Supervised Learning :

**##1. What is supervised learning?**

\_Supervised learning\_ is where you have input data (X) and their corresponding output variables.

**##2. What is linear regression ?**

\_Linear Regression\_ is a parametric, discriminative supervised learning algorithm to predict continuous values of a target variable by fitting the best linear relationship between the dependent & independent variable.

**##3. What is gradient descent ?**

\_Gradient Descent\_ is a first order optimization algorithm which is used for finding the local minima of an objective function. It starts with intial set of parameter values and iteratively moves towards a set of values that minimize the function. This iterative minimization is done by taking steps towards the negative direction of the function gradient.

**##4. What is logistic regression ?**

\_Logistic regression\_ is a parametric, discriminative supervised learning algorithm for classification, i.e used where the response variable is categorical by applying a sigmoid function to a linear prediction.The idea of logistic regression is to find a relationship between features and probability of particular outcome.

**##5. What is maximum likelihood estimation ?**

The principle of maximum likelihood states that we should choose parameters so as to make the data as high probability as possible. i.e we should choose parameters to maximize likelihood function.

Note : \*\*Probability\*\* in this mathematical context describes the plausibility of a random outcome, given a model parameter value, without reference to any observed data. \*\*Likelihood\*\* describes the plausibility of a model parameter value, given specific observed data.

**##6. What is K- Nearest Neighbors ?**

It's a non-parametric supervised learning algorithm in which we assign a label to new data based on the labels of training examples which are most near to it. It's a lazy learning technique because it goes through complete training data every time it needs to predict a test sample.

**##7. How is KNN different from k-means clustering ?**

K-Nearest Neighbors is a supervised classication algorithm, while k-means clustering is an unsupervised clustering algorithm. In order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled point into (thus the nearest neighbor part).

K-means clustering requires only a set of unlabeled points and a threshold: the algorithm will take unlabeled points and gradually learn how to cluster them into groups by computing the mean of the distance between different points.

The critical difference here is that KNN needs labeled points and is thus supervised learning, while k-means doesn’t — and is thus unsupervised learning.

**## 8. What is Naive Bayes ?**

It is a supervised learning algorithm based on bayes theorem. It classifies different instances into predefined classes, assuming there is no interdependency of features.

* \*\*Pros\*\*:
  + Easy and fast to predict class of test data set. Also, performs well in multi-class prediction.
* \*\*Cons\*\*:
  + Bad estimator: Probability outputs from predict\_proba are not to be taken too seriously.

- Assumption of independent predictors: In real life, it is almost impossible that we get a set of predictors which are completely independent.

**##9. What is Support Vector Machines ?**

\_Support Vector Machines\_ is an non-parametric, discriminative supervised learning algorithm which identifies optimal separating hyperplane which maximizes the margin between different classes of the training data.

**###10. Kernel Functions ?**

Kernel methods owe their name to the use of kernel functions, which enable them to operate in a high-dimensional feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space.\*\*Note\*\* : Any linear model can be turned into a non-linear model by applying the kernel trick to the model: replacing its features (predictors) by a kernel function

**###11. Examples of SVM Kernels.**

* Polynomial kernel
* Gaussian radial basis function (RBF)
* Sigmoid kernel

\*\*Pros\*\*:

* It is really effective in higher dimension. If you have more features than training examples, most of the algorithms perform very bad, but SVM is the only algorithm which can saves you in this situation.
* Best algorithm if you data are separable. That two classes are not mixed.
* Only support vectors affect the optimally spaced hyperplane. So, it is less affected by outliers.

\*\*Cons\*\*:

* On large dataset it takes too much time. Mainly because of kernel function calculations and finding optimal hyperplane in higher dimensions.
* Can not perform well in case of overlapping classes.
* Can only give you 0–1 classification. Probably estimates computation are really expensive.

**## 12. What is bagging ?**

\_Bagging\_ is an ensemble technique mainly used to reduce the variance of our predictions by combining the results of multiple classifiers modelled on different sub-samples of the same dataset. In Bagging, individual learner are trained in parallel.

**##13. What is decision trees ?**

Decision trees are non-parametric supervised learning algorithm.

Given the training data, a decision tree algorithm divides the feature space into regions. For inference, we first see which region does the test data point fall in and take the mean label value (regression) or the majority label value ( classification ).There are couple of algorithms there to build a decision tree , we only talk about a few which are:

* CART (Classification and Regression Trees) → uses Gini Index(Classification) as metric.
* ID3 (Iterative Dichotomiser 3) → uses Entropy function and Information gain as metrics.

**#### 14. Finding the variable/feature for best split**.

**Gini Index:** Measure of variance across all classes of the data. Measures the impurity of the data.Ex. Given a binary classi cation problem, the number of positive cases equals the negative ones.

GI = 1/2\*(1–1/2)+1/2\*(1–1/2) = 1/2 </br>

This is maximum GI possible. As we split data, and move towards subtree, GI decreases to zero with increase in depth of tree.

**Entropy:** Measure of randomness. More the random data, higher the entropy.

E = -p\*log(p) ; p - probability

**Information Gain:** Decrease in entropy. The difference between the entropy before the split and the average entropy after split is obtained to decide when to split.

The variable which provides maximum entropy gain is chosen!

* Pros:
  + Easy to understand and visualise.
  + Can be used for feature engineering.
  + Very little data preparation needed for algorithm.
* Cons:
  + If not tuned well, may lead to overfitting.
  + Unstable. Small variation in data leads to completely different tree formation.
  + In case of imbalanced dataset, decision trees are biased.However, by using proper splitting criteria, this issue can be resolved.
* Important Parameters:
* \*\*Minimum Samples Split\*\*: Minimum number of sample required to split a node. This parameter helps in reducing over tting. High value: Underfitting, Low value: overfitting.
* \*\*Maximum Depth of a Tree\*\*: Most in uential parameter. Gives limit on vertical depth decide upto which level pruning is required. Higher value: overfitting, Lower value: Underfitting
* \*\*Maximum Features\*\*: At each node, while splitting either we can chose best feature from pool of all the features or limited number of random features. This parameter adds a little randomness – good generalised model.

**##15. What is random forest ?**

Random forest improves bagging further by adding some randomness. In random forest, only a subset of features are selected at random to construct a tree (while often not subsample instances). The benefit is that random forest decorrelates the trees.

For example, suppose we have a dataset. There is one very predicative feature, and a couple of moderately predicative features. In bagging trees, most of the trees will use this very predicative feature in the top split, and therefore making most of the trees look similar, and highly correlated. Averaging many highly correlated results won't lead to a large reduction in variance compared with uncorrelated results. In random forest for each split we only consider a subset of the features and therefore reduce the variance even further by introducing more uncorrelated trees.

* \*\*Pros\*\*:
  + As it predicts by aggregating the predictions from smaller predictors the variance decreases. Less overfitting.
  + Useful when missing data is huge.
* \*\*Cons\*\*:
  + Better with classification than regression.
  + Black box approach: Many factors are random.
  + Slight increase in Bias
* \*\*Parameters\*\*:
  + \*\*n\_estimators\*\*: Number of trees in the model. The larger the better, but the longer it will take to compute.
  + \*\*max\_features\*\*: Size of the random subsets of features to consider when splitting a node. Lower the #features, greater the reduction in variance, but greater the increase in bias.
  + \*\*feature\_importances\_\*\*: The relative importances of each feature to the model. Features used in the tree at the top nodes are relatively more important as more data points are dependent on that feature.

**##16. What is boosting ?**

Boosting builds on weak learners, and in an iterative fashion. In each iteration, a new learner is added, while all existing learners are kept unchanged. All learners are weighted based on their performance (e.g., accuracy), and after a weak learner is added, the data are re-weighted: examples that are misclassified gain more weights, while examples that are correctly classified lose weights. Thus, future weak learners focus more on examples that previous weak learners misclassified.

**### 17. Types of Boosting Algorithms**

* AdaBoost (Adaptive Boosting)\*\*
* Gradient Tree Boosting\*\*
* XGBoost
* \*\*Pros\*\*:
  + Automatically do feature engineering.
  + Very little data preparation needed for algorithm.
* \*\*Cons\*\*:
  + Time and computation expensive.
  + Complexity of the classiffication increases.
  + Hard to implement in real time platform.

**## 18.Metrics**

**Confusion Matrix** : An NxN matrix where N is the no. of classes, that summarizes how successful a classification model's predictions are.

**Accuracy :** Accuracy is the fraction of predictions our model got right. Suppose you build a model that classified 100 tumors as malignant (the positive class) or benign (the negative class):

**Recall or Sensitivity or True Positive Rate:** Number of items correctly identified as positive out of total true positives. High recall means you’re not missing many positives.Our model has a recall of 0.11—in other words, it correctly identifies 11% of all malignant tumors.

**Precision:** Number of items correctly identified as positive out of total items identified as positive. High precision means low “false alarm rate” (if you test positive, you’re probably positive). Our model has a precision of 0.5—in other words, when it predicts a tumor is malignant, it is correct 50% of the time.

**True Negative Rate or Specificity :** Number of items correctly identified as negative out of total true negatives.

**Type 1 Error or False Positive Rate or false alarm rate :** Number of items wrongly identified as positive out of total true negatives.

**Type 2 Error or False Negative Rate or miss rate :** Number of items wrongly identified as negative out of total true positives.

**RMSE (Root Mean Square Error):** It represents the sample standard deviation of the differences between predicted values and observed values (called residuals).

**MAE**: MAE is the average of the absolute difference between the predicted values and observed value. **R Squared (R²) and Adjusted R Square :** R Squared & Adjusted R Squared (\_goodness of fit measure\_) are often used for explanatory purposes and explains how well your selected independent variable(s) explain the variability in your dependent variable(s). Note\*\* : Higher the MSE, smaller the R\_squared and poorer is the model.

Just like R², adjusted R² also shows how well terms fit a curve or line but adjusts for the number of terms in a model.Note\*\* : The more predictors you add the higher R^2 become hence use adjusted R^2 which adjusts for the degrees of freedom.

**## 19. Explain how a ROC curve works?**

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It’s often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

AUC ROC = area under the ROC curve.

# Unsupervised Learning

**## 20. What is unsupervised Learning?**

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

**##21. What is clustering?**

It's an machine learning technique which segregate the various data points into different groups called clusters such that entities in a particular group comparatively have more similar traits than entities in another group.

**## 22. Top 5 Clustering Algorithms to know:**

* K-Means Clustering
* Agglomerative Hierarchical Clustering
* Mean-Shift Clustering
* Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
* Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM)

**## 23. K- means Clustering ?**

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

**###24. K-Means Clustering Algorithm?**

Pros

* K-Means has the advantage that it’s pretty fast, as all we’re really doing is computing the distances between points and group centers; very few computations! It thus has a linear complexity O(n).

Cons

* You have to select how many groups/classes there are.
* K-means also starts with a random choice of cluster centers and therefore it may yield different clustering results on different runs of the algorithm. Thus, the results may not be repeatable and lack consistency.

**##25. Determining The Optimal Number Of Clusters: 3 Must Know Methods?**

* Elbow method (within-cluster sum of square vs number of clusters) : We want it to be as small as possible.
* Average silhouette method (average silhouette of observations (avg.sil) vs number of clusters) .high average silhouette width indicates a good clustering.
* Gap statistic method The gap statistic compares the total intracluster variation for different values of k with their expected values under null reference distribution of the data i.e. a distribution with no obvious clustering.

**##26. Hierarchical Clustering.**

It is a type of connectivity model clustering which is based on the fact that data points that are closer to each other are more similar than the data points lying far away in a data space. As the name speaks for itself, the hierarchical clustering forms the hierarchy of the clusters that can be studied by visualising dendogram.

Pros : Hierarchical clustering does not require us to specify the number of clusters and we can even select which number of clusters looks best since we are building a tree.

Cons : Lower efficiency, as it has a time complexity of O(n³)

**##27. Dimesionality Reduction Technique:**

Mainly there are two types of dimesion reduction technique:

1. Matrix Factorization(PCA)

2. Neighbour Graphs (UMAP, T-SNE)

**###28. PCA (Principal Component Analysis):**

* Principal Component Analysis (PCA) is a dimension reduction technique that projects the data into a lower dimensional space.
* PCA uses Singular Value Decomposition (SVD), which is a matrix factorization method that decomposes a matrix into three smaller matrices.
* PCA finds top N principal components, which are dimensions along which the data vary (spread out) the most. Intuitively, the more spread out the data along a specific dimension,the more information is contained, thus the more important this dimension is for the pattern recognition of the dataset
* PCA can be used as pre-step for data visualization: reducing high dimensional data into 2D or 3D. Note : PCA is interpretable dimension reduction.

**###29. T-SNE(t-distributed stochastic neighbor embedding):**

\* It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.

\* The t-SNE algorithm comprises two main stages.First, t-SNE constructs a probability distribution over pairs of high-dimensional objects in such a way that similar objects have a high probability of being picked, whilst dissimilar points have an extremely small probability of being picked.

\* Second, t-SNE defines a similar probability distribution over the points in the low-dimensional map, and it minimizes the Kullback–Leibler divergence between the two distributions with respect to the locations of the points in the map.

**###30. UMAP(Uniform Manifold Approximation and Projection):**

\* UMAP (Uniform Manifold Approximation and Projection) is a novel manifold learning technique for dimension reduction

\* The UMAP algorithm is competitive with t-SNE for visualization quality, and arguably preserves more of the global structure with superior run time performance.

\* UMAP’s topological foundations allow it to scale to signicantly larger dataset sizes than are feasible for t-SNE.

**##31 : Cross Entropy or Log Loss**

Cross-entropy is commonly used to quantify the difference between two probability distributions.Cross-entropy loss measures how close is the predicted distribution to the true distribution.

**##32 : Why the Negative Sign?**

Log Loss uses negative log to provide an easy metric for comparison. It takes this approach because the positive log of numbers < 1 returns negative values, which is confusing to work with when comparing the performance of two models.

**##33. Explain Bias-Variance Tradeoff**

**Bias:** is the simplifying assumptions made by the model to make the target function easier to approximate.

**Variance :** is the amount that the estimate of the target function will change given different training data.

**The bias–variance trade-off:** is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.

# Statistics for Data Science

* **Population :** The entire group one desires information about.
* **Sample:** A subset of the population taken because the entire population is usually too large to analyze. It's characteristics are taken to be representative of the population.
* **Mean:** The sum of all the values in the sample divided by the number of values in the sample/population.
* **Median:** The median is the value separating the higher half of a data sample from the lower half.
* **Standard Deviation:** Square root of the variance. It measures the dispersion around the mean.
* **Percentiles:** An extension of median to values other than 50%.
* **Interquartile range (IQR):** the difference between the 75th and 25th percentile
* **Mode :** The most frequently occuring value
* Range: Difference between the maximum value and the minimum value.

Notice that most of these fall into one of two categories: they capture either the center of the distribution (e.g., mean, median, mode), or its spread (e.g., variance, IQR, range). These two categories are often called \*\*measures of central tendency\*\* and \*\*measures of dispersion\*\*, respectively.

**##34.Important Distributions :**

**Gaussian/Normal :** We say x ∼ N (µ, σ2) to mean that x is drawn from a Gaussian (or Normal) distribution with mean µ and variance σ2 (or equivalently standard deviation σ). We’ll often use the standard normal distribution, or N (0, 1) (i.e., mean 0 and variance 1).

The probability of getting a value within 1 standard deviation of the mean is about 68%. For 2 standard deviations, it’s about 95%, and for 3 standard deviations it’s about 99%. This is sometimes called the \*\*“68-95-99 rule”\*\*.

**Bernoulli :** A Bernoulli random variable can be thought of as the outcome of flipping a biased coin, where the probability of heads is p. To be more precise, a Bernoulli random variable takes on value 1 with probability p and value 0 with probability 1−p. Its expectation is p, and its variance is p(1 − p). Bernoulli variables are typically used to model binary random variables.

**##35 : Central Limit Theorem :**

The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30.

**##36: Confidence Intervals :**

A confidence interval is how much uncertainty there is with any particular statistic. Confidence intervals are often used with a margin of error. It tells you how confident you can be that the results from a poll or survey reflect what you would expect to find if it were possible to survey the entire population.

**##37 : Hypothesis Testing :** Hypothesis testing in statistics is a way for you to test the results of a survey or experiment to see if you have meaningful results.

Hypothesis testing steps:

* Figure out your null hypothesis,
* State your null hypothesis,
* Choose what kind of test you need to perform,
* Either support or reject the null hypothesis.

**##38 : Significance Level :**

The significance level α is the probability of making the wrong decision when the null hypothesis is true. Alpha levels (sometimes just called “significance levels”) are used in hypothesis tests. Usually, these tests are run with an alpha level of .05 (5%), but other levels commonly used are .01 and .10.

**## 39 : P-value :**

A p value is used in hypothesis testing to help you support or reject the null hypothesis. The p value is the evidence against a null hypothesis. The smaller the p-value, the strong the evidence that you should reject the null hypothesis.

**## 40 : A/B testing:**

A statistical way of comparing two (or more) techniques, typically an incumbent against a new rival. A/B testing aims to determine not only which technique performs better but also to understand whether the difference is statistically significant. A/B testing usually considers only two techniques using one measurement, but it can be applied to any finite number of techniques and measures.

**## 41 : Correlation:** Correlation is a statistical measure that describes the association between random variables.

**Types of Correlation :**

* Pearson Correlation Coefficient (measures the linear association between continuous variables)
* Spearman's Correlation (special case of Pearson ρ applied to ranked (sorted) variables. appropriate to use with both continuous and discrete data.)
* Kendall's Tau (more appropriate for discrete data.)

**## 42 : Statistical Hypothesis Tests :**

**Normality Tests** (Statistical tests that you can use to check if your data has a Gaussian distribution.)

Shapiro-Wilk Test: Tests whether a data sample has a Gaussian distribution.

Correlation Tests (Statistical tests that you can use to check if two samples are related)

Pearson’s Correlation Coefficient : Tests whether two samples have a monotonic relationship.

Spearman’s Rank Correlation : Tests whether two samples have a monotonic relationship.

**Chi-Squared Test :** Tests whether two categorical variables are related or independent.

Parametric Statistical Hypothesis Tests (Statistical tests that you can use to compare data samples.)

**Student’s t-test:** Tests whether the means of two independent samples are significantly different.

**Paired Student’s t-test:** Tests whether the means of two paired samples are significantly different.

**Analysis of Variance Test (ANOVA) :** Tests whether the means of two or more independent samples are significantly different.

**Repeated Measures ANOVA Test**: Tests whether the means of two or more paired samples are significantly different.

**### 43: Nonparametric Statistical Hypothesis Tests:**

* Mann-Whitney U Test : Tests whether the distributions of two independent samples are equal or not.
* Wilcoxon Signed-Rank Test: Tests whether the distributions of two paired samples are equal or not.
* Kruskal-Wallis H Test: Tests whether the distributions of two or more independent samples are equal or not.
* Friedman Test : Tests whether the distributions of two or more paired samples are equal or not.

**### 44: AI Learning Models: Knowledge-Based Classification -**

* Inductive Learning: This type of AI learning model is based on inferring a general rule from datasets of input-output pairs.
* Deductive Learning: This type of AI learning technique starts with a series of rules and infers new rules that are more efficient in the context of a specific AI algorithm.

**### 45: AI Learning Models: Feedback-Based Classification -**

Based on the feedback characteristics, AI learning models can be classified as supervised, unsupervised, semi-supervised or reinforced.

Unsupervised Learning: Unsupervised models focus on learning a pattern in the input data without any external feedback. Clustering is a classic example of unsupervised learning models.

Supervised Learning: Supervised learning models use external feedback to learning functions that map inputs to output observations. In those models the external environment acts as a “teacher” of the AI algorithms.

Semi-supervised Learning: Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for training – typically a small amount of labeled data with a large amount of unlabeled data.The goal of a semi-supervised model is to classify some of the unlabeled data using the labeled information set.

Reinforcement Learning: Reinforcement learning models use opposite dynamics such as rewards and punishment to “reinforce” different types of knowledge. This type of learning technique is becoming really popular in modern AI solutions.

**##46: Data Mining Vs Machine Learning -**

Machine learning focuses on prediction, based on known properties learned from the training data.

Data mining focuses on the discovery of (previously) unknown properties in the data. This is the analysis step of Knowledge Discovery in Databases.

**##47: Workflow of Data Science Project –**

Given a data science / machine learning project, what steps should we follow? Here's how we should tackle it:

**Specify business objective-** Are we trying to win more customers, achieve higher satisfaction, or gain more revenues?

**Define problem-** What is the specific gap in your ideal world and the real one that requires machine learning to fill? Ask questions that can be addressed using your data and predictive modeling (ML algorithms).

**Create a common sense baseline -** But before you resort to ML, set up a baseline to solve the problem as if you know zero data science. You may be amazed at how effective this baseline is. It can be as simple as recommending the top N popular items or other rule-based logic. This baseline can also server as a good benchmark for ML algorithms.

**Review ML literatures -** To avoid reinventing the wheel and get inspired on what techniques / algorithms are good at addressing the questions using our data.

**Set up a single-number metric -** What it means to be successful - high accuracy, lower error, or bigger AUC - and how do you measure it? The metric has to align with high-level goals, most often the success of your business. Set up a single-number against which all models are measured.

**Do exploratory data analysis (EDA) -** Play with the data to get a general idea of data type, distribution, variable correlation, facets etc. This step would involve a lot of plotting.

**Partition data -** Validation set should be large enough to detect differences between the models you are training; test set should be large enough to indicate the overall performance of the final model; training set, needless to say, the larger the merrier.

**Preprocess -** This would include data integration, cleaning, transformation, reduction, discretization and more.

**Engineer features -** Coming up with features is difficult, time-consuming, requires expert knowledge. Applied machine learning is basically feature engineering. This step usually involves feature selection and creation, using domain knowledge. Can be minimal for deep learning projects.

**Develop models -** Choose which algorithm to use, what hyperparameters to tune, which architecture to use etc.

**Ensemble -** Ensemble can usually boost performance, depending on the correlations of the models/features. So it’s always a good idea to try out. But be open-minded about making tradeoff - some ensemble are too complex/slow to put into production.

**Deploy model -** Deploy models into production for inference.

**Monitor model -** Monitor model performance, and collect feedbacks.

**Iterate -** Iterate the previous steps. Data science tends to be an iterative process, with new and improved models being developed over time.

**## 48 : Parametric v/s Nonparametric ?**

A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.

A learning model where the number of parameters is not determined prior to training. On the contrary, nonparametric models (can) become more and more complex with an increasing amount of data.

**## 49 : Discriminative vs Generative Learning Algorithm ?**

Discriminative algorithms model p(y|x; w), that is, given the dataset and learned parameter, what is the probability of y belonging to a specific class. A discriminative algorithm doesn't care about how the data was generated.Ex: Linear Regression, Logistic Regression, Support Vector Machines etc.

Generative algorithms model p(x|y), that is, the distribution of features given that it belongs to a certain class. A generative algorithm models how the data was generated.Ex: Naive Bayes, Hidden Markov Models etc.

Given a training set, an algorithm like logistic regression or the perceptron algorithm (basically) tries to find a straight line—that is, a decision boundary—that separates the elephants and dogs. Then, to classify a new animal as either an elephant or a dog, it checks on which side of the decision boundary it falls, and makes its prediction accordingly.

First, looking at elephants, we can build a model of what elephants look like. Then, looking at dogs, we can build a separate model of what dogs look like. Finally, to classify a new animal, we can match the new animal against the elephant model, and match it against the dog model, to see whether the new animal looks more like the elephants or more like the dogs we had seen in the training set.

**## 50. What is cross validation ?**

Cross Validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a validation set to evaluate it. For ex: K fold CV divides the data into k folds, train on each k-1 folds and evaluate it on remaining 1 fold. The result of k models can be averaged to get a overall model performance.

**Time - Series Cross Validation** : Experimenters cannot cut out a piece in the middle, and train on data before and after this portion. Instead, they need to train on a set of data that is older than the test data.

**##51. What is overfitting?**

Overfitting or High Variance is a modeling error which is caused by a hypothesis function that fits the training data too close but does not generalise well to predict new data.

**## 52. What is regularization?**

Regulariztion is a technique to prevent overfitting by penalizing the coefficients of the cost function.

**### 53 : Ridge Regression :**

It performs ‘L2 regularization’, i.e. adds penalty equivalent to square of the magnitude of coefficients. `L 2 regularizer` is also called a gaussian prior or weight decay . Thus, it optimises the following:

Objective = RSS + α \* (sum of square of coefficients)

**### 54: Lasso Regression:**

LASSO stands for Least Absolute Shrinkage and Selection Operator.Lasso regression performs L1 regularization, i.e. it adds a factor of sum of absolute value of coefficients in the optimisation objective.

Objective = RSS + α \* (sum of absolute value of coefficients)

**### 55 : Elastic nets :**

A technique known as Elastic Nets, which is a combination of Lasso and Ridge regression is used to tackle the limitations of both Ridge and Lasso Regression.

**##56: Loss Functions for Regression and Classification?**

* Regression Loss Function
  + - * Square or l2 loss (not robust)
      * Absolute or Laplace loss (not differentiable)
      * Huber Loss (robust and differentiable)
* Classification Loss Function
  + - * SVM/Hinge loss
      * log loss

**## 57 : How do you handle missing or corrupted data in a dataset?**

Before jumping to the methods of data imputation, we have to understand the reason why data goes missing.

Missing Completely at Random (MCAR): The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.

Missing at Random (MAR) - a weaker assumption than MCAR: Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data.

Missing not at Random (MNAR): Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable’s value (e.g. Let’s assume that females generally don’t want to reveal their ages! Here the missing value in age variable is impacted by gender variable).

Methods :

* Listwise Deletion : In the listwise deletion method, all rows that have one or more column values missing are deleted.
* Mean, Median and Mode Imputation: In the mean/median/mode imputation method, all missing values in a particular column are substituted with the mean/median/mode, which is calculated using all the values available in that column.
* Multiple Imputation
* Last Observation Carried Forward (LOCF)
* KNN (K Nearest Neighbors)

**## 58 : How would you handle an imbalanced dataset?**

* Using a better metrics like AUROC, Precision, Recall etc.
* Cost-sensitive Learning
* Over sampling of the minority class or Under sampling of the majority class.
* SMOTE (Synthetic Minority Over-sampling Technique.)
* Anomaly Detection

**##59: how do you detect outliers?**

Outliers are extreme values that deviate from other observations on data , they may indicate a variability in a measurement, experimental errors or a novelty.

**### 60 : How to find outliers ?**

Visualize the Data -

* Histogram: A histogram is the best way to check univariate data — data containing a single variable — for outliers
* Scatter Plot: A scatter plot is useful to find outliers in bivariate data (data with two variables). You can easily spot the outliers because they will be far away from the majority of points on the scatter plot.
* Box Plot

**## 61 :What is regression? Which models can you use to solve a regression problem?**

Regression is a part of supervised ML. Regression models investigate the relationship between a dependent (target) and independent variable (s) (predictor).

Here are some common regression models :

Linear Regression- establishes a linear relationship between target and predictor (s). It predicts a numeric value and has a shape of a straight line.

Polynomial Regression - has a regression equation with the power of independent variable more than 1. It is a curve that fits into the data points.

Ridge Regression -helps when predictors are highly correlated (multicollinearity problem). It penalizes the squares of regression coefficients but doesn’t allow the coefficients to reach zeros (uses L2 regularization).

Lasso Regression -penalizes the absolute values of regression coefficients and allows some of the coefficients to reach absolute zero (thereby allowing feature selection).

**## 67 :How do we check if a variable follows the normal distribution? ‍**

1. Plot a histogram out of the sampled data. If you can fit the bell-shaped "normal" curve to the histogram, then the hypothesis that the underlying random variable follows the normal distribution can not be rejected.

2. Check Skewness and Kurtosis of the sampled data. Skewness = 0 and kurtosis = 3 are typical for a normal distribution, so the farther away they are from these values, the more non-normal the distribution.

3. Use Kolmogorov-Smirnov or/and Shapiro-Wilk tests for normality. They take into account both Skewness and Kurtosis simultaneously.

4. Check for Quantile-Quantile plot. It is a scatterplot created by plotting two sets of quantiles against one another. Normal Q-Q plot place the data points in a roughly straight line.

**##68: What if we want to build a model for predicting prices? Are prices distributed normally? Do we need to do any pre-processing for prices?**

Data is not normal. Specially, real-world datasets or uncleaned datasets always have certain skewness. Same goes for the price prediction. Price of houses or any other thing under consideration depends on a number of factors. So, there's a great chance of presence of some skewed values i.e outliers if we talk in data science terms. Yes, you may need to do pre-processing. Most probably, you will need to remove the outliers to make your distribution near-to-normal.

**## 69: What is the normal equation? ‍**

Normal equations are equations obtained by setting equal to zero the partial derivatives of the sum of squared errors (least squares); normal equations allow one to estimate the parameters of a multiple linear regression.

**## 70 : What is SGD  —  stochastic gradient descent? What’s the difference with the usual gradient descent? ‍**

In both gradient descent (GD) and stochastic gradient descent (SGD), you update a set of parameters in an iterative manner to minimize an error function.

While in GD, you have to run through ALL the samples in your training set to do a single update for a parameter in a particular iteration, in SGD, on the other hand, you use ONLY ONE or SUBSET of training sample from your training set to do the update for a parameter in a particular iteration. If you use SUBSET, it is called Minibatch Stochastic gradient Descent.

**## 71 : Which metrics for evaluating regression models do you know?**

1. Mean Squared Error(MSE)

2. Root Mean Squared Error(RMSE)

3. Mean Absolute Error(MAE)

4. R² or Coefficient of Determination

5. Adjusted R²

**## 72 : Why do we need to split our data into three parts: train, validation, and test?**

The training set is used to fit the model, i.e. to train the model with the data. The validation set is then used to provide an unbiased evaluation of a model while fine-tuning hyperparameters. This improves the generalization of the model. Finally, a test data set which the model has never "seen" before should be used for the final evaluation of the model. This allows for an unbiased evaluation of the model. The evaluation should never be performed on the same data that is used for training. Otherwise the model performance would not be representative.

**## 73 : Can you explain how cross-validation works?**

Cross-validation is the process to separate your total training set into two subsets: training and validation set, and evaluate your model to choose the hyperparameters. But you do this process iteratively, selecting differents training and validation set, in order to reduce the bias that you would have by selecting only one validation set.

**## 74 : What is K-fold cross-validation?**

K fold cross validation is a method of cross validation where we select a hyperparameter k. The dataset is now divided into k parts. Now, we take the 1st part as validation set and remaining k-1 as training set. Then we take the 2nd part as validation set and remaining k-1 parts as training set. Like this, each part is used as validation set once and the remaining k-1 parts are taken together and used as training set.It should not be used in a time series data.

**### 75 :How do we choose K in K-fold cross-validation? What’s your favorite K?**

There are two things to consider while deciding K: the number of models we get and the size of validation set. We do not want the number of models to be too less, like 2 or 3. At least 4 models give a less biased decision on the metrics. On the other hand, we would want the dataset to be at least 20-25% of the entire data. So that at least a ratio of 3:1 between training and validation set is maintained. I tend to use 4 for small datasets and 5 for large ones as K.

**## 76 : How do we evaluate classification models?**

1. Accuracy

2. Precision

3. Recall

4. F1 Score

5. Logistic loss (also known as Cross-entropy loss)

6. Jaccard similarity coefficient score

**## 78 : What is accuracy?**

Accuracy is a metric for evaluating classification models. It is calculated by dividing the number of correct predictions by the number of total predictions.

**##79: Is accuracy always a good metric?**

Accuracy is not a good performance metric when there is imbalance in the dataset. For example, in binary classification with 95% of A class and 5% of B class, a constant prediction of A class would have an accuracy of 95%. In case of imbalance dataset, we need to choose Precision, recall, or F1 Score .

**## 80 : Precision-recall trade-off ‍??**

Tradeoff means increasing one parameter would lead to decreasing of other. Precision-recall tradeoff occur due to increasing one of the parameter(precision or recall) while keeping the model same.

In an ideal scenario where there is a perfectly separable data, both precision and recall can get maximum value of 1.0. But in most of the practical situations, there is noise in the dataset and the dataset is not perfectly separable. There might be some points of positive class closer to the negative class and vice versa. In such cases, shifting the decision boundary can either increase the precision or recall but not both. Increasing one parameter leads to decreasing of the other.

**## 81 : What is the ROC curve? When to use it? ‍**

ROC stands for \*Receiver Operating Characteristics\*. The diagrammatic representation that shows the contrast between true positive rate vs true negative rate. It is used when we need to predict the probability of the binary outcome.

**## 82 : What is AUC (AU ROC)? When to use it? ‍**

AUC stands for Area Under the ROC Curve. ROC is a probability curve and AUC represents degree or measure of separability. It's used when we need to value how much model is capable of distinguishing between classes. The value is between 0 and 1, the higher the better.

**83 . How to interpret the AU ROC score?** ‍

An excellent model has AUC near to the 1 which means it has good measure of separability. A poor model has AUC near to the 0 which means it has worst measure of separability. When AUC score is 0.5, it means model has no class separation capacity whatsoever.

**83. What is the PR (precision-recall) curve? ‍**

A \*precision\*-\*recall curve\* (or PR Curve) is a plot of the precision (y-axis) and the recall (x-axis) for different probability thresholds. Precision-recall curves (PR curves) are recommended for highly skewed **85. In which cases AU PR is better than AU ROC? ‍**

domains where ROC curves may provide an excessively optimistic view of the performance.

**84. What is the area under the PR curve? Is it a useful metric? ‍**

The Precision-Recall AUC is just like the ROC AUC, in that it summarizes the curve with a range of threshold values as a single score.A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate.

What is different however is that AU ROC looks at a true positive rate TPR and false positive rate FPR while AU PR looks at positive predictive value PPV and true positive rate TPR.

Typically, if true negatives are not meaningful to the problem or you care more about the positive class, AU PR is typically going to be more useful; otherwise, If you care equally about the positive and negative class or your dataset is quite balanced, then going with AU ROC is a good idea.

**86: What do we do with categorical variables? ‍**

Categorical variables must be encoded before they can be used as features to train a machine learning model. There are various encoding techniques, including:

* One-hot encoding
* Label encoding
* Ordinal encoding
* Target encoding

**87: Why do we need one-hot encoding? ‍**

If we simply encode categorical variables with a Label encoder, they become ordinal which can lead to undesirable consequences. In this case, linear models will treat category with id 4 as twice better than a category with id 2. One-hot encoding allows us to represent a categorical variable in a numerical vector space which ensures that vectors of each category have equal distances between each other. This approach is not suited for all situations, because by using it with categorical variables of high cardinality (e.g. customer id) we will encounter problems that come into play because of the curse of dimensionality.

**88.How do we interpret weights in linear models? ‍**

Without normalizing weights or variables, if you increase the corresponding predictor by one unit, the coefficient represents on average how much the output changes. By the way, this interpretation still works for logistic regression - if you increase the corresponding predictor by one unit, the weight represents the change in the log of the odds. If the variables are normalized, we can interpret weights in linear models like the importance of this variable in the predicted result.

**89. If a weight for one variable is higher than for another  —  can we say that this variable is more important?** ‍

Yes - if your predictor variables are normalized. Without normalization, the weight represents the change in the output per unit change in the predictor. If you have a predictor with a huge range and scale that is used to predict an output with a very small range - for example, using each nation's GDP to predict maternal mortality rates - your coefficient should be very small. That does not necessarily mean

**90. What are the main parameters of the random forest model? ‍**

* `max\_depth`: Longest Path between root node and the leaf
* `min\_sample\_split`: The minimum number of observations needed to split a given node
* `max\_leaf\_nodes`: Conditions the splitting of the tree and hence, limits the growth of the trees
* `min\_samples\_leaf`: minimum number of samples in the leaf node
* `n\_estimators`: Number of trees
* `max\_sample`: Fraction of original dataset given to any individual tree in the given model
* `max\_features`: Limits the maximum number of features provided to trees in random forest model

**91. How do we select the depth of the trees in random forest? ‍**

The greater the depth, the greater amount of information is extracted from the tree, however, there is a limit to this, and the algorithm even if defensive against overfitting may learn complex features of noise present in data and as a result, may overfit on noise. Hence, there is no hard thumb rule in deciding the depth, but literature suggests a few tips on tuning the depth of the tree to prevent overfitting:

* + limit the maximum depth of a tree
  + limit the number of test nodes
  + limit the minimum number of objects at a node required to split
  + do not split a node when, at least, one of the resulting subsample sizes is below a given threshold
  + stop developing a node if it does not sufficiently improve the fit.

**92. How do we know how many trees we need in random forest? ‍**

The number of trees in random forest is worked by n\_estimators, and a random forest reduces overfitting by increasing the number of trees. There is no fixed thumb rule to decide the number of trees in a random forest, it is rather fine tuned with the data, typically starting off by taking the square of the number of features (n) present in the data followed by tuning until we get the optimal results.

**93. What happens when we have correlated features in our data? ‍**

In random forest, since random forest samples some features to build each tree, the information contained in correlated features is twice as much likely to be picked than any other information contained in other features. In general, when you are adding correlated features, it means that they linearly contains the same information and thus it will reduce the robustness of your model. Each time you train your model, your model might pick one feature or the other to "do the same job" i.e. explain some variance, reduce entropy, etc.

**94. What is collaborative filtering? ‍**

* Collaborative filtering is the most prominent approach to generate recommendations.
* It uses the wisdom of the crowd, i.e. it gives recommendations based on the experience of others.
* A recommendation is calculated as the average of other experiences.
* Say we want to give a score that indicates how much user u will like an item i. Then we can calculate it with the experience of N other users U as r\_ui = 1/N \* sum(v in U) r\_vi.
* In order to rate similar experiences with a higher weight, we can introduce a similarity between users that we use as a multiplier for each rating.
* Also, as users have an individual profile, one user may have an average rating much larger than another user, so we use normalization techniques (e.g. centering or Z-score normalization) to remove the users' biases.
* Collaborative filtering does only need a rating matrix as input and improves over time. However, it does not work well on sparse data, does not work for cold starts (see below) and usually tends to overfit.

**95. How we can incorporate implicit feedback (clicks, etc) into our recommender systems? ‍**

In comparison to explicit feedback, implicit feedback datasets lack negative examples. For example, explicit feedback can be a positive or a negative rating, but implicit feedback may be the number of purchases or clicks. One popular approach to solve this problem is named weighted alternating least squares (wALS) [Hu, Y., Koren, Y., & Volinsky, C. (2008, December). Collaborative filtering for implicit feedback datasets. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 263-272). IEEE.]. Instead of modeling the rating matrix directly, the numbers (e.g. amount of clicks) describe the strength in observations of user actions. The model tries to find latent factors that can be used to predict the expected preference of a user for an item.

**96. What is the cold start problem? ‍**

Collaborative filterung incorporates crowd knowledge to give recommendations for certain items. Say we want to recommend how much a user will like an item, we then will calculate the score using the recommendations of other users for this certain item. We can distinguish between two different ways of a cold start problem now. First, if there is a new item that has not been rated yet, we cannot give any recommendation. Also, when there is a new user, we cannot calculate a similarity to any other user.

**97. Possible approaches to solving the cold start problem? ‍**

\* Content-based filtering incorporates features about items to calculate a similarity between them. In this way, we can recommend items that have a high similarity to items that a user liked already. In this way, we are not dependant on the ratings of other users for a given item anymore and solve the cold start problem for new items.

\* Demographic filtering incorporates user profiles to calculate a similarity between them and solves the cold start problem for new users.

\*\*Which models do you know for solving time series problems? ‍⭐️\*\*

\* Simple Exponential Smoothing: approximate the time series with an exponentional function

\* Trend-Corrected Exponential Smoothing (Holt‘s Method): exponential smoothing that also models the trend

\* Trend- and Seasonality-Corrected Exponential Smoothing (Holt-Winter‘s Method): exponential smoothing that also models trend and seasonality

\* Time Series Decomposition: decomposed a time series into the four components trend, seasonal variation, cycling varation and irregular component

\* Autoregressive models: similar to multiple linear regression, except that the dependent variable y\_t depends on its own previous values rather than other independent variables.

\* Deep learning approaches (RNN, LSTM, etc.)

1. What is Machine learning?

Machine learning is a branch of computer science which deals with system programming in order to automatically learn and improve with experience. For example: Robots are programed so that they can perform the task based on data they gather from sensors. It automatically learns programs from data.

1. Mention the difference between Data Mining and Machine learning?

Machine learning relates with the study, design and development of the algorithms that give computers the capability to learn without being explicitly programmed. While, data mining can be defined as the process in which the unstructured data tries to extract knowledge or unknown interesting patterns. During this process machine, learning algorithms are used.

1. What is ‘Overfitting’ in Machine learning?

In machine learning, when a statistical model describes random error or noise instead of underlying relationship ‘overfitting’ occurs. When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.

1. Why overfitting happens?

The possibility of overfitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.

1. How can you avoid overfitting ?

By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it. But if you have a small database and you are forced to come with a model based on that. In such situation, you can use a technique known as cross validation. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model. In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to “test” the model in the training phase.

6) What is inductive machine learning?

The inductive machine learning involves the process of learning by examples, where a system, from a set of observed instances tries to induce a general rule.

7) What are the five popular algorithms of Machine Learning?

a) Decision Trees b) Neural Networks (back propagation) c) Probabilistic networks d) Nearest Neighbor e) Support vector machines

8) What are the different Algorithm techniques in Machine Learning?

The different types of techniques in Machine Learning are a) Supervised Learning b) Unsupervised Learning c) Semi-supervised Learning d) Reinforcement Learning e) Transduction f) Learning to Learn

9) What are the three stages to build the hypotheses or model in machine learning?

a) Model building b) Model testing c) Applying the model

10) What is the standard approach to supervised learning?

The standard approach to supervised learning is to split the set of example into the training set and the test.

11) What is ‘Training set’ and ‘Test set’?

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as ‘Training Set’. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

12) List down various approaches for machine learning?

The different approaches in Machine Learning are a) Concept Vs Classification Learning b) Symbolic Vs Statistical Learning c) Inductive Vs Analytical Learning

13) What is not Machine Learning?

a) Artificial Intelligence b) Rule based inference

14) Explain what is the function of ‘Unsupervised Learning’?

a) Find clusters of the data b) Find low-dimensional representations of the data c) Find interesting directions in data d) Interesting coordinates and correlations e) Find novel observations/ database cleaning

15) Explain what is the function of ‘Supervised Learning’?

a) Classifications b) Speech recognition c) Regression d) Predict time series e) Annotate strings

16) What is algorithm independent machine learning?

Machine learning in where mathematical foundations is independent of any particular classifier or learning algorithm is referred as algorithm independent machine learning?

17) What is the difference between artificial learning and machine learning?

Designing and developing algorithms according to the behaviours based on empirical data are known as Machine Learning. While artificial intelligence in addition to machine learning, it also covers other aspects like knowledge representation, natural language processing, planning, robotics etc.

18) What is classifier in machine learning?

A classifier in a Machine Learning is a system that inputs a vector of discrete or continuous feature values and outputs a single discrete value, the class.

19) What are the advantages of Naive Bayes?

In Naïve Bayes classifier will converge quicker than discriminative models like logistic regression, so you need less training data. The main advantage is that it can’t learn interactions between features.

20) In what areas Pattern Recognition is used?

Pattern Recognition can be used in a) Computer Vision b) Speech Recognition c) Data Mining d) Statistics e) Informal Retrieval f) Bio-Informatics

21) What is Genetic Programming?

Genetic programming is one of the two techniques used in machine learning. The model is based on the testing and selecting the best choice among a set of results.

22) What is Inductive Logic Programming in Machine Learning?

Inductive Logic Programming (ILP) is a subfield of machine learning which uses logical programming representing background knowledge and examples.

23) What is Model Selection in Machine Learning?

The process of selecting models among different mathematical models, which are used to describe the same data set is known as Model Selection. Model selection is applied to the fields of statistics, machine learning and data mining.

24) What are the two methods used for the calibration in Supervised Learning?

The two methods used for predicting good probabilities in Supervised Learning are a) Platt Calibration b) Isotonic Regression These methods are designed for binary classification, and it is not trivial.

25) Which method is frequently used to prevent overfitting?

When there is sufficient data ‘Isotonic Regression’ is used to prevent an overfitting issue.

26) What is the difference between heuristic for rule learning and heuristics for decision trees?

The difference is that the heuristics for decision trees evaluate the average quality of a number of disjointed sets while rule learners only evaluate the quality of the set of instances that is covered with the candidate rule.

27) What is Perceptron in Machine Learning?

In Machine Learning, Perceptron is an algorithm for supervised classification of the input into one of several possible non-binary outputs.

28) Explain the two components of Bayesian logic program?

Bayesian logic program consists of two components. The first component is a logical one ; it consists of a set of Bayesian Clauses, which captures the qualitative structure of the domain. The second component is a quantitative one, it encodes the quantitative information about the domain.

29) What are Bayesian Networks (BN) ?

Bayesian Network is used to represent the graphical model for probability relationship among a set of variables .

30) Why instance based learning algorithm sometimes referred as Lazy learning algorithm?

Instance based learning algorithm is also referred as Lazy learning algorithm as they delay the induction or generalization process until classification is performed.

31) What are the two classification methods that SVM ( Support Vector Machine) can handle?

a) Combining binary classifiers b) Modifying binary to incorporate multiclass learning

32) What is ensemble learning?

To solve a particular computational program, multiple models such as classifiers or experts are strategically generated and combined. This process is known as ensemble learning.

33) Why ensemble learning is used?

Ensemble learning is used to improve the classification, prediction, function approximation etc of a model.

34) When to use ensemble learning?

Ensemble learning is used when you build component classifiers that are more accurate and independent from each other.

35) What are the two paradigms of ensemble methods?

The two paradigms of ensemble methods are a) Sequential ensemble methods b) Parallel ensemble methods

36) What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. Bagging is a method in ensemble for improving unstable estimation or classification schemes. While boosting method are used sequentially to reduce the bias of the combined model. Boosting and Bagging both can reduce errors by reducing the variance term.

37) What is bias-variance decomposition of classification error in ensemble method?

The expected error of a learning algorithm can be decomposed into bias and variance. A bias term measures how closely the average classifier produced by the learning algorithm matches the target function. The variance term measures how much the learning algorithm’s prediction fluctuates for different training sets.

38) What is an Incremental Learning algorithm in ensemble?

Incremental learning method is the ability of an algorithm to learn from new data that may be available after classifier has already been generated from already available dataset.

39) What is PCA, KPCA and ICA used for?

PCA (Principal Components Analysis), KPCA ( Kernel based Principal Component Analysis) and ICA ( Independent Component Analysis) are important feature extraction techniques used for dimensionality reduction.

40) What is dimension reduction in Machine Learning?

In Machine Learning and statistics, dimension reduction is the process of reducing the number of random variables under considerations and can be divided into feature selection and feature extraction.

41) What are support vector machines?

Support vector machines are supervised learning algorithms used for classification and regression analysis.

42) What are the components of relational evaluation techniques?

The important components of relational evaluation techniques are a) Data Acquisition b) Ground Truth Acquisition c) Cross Validation Technique d) Query Type e) Scoring Metric f) Significance Test

43) What are the different methods for Sequential Supervised Learning?

The different methods to solve Sequential Supervised Learning problems are a) Sliding-window methods b) Recurrent sliding windows c) Hidden Markow models d) Maximum entropy Markow models e) Conditional random fields f) Graph transformer networks

44) What are the areas in robotics and information processing where sequential prediction problem arises?

The areas in robotics and information processing where sequential prediction problem arises are a) Imitation Learning b) Structured prediction c) Model based reinforcement learning

45) What is batch statistical learning?

Statistical learning techniques allow learning a function or predictor from a set of observed data that can make predictions about unseen or future data. These techniques provide guarantees on the performance of the learned predictor on the future unseen data based on a statistical assumption on the data generating process.

46) What is PAC Learning? P

AC (Probably Approximately Correct) learning is a learning framework that has been introduced to analyze learning algorithms and their statistical efficiency.

47) What are the different categories you can categorized the sequence learning process?

a) Sequence prediction b) Sequence generation c) Sequence recognition d) Sequential decision

48) What is sequence learning?

Sequence learning is a method of teaching and learning in a logical manner.

49) What are two techniques of Machine Learning ?

The two techniques of Machine Learning are a) Genetic Programming b) Inductive Learning

50) Give a popular application of machine learning that you see on day to day basis?

The recommendation engine implemented by major ecommerce websites uses Machine Learning

**1. Why do you use feature selection?**

Feature selection is the process of selecting a subset of relevant features for use in model construction. Feature selection is itself useful, but it mostly acts as a filter, muting out features that aren’t useful in addition to your existing features.

Feature selection methods aid you in your mission to create an accurate predictive model. They help you by choosing features that will give you as good or better accuracy whilst requiring less data.

Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model.

Fewer attributes is desirable because it reduces the complexity of the model, and a simpler model is simpler to understand and explain.

**#### Filter Methods**

Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable.

Some examples of some filter methods include the Chi squared test, information gain and correlation coefficient scores.

**#### Embedded Methods**

Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods.

Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients).

Examples of regularization algorithms are the LASSO, Elastic Net and Ridge Regression.

**#### Misleading**

Including redundant attributes can be misleading to modeling algorithms. Instance-based methods such as k-nearest neighbor use small neighborhoods in the attribute space to determine classification and regression predictions. These predictions can be greatly skewed by redundant attributes.

**#### Overfitting**

Keeping irrelevant attributes in your dataset can result in overfitting. Decision tree algorithms like C4.5 seek to make optimal spits in attribute values. Those attributes that are more correlated with the prediction are split on first. Deeper in the tree less relevant and irrelevant attributes are used to make prediction decisions that may only be beneficial by chance in the training dataset. This overfitting of the training data can negatively affect the modeling power of the method and cripple the predictive accuracy.

**## 27. What is a neural network?**

Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output as shown in the graphic below.

Although there are many different kinds of learning rules used by neural networks, this demonstration is concerned only with one: the delta rule. The delta rule is often utilized by the most common class of ANNs called 'backpropagation neural networks' (BPNNs). Backpropagation is an abbreviation for the backwards propagation of error. With the delta rule, as with other types of back propagation, 'learning' is a supervised process that occurs with each cycle or 'epoch' (i.e. each time the network is presented with a new input pattern) through a forward activation flow of outputs, and the backwards error propagation of weight adjustments. More simply, when a neural network is initially presented with a pattern it makes a random 'guess' as to what it might be. It then sees how far its answer was from the actual one and makes an appropriate adjustment to its connection weights. More graphically, the process looks something like this:

Backpropagation performs a gradient descent within the solution's vector space towards a 'global minimum' along the steepest vector of the error surface. The global minimum is that theoretical solution with the lowest possible error. The error surface itself is a hyperparaboloid but is seldom 'smooth'. Indeed, in most problems, the solution space is quite irregular with numerous 'pits' and 'hills' which may cause the network to settle down in a 'local minimum' which is not the best overall solution.

Since the nature of the error space can not be known a priori, neural network analysis often requires a large number of individual runs to determine the best solution. Most learning rules have built-in mathematical terms to assist in this process which control the 'speed' (Beta-coefficient) and the 'momentum' of the learning. The speed of learning is actually the rate of convergence between the current solution and the global minimum. Momentum helps the network to overcome obstacles (local minima) in the error surface and settle down at or near the global minimum.

Once a neural network is 'trained' to a satisfactory level it may be used as an analytical tool on other data. To do this, the user no longer specifies any training runs and instead allows the network to work in forward propagation mode only. New inputs are presented to the input pattern where they filter into and are processed by the middle layers as though training were taking place, however, at this point the output is retained and no backpropagation occurs. The output of a forward propagation run is the predicted model for the data which can then be used for further analysis and interpretation.

**## 28. How do you deal with sparse data?**

We could take a look at L1 regularization since it best fits the sparse data and does feature selection. If linear relationship - linear regression either - svm. Also it would be nice to use one-hot-encoding or bag-of-words. A one hot encoding is a representation of categorical variables as binary vectors.This first requires that the categorical values be mapped to integer values.Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

**## 30. Pseudo Labeling**

Pseudo-labeling is a technique that allows you to use predicted with \*\*confidence\*\* test data in your training process. This effectivey works by allowing your model to look at more samples, possibly varying in distributions.

##**Knowledge Distillation**

It is the process by which a considerably larger model is able to transfer its knowledge to a smaller one. Applications include NLP and object detection allowing for less powerful hardware to make good inferences without significant loss of accuracy.

Example: model compression which is used to compress the knowledge of multiple models into a single neural network.

#### Q77 What is the advantage of performing dimensionality reduction before fitting an SVM?

Support Vector Machine Learning Algorithm performs better in the reduced space. It is beneficial to perform dimensionality reduction before fitting an SVM if the number of features is large when compared to the number of observations.

#### Q78 How will you assess the statistical significance of an insight whether it is a real insight or just by chance?

Statistical importance of an insight can be accessed using Hypothesis Testing.

#### Q82 Explain how a ROC curve works.

The ROC curve is a graphical representation of the contrast between true positive rates and the false positive rate at various thresholds. It’s often used as a proxy for the trade-off between the sensitivity of the model (true positives) vs the fall-out or the probability it will trigger a false alarm (false positives).

#### Q83 Define precision and recall.

Recall is also known as the true positive rate: the amount of positives your model claims compared to the actual number of positives there are throughout the data. Precision is also known as the positive predictive value, and it is a measure of the amount of accurate positives your model claims compared to the number of positives it actually claims. It can be easier to think of recall and precision in the context of a case where you’ve predicted that there were 10 apples and 5 oranges in a case of 10 apples. You’d have perfect recall (there are actually 10 apples, and you predicted there would be 10) but 66.7% precision because out of the 15 events you predicted, only 10 (the apples) are correct.

#### Q84 What is Bayes’ Theorem? How is it useful in a machine learning context?

Bayes’ Theorem gives you the posterior probability of an event given what is known as prior knowledge. Mathematically, it’s expressed as the true positive rate of a condition sample divided by the sum of the false positive rate of the population and the true positive rate of a condition. Say you had a 60% chance of actually having the flu after a flu test, but out of people who had the flu, the test will be false 50% of the time, and the overall population only has a 5% chance of having the flu.Would you actually have a 60% chance of having the flu after having a positive test?

Bayes’ Theorem says no. It says that you have a (.6 \* 0.05) (True Positive Rate of a Condition Sample) / (.6\*0.05)(True Positive Rate of a Condition Sample) + (.5\*0.95) (False Positive Rate of a Population) = 0.0594 or 5.94% chance of getting a flu.

Bayes’ Theorem is the basis behind a branch of machine learning that most notably includes theNaive Bayes classifier

#### Q87 What’s your favorite algorithm, and can you explain it to me in less than a minute?

This type of question tests your understanding of how to communicate complex and technical nuances with poise and the ability to summarize quickly and efficiently. Make sure you have a choice and make sure you can explain different algorithms so simply and effectively that a five-year-old could grasp the basics!

#### Q88 What’s the difference between Type I and Type II error?

Type I error is a false positive, while Type II error is a false negative. Briefly stated, Type I error means claiming something has happened when it hasn’t, while Type II error means that you claim nothing is happening when in fact something is.

A clever way to think about this is to think of Type I error as telling a man he is pregnant, while Type II error means you tell a pregnant woman she isn’t carrying a baby.

#### Q89 What’s a Fourier transform?

A Fourier transform is a generic method to decompose generic functions into a superposition of symmetric functions. Or as this more intuitive tutorial puts it, given a smoothie, it’s how we find the recipe. The Fourier transform finds the set of cycle speeds, amplitudes and phases to match any time signal. A Fourier transform converts a signal from time to frequency domain — it’s a very common way to extract features from audio signals or other time series such as sensor data.

#### Q91 What is deep learning, and how does it contrast with other machine learning algorithms?

Deep learning is a subset of machine learning that is concerned with neural networks: how to use backpropagation and certain principles from neuroscience to more accurately model large sets of unlabelled or semi-structured data. In that sense, deep learning represents an unsupervised learningalgorithm that learns representations of data through the use of neural nets.

#### Q92 What’s the difference between a generative and discriminative model?

A generative model will learn categories of data while a discriminative model will simply learn the distinction between different categories of data. Discriminative models will generally outperform generative models on classification tasks.

#### Q93 What cross-validation technique would you use on a time series dataset?

Instead of using standard k-folds cross-validation, you have to pay attention to the fact that a time series is not randomly distributed data — it is inherently ordered by chronological order. If a pattern emerges in later time periods for example, your model may still pick up on it even if that effect doesn’t hold in earlier years! You’ll want to do something like forward chaining where you’ll be able to model on past data then look at forward-facing data.

fold 1 : training [1], test [2]

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

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

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

fold 5 : training [1 2 3 4 5], test [6]

#### Q94 How is a decision tree pruned?

Pruning is what happens in decision trees when branches that have weak predictive power are removed in order to reduce the complexity of the model and increase the predictive accuracy of a decision tree model. Pruning can happen bottom-up and top-down, with approaches such as reduced error pruning and cost complexity pruning. Reduced error pruning is perhaps the simplest version: replace each node. If it doesn’t decrease predictive accuracy, keep it pruned. While simple, this heuristic actually comes pretty close to an approach that would optimize for maximum accuracy.

#### Q95 Which is more important to you– model accuracy, or model performance.

Well, it has everything to do with how model accuracy is only a subset of model performance, and at that, a sometimes misleading one. For example, if you wanted to detect fraud in a massive dataset with a sample of millions, a more accurate model would most likely predict no fraud at all if only a vast minority of cases were fraud. However, this would be useless for a predictive model — a model designed to find fraud that asserted there was no fraud at all! Questions like this help you demonstrate that you understand model accuracy isn’t the be-all and end-all of model performance.

#### Q96 What’s the F1 score? How would you use it?

The F1 score is a measure of a model’s performance. It is a weighted average of the precision and recall of a model, with results tending to 1 being the best, and those tending to 0 being the worst. You would use it in classification tests where true negatives don’t matter much.

#### Q97 How would you handle an imbalanced dataset?

An imbalanced dataset is when you have, for example, a classification test and 90% of the data is in one class. That leads to problems: an accuracy of 90% can be skewed if you have no predictive power on the other category of data! Here are a few tactics to get over the hump:

1- Collect more data to even the imbalances in the dataset. 2- Resample the dataset to correct for imbalances. 3- Try a different algorithm altogether on your dataset.

What’s important here is that you have a keen sense for what damage an unbalanced dataset can cause, and how to balance that.

#### Q98 When should you use classification over regression?

Classification produces discrete values and dataset to strict categories, while regression gives you continuous results that allow you to better distinguish differences between individual points. You would use classification over regression if you wanted your results to reflect the belongingness of data points in your dataset to certain explicit categories (ex: If you wanted to know whether a name was male or female rather than just how correlated they were with male and female names.)

#### Q99 Name an example where ensemble techniques might be useful.

Ensemble techniques use a combination of learning algorithms to optimize better predictive performance. They typically reduce overfitting in models and make the model more robust (unlikely to be influenced by small changes in the training data). You could list some examples of ensemble methods, from bagging to boosting to a “bucket of models” method and demonstrate how they could increase predictive power.

#### Q101 What evaluation approaches would you work to gauge the effectiveness of a machine learning model?

You would first split the dataset into training and test sets, or perhaps use cross-validation techniques to further segment the dataset into composite sets of training and test sets within the data. You should then implement a choice selection of performance metrics: here is a fairly comprehensive list. You could use measures such as the F1 score, the accuracy, and the confusion matrix. What’s important here is to demonstrate that you understand the nuances of how a model is measured and how to choose the right performance measures for the right situations.

#### Q103 What’s the “kernel trick” and how is it useful?

The Kernel trick involves kernel functions that can enable in higher-dimension spaces without explicitly calculating the coordinates of points within that dimension: instead, kernel functions compute the inner products between the images of all pairs of data in a feature space. This allows them the very useful attribute of calculating the coordinates of higher dimensions while being computationally cheaper thanthe explicit calculation of said coordinates. Many algorithms can be expressed in terms of inner products.Using the kernel trick enables us effectively run algorithms in a high-dimensional space with lower-dimensional data.

# ## Machine Learning Interview Questions: Company/Industry Specific

#### Q110 How would you implement a recommendation system for our company’s users?

A lot of machine learning interview questions of this type will involve implementation of machine learning models to a company’s problems. You’ll have to research the company and its industry in-depth, especially the revenue drivers the company has, and the types of users the company takes on in the context of the industry it’s in.

#### Q111 How can we use your machine learning skills to generate revenue?

This is a tricky question. The ideal answer would demonstrate knowledge of what drives the business and how your skills could relate. For example, if you were interviewing for music-streaming startup Spotify, you could remark that your skills at developing a better recommendation model would increase user retention, which would then increase revenue in the long run.The startup metrics Slideshare linked above will help you understand exactly what performance indicators are important for startups and tech companies as they think about revenue and growth.

#### Q112 What do you think of our current data process?

This kind of question requires you to listen carefully and impart feedback in a manner that is constructive and insightful. Your interviewer is trying to gauge if you’d be a valuable member of their team and whether you grasp the nuances of why certain things are set the way they are in the company’s data process based on company- or industry-specific conditions. They’re trying to see if you can be an intellectual peer. Act accordingly.

#### Q116 How would you approach the “Netflix Prize” competition?

The Netflix Prize was a famed competition where Netflix offered $1,000,000 for a better collaborative filtering algorithm. The team that won called BellKor had a 10% improvement and used an ensemble of different methods to win. Some familiarity with the case and its solution will help demonstrate you’ve paid attention to machine learning for a while.

#### Q117 Where do you usually source datasets?

Machine learning interview questions like these try to get at the heart of your machine learning interest. Somebody who is truly passionate about machine learning will have gone off and done side projects on their own, and have a good idea of what great datasets are out there. If you’re missing any, check out Quandl for economic and financial data, and Kaggle’s Datasets collection for another great list.

#### Q118 How do you think Google is training data for self-driving cars?

Machine learning interview questions like this one really test your knowledge of different machine learning methods, and your inventiveness if you don’t know the answer. Google is currently using recaptcha to source labelled data on storefronts and traffic signs. They are also building on training data collected by Sebastian Thrun at GoogleX — some of which was obtained by his grad students driving buggies on desert dunes!

#### Q119 How would you simulate the approach AlphaGo took to beat Lee Sidol at Go?

AlphaGo beating Lee Sidol, the best human player at Go, in a best-of-five series was a truly seminal event in the history of machine learning and deep learning. The Nature paper above describes how this was accomplished with “Monte-Carlo tree search with deep neural networks that have been trained by supervised learning, from human expert games, and by reinforcement learning from games of self-play.”

### ****1. Explain the terms Artificial Intelligence (AI), Machine Learning (ML and Deep Learning?****

[Artificial Intelligence](https://www.mygreatlearning.com/blog/what-is-artificial-intelligence/) (AI) is the domain of producing intelligent machines. ML refers to systems that can assimilate from experience (training data) and Deep Learning (DL) states to systems that learn from experience on large data sets. ML can be considered as a subset of AI. Deep Learning (DL) is ML but useful to large data sets. In summary, DL is a subset of ML & both were the subsets of AI. Additional Information: ASR (Automatic Speech Recognition) & [NLP](https://www.mygreatlearning.com/blog/natural-language-processing-tutorial/) (Natural Language Processing) fall under AI and overlay with ML & DL as ML is often utilized for NLP and ASR tasks.

### ****2. What are the different types of Learning/ Training models in ML?****

ML algorithms can be primarily classified depending on the presence/absence of target variables.

**A. Supervised learning:** [Target is present]  
The machine learns using labelled data. The model is trained on an existing data set before it starts making decisions with the new data.  
The target variable is continuous: Linear Regression, polynomial Regression, quadratic Regression.  
The target variable is categorical: Logistic regression, Naive Bayes, KNN, SVM, Decision Tree, Gradient Boosting, ADA boosting, Bagging, Random forest etc.

***B. Unsupervised learning:*** [Target is absent]  
The machine is trained on unlabelled data and without any proper guidance. It automatically infers patterns and relationships in the data by creating clusters. The model learns through observations and deduced structures in the data.  
Principal component Analysis, Factor analysis, Singular Value Decomposition etc.

***C. Reinforcement Learning:***  
The model learns through a trial and error method. This kind of learning involves an agent that will interact with the environment to create actions and then discover errors or rewards of that action.

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

Machine Learning involves algorithms that learn from patterns of data and then apply it to decision making. Deep Learning, on the other hand, is able to learn through processing data on its own and is quite similar to the human brain where it identifies something, analyse it, and makes a decision.  
The key differences are as follow:

* The manner in which data is presented to the system.
* Machine learning algorithms always require structured data and deep learning networks rely on layers of artificial neural networks.

### ****4. What is the main key difference between supervised and unsupervised machine learning?****

Supervised learning technique needs labeled data to train the model. For example, to solve a classification problem (a supervised learning task), you need to have label data to train the model and to classify the data into your labeled groups. Unsupervised learning does not  need any labelled dataset. This is the main key difference between supervised learning and unsupervised learning.

### 5. How do you select important variables while working on a data set?

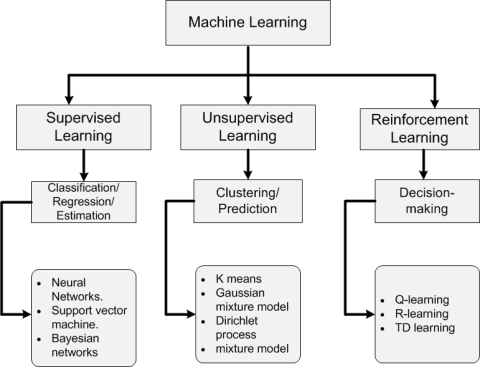
There are various means to select important variables from a data set that include the following:

* Identify and discard correlated variables before finalizing on important variables
* The variables could be selected based on ‘p’ values from Linear Regression
* Forward, Backward, and Stepwise selection
* Lasso Regression
* Random Forest and plot variable chart
* Top features can be selected based on information gain for the available set of features.

### ****6. There are many machine learning algorithms till now. If given a data set, how can one determine which algorithm to be used for that?****

Machine Learning algorithm to be used purely depends on the type of data in a given dataset. If data is linear then, we use linear regression. If data shows non-linearity then, the bagging algorithm would do better. If the data is to be analyzed/interpreted for some business purposes then we can use decision trees or SVM. If the dataset consists of images, videos, audios then, neural networks would be helpful to get the solution accurately.

So, there is no certain metric to decide which algorithm to be used for a given situation or a data set. We need to explore the data using EDA (Exploratory Data Analysis) and understand the purpose of using the dataset to come up with the best fit algorithm. So, it is important to study all the algorithms in detail.



### 7. How are covariance and correlation different from one another?

Covariance measures how two variables are related to each other and how one would vary with respect to changes in the other variable. If the value is positive it means there is a direct relationship between the variables and one would increase or decrease with an increase or decrease in the base variable respectively, given that all other conditions remain constant.

Correlation quantifies the relationship between two random variables and has only three specific values, i.e., 1, 0, and -1.

1 denotes a positive relationship, -1 denotes a negative relationship, and 0 denotes that the two variables are independent of each other.

### ****8. State the differences between causality and correlation?****

Causality applies to situations where one action, say X, causes an outcome, say Y, whereas Correlation is just relating one action (X) to another action(Y) but X does not necessarily cause Y.

### ****9. We look at machine learning software almost all the time. How do we apply Machine Learning to Hardware?****

We have to build ML algorithms in System Verilog which is a Hardware development Language and then program it onto an FPGA to apply Machine Learning to hardware.

### ****10. Explain One-hot encoding and Label Encoding. How do they affect the dimensionality of the given dataset?****

One-hot encoding is the representation of categorical variables as binary vectors. Label Encoding is converting labels/words into numeric form. Using one-hot encoding increases the dimensionality of the data set. Label encoding doesn’t affect the dimensionality of the data set. One-hot encoding creates a new variable for each level in the variable whereas, in Label encoding, the levels of a variable get encoded as 1 and 0.

**Deep Learning Interview Questions**

Deep Learning is a part of machine learning that works with neural networks. It involves a hierarchical structure of networks that set up a process to help machines learn the human logics behind any action. We have compiled a list of the frequently asked deep leaning interview questions to help you prepare.

**What is overfitting?**

Overfitting is a type of modelling error which results in the failure to predict future observations effectively or fit additional data in the existing model. It occurs when a function is too closely fit to a limited set of data points and usually ends with more parameters [**read more…**](https://www.mygreatlearning.com/blog/deep-learning-interview-questions/)

**What is Multilayer Perceptron and Boltzmann Machine?**

The Boltzmann machine is a simplified version of the multilayer perceptron. This is a two layer model with a visible input layer and a hidden layer which makes stochastic decisions for the [**read more…**](https://www.mygreatlearning.com/blog/deep-learning-interview-questions/)

### 11. When does regularization come into play in Machine Learning?

At times when the model begins to underfit or overfit, regularization becomes necessary. It is a regression that diverts or regularizes the coefficient estimates towards zero. It reduces flexibility and discourages learning in a model to avoid the risk of overfitting. The model complexity is reduced and it becomes better at predicting.

### ****12. What is Bias, Variance and what do you mean by Bias-Variance Tradeoff?****

Both are errors in Machine Learning Algorithms. When the algorithm has limited flexibility to deduce the correct observation from the dataset, it results in bias. On the other hand, variance occurs when the model is extremely sensitive to small fluctuations.

If one adds more features while building a model, it will add more complexity and we will lose bias but gain some variance. In order to maintain the optimal amount of error, we perform a tradeoff between bias and variance based on the needs of a business.

Bias stands for the error because of the erroneous or overly simplistic assumptions in the learning algorithm . This  assumption can lead to the model underfitting the data, making it hard for it to have high predictive accuracy and for you to generalize your knowledge from the training set to the test set.

Variance is also an error because of  too much complexity in the learning algorithm. This can be the reason for the algorithm being highly sensitive to high degrees of variation in training data, which can lead your model to overfit the data. Carrying too much noise from the training data for your model to be very useful for your test data.

The bias-variance decomposition essentially decomposes the learning error from any algorithm by adding the bias, the variance and a bit of irreducible error due to noise in the underlying dataset. Essentially, if you make the model more complex and add more variables, you’ll lose bias but gain some variance — in order to get the optimally reduced amount of error, you’ll have to trade off bias and variance. You don’t want either high bias or high variance in your model.

### ****13. How can we relate standard deviation and variance?****

*Standard deviation* refers to the spread of your data from the mean. *Variance* is the average degree to which each point differs from the mean i.e. the average of all data points. We can relate Standard deviation and Variance because it is the square root of Variance.

### ****14. A data set is given to you and it has missing values which spread along 1standard deviation from the mean. How much of the data would remain untouched?****

It is given that the data is spread across mean that is the data is spread across an average. So, we can presume that it is a normal distribution. In a normal distribution, about 68% of data lies in 1 standard deviation from averages like mean, mode or median. That means about 32% of the data remains uninfluenced by missing values.

### ****15. Is a high variance in data good or bad?****

Higher variance directly means that the data spread is big and the feature has a variety of data. Usually, high variance in a feature is seen as not so good quality.

### ****16. If your dataset is suffering from high variance, how would you handle it?****

For datasets with high variance, we could use the bagging algorithm to handle it. Bagging algorithm splits the data into subgroups with sampling replicated from random data. After the data is split, random data is used to create rules using a training algorithm. Then we use polling technique to combine all the predicted outcomes of the model.

### ****17. A data set is given to you about utilities fraud detection. You have built aclassifier model and achieved a performance score of 98.5%. Is this a goodmodel? If yes, justify. If not, what can you do about it?****

Data set about utilities fraud detection is not balanced enough i.e. imbalanced. In such a data set, accuracy score cannot be the measure of performance as it may only be predict the majority class label correctly but in this case our point of interest is to predict the minority label. But often minorities are treated as noise and ignored. So, there is a high probability of misclassification of the minority label as compared to the majority label. For evaluating the model performance in case of imbalanced data sets, we should use Sensitivity (True Positive rate) or Specificity (True Negative rate) to determine class label wise performance of the classification model. If the minority class label’s performance is not so good, we could do the following:

1. We can use under sampling or over sampling to balance the data.
2. We can change the prediction threshold value.
3. We can assign weights to labels such that the minority class labels get larger weights.
4. We could detect anomalies.

### ****18. Explain the handling of missing or corrupted values in the given dataset.****

An easy way to handle missing values or corrupted values is to drop the corresponding rows or columns. If there are too many rows or columns to drop then we consider replacing the missing or corrupted values with some new value.

Identifying missing values and dropping the rows or columns can be done by using IsNull() and dropna( ) functions in Pandas. Also, the Fillna() function in Pandas replaces the incorrect values with the placeholder value.

### ****19. What is Time series?****

A Time series is a sequence of numerical data points in successive order. It tracks the movement of the chosen data points, over a specified period of time and records the data points at regular intervals. Time series doesn’t require any minimum or maximum time input. Analysts often use Time series to examine data according to their specific requirement.

### ****20. What is a Box-Cox transformation?****

Box-Cox transformation is a power transform which transforms non-normal dependent variables into normal variables as normality is the most common assumption made while using many statistical techniques. It has a lambda parameter which when set to 0 implies that this transform is equivalent to log-transform. It is used for variance stabilization and also to normalize the distribution.

### 21. What is the difference between stochastic gradient descent (SGD) and gradient descent (GD)?

Gradient Descent and Stochastic Gradient Descent are the algorithms that find the set of parameters that will minimize a loss function.  
The difference is that in Gradient Descend, all training samples are evaluated for each set of parameters. While in Stochastic Gradient Descent only one training sample is evaluated for the set of parameters identified.

### ****22. What is the exploding gradient problem while using back propagation technique?****

When large error gradients accumulate and result in large changes in the neural network weights during training, it is called the exploding gradient problem. The values of weights can become so large as to overflow and result in NaN values. This makes the model unstable and the learning of the model to stall just like the vanishing gradient problem.

### 23. Can you mention some advantages and disadvantages of decision trees?

The advantages of decision trees are that they are easier to interpret, are nonparametric and hence robust to outliers, and have relatively few parameters to tune.  
On the other hand, the disadvantage is that they are prone to overfitting.

### ****24. Explain the differences between Random Forest and Gradient Boosting machines.****

Random forests are a significant number of decision trees pooled using averages or majority rules at the end. Gradient boosting machines also combine decision trees but at the beginning of the process unlike Random forests. Random forest creates each tree independent of the others while gradient boosting develops one tree at a time. Gradient boosting yields better outcomes than random forests if parameters are carefully tuned but it’s not a good option if the data set contains a lot of outliers/anomalies/noise as it can result in overfitting of the model.Random forests perform well for [multiclass object detection](http://www.svcl.ucsd.edu/publications/conference/2014/nips2014.pdf). Gradient Boosting performs well when there is data which is not balanced such as in [real time risk assessment.](https://www.academia.edu/7707785/Application_of_Stochastic_Gradient_Boosting_SGB_Technique_to_Enhance_the_Reliability_of_Real-Time_Risk_Assessment_Using_AVI_and_RTMS_Data)

### ****25. What is a confusion matrix and why do you need it?****

Confusion matrix (also called the error matrix) is a table that is frequently used to illustrate the performance of a classification model i.e. classifier on a set of test data for which the true values are well-known.

It allows us to visualize the performance of an algorithm/model. It allows us to easily identify the confusion between different classes. It is used as a performance measure of a model/algorithm.

A confusion matrix is known as a summary of predictions on a classification model. The number of right and wrong predictions were summarized with count values and broken down by each class label. It gives us information about the errors made through the classifier and also the types of errors made by a classifier.

### ****26. What’s a Fourier transform?****

Fourier Transform is a mathematical technique that transforms any function of time to a function of frequency. Fourier transform is closely related to Fourier series. It takes any time-based pattern for input and calculates the overall cycle offset, rotation speed and strength for all possible cycles. Fourier transform is best applied to waveforms since it has functions of time and space. Once a Fourier transform applied on a waveform, it gets decomposed into a sinusoid.

### ****27. What do you mean by Associative Rule Mining (ARM)?****

Associative Rule Mining is one of the techniques to discover patterns in data like features (dimensions) which occur together and features (dimensions) which are correlated. It is mostly used in Market-based Analysis to find how frequently an itemset occurs in a transaction. Association rules have to satisfy minimum support and minimum confidence at the very same time. Association rule generation generally comprised of two different steps:

* “A min support threshold is given to obtain all frequent item-sets in a database.”
* “A min confidence constraint is given to these frequent item-sets in order to form the association rules.”

Support is a measure of how often the “item set” appears in the data set and Confidence is a measure of how often a particular rule has been found to be true.

### ****28. What is Marginalisation? Explain the process.****

Marginalisation is summing the probability of a random variable X given joint probability distribution of X with other variables. It is an application of the law of total probability.

P(X=x) = ∑YP(X=x,Y)

Given the joint probability P(X=x,Y), we can use marginalization to find P(X=x). So, it is to find distribution of one random variable by exhausting cases on other random variables.

### ****29. Explain the phrase “Curse of Dimensionality”.****

The Curse of Dimensionality refers to the situation when your data has too many features.

The phrase is used to express the difficulty of using brute force or grid search to optimize a function with too many inputs.

It can also refer to several other issues like:

* If we have more features than observations, we have a risk of overfitting the model.
* When we have too many features, observations become harder to cluster. Too many dimensions cause every observation in the dataset to appear equidistant from all others and no meaningful clusters can be formed.

Dimensionality reduction techniques like PCA come to the rescue in such cases.

### 30. What is the Principle Component Analysis?

The idea here is to reduce the dimensionality of the data set by reducing the number of variables that are correlated with each other. Although the variation needs to be retained to the maximum extent.

The variables are transformed into a new set of variables that are known as Principal Components’. These PCs are the eigenvectors of a covariance matrix and therefore are orthogonal.

**NLP Interview Questions**

NLP or Natural Language Processing helps machines analyse natural languages with the intention of learning them. It extracts information from data by applying machine learning algorithms. Apart from learning the basics of NLP, it is important to prepare specifically for the interviews.

**Explain Dependency Parsing in NLP?**

Dependency Parsing, also known as Syntactic parsing in NLP is a process of assigning syntactic structure to a sentence and identifying its dependency parses. This process is crucial to understand the correlations between the “head” words in the syntactic [**read more…**](https://www.mygreatlearning.com/blog/nlp-interview-questions/)

Which of the following architecture can be trained faster and needs less amount of training data

a. LSTM based Language Modelling

b. Transformer architecture

### ****31. Why is rotation of components so important in Principle Component Analysis (PCA)?****

Rotation in PCA is very important as it maximizes the separation within the variance obtained by all the components because of which interpretation of components would become easier. If the components are not rotated, then we need extended components to describe variance of the components.

### 32. What are outliers? Mention three methods to deal with outliers.

A data point that is considerably distant from the other similar data points is known as an outlier. They may occur due to experimental errors or variability in measurement. They are problematic and can mislead a training process, which eventually results in longer training time, inaccurate models, and poor results.

The three methods to deal with outliers are:  
**Univariate method** – looks for data points having extreme values on a single variable  
**Multivariate method** – looks for unusual combinations on all the variables  
**Minkowski error** – reduces the contribution of potential outliers in the training process.

### 33. What is the difference between regularization and normalisation?

Normalisation adjusts the data; regularisation adjusts the prediction function. If your data is on very different scales (especially low to high), you would want to normalise the data. Alter each column to have compatible basic statistics. This can be helpful to make sure there is no loss of accuracy. One of the goals of model training is to identify the signal and ignore the noise if the model is given free rein to minimize error, there is a possibility of suffering from overfitting. Regularization imposes some control on this by providing simpler fitting functions over complex ones.

### ****34. Explain the difference between Normalization and Standardization.****

Normalization and Standardization are the two very popular methods used for feature scaling. Normalization refers to re-scaling the values to fit into a range of [0,1]. Standardization refers to re-scaling data to have a mean of 0 and a standard deviation of 1 (Unit variance). Normalization is useful when all parameters need to have the identical positive scale however the outliers from the data set are lost. Hence, standardization is recommended for most applications.

### 35. List the most popular distribution curves along with scenarios where you will use them in an algorithm.

The most popular distribution curves are as follows- Bernoulli Distribution, Uniform Distribution, Binomial Distribution, Normal Distribution, Poisson Distribution, and Exponential Distribution.  
Each of these distribution curves is used in various scenarios.

Bernoulli Distribution can be used to check if a team will win a championship or not, a newborn child is either male or female, you either pass an exam or not, etc.

**Uniform distribution** is a probability distribution that has a constant probability. Rolling a single dice is one example because it has a fixed number of outcomes.

***Binomial distribution*** is a probability with only two possible outcomes, the prefix ‘bi’ means two or twice. An example of this would be a coin toss. The outcome will either be heads or tails.

**Normal distribution** describes how the values of a variable are distributed. It is typically a symmetric distribution where most of the observations cluster around the central peak. The values further away from the mean taper off equally in both directions. An example would be the height of students in a classroom.

**Poisson distribution** helps predict the probability of certain events happening when you know how often that event has occurred. It can be used by businessmen to make forecasts about the number of customers on certain days and allows them to adjust supply according to the demand.

**Exponential distribution** is concerned with the amount of time until a specific event occurs. For example, how long a car battery would last, in months.

### 36. How do we check the normality of a data set or a feature?

Visually, we can check it using plots. There is a list of Normality checks, they are as follow:

* Shapiro-Wilk W Test
* Anderson-Darling Test
* Martinez-Iglewicz Test
* Kolmogorov-Smirnov Test
* D’Agostino Skewness Test

### ****37. What is****[Linear Regression](https://www.mygreatlearning.com/blog/linear-regression-for-beginners-machine-learning/)****?****

Linear Function can be defined as a Mathematical function on a 2D plane as,  Y =Mx +C, where Y is a dependent variable and X is Independent Variable, C is Intercept and M is slope and same can be expressed as Y is a Function of X or Y = F(x).

At any given value of X, one can compute the value of Y, using the equation of Line. This relation between Y and X, with a degree of the polynomial as 1 is called Linear Regression.

In Predictive Modeling, LR is represented as Y = Bo + B1x1 + B2x2  
The value of B1 and B2 determines the strength of the correlation between features and the dependent variable.

Example: Stock Value in $ = Intercept + (+/-B1)\*(Opening value of Stock) + (+/-B2)\*(Previous Day Highest value of Stock)

### 38. Differentiate between regression and classification.

Regression and classification are categorized under the same umbrella of supervised machine learning. The main difference between them is that the output variable in the regression is numerical (or continuous) while that for classification is categorical (or discrete).

Example: To predict the definite Temperature of a place is Regression problem whereas predicting whether the day will be Sunny cloudy or there will be rain is a case of classification.

### 39. What is target imbalance? How do we fix it? A scenario where you have performed target imbalance on data. Which metrics and algorithms do you find suitable to input this data onto?

If you have categorical variables as the target when you cluster them together or perform a frequency count on them if there are certain categories which are more in number as compared to others by a very significant number. This is known as the target imbalance.

Example: Target column – 0,0,0,1,0,2,0,0,1,1 [0s: 60%, 1: 30%, 2:10%] 0 are in majority. To fix this, we can perform up-sampling or down-sampling. Before fixing this problem let’s assume that the performance metrics used was confusion metrics. After fixing this problem we can shift the metric system to AUC: ROC. Since we added/deleted data [up sampling or downsampling], we can go ahead with a stricter algorithm like SVM, Gradient boosting or ADA boosting.

### 40. List all assumptions for data to be met before starting with linear regression.

Before starting linear regression, the assumptions to be met are as follow:

* Linear relationship
* Multivariate normality
* No or little multicollinearity
* No auto-correlation
* Homoscedasticity

### 41. When does the linear regression line stop rotating or finds an optimal spot where it is fitted on data?

A place where the highest RSquared value is found, is the place where the line comes to rest. RSquared represents the amount of variance captured by the virtual linear regression line with respect to the total variance captured by the dataset.

### 42. Why is logistic regression a type of classification technique and not a regression? Name the function it is derived from?

Since the target column is categorical, it uses linear regression to create an odd function that is wrapped with a log function to use regression as a classifier. Hence, it is a type of classification technique and not a regression. It is derived from cost function.

### ****43. What could be the issue when the beta value for a certain variable varies way too much in each subset when regression is run on different subsets of the given dataset?****

Variations in the beta values in every subset implies that the dataset is heterogeneous. To overcome this problem, we can use a different model for each of the clustered subsets of the dataset or use a non-parametric model such as decision trees.

### ****44. What does the term Variance Inflation Factor mean?****

Variation Inflation Factor (VIF) is the ratio of variance of the model to variance of the model with only one independent variable. VIF gives the estimate of volume of multicollinearity in a set of many regression variables.

VIF = Variance of model Variance of model with one independent variable

### 45. Which machine learning algorithm is known as the lazy learner and why is it called so?

KNN is a Machine Learning algorithm known as a lazy learner. K-NN is a lazy learner because it doesn’t learn any machine learnt values or variables from the training data but dynamically calculates distance every time it wants to classify, hence memorises the training dataset instead.

**Python Interview Questions**

Here’s a list of the top 101 interview questions with answers to help you prepare. The first set of questions and answers are curated for freshers while the second set is designed for advanced users.

**What are functions in Python?**

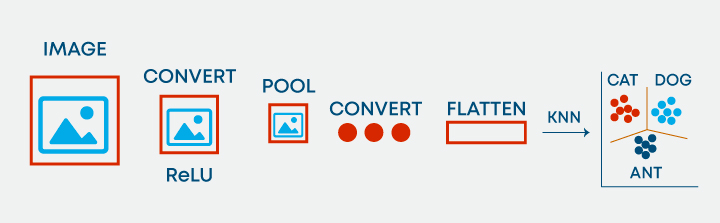
Functions in Python refer to blocks that have organised, and reusable codes to perform single, and related events. Functions are important to create better modularity for applications which reuse high degree of coding. Python has a number of built-in functions[**read more…**](https://www.mygreatlearning.com/blog/python-interview-questions/)

What are dataframes?

A pandas dataframe is a data structure in pandas which is mutable. Pandas has support for heterogeneous data which is arranged across two axes.( rows and columns).

Reading files into pandas:- [**Read more…**](https://www.mygreatlearning.com/blog/python-interview-questions/)

### 46. Is it possible to use KNN for image processing?



Yes, it is possible to use KNN for image processing. It can be done by converting the 3-dimensional image into a single-dimensional vector and using the same as input to KNN.

### ****47. Differentiate between K-Means and KNN algorithms?****

KNN is Supervised Learning where-as K-Means is Unsupervised Learning. With KNN, we predict the label of the unidentified element based on its nearest neighbour and further extend this approach for solving classification/regression-based problems.

K-Means is Unsupervised Learning, where we don’t have any Labels present, in other words, no Target Variables and thus we try to cluster the data based upon their coordinates and try to establish the nature of the cluster based on the elements filtered for that cluster.

### 48. How does the SVM algorithm deal with self-learning?

SVM has a learning rate and expansion rate which takes care of this. The learning rate compensates or penalises the hyperplanes for making all the wrong moves and expansion rate deals with finding the maximum separation area between classes.

### 49. What are Kernels in SVM? List popular kernels used in SVM along with a scenario of their applications.

The function of kernel is to take data as input and transform it into the required form. A few popular Kernels used in SVM are as follows: RBF, Linear, Sigmoid, Polynomial, Hyperbolic, Laplace, etc.

### ****50. What is Kernel Trick in an SVM Algorithm?****

Kernel Trick is a mathematical function which when applied on data points, can find the region of classification between two different classes. Based on the choice of function, be it linear or radial, which purely depends upon the distribution of data, one can build a classifier.

[**Enroll to Machine Learning Course for Free**](https://www.greatlearning.in/academy/learn-for-free/courses/machine-learning-with-python?utm_medium=post&utm_source=glblog&utm_campaign=gla_course)

### 51. What are ensemble models? Explain how ensemble techniques yield better learning as compared to traditional classification ML algorithms?

Ensemble is a group of models that are used together for prediction both in classification and regression class. Ensemble learning helps improve ML results because it combines several models. By doing so, it allows a better predictive performance compared to a single model.   
They are superior to individual models as they reduce variance, average out biases, and have lesser chances of overfitting.

### 52. What are overfitting and underfitting? Why does the decision tree algorithm suffer often with overfitting problem?

Overfitting is a statistical model or machine learning algorithm which captures the noise of the data. Underfitting is a model or machine learning algorithm which does not fit the data well enough and occurs if the model or algorithm shows low variance but high bias.

In decision trees, overfitting occurs when the tree is designed to perfectly fit all samples in the training data set. This results in branches with strict rules or sparse data and affects the accuracy when predicting samples that aren’t part of the training set.

Also Read: [Overfitting and Underfitting in Machine Learning](https://www.mygreatlearning.com/blog/overfitting-and-underfitting-in-machine-learning/)

### 53. What is OOB error and how does it occur?

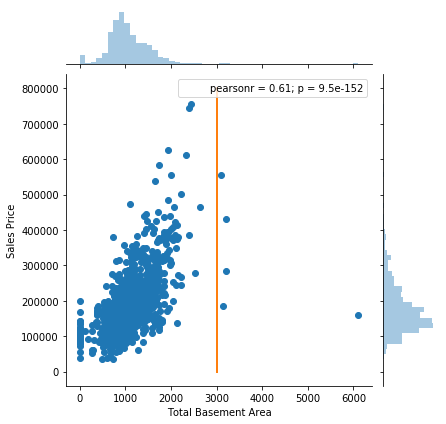
For each bootstrap sample, there is one-third of data that was not used in the creation of the tree, i.e., it was out of the sample. This data is referred to as out of bag data. In order to get an unbiased measure of the accuracy of the model over test data, out of bag error is used. The out of bag data is passed for each tree is passed through that tree and the outputs are aggregated to give out of bag error. This percentage error is quite effective in estimating the error in the testing set and does not require further cross-validation.

### 54. Why boosting is a more stable algorithm as compared to other ensemble algorithms?

Boosting focuses on errors found in previous iterations until they become obsolete. Whereas in bagging there is no corrective loop. This is why boosting is a more stable algorithm compared to other ensemble algorithms.

### ****55. How do you handle outliers in the data?****

Outlier is an observation in the data set that is far away from other observations in the data set. We can discover outliers using tools and functions like box plot, scatter plot, Z-Score, IQR score etc. and then handle them based on the visualization we have got. To handle outliers, we can cap at some threshold, use transformations to reduce skewness of the data and remove outliers if they are anomalies or errors.



### 56. List popular cross validation techniques.

There are mainly six types of cross validation techniques. They are as follow:

* **K fold**
* **Stratified k fold**
* **Leave one out**
* **Bootstrapping**
* **Random search cv**
* **Grid search cv**

### 57. Is it possible to test for the probability of improving model accuracy without cross-validation techniques? If yes, please explain.

Yes, it is possible to test for the probability of improving model accuracy without cross-validation techniques. We can do so by running the ML model for say **n** number of iterations, recording the accuracy. Plot all the accuracies and remove the 5% of low probability values. Measure the left [low] cut off and right [high] cut off. With the remaining 95% confidence, we can say that the model can go as low or as high [as mentioned within cut off points].

### 58. Name a popular dimensionality reduction algorithm.

Popular dimensionality reduction algorithms are Principal Component Analysis and Factor Analysis.  
Principal Component Analysis creates one or more index variables from a larger set of measured variables. Factor Analysis is a model of the measurement of a latent variable. This latent variable cannot be measured with a single variable and is seen through a relationship it causes in a set of**y** variables.

### 59. How can we use a dataset without the target variable into supervised learning algorithms?

Input the data set into a clustering algorithm, generate optimal clusters, label the cluster numbers as the new target variable. Now, the dataset has independent and target variables present. This ensures that the dataset is ready to be used in supervised learning algorithms.

### 60. List all types of popular [recommendation systems](https://www.mygreatlearning.com/blog/masterclass-on-movie-recommendation-system/)? Name and explain two personalized recommendation systems along with their ease of implementation.

Popularity based recommendation, content-based recommendation, user-based collaborative filter, and item-based recommendation are the popular types of recommendation systems.  
Personalised Recommendation systems are- Content-based recommendation, user-based collaborative filter, and item-based recommendation. User-based collaborative filter and item-based recommendations are more personalised. Ease to maintain: Similarity matrix can be maintained easily with Item-based recommendation.

### 61. How do we deal with sparsity issues in recommendation systems? How do we measure its effectiveness? Explain.

Singular value decomposition can be used to generate the prediction matrix. RMSE is the measure that helps us understand how close the prediction matrix is to the original matrix.

### 62. Name and define techniques used to find similarities in the recommendation system.

Pearson correlation and Cosine correlation are techniques used to find similarities in recommendation systems.

### ****63. State the limitations of Fixed Basis Function.****

Linear separability in feature space doesn’t imply linear separability in input space. So, Inputs are non-linearly transformed using vectors of basic functions with increased dimensionality. Limitations of Fixed basis functions are:

1. Non-Linear transformations cannot remove overlap between two classes but they can increase overlap.
2. Often it is not clear which basis functions are the best fit for a given task. So, learning the basic functions can be useful over using fixed basis functions.
3. If we want to use only fixed ones, we can use a lot of them and let the model figure out the best fit but that would lead to overfitting the model thereby making it unstable.

### ****64. Define and explain the concept of Inductive Bias with some examples.****

Inductive Bias is a set of assumptions that humans use to predict outputs given inputs that the learning algorithm has not encountered yet. When we are trying to learn Y from X and the hypothesis space for Y is infinite, we need to reduce the scope by our beliefs/assumptions about the hypothesis space which is also called inductive bias. Through these assumptions, we constrain our hypothesis space and also get the capability to incrementally test and improve on the data using hyper-parameters. Examples:

1. We assume that Y varies linearly with X while applying Linear regression.
2. We assume that there exists a hyperplane separating negative and positive examples.

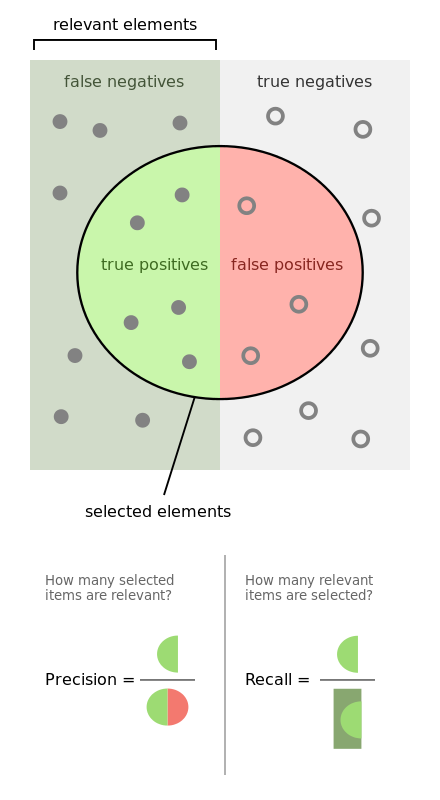
### ****65. Explain the term instance-based learning.****

Instance Based Learning is a set of procedures for regression and classification which produce a class label prediction based on resemblance to its nearest neighbors in the training data set. These algorithms just collects all the data and get an answer when required or queried. In simple words they are a set of procedures for solving new problems based on the solutions of already solved problems in the past which are similar to the current problem.

### 66. Keeping train and test split criteria in mind, is it good to perform scaling before the split or after the split?

Scaling should be done post-train and test split ideally. If the data is closely packed, then scaling post or pre-split should not make much difference.

### ****67. Define precision, recall and F1 Score?****



The metric used to access the performance of the classification model is Confusion Metric. Confusion Metric can be further interpreted with the following terms:-

**True Positives (TP)** – These are the correctly predicted positive values. It implies that the value of the actual class is yes and the value of the predicted class is also yes.

**True Negatives (TN)** – These are the correctly predicted negative values. It implies that the value of the actual class is no and the value of the predicted class is also no.

**False positives and false negatives**, these values occur when your actual class contradicts with the predicted class.

**Now,**  
**Recall,** also known as Sensitivity is the ratio of true positive rate (TP), to all observations in actual class – yes  
Recall = TP/(TP+FN)

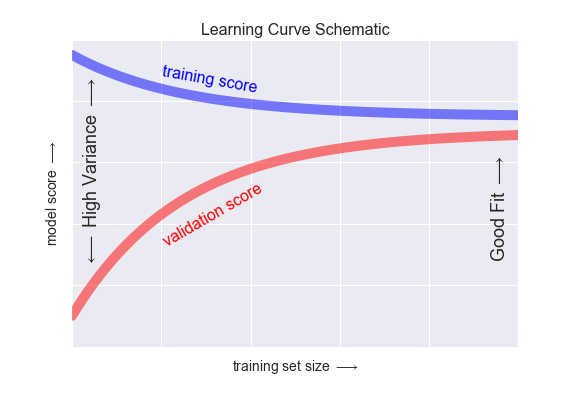
**Precision** is the ratio of positive predictive value, which measures the amount of accurate positives model predicted viz a viz number of positives it claims.  
Precision = TP/(TP+FP)

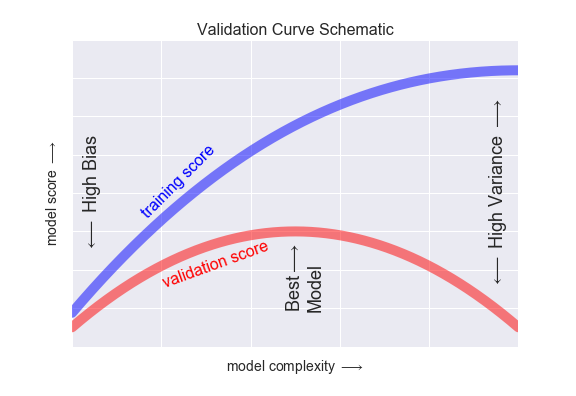
**Accuracy** is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.  
Accuracy = (TP+TN)/(TP+FP+FN+TN)

**F1 Score** is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. Accuracy works best if false positives and false negatives have a similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall.

### ****68. Plot validation score and training score with data set size on the x-axis and another plot with model complexity on the x-axis.****

For high bias in the models, the performance of the model on the validation data set is similar to the performance on the training data set. For high variance in the models, the performance of the model on the validation set is worse than the performance on the training set.





### ****69. What is Bayes’ Theorem? State at least 1 use case with respect to the machine learning context?****

Bayes’ Theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes’ theorem, a person’s age can be used to more accurately assess the probability that they have cancer than can be done without the knowledge of the person’s age.  
Chain rule for Bayesian probability can be used to predict the likelihood of the next word in the sentence.

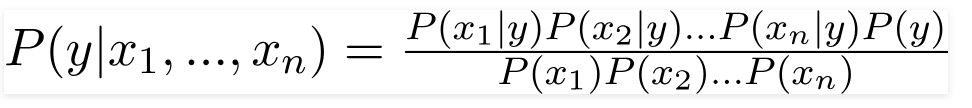
### ****70. What is Naive Bayes? Why is it Naive?****

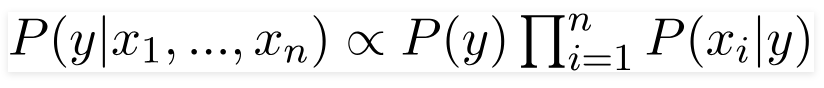
Naive Bayes classifiers are a series of classification algorithms that are based on the Bayes theorem. This family of algorithm shares a common principle which treats every pair of features independently while being classified.

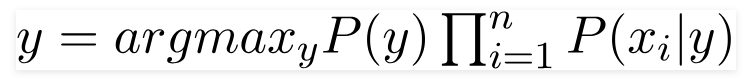
Naive Bayes is considered Naive because the attributes in it (for the class) is independent of others in the same class.  This lack of dependence between two attributes of the same class creates the quality of naiveness.

### ****71. Explain how a Naive Bayes Classifier works.****

Naive Bayes classifiers are a family of algorithms which are derived from the Bayes theorem of probability. It works on the fundamental assumption that every set of two features that is being classified is independent of each other and every feature makes an equal and independent contribution to the outcome.



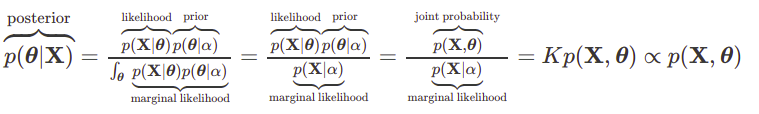




### ****72. What do the terms prior probability and marginal likelihood in context of Naive Bayes theorem mean?****

Prior probability is the percentage of dependent binary variables in the data set. If you are given a dataset and dependent variable is either 1 or 0 and percentage of 1 is 65% and percentage of 0 is 35%. Then, the probability that any new input for that variable of being 1 would be 65%.

Marginal likelihood is the denominator of the Bayes equation and it makes sure that the posterior probability is valid by making its area 1.



### ****73. Explain the difference between Lasso and Ridge?****

Lasso(L1) and Ridge(L2) are the regularization techniques where we penalize the coefficients to find the optimum solution. In ridge, the penalty function is defined by the sum of the squares of the coefficients and for the Lasso, we penalize the sum of the absolute values of the coefficients. Another type of regularization method is ElasticNet, it is a hybrid penalizing function of both lasso and ridge.

### ****74. What’s the difference between probability and likelihood?****

Probability is the measure of the likelihood that an event will occur that is, what is the certainty that a specific event will occur? Where-as a likelihood function is a function of parameters within the parameter space that describes the probability of obtaining the observed data.  
So the fundamental difference is, Probability attaches to possible results; likelihood attaches to hypotheses.

### ****75. Why would you Prune your tree?****

In the context of data science or AIML, pruning refers to the process of reducing redundant branches of a decision tree. Decision Trees are prone to overfitting, pruning the tree helps to reduce the size and minimizes the chances of overfitting. Pruning involves turning branches of a decision tree into leaf nodes and removing the leaf nodes from the original branch. It serves as a tool to perform the tradeoff.

### ****76. Model accuracy or Model performance? Which one will you prefer and why?****

This is a trick question, one should first get a clear idea, what is Model Performance? If Performance means speed, then it depends upon the nature of the application, any application related to the real-time scenario will need high speed as an important feature. Example: The best of Search Results will lose its virtue if the Query results do not appear fast.

If Performance is hinted at Why Accuracy is not the most important virtue – For any imbalanced data set, more than Accuracy, it will be an F1 score than will explain the business case and in case data is imbalanced, then Precision and Recall will be more important than rest.

### ****77. List the advantages and limitations of the Temporal Difference Learning Method.****

Temporal Difference Learning Method is a mix of Monte Carlo method and Dynamic programming method. Some of the advantages of this method include:

1. It can learn in every step online or offline.
2. It can learn from a sequence which is not complete as well.
3. It can work in continuous environments.
4. It has lower variance compared to MC method and is more efficient than MC method.

Limitations of TD method are:

1. It is a biased estimation.
2. It is more sensitive to initialization.

### ****78. How would you handle an imbalanced dataset?****

Sampling Techniques can help with an imbalanced dataset. There are two ways to perform sampling, Under Sample or Over Sampling.

In Under Sampling, we reduce the size of the majority class to match minority class thus help by improving performance w.r.t storage and run-time execution, but it potentially discards useful information.

For Over Sampling, we upsample the Minority class and thus solve the problem of information loss, however, we get into the trouble of having Overfitting.

There are other techniques as well –  
**Cluster-Based Over Sampling**– In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size

**Synthetic Minority Over-sampling Technique (SMOTE) –**A subset of data is taken from the minority class as an example and then new synthetic similar instances are created which are then added to the original dataset. This technique is good for Numerical data points.

### ****79. Mention some of the EDA Techniques?****

Exploratory Data Analysis (EDA) helps analysts to understand the data better and forms the foundation of better models.

**Visualization**

* Univariate visualization
* Bivariate visualization
* Multivariate visualization

**Missing Value Treatment** – Replace missing values with Either Mean/Median

**Outlier Detection** – Use Boxplot to identify the distribution of Outliers, then Apply IQR to set the boundary for IQR

**Transformation** – Based on the distribution, apply a transformation on the features

**Scaling the Dataset** – Apply MinMax, Standard Scaler or Z Score Scaling mechanism to scale the data.

**Feature Engineering** – Need of the domain, and SME knowledge helps Analyst find derivative fields which can fetch more information about the nature of the data

**Dimensionality reduction** — Helps in reducing the volume of data without losing much information

### ****80. Mention why feature engineering is important in model building and list out some of the techniques used for feature engineering.****

Algorithms necessitate features with some specific characteristics to work appropriately. The data is initially in a raw form. You need to extract features from this data before supplying it to the algorithm. This process is called feature engineering. When you have relevant features, the complexity of the algorithms reduces. Then, even if a non-ideal algorithm is used, results come out to be accurate.

Feature engineering primarily has two goals:

* Prepare the suitable input data set to be compatible with the machine learning algorithm constraints.
* Enhance the performance of machine learning models.

Some of the techniques used for feature engineering include Imputation, Binning, Outliers Handling, Log transform, grouping operations, One-Hot encoding, Feature split, Scaling, Extracting date.

### ****81. Differentiate between Statistical Modeling and Machine Learning?****

Machine learning models are about making accurate predictions about the situations, like Foot Fall in restaurants, Stock-Price, etc. where-as, Statistical models are designed for inference about the relationships between variables, as What drives the sales in a restaurant, is it food or Ambience.

### ****82. Differentiate between Boosting and Bagging?****

Bagging and Boosting are variants of Ensemble Techniques.

**Bootstrap Aggregation or bagging** is a method that is used to reduce the variance for algorithms having very high variance. Decision trees are a particular family of classifiers which are susceptible to having high bias.

Decision trees have a lot of sensitiveness to the type of data they are trained on. Hence generalization of results is often much more complex to achieve in them despite very high fine-tuning. The results vary greatly if the training data is changed in decision trees.

Hence bagging is utilised where multiple decision trees are made which are trained on samples of the original data and the final result is the average of all these individual models.

**Boosting**is the process of using an n-weak classifier system for prediction such that every weak classifier compensates for the weaknesses of its classifiers. By weak classifier, we imply a classifier which performs poorly on a given data set.

It’s evident that boosting is not an algorithm rather it’s a process. Weak classifiers used are generally logistic regression, shallow decision trees etc.

There are many algorithms which make use of boosting processes but two of them are mainly used: Adaboost and Gradient Boosting and XGBoost.

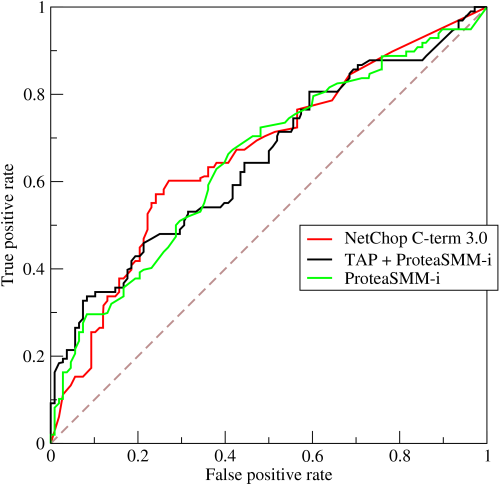
### ****83. What is the significance of Gamma and Regularization in SVM?****

The gamma defines influence. Low values meaning ‘far’ and high values meaning ‘close’.  If gamma is too large, the radius of the area of influence of the support vectors only includes the support vector itself and no amount of regularization with C will be able to prevent overfitting.  If gamma is very small, the model is too constrained and cannot capture the complexity of the data.

The regularization parameter (lambda) serves as a degree of importance that is given to miss-classifications. This can be used to draw the tradeoff with OverFitting.

### ****84. Define ROC curve work****

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



### ****85. What is the difference between a generative and discriminative model?****

A generative model learns the different categories of data. On the other hand, a discriminative model will only learn the distinctions between different categories of data. Discriminative models perform much better than the generative models when it comes to classification tasks.

### ****86. What are hyperparameters and how are they different from parameters?****

A parameter is a variable that is internal to the model and whose value is estimated from the training data. They are often saved as part of the learned model. Examples include weights, biases etc.

A hyperparameter is a variable that is external to the model whose value cannot be estimated from the data. They are often used to estimate model parameters. The choice of parameters is sensitive to implementation. Examples include learning rate, hidden layers etc.

### ****87. What is shattering a set of points? Explain VC dimension.****

In order to shatter a given configuration of points, a classifier must be able to, for all possible assignments of positive and negative for the points, perfectly partition the plane such that positive points are separated from negative points. For a configuration of **n** points, there are **2n**possible assignments of positive or negative.

When choosing a classifier, we need to consider the type of data to be classified and this can be known by VC dimension of a classifier. It is defined as cardinality of the largest set of points that the classification algorithm i.e. the classifier can shatter. In order to have a VC dimension of at least **n**, a classifier must be able to shatter a single given configuration of **n** points.

### ****88. What are some differences between a linked list and an array?****

Arrays and Linked lists are both used to store linear data of similar types. However, there are a few difference between them.

|  |  |
| --- | --- |
| **Array** | **Linked List** |
| Elements are well-indexed, making specific element accessing easier | Elements need to be accessed in a cumulative manner |
| Operations (insertion, deletion) are faster in array | Linked list takes linear time, making operations a bit slower |
| Arrays are of fixed size | Linked lists are dynamic and flexible |
| Memory is assigned during compile time in an array | Memory is allocated during execution or runtime in Linked list. |
| Elements are stored consecutively in arrays. | Elements are stored randomly in Linked list |
| Memory utilization is inefficient in the array | Memory utilization is efficient in the linked list. |

### ****89. What is the meshgrid () method and the contourf () method? State some usesof both.****

The meshgrid( ) function in numpy takes two arguments as input : range of x-values in the grid, range of y-values in the grid whereas meshgrid needs to be built before the contourf( ) function in matplotlib is used which takes in many inputs : x-values, y-values, fitting curve (contour line) to be plotted in grid, colours etc.

 Meshgrid () function is used to create a grid using 1-D arrays of x-axis inputs and y-axis inputs to represent the matrix indexing. Contourf () is used to draw filled contours using the given x-axis inputs, y-axis inputs, contour line, colours etc.

### ****90. Describe a hash table.****

Hashing is a technique for identifying unique objects from a group of similar objects. Hash functions are large keys converted into small keys in hashing techniques. The values of hash functions are stored in data structures which are known hash table.

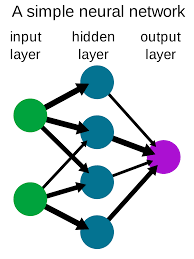
### ****91. List the advantages and disadvantages of using****[neural networks](https://www.mygreatlearning.com/blog/types-of-neural-networks/)****.****

Advantages:

We can store information on the entire network instead of storing it in a database. It has the ability to work and give a good accuracy even with inadequate information. A neural network has parallel processing ability and distributed memory.

Disadvantages:

Neural Networks requires processors which are capable of parallel processing. It’s unexplained functioning of the network is also quite an issue as it reduces the trust in the network in some situations like when we have to show the problem we noticed to the network. Duration of the network is mostly unknown. We can only know that the training is finished by looking at the error value but it doesn’t give us optimal results.



### ****92. You have to train a 12GB dataset using a neural network with a machine which has only 3GB RAM. How would you go about it?****

We can use NumPy arrays to solve this issue. Load all the data into an array. In NumPy, arrays have a property to map the complete dataset without loading it completely in memory. We can pass the index of the array, dividing data into batches, to get the data required and then pass the data into the neural networks. But be careful about keeping the batch size normal.

### ****93. Write a simple code to binarize data.****

Conversion of data into binary values on the basis of certain threshold is known as binarizing of data. Values below the threshold are set to 0 and those above the threshold are set to 1 which is useful for feature engineering.

Code:

from sklearn.preprocessing import Binarizer

import pandas

import numpy

names\_list = ['Alaska', 'Pratyush', 'Pierce', 'Sandra', 'Soundarya', 'Meredith', 'Richard', 'Jackson', 'Tom',’Joe’]

data\_frame = pandas.read\_csv(url, names=names\_list)

array = dataframe.values

# Splitting the array into input and output

A = array [: 0:7]

B = array [:7]

binarizer = Binarizer(threshold=0.0). fit(X)

binaryA = binarizer.transform(A)

numpy.set\_printoptions(precision=5)

print (binaryA [0:7:])

### ****94. What is an Array?****

The array is defined as a collection of similar items, stored in a contiguous manner. Arrays is an intuitive concept as the need to group similar objects together arises in our day to day lives. Arrays satisfy the same need. How are they stored in the memory? Arrays consume blocks of data, where each element in the array consumes one unit of memory. The size of the unit depends on the type of data being used. For example, if the data type of elements of the array is int, then 4 bytes of data will be used to store each element. For character data type, 1 byte will be used. This is implementation specific, and the above units may change from computer to computer.

Example:

fruits = [‘apple’, banana’, pineapple’]

In the above case, fruits is a list that comprises of three fruits. To access them individually, we use their indexes. Python and C are 0- indexed languages, that is, the first index is 0. MATLAB on the contrary starts from 1, and thus is a 1-indexed language.

### ****95. What are the advantages and disadvantages of using an Array?****

Advantages:

1. Random access is enabled
2. Saves memory
3. Cache friendly
4. Predictable compile timing
5. Helps in re-usability of code

Disadvantages:

1. Addition and deletion of records is time consuming even though we get the element of interest immediately through random access. This is due to the fact that the elements need to be reordered after insertion or deletion.
2. If contiguous blocks of memory are not available in the memory, then there is an overhead on the CPU to search for the most optimal contiguous location available for the requirement.

Now that we know what arrays are, we shall understand them in detail by solving some interview questions. Before that, let us see the functions that Python as a language provides for arrays, also known as, lists.

append() – Adds an element at the end of the list  
copy() – returns a copy of a list.  
reverse() – reverses the elements of the list  
sort() – sorts the elements in ascending order by default.

### ****96. What is Lists in Python?****

Lists is an effective data structure provided in python. There are various functionalities associated with the same. Let us consider the scenario where we want to copy a list to another list. If the same operation had to be done in C programming language, we would have to write our own function to implement the same.

On the contrary, Python provides us with a function called copy. We can copy a list to another just by calling the copy function.

new\_list = old\_list.copy()

We need to be careful while using the function. copy() is a shallow copy function, that is, it only stores the references of the original list in the new list. If the given argument is a compound data structure like a list then python creates another object of the same type (in this case, a new list) but for everything inside old list, only their reference is copied. Essentially, the new list consists of references to the elements of the older list.

Hence, upon changing the original list, the new list values also change. This can be dangerous in many applications. Therefore, Python provides us with another functionality called as deepcopy.  Intuitively, we may consider that deepcopy() would follow the same paradigm, and the only difference would be that for each element we will recursively call deepcopy. Practically, this is not the case.

deepcopy() preserves the graphical structure of the original compound data. Let us understand this better with the help of an example:

import copy.deepcopy

a = [1,2]

b = [a,a] # there's only 1 object a

c = deepcopy(b)

# check the result by executing these lines

c[0] is a # return False, a new object a' is created

c[0] is c[1] # return True, c is [a',a'] not [a',a'']

This is the tricky part, during the process of deepcopy() a hashtable implemented as a dictionary in python is used to map: old\_object reference onto new\_object reference.

Therefore, this prevents unnecessary duplicates and thus preserves the structure of the copied compound data structure. Thus, in this case, c[0] is not equal to a, as internally their addresses are different.

Normal copy

>>> a = [[1, 2, 3], [4, 5, 6]]

>>> b = list(a)

>>> a

[[1, 2, 3], [4, 5, 6]]

>>> b

[[1, 2, 3], [4, 5, 6]]

>>> a[0][1] = 10

>>> a

[[1, 10, 3], [4, 5, 6]]

>>> b # b changes too -> Not a deepcopy.

[[1, 10, 3], [4, 5, 6]]

Deep copy

>>> import copy

>>> b = copy.deepcopy(a)

>>> a

[[1, 10, 3], [4, 5, 6]]

>>> b

[[1, 10, 3], [4, 5, 6]]

>>> a[0][1] = 9

>>> a

[[1, 9, 3], [4, 5, 6]]

>>> b # b doesn't change -> Deep Copy

[[1, 10, 3], [4, 5, 6]]

Now that we have understood the concept of lists, let us solve interview questions to get better exposure on the same.

### ****97. Given an array of integers where each element represents the max number of steps that can be made forward from that element. The task is to find the minimum number of jumps to reach the end of the array (starting from the first element). If an element is 0, then cannot move through that element.****

Solution: This problem is famously called as end of array problem. We want to determine the minimum number of jumps required in order to reach the end. The element in the array represents the maximum number of jumps that, that particular element can take.

Let us understand how to approach the problem initially.

We need to reach the end. Therefore, let us have a count that tells us how near we are to the end. Consider the array A=[1,2,3,1,1]

In the above example we can go from

> 2 - >3 - > 1 - > 1 - 4 jumps

1 - > 2 - > 1 - > 1 - 3 jumps

1 - > 2 - > 3 - > 1 - 3 jumps

Hence, we have a fair idea of the problem. Let us come up with a logic for the same.

Let us start from the end and move backwards as that makes more sense intuitionally. We will use variables right and prev\_r denoting previous right to keep track of the jumps.

Initially, right = prev\_r = the last but one element. We consider the distance of an element to the end, and the number of jumps possible by that element. Therefore, if the sum of the number of jumps possible and the distance is greater than the previous element, then we will discard the previous element and use the second element’s value to jump. Try it out using a pen and paper first. The logic will seem very straight forward to implement. Later, implement it on your own and then verify with the result.

def min\_jmp(arr):

n = len(arr)

right = prev\_r = n-1

count = 0

# We start from rightmost index and travesre array to find the leftmost index

# from which we can reach index 'right'

while True:

for j in (range(prev\_r-1,-1,-1)):

if j + arr[j] >= prev\_r:

right = j

if prev\_r != right:

prev\_r = right

else:

break

count += 1

return count if right == 0 else -1

# Enter the elements separated by a space

arr = list(map(int, input().split()))

print(min\_jmp(n, arr))

### ****98. Given a string S consisting only ‘a’s and ‘b’s, print the last index of the ‘b’ present in it.****

When we have are given a string of a’s and b’s, we can immediately find out the first location of a character occurring. Therefore, to find the last occurrence of a character, we reverse the string and find the first occurrence, which is equivalent to the last occurrence in the original string.

Here, we are given input as a string. Therefore, we begin by splitting the characters element wise using the function split. Later, we reverse the array, find the first occurrence position value, and get the index by finding the value len – position -1, where position is the index value.

def split(word):

return [(char) for char in word]

a = input()

a= split(a)

a\_rev = a[::-1]

pos = -1

for i in range(len(a\_rev)):

if a\_rev[i] == ‘b’:

pos = len(a\_rev)- i -1

print(pos)

break

else:

continue

if pos==-1:

print(-1)

99. Rotate the elements of an array by d positions to the left. Let us initially look at an example.

A = [1,2,3,4,5]

A <<2

[3,4,5,1,2]

A<<3

[4,5,1,2,3]

There exists a pattern here, that is, the first d elements are being interchanged with last n-d +1 elements. Therefore we can just swap the elements. Correct? What if the size of the array is huge, say 10000 elements. There are chances of memory error, run-time error etc. Therefore, we do it more carefully. We rotate the elements one by one in order to prevent the above errors, in case of large arrays.

# Rotate all the elements left by 1 position

def rot\_left\_once ( arr):

n = len( arr)

tmp = arr [0]

for i in range ( n-1): #[0,n-2]

arr[i] = arr[i + 1]

arr[n-1] = tmp

# Use the above function to repeat the process for d times.

def rot\_left (arr, d):

n = len (arr)

for i in range (d):

rot\_left\_once ( arr, n)

arr = list( map( int, input().split()))

rot =int( input())

leftRotate ( arr, rot)

for i in range( len(arr)):

print( arr[i], end=' ')

### ****100. Water Trapping Problem:****

Given an array arr[] of N non-negative integers which represents the height of blocks at index I, where the width of each block is 1. Compute how much water can be trapped in between blocks after raining.

#  Structure is like below:

# | |

# |\_|

# answer is we can trap two units of water.

Solution: We are given an array, where each element denotes the height of the block. One unit of height is equal to one unit of water, given there exists space between the 2 elements to store it. Therefore, we need to find out all such pairs that exist which can store water. We need to take care of the possible cases:

1. There should be no overlap of water saved
2. Water should not overflow

Therefore, let us find start with the extreme elements, and move towards the centre.

n = int(input())

arr = [int(i) for i in input().split()]

left, right = [arr[0]], [0] \* n

# left =[arr[0]]

#right = [ 0 0 0 0…0] n terms

right[n-1] = arr[-1] # right most element

# we use two arrays left[ ] and right[ ], which keep track of elements greater than all  
# elements the order of traversal respectively.

for elem in arr[1 : ]:

left.append(max(left[-1], elem) )

for i in range( len( arr)-2, -1, -1):

right[i] = max( arr[i] , right[i+1] )

water = 0

# once we have the arrays left, and right, we can find the water capacity between these arrays.

for i in range( 1, n - 1):

add\_water = min( left[i - 1], right[i]) - arr[i]

if add\_water > 0:

water += add\_water

print(water)

### ****101. Explain Eigenvectors and Eigenvalues.****

**Ans.** Linear transformations are helpful to understand using eigenvectors. They find their prime usage in the creation of covariance and correlation matrices in data science.

Simply put, eigenvectors are directional entities along which linear transformation features like compression, flip etc. can be applied.

Eigenvalues are the magnitude of the linear transformation features along each direction of an Eigenvector.

### ****102.**** ****How would you define the number of clusters in a clustering algorithm?****

**Ans.**The number of clusters can be determined by finding the silhouette score. Often we aim to get some inferences from data using clustering techniques so that we can have a broader picture of a number of classes being represented by the data. In this case, the silhouette score helps us determine the number of cluster centres to cluster our data along.

Another technique that can be used is the elbow method.

### ****103. What are the performance metrics that can be used to estimate the efficiency of a linear regression model?****

**Ans.** The performance metric that is used in this case is:

1. Mean Squared Error
2. R2score
3. Adjusted  R2 score
4. Mean Absolute score

### ****104. What is the default method of splitting in decision trees?****

The default method of splitting in decision trees is the Gini Index. Gini Index is the measure of impurity of a particular node.

This can be changed by making changes to classifier parameters.

### ****105. How is p-value useful?****

**Ans.** The p-value gives the probability of the null hypothesis is true. It gives us the statistical significance of our results. In other words, p-value determines the confidence of a model in a particular output.

### ****106. Can logistic regression be used for classes more than 2?****

**Ans.** No, logistic regression cannot be used for classes more than 2 as it is a binary classifier. For multi-class classification algorithms like Decision Trees, Naïve Bayes’ Classifiers are better suited.

### ****107. What are the hyperparameters of a logistic regression model?****

**Ans.** Classifier penalty, classifier solver and classifier C are the trainable hyperparameters of a Logistic Regression Classifier. These can be specified exclusively with values in Grid Search to hyper tune a Logistic Classifier.

### ****108. Name a few hyper-parameters of decision trees?****

**Ans.**The most important features which one can tune in decision trees are:

1. Splitting criteria
2. Min\_leaves
3. Min\_samples
4. Max\_depth

### ****109. How to deal with multicollinearity?****

**Ans.** Multi collinearity can be dealt with by the following steps:

* Remove highly correlated predictors from the model.
* Use [Partial Least Squares Regression (PLS)](https://blog.minitab.com/blog/statistics-and-quality-data-analysis/giving-thanks-for-the-regression-menu-v2) or [Principal Components Analysis](https://support.minitab.com/minitab/19/help-and-how-to/modeling-statistics/multivariate/how-to/principal-components/before-you-start/overview/),

### ****110. What is Heteroscedasticity?****

**Ans.** It is a situation in which the variance of a variable is unequal across the range of values of the predictor variable.

It should be avoided in regression as it introduces unnecessary variance.

### ****111. Is ARIMA model a good fit for every time series problem?****

**Ans.** No, ARIMA model is not suitable for every type of time series problem. There are situations where ARMA model and others also come in handy.

ARIMA is best when different standard temporal structures require to be captured for time series data.

### ****112. How do you deal with the class imbalance in a classification problem?****

**Ans.** Class imbalance can be dealt with in the following ways:

1. Using class weights
2. Using Sampling
3. Using SMOTE
4. Choosing loss functions like Focal Loss

### ****113. What is the role of cross-validation?****

**Ans.** Cross-validation is a technique which is used to increase the performance of a machine learning algorithm, where the machine is fed sampled data out of the same data for a few times. The sampling is done so that the dataset is broken into small parts of the equal number of rows, and a random part is chosen as the test set, while all other parts are chosen as train sets.

### ****114. What is a voting model?****

**Ans.** A voting model is an ensemble model which combines several classifiers but to produce the final result, in case of a classification-based model, takes into account, the classification of a certain data point of all the models and picks the most vouched/voted/generated option from all the given classes in the target column.

### ****115. How to deal with very few data samples? Is it possible to make a model out of it?****

**Ans.** If very few data samples are there, we can make use of oversampling to produce new data points. In this way, we can have new data points.

### ****116. What are the hyperparameters of an SVM?****

**Ans.** The gamma value, c value and the type of kernel are the hyperparameters of an SVM model.

### ****117. What is Pandas Profiling?****

**Ans.** Pandas profiling is a step to find the effective number of usable data. It gives us the statistics of NULL values and the usable values and thus makes variable selection and data selection for building models in the preprocessing phase very effective.

### ****118. What impact does correlation have on PCA?****

**Ans.** If data is correlated PCA does not work well. Because of the correlation of variables the effective variance of variables decreases. Hence correlated data when used for PCA does not work well.

### ****119. How is PCA different from LDA?****

**Ans.**PCA is unsupervised. LDA is unsupervised.

PCA takes into consideration the variance. LDA takes into account the distribution of classes.

### ****120. What distance metrics can be used in KNN?****

**Ans.** Following distance metrics can be used in KNN.

* Manhattan
* Minkowski
* Tanimoto
* Jaccard
* Mahalanobis

### ****121. Which metrics can be used to measure correlation of categorical data?****

**Ans.** Chi square test can be used for doing so. It gives the measure of correlation between categorical predictors.

### ****122. Which algorithm can be used in value imputation in both categorical and continuous categories of data?****

**Ans.** KNN is the only algorithm that can be used for imputation of both categorical and continuous variables.

### ****123. When should ridge regression be preferred over lasso?****

**Ans.** We should use ridge regression when we want to use all predictors and not remove any as it reduces the coefficient values but does not nullify them.

### ****124. Which algorithms can be used for important variable selection?****

**Ans.** Random Forest, Xgboost and plot variable importance charts can be used for variable selection.

### ****125. What ensemble technique is used by Random forests?****

**Ans.** Bagging is the technique used by Random Forests. Random forests are a collection of trees which work on sampled data from the original dataset with the final prediction being a voted average of all trees.

### ****126. What ensemble technique is used by gradient boosting trees?****

**Ans.**Boosting is the technique used by GBM.

### ****127. If we have a high bias error what does it mean? How to treat it?****

**Ans.** High bias error means that that model we are using is ignoring all the important trends in the model and the model is underfitting.

To reduce underfitting:

* We need to increase the complexity of the model
* Number of features need to be increased

Sometimes it also gives the impression that the data is noisy. Hence noise from data should be removed so that most important signals are found by the model to make effective predictions.

Increasing the number of epochs results in increasing the duration of training of the model. It’s helpful in reducing the error.

### ****128. Which type of sampling is better for a classification model and why?****

**Ans.** Stratified sampling is better in case of classification problems because it takes into account the balance of classes in train and test sets. The proportion of classes is maintained and hence the model performs better. In case of random sampling of data, the data is divided into two parts without taking into consideration the balance classes in the train and test sets. Hence some classes might be present only in tarin sets or validation sets. Hence the results of the resulting model are poor in this case.

### ****129. What is a good metric for measuring the level of multicollinearity?****

**Ans.** VIF or 1/tolerance is a good measure of measuring multicollinearity in models. VIF is the percentage of the variance of a predictor which remains unaffected by other predictors. So higher the VIF value, greater is the multicollinearity amongst the predictors.

A **rule of thumb** for interpreting the variance inflation factor:

* 1 = not correlated.
* Between 1 and 5 = moderately correlated.
* Greater than 5 = highly correlated.

### ****130. When can be a categorical value treated as a continuous variable and what effect does it have when done so?****

**Ans.** A categorical predictor can be treated as a continuous one when the nature of data points it represents is ordinal. If the predictor variable is having ordinal data then it can be treated as continuous and its inclusion in the model increases the performance of the model.

### ****131. What is the role of maximum likelihood in logistic regression.****

**Ans.** Maximum likelihood equation helps in estimation of most probable values of the estimator’s predictor variable coefficients which produces results which are the most likely or most probable and are quite close to the truth values.

### ****132. Which distance do we measure in the case of KNN?****

**Ans.** The hamming distance is measured in case of KNN for the determination of nearest neighbours. Kmeans uses euclidean distance.

### ****133. What is a pipeline?****

**Ans.** A pipeline is a sophisticated way of writing software such that each intended action while building a model can be serialized and the process calls the individual functions for the individual tasks. The tasks are carried out in sequence for a given sequence of data points and the entire process can be run onto n threads by use of composite estimators in scikit learn.

### ****134. Which sampling technique is most suitable when working with time-series data?****

**Ans.** We can use a custom iterative sampling such that we continuously add samples to the train set. We only should keep in mind that the sample used for validation should be added to the next train sets and a new sample is used for validation.

### ****135. What are the benefits of pruning?****

**Ans.**Pruning helps in the following:

1. Reduces overfitting
2. Shortens the size of the tree
3. Reduces complexity of the model
4. Increases bias

### ****136. What is normal distribution?****

**Ans.** The distribution having the below properties is called normal distribution.

* The [mean, mode and median](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/) are all equal.
* The curve is symmetric at the center (i.e. around the mean, μ).
* Exactly half of the values are to the left of center and exactly half the values are to the right.
* The total area under the curve is 1.

### ****137. What is the 68 per cent rule in normal distribution?****

**Ans.** The normal distribution is a bell-shaped curve. Most of the data points are around the median. Hence approximately 68 per cent of the data is around the median. Since there is no skewness and its bell-shaped.

### ****138. What is a chi-square test?****

**Ans.** A chi-square determines if a sample data matches a population.

A chi-square test for independence compares two variables in a contingency table to see if they are related.

A very small chi-square test statistics implies observed data fits the expected data extremely well.

### ****139. What is a random variable****?

**Ans.** A Random Variable is a set of possible values from a random experiment. Example: Tossing a coin: we could get Heads or Tails. Rolling of a dice: we get 6 values

### ****140. What is the degree of freedom?****

**Ans.** It is the number of independent values or quantities which can be assigned to a statistical distribution. It is used in Hypothesis testing and chi-square test.

### ****141. Which kind of recommendation system is used by amazon to recommend similar items?****

**Ans.** Amazon uses a collaborative filtering algorithm for the recommendation of similar items. It’s a user to user similarity based mapping of user likeness and susceptibility to buy.

### ****142. What is a false positive?****

**Ans.** It is a test result which wrongly indicates that a particular condition or attribute is present.

Example – “Stress testing, a routine diagnostic tool used in detecting heart disease, results in a significant number of false positives in women”

### ****143. What is a false negative?****

**Ans.** A test result which wrongly indicates that a particular condition or attribute is absent.

Example – “it’s possible to have a false negative—the test says you aren’t pregnant when you are”

### ****144. What is the error term composed of in regression?****

**Ans.** Error is a sum of bias error+variance error+ irreducible error in regression. Bias and variance error can be reduced but not the irreducible error.

### ****145. Which performance metric is better R2 or adjusted R2?****

**Ans.** Adjusted R2 because the performance of predictors impacts it. R2 is independent of predictors and shows performance improvement through increase if the number of predictors is increased.

### ****146. What’s the difference between Type I and Type II error?****

Type I and Type II error in machine learning refers to false values. Type I is equivalent to a False positive while Type II is equivalent to a False negative. In Type I error, a hypothesis which ought to be accepted doesn’t get accepted. Similarly, for Type II error, the hypothesis gets rejected which should have been accepted in the first place.

### ****147. What do you understand by L1 and L2 regularization?****

L2 regularization: It tries to spread error among all the terms. L2 corresponds to a Gaussian prior.

L1 regularization: It is more binary/sparse, with many variables either being assigned a 1 or 0 in weighting. L1 corresponds to setting a Laplacean prior on the terms.

### ****148. Which one is better, Naive Bayes Algorithm or Decision Trees?****

Although it depends on the problem you are solving, but some general advantages are following:

**Naive Bayes:**

* Work well with small dataset compared to DT which need more data
* Lesser overfitting
* Smaller in size and faster in processing

**Decision Trees:**

* Decision Trees are very flexible, easy to understand, and easy to debug
* No preprocessing or transformation of features required
* Prone to overfitting but you can use pruning or Random forests to avoid that.

### ****149. What do you mean by the ROC curve?****

Receiver operating characteristics (ROC curve): ROC curve illustrates the diagnostic ability of a binary classifier. It is calculated/created by plotting True Positive against False Positive at various threshold settings. The performance metric of ROC curve is AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

### ****150. What do you mean by AUC curve?****

AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

### ****151. What is log likelihood in logistic regression?****

It is the sum of the likelihood residuals. At record level, the natural log of the error (residual) is calculated for each record, multiplied by minus one, and those values are totaled. That total is then used as the basis for deviance (2 x ll) and likelihood (exp(ll)).

The same calculation can be applied to a naive model that assumes absolutely no predictive power, and a saturated model assuming perfect predictions.

The likelihood values are used to compare different models, while the deviances (test, naive, and saturated) can be used to determine the predictive power and accuracy. Logistic regression accuracy of the model will always be 100 percent for the development data set, but that is not the case once a model is applied to another data set.

### ****152. How would you evaluate a logistic regression model?****

Model Evaluation is a very important part in any analysis to answer the following questions,

How well does the model fit the data?, Which predictors are most important?, Are the predictions accurate?

So the following are the criterion to access the model performance,

1. **Akaike Information Criteria (AIC)**: In simple terms, AIC estimates the relative amount of information lost by a given model. So the less information lost the higher the quality of the model. Therefore, we always prefer models with minimum AIC.

2. **Receiver operating characteristics (ROC curve)**: ROC curve illustrates the diagnostic ability of a binary classifier. It is calculated/ created by plotting True Positive against False Positive at various threshold settings. The performance metric of ROC curve is AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

3. **Confusion Matrix**: In order to find out how well the model does in predicting the target variable, we use a confusion matrix/ classification rate. It is nothing but a tabular representation of actual Vs predicted values which helps us to find the accuracy of the model.

### ****153. What are the advantages of SVM algorithms?****

SVM algorithms have basically advantages in terms of complexity. First I would like to clear that both Logistic regression as well as SVM can form non linear decision surfaces and can be coupled with the kernel trick. If Logistic regression can be coupled with kernel then why use SVM?

● SVM is found to have better performance practically in most cases.

● SVM is computationally cheaper O(N^2\*K) where K is no of support vectors (support vectors are those points that lie on the class margin) where as logistic regression is O(N^3)

● Classifier in SVM depends only on a subset of points . Since we need to maximize distance between closest points of two classes (aka margin) we need to care about only a subset of points unlike logistic regression.

### ****154. Why does XGBoost perform better than SVM?****

First reason is that XGBoos is an ensemble method that uses many trees to make a decision so it gains power by repeating itself.

SVM is a linear separator, when data is not linearly separable SVM needs a Kernel to project the data into a space where it can separate it, there lies its greatest strength and weakness, by being able to project data into a high dimensional space SVM can find a linear separation for almost any data but at the same time it needs to use a Kernel and we can argue that there’s not a perfect kernel for every dataset.

### ****155. What is the difference between SVM Rank and SVR (Support Vector Regression)?****

One is used for ranking and the other is used for regression.

There is a crucial difference between regression and ranking. In regression, the absolute value is crucial. A real number is predicted.

In ranking, the only thing of concern is the ordering of a set of examples. We only want to know which example has the highest rank, which one has the second-highest, and so on. From the data, we only know that example 1 should be ranked higher than example 2, which in turn should be ranked higher than example 3, and so on. We do not know by how much example 1 is ranked higher than example 2, or whether this difference is bigger than the difference between examples 2 and 3.

### ****156. What is the difference between the normal soft margin SVM and SVM with a linear kernel?****

**Hard-margin**

You have the basic SVM – hard margin. This assumes that data is very well behaved, and you can find a perfect classifier – which will have 0 error on train data.

**Soft-margin**

Data is usually not well behaved, so SVM hard margins may not have a solution at all. So we allow for a little bit of error on some points. So the training error will not be 0, but average error over all points is minimized.

**Kernels**

The above assume that the best classifier is a straight line. But what is it is not a straight line. (e.g. it is a circle, inside a circle is one class, outside is another class). If we are able to map the data into higher dimensions – the higher dimension may give us a straight line.

### ****157. How is linear classifier relevant to SVM?****

An svm is a type of linear classifier. If you don’t mess with kernels, it’s arguably the most simple type of linear classifier.

Linear classifiers (all?) learn linear fictions from your data that map your input to scores like so: scores = Wx + b. Where W is a matrix of learned weights, b is a learned bias vector that shifts your scores, and x is your input data. This type of function may look familiar to you if you remember y = mx + b from high school.

A typical svm loss function ( the function that tells you how good your calculated scores are in relation to the correct labels ) would be hinge loss. It takes the form: Loss = sum over all scores except the correct score of max(0, scores – scores(correct class) + 1).

### ****158. What are the advantages of using a naive Bayes for classification?****

* Very simple, easy to implement and fast.
* If the NB conditional independence assumption holds, then it will converge quicker than discriminative models like logistic regression.
* Even if the NB assumption doesn’t hold, it works great in practice.
* Need less training data.
* Highly scalable. It scales linearly with the number of predictors and data points.
* Can be used for both binary and mult-iclass classification problems.
* Can make probabilistic predictions.
* Handles continuous and discrete data.
* Not sensitive to irrelevant features.

### ****159. Are Gaussian Naive Bayes the same as binomial Naive Bayes?****

Binomial Naive Bayes: It assumes that all our features are binary such that they take only two values. Means 0s can represent “word does not occur in the document” and 1s as “word occurs in the document”.

Gaussian Naive Bayes: Because of the assumption of the normal distribution, Gaussian Naive Bayes is used in cases when all our features are continuous. For example in Iris dataset features are sepal width, petal width, sepal length, petal length. So its features can have different values in the data set as width and length can vary. We can’t represent features in terms of their occurrences. This means data is continuous. Hence we use Gaussian Naive Bayes here.

### ****160. What is the difference between the Naive Bayes Classifier and the Bayes classifier?****

Naive Bayes assumes conditional independence, P(X|Y, Z)=P(X|Z)

P(X|Y,Z)=P(X|Z)

P(X|Y,Z)=P(X|Z), Whereas more general Bayes Nets (sometimes called Bayesian Belief Networks), will allow the user to specify which attributes are, in fact, conditionally independent.

For the Bayesian network as a classifier, the features are selected based on some scoring functions like Bayesian scoring function and minimal description length(the two are equivalent in theory to each other given that there is enough training data). The scoring functions mainly restrict the structure (connections and directions) and the parameters(likelihood) using the data. After the structure has been learned the class is only determined by the nodes in the Markov blanket(its parents, its children, and the parents of its children), and all variables given the Markov blanket are discarded.

### ****161. In what real world applications is Naive Bayes classifier used?****

Some of real world examples are as given below

* To mark an email as spam, or not spam?
* Classify a news article about technology, politics, or sports?
* Check a piece of text expressing positive emotions, or negative emotions?
* Also used for face recognition software

### ****162. Is naive Bayes supervised or unsupervised?****

First, Naive Bayes is not one algorithm but a family of Algorithms that inherits the following attributes:

1.Discriminant Functions

2.Probabilistic Generative Models

3.Bayesian Theorem

4.Naive Assumptions of Independence and Equal Importance of feature vectors.

Moreover, it is a special type of Supervised Learning algorithm that could do simultaneous multi-class predictions (as depicted by standing topics in many news apps).

Since these are generative models, so based upon the assumptions of the random variable mapping of each feature vector these may even be classified as Gaussian Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes, etc.

### ****164. What do you understand by Precision and Recall?****

In pattern recognition, The information retrieval and classification in machine learning are part of **precision**. It is also called as positive predictive value which is the fraction of relevant instances among the retrieved instances.

**Recall** is also known as sensitivity and the fraction of the total amount of relevant instances which  were actually retrieved.

Both precision and recall are therefore based on an understanding and measure of relevance.

### ****165. What Are the Three Stages of Building a Model in Machine Learning?****

To build a model in machine learning, you need to follow few steps:

1. Understand the business model
2. Data acquisitions
3. Data cleaning
4. Exploratory data analysis
5. Use machine learning algorithms to make a model
6. Use unknown dataset to check the accuracy of the model

### ****166. How Do You Design an Email Spam Filter in Machine Learning?****

1. Understand the business model: Try to understand the related attributes for the spam mail
2. Data acquisitions: Collect the spam mail to read the hidden pattern from them
3. Data cleaning: Clean the unstructured or semi structured data
4. Exploratory data analysis: Use statistical concepts to understand the data like spread, outlier, etc.
5. Use machine learning algorithms to make a model: can use naive bayes or some other algorithms as well
6. Use unknown dataset to check the accuracy of the model

### ****169. What is Kernel SVM?****

SVM algorithms have basically advantages in terms of complexity. First I would like to clear that both Logistic regression as well as SVM can form non linear decision surfaces and can be coupled with the kernel trick. If Logistic regression can be coupled with kernel then why use SVM?

● SVM is found to have better performance practically in most cases.

● SVM is computationally cheaper O(N^2\*K) where K is no of support vectors (support vectors are those points that lie on the class margin) where as logistic regression is O(N^3)

● Classifier in SVM depends only on a subset of points . Since we need to maximize distance between closest points of two classes (aka margin) we need to care about only a subset of points unlike logistic regression.

### ****170. What is the process of carrying out a linear regression?****

**Linear Regression** Analysis consists of more than just fitting a **linear** line through a cloud of data points. It consists of 3 stages–

 (1) analyzing the correlation and directionality of the data,

 (2) estimating the **model**, i.e., fitting the line,

(3) evaluating the validity and usefulness of the **model**.