights until the error can't go any lower. Remember, the purpose of recurrent nets is to classify sequential input accurately. We rely on the backpropagation of error and gradient descent to do so.

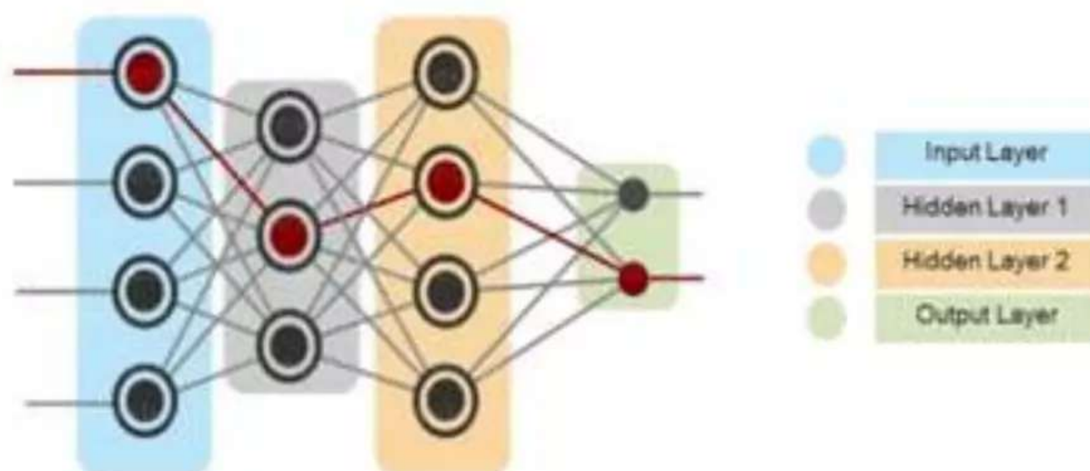
## 103. How Does an LSTM Network Work?

*Long-Short-Term Memory (LSTM)* is a special kind of recurrent neural network capable of learning long-term dependencies, remembering information for long periods as its default behavior. There are three steps in an LSTM network:

- Step 1: The network decides what to forget and what to remember.
- Step 2: It selectively updates cell state values.
- Step 3: The network decides what part of the current state makes it to the output.

## 104. What Is a Multi-layer Perceptron(MLP)?

As in *Neural Networks, MLPs* have an input layer, a hidden layer, and an output layer. It has the same structure as a single-layer *perceptron* with one or more hidden layers. A single-layer perceptron can classify only linear separable classes with binary output (0,1), but MLP can classify nonlinear classes.




Except for the input layer, each node in the other layers uses a nonlinear activation function. This means the input layers, the data coming in, and the activation function is based upon all nodes and weights being added together, producing the output. MLP uses a supervised learning method called "backpropagation." In *backpropagation*, the neural network calculates the error with the help of the cost function. It propagates this error backward from where it came from (adjusts the weights to train the model more accurately).