**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.

**2)      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.

**3)      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.

**4)      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.

**5)      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?**

PAC (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.

**Q1- What’s the trade-off between bias and variance?**

*More reading:*[*Bias-Variance Tradeoff (Wikipedia)*](https://en.wikipedia.org/wiki/Bias-variance_tradeoff)

Bias is error due to erroneous or overly simplistic assumptions in the learning algorithm you’re using. This can lead to the model underfitting your 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 error due to too much complexity in the learning algorithm you’re using. This leads to the algorithm being highly sensitive to high degrees of variation in your training data, which can lead your model to overfit the data. You’ll be carrying too much noise from your 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 tradeoff bias and variance. You don’t want either high bias or high variance in your model.

**Q2- What is the difference between supervised and unsupervised machine learning?**

*More reading:*[*What is the difference between supervised and unsupervised machine learning? (Quora)*](https://www.quora.com/What-is-the-difference-between-supervised-and-unsupervised-learning-algorithms)

Supervised learning requires training labeled data. For example, in order to do classification (a supervised learning task), you’ll need to first label the data you’ll use to train the model to classify data into your labeled groups. Unsupervised learning, in contrast, does not require labeling data explicitly.

**Q3- How is KNN different from k-means clustering?**

*More reading:*[*How is the k-nearest neighbor algorithm different from k-means clustering? (Quora)*](https://www.quora.com/How-is-the-k-nearest-neighbor-algorithm-different-from-k-means-clustering)

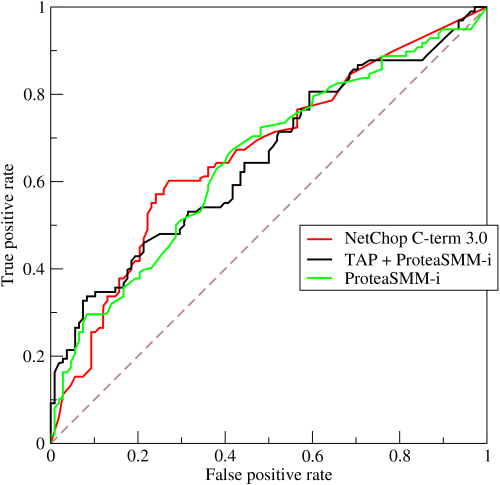
K-Nearest Neighbors is a supervised classification algorithm, while k-means clustering is an unsupervised clustering algorithm. While the mechanisms may seem similar at first, what this really means is that in order for K-Nearest Neighbors to work, you need labeled data you want to classify an unlabeled 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.

**Q4- Explain how a ROC curve works.**

*More reading:*[*Receiver operating characteristic (Wikipedia)*](https://en.wikipedia.org/wiki/Receiver_operating_characteristic)

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).



**Q5- Define precision and recall.**

*More reading:*[*Precision and recall (Wikipedia)*](https://en.wikipedia.org/wiki/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.

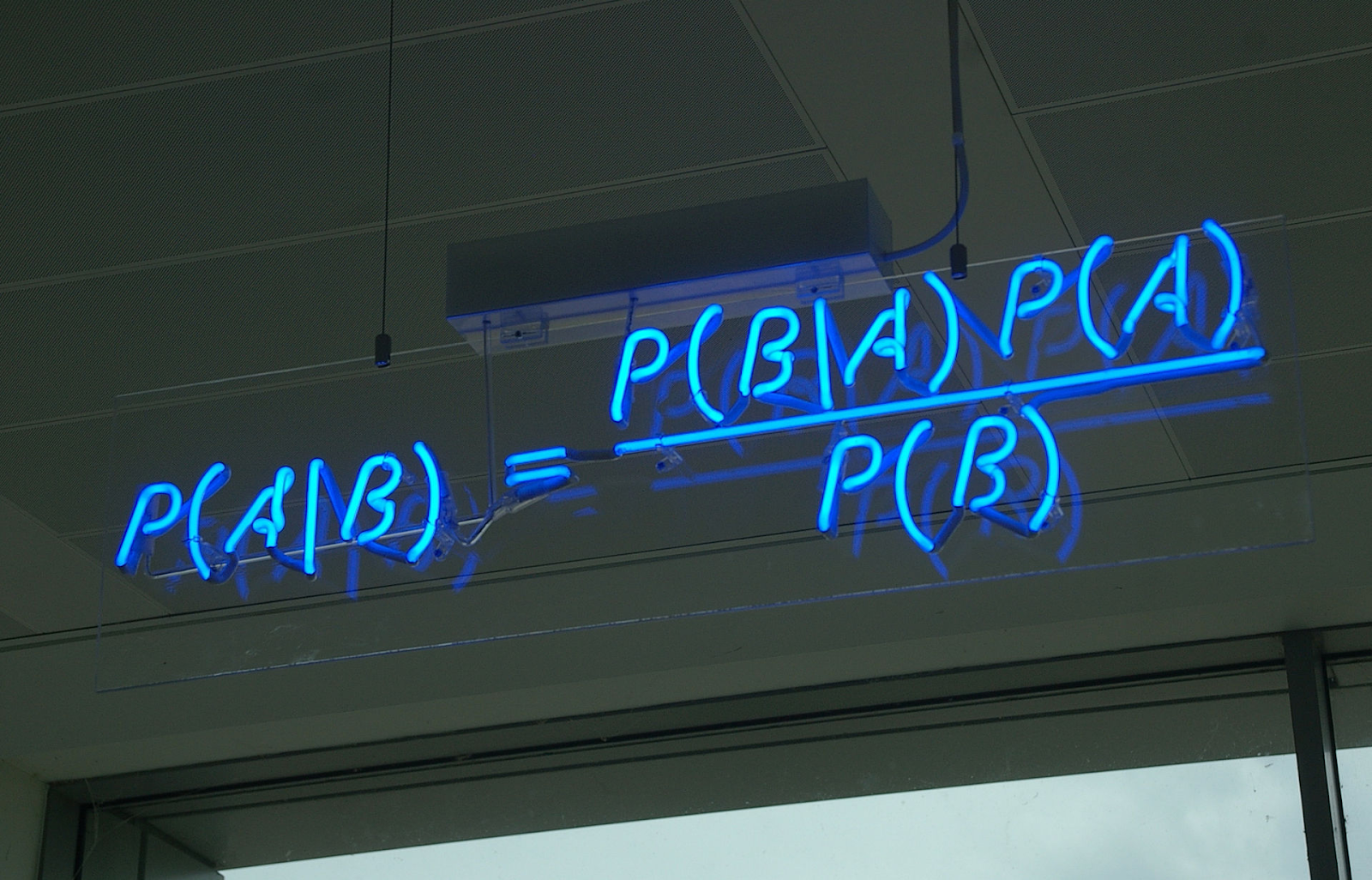
**Q6- What is Bayes’ Theorem? How is it useful in a machine learning context?**

*More reading:*[*An Intuitive (and Short) Explanation of Bayes’ Theorem (BetterExplained)*](https://betterexplained.com/articles/an-intuitive-and-short-explanation-of-bayes-theorem/)

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 the Naive Bayes classifier. That’s something important to consider when you’re faced with machine learning interview questions.

**Q7- Why is “Naive” Bayes naive?**

*More reading:*[*Why is “naive Bayes” naive? (Quora)*](https://www.quora.com/Why-is-naive-Bayes-naive?share=1)

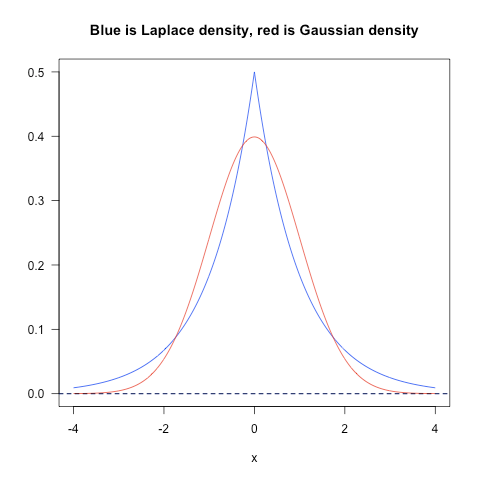
Despite its practical applications, especially in text mining, Naive Bayes is considered “Naive” because it makes an assumption that is virtually impossible to see in real-life data: the conditional probability is calculated as the pure product of the individual probabilities of components. This implies the absolute independence of features — a condition probably never met in real life.

As a Quora commenter put it whimsically, a Naive Bayes classifier that figured out that you liked pickles and ice cream would probably naively recommend you a pickle ice cream.

**Q8- Explain the difference between L1 and L2 regularization.**

*More reading:*[*What is the difference between L1 and L2 regularization? (Quora)*](https://www.quora.com/What-is-the-difference-between-L1-and-L2-regularization)

L2 regularization tends to spread error among all the terms, while L1 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, while L2 corresponds to a Gaussian prior.



**Q9- 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!

**Q10- What’s the difference between Type I and Type II error?**

*More reading:*[*Type I and type II errors (Wikipedia)*](https://en.wikipedia.org/wiki/Type_I_and_type_II_errors)

Don’t think that this is a trick question! Many machine learning interview questions will be an attempt to lob basic questions at you just to make sure you’re on top of your game and you’ve prepared all of your bases.

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.

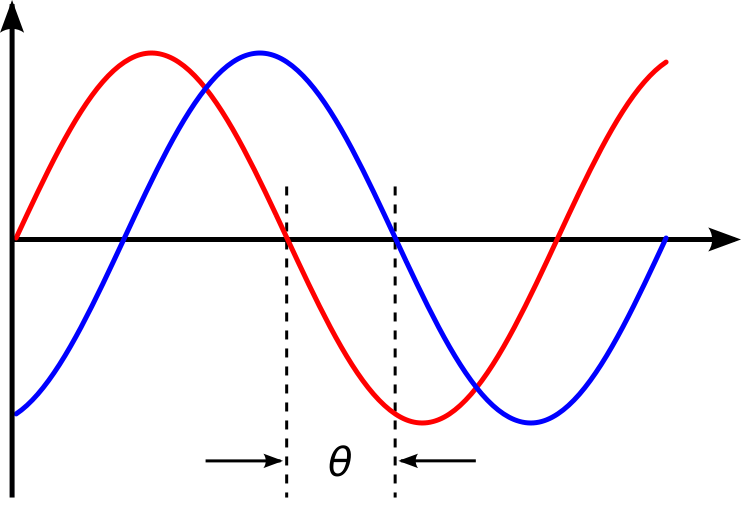
**Q11- What’s a Fourier transform?**

*More reading:*[*Fourier transform (Wikipedia)*](https://en.wikipedia.org/wiki/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](https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/) 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.

**Q12- What’s the difference between probability and likelihood?**

*More reading:*[*What is the difference between “likelihood” and “probability”? (Cross Validated)*](https://stats.stackexchange.com/questions/2641/what-is-the-difference-between-likelihood-and-probability#2647)



**Q13- What is deep learning, and how does it contrast with other machine learning algorithms?**

*More reading:*[*Deep learning (Wikipedia)*](https://en.wikipedia.org/wiki/Deep_learning)

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 learning algorithm that learns representations of data through the use of neural nets.

**Q14- What’s the difference between a generative and discriminative model?**

*More reading:*[*What is the difference between a Generative and Discriminative Algorithm? (Stack Overflow)*](https://stackoverflow.com/questions/879432/what-is-the-difference-between-a-generative-and-discriminative-algorithm)

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.

**Q15- What cross-validation technique would you use on a time series dataset?**

*More reading:*[*Using k-fold cross-validation for time-series model selection (CrossValidated)*](https://stats.stackexchange.com/questions/14099/using-k-fold-cross-validation-for-time-series-model-selection)

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]

**Q16- How is a decision tree pruned?**

*More reading:*[*Pruning (decision trees)*](https://en.wikipedia.org/wiki/Pruning_%28decision_trees%29)

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.

**Q17- Which is more important to you– model accuracy, or model performance?**

*More reading:*[*Accuracy paradox (Wikipedia)*](https://en.wikipedia.org/wiki/Accuracy_paradox)

This question tests your grasp of the nuances of machine learning model performance! Machine learning interview questions often look towards the details. There are models with higher accuracy that can perform worse in predictive power — how does that make sense?

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.

**Q18- What’s the F1 score? How would you use it?**

*More reading:*[*F1 score (Wikipedia)*](https://en.wikipedia.org/wiki/F1_score)

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.

**Q19- How would you handle an imbalanced dataset?**

*More reading:*[*8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset (Machine Learning Mastery)*](http://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-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.

**Q20- When should you use classification over regression?**

*More reading:*[*Regression vs Classification (Math StackExchange)*](https://math.stackexchange.com/questions/141381/regression-vs-classification)

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.)

**Q21- Name an example where ensemble techniques might be useful.**

*More reading:*[*Ensemble learning (Wikipedia)*](https://en.wikipedia.org/wiki/Ensemble_learning)

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.

**Q22- How do you ensure you’re not overfitting with a model?**

*More reading:*[*How can I avoid overfitting? (Quora)*](https://www.quora.com/How-can-I-avoid-overfitting)

This is a simple restatement of a fundamental problem in machine learning: the possibility of overfitting training data and carrying the noise of that data through to the test set, thereby providing inaccurate generalizations.

There are three main methods to avoid overfitting:

**1-** Keep the model simpler: reduce variance by taking into account fewer variables and parameters, thereby removing some of the noise in the training data.

**2-** Use cross-validation techniques such as k-folds cross-validation.

**3-** Use regularization techniques such as LASSO that penalize certain model parameters if they’re likely to cause overfitting.

**Q23- What evaluation approaches would you work to gauge the effectiveness of a machine learning model?**

*More reading:*[*How to Evaluate Machine Learning Algorithms (Machine Learning Mastery)*](http://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/)

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](http://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/). 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.

**Q24- How would you evaluate a logistic regression model?**

*More reading:*[*Evaluating a logistic regression (CrossValidated)*](https://stats.stackexchange.com/questions/71517/evaluating-a-logistic-regression#71522)

A subsection of the question above. You have to demonstrate an understanding of what the typical goals of a logistic regression are (classification, prediction etc.) and bring up a few examples and use cases.

**Q25- What’s the “kernel trick” and how is it useful?**

*More reading:*[*Kernel method (Wikipedia)*](https://en.wikipedia.org/wiki/Kernel_method)

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 than the 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: Programming

These machine learning interview questions test your knowledge of programming principles you need to implement machine learning principles in practice. Machine learning interview questions tend to be technical questions that test your logic and programming skills: this section focuses more on the latter.

**Q26- How do you handle missing or corrupted data in a dataset?**

*More reading:*[*Handling missing data (O’Reilly)*](https://www.oreilly.com/learning/handling-missing-data)

You could find missing/corrupted data in a dataset and either drop those rows or columns, or decide to replace them with another value.

In Pandas, there are two very useful methods: isnull() and dropna() that will help you find columns of data with missing or corrupted data and drop those values. If you want to fill the invalid values with a placeholder value (for example, 0), you could use the fillna() method.

**Q27- Do you have experience with Spark or big data tools for machine learning?**

*More reading:*[*50 Top Open Source Tools for Big Data (Datamation)*](http://www.datamation.com/data-center/50-top-open-source-tools-for-big-data-1.html)

You’ll want to get familiar with the meaning of big data for different companies and the different tools they’ll want. Spark is the big data tool most in demand now, able to handle immense datasets with speed. Be honest if you don’t have experience with the tools demanded, but also take a look at job descriptions and see what tools pop up: you’ll want to invest in familiarizing yourself with them.

**Q28- Pick an algorithm. Write the psuedo-code for a parallel implementation.**

*More reading:*[*Writing pseudocode for parallel programming (Stack Overflow)*](https://stackoverflow.com/questions/5583257/writing-pseudocode-for-parallel-programming)

This kind of question demonstrates your ability to think in parallelism and how you could handle concurrency in programming implementations dealing with big data. Take a look at pseudocode frameworks such as [Peril-L](http://www.eng.utah.edu/~cs4960-01/lecture4.pdf) and visualization tools such as [Web Sequence Diagrams](https://www.websequencediagrams.com/) to help you demonstrate your ability to write code that reflects parallelism.

**Q29- What are some differences between a linked list and an array?**

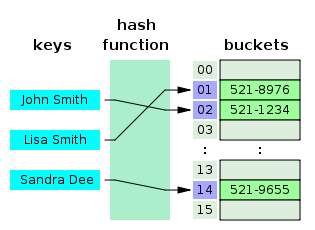
*More reading:*[*Array versus linked list (Stack Overflow)*](https://stackoverflow.com/questions/166884/array-versus-linked-list#167016)

An array is an ordered collection of objects. A linked list is a series of objects with pointers that direct how to process them sequentially. An array assumes that every element has the same size, unlike the linked list. A linked list can more easily grow organically: an array has to be pre-defined or re-defined for organic growth. Shuffling a linked list involves changing which points direct where — meanwhile, shuffling an array is more complex and takes more memory.

**Q30- Describe a hash table.**

*More reading:*[*Hash table (Wikipedia)*](https://en.wikipedia.org/wiki/Hash_table)

A hash table is a data structure that produces an associative array. A key is mapped to certain values through the use of a hash function. They are often used for tasks such as database indexing.



**Q31- Which data visualization libraries do you use? What are your thoughts on the best data visualization tools?**

*More reading:*[*31 Free Data Visualization Tools (Springboard)*](https://www.springboard.com/blog/31-free-data-visualization-tools/)

What’s important here is to define your views on how to properly visualize data and your personal preferences when it comes to tools. Popular tools include R’s ggplot, Python’s seaborn and matplotlib, and tools such as Plot.ly and Tableau.

**Q1.** **You are given a train data set having 1000 columns and 1 million rows. The data set is based on a classification problem. Your manager has asked you to reduce the dimension of this data so that model computation time can be reduced. Your machine has memory constraints. What would you do? (You are free to make practical assumptions.)**

**Answer:** Processing a high dimensional data on a limited memory machine is a strenuous task, your interviewer would be fully aware of that. Following are the methods you can use to tackle such situation:

1. Since we have lower RAM, we should close all other applications in our machine, including the web browser, so that most of the memory can be put to use.
2. We can randomly sample the data set. This means, we can create a smaller data set, let’s say, having 1000 variables and 300000 rows and do the computations.
3. To reduce dimensionality, we can separate the numerical and categorical variables and remove the correlated variables. For numerical variables, we’ll use correlation. For categorical variables, we’ll use chi-square test.
4. Also, we can use [PCA](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/) and pick the components which can explain the maximum variance in the data set.
5. Using online learning algorithms like Vowpal Wabbit (available in Python) is a possible option.
6. Building a linear model using Stochastic Gradient Descent is also helpful.
7. We can also apply our business understanding to estimate which all predictors can impact the response variable. But, this is an intuitive approach, failing to identify useful predictors might result in significant loss of information.

**Note:** For point 4 & 5, make sure you read about [online learning algorithms](https://www.analyticsvidhya.com/blog/2015/01/introduction-online-machine-learning-simplified-2/) & [Stochastic Gradient Descent](https://www.quora.com/Whats-the-difference-between-gradient-descent-and-stochastic-gradient-descent). These are advanced methods.

**Q2.** **Is rotation necessary in PCA? If yes, Why? What will happen if you don’t rotate the components?**

**Answer:** Yes, rotation (orthogonal) is necessary because it maximizes the difference between variance captured by the component. This makes the components easier to interpret. Not to forget, that’s the motive of doing PCA where, we aim to select fewer components (than features) which can explain the maximum variance in the data set. By doing rotation, the relative location of the components doesn’t change, it only changes the actual coordinates of the points.

If we don’t rotate the components, the effect of PCA will diminish and we’ll have to select more number of components to explain variance in the data set.

Know more: [PCA](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/)

**Q3.** **You are given a data set. The data set has missing values which spread along 1 standard deviation from the median. What percentage of data would remain unaffected? Why?**

**Answer:** This question has enough hints for you to start thinking! Since, the data is spread across median, let’s assume it’s a normal distribution. We know, in a normal distribution, ~68% of the data lies in 1 standard deviation from mean (or mode, median), which leaves ~32% of the data unaffected. Therefore, ~32% of the data would remain unaffected by missing values.

**Q4. You are given a data set on cancer detection. You’ve build a classification model and achieved an accuracy of 96%. Why shouldn’t you be happy with your model performance? What can you do about it?**

**Answer:** If you have worked on enough data sets, you should deduce that cancer detection results in imbalanced data. In an imbalanced data set, accuracy should not be used as a measure of performance because 96% (as given) might only be predicting majority class correctly, but our class of interest is minority class (4%) which is the people who actually got diagnosed with cancer. Hence, in order to evaluate model performance, we should use Sensitivity (True Positive Rate), Specificity (True Negative Rate), F measure to determine class wise performance of the classifier. If the minority class performance is found to to be poor, we can undertake the following steps:

1. We can use undersampling, oversampling or SMOTE to make the data balanced.
2. We can alter the prediction threshold value by doing [probability caliberation](https://www.analyticsvidhya.com/blog/2016/07/platt-scaling-isotonic-regression-minimize-logloss-error/) and finding a optimal threshold using AUC-ROC curve.
3. We can assign weight to classes such that the minority classes gets larger weight.
4. We can also use anomaly detection.

Know more: [Imbalanced Classification](https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/)

**Q5**. **Why is naive Bayes so ‘naive’ ?**

**Answer:** naive Bayes is so ‘naive’ because it assumes that all of the features in a data set are equally important and independent. As we know, these assumption are rarely true in real world scenario.

**Q6. Explain prior probability, likelihood and marginal likelihood in context of naiveBayes algorithm?**

**Answer:** Prior probability is nothing but, the proportion of dependent (binary) variable in the data set. It is the closest guess you can make about a class, without any further information. For example: In a data set, the dependent variable is binary (1 and 0). The proportion of 1 (spam) is 70% and 0 (not spam) is 30%. Hence, we can estimate that there are 70% chances that any new email would  be classified as spam.

Likelihood is the probability of classifying a given observation as 1 in presence of some other variable. For example: The probability that the word ‘FREE’ is used in previous spam message is likelihood. Marginal likelihood is, the probability that the word ‘FREE’ is used in any message.

**Q7.** **You are working on a time series data set. You manager has asked you to build a high accuracy model. You start with the decision tree algorithm, since you know it works fairly well on all kinds of data. Later, you tried a time series regression model and got higher accuracy than decision tree model. Can this happen? Why?**

**Answer:** Time series data is known to posses linearity. On the other hand, a decision tree algorithm is known to work best to detect non – linear interactions. The reason why decision tree failed to provide robust predictions because it couldn’t map the linear relationship as good as a regression model did. Therefore, we learned that, a linear regression model can provide robust prediction given the data set satisfies its [linearity assumptions](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/).

**Q8.** **You are assigned a new project which involves helping a food delivery company save more money. The problem is, company’s delivery team aren’t able to deliver food on time. As a result, their customers get unhappy. And, to keep them happy, they end up delivering food for free. Which machine learning algorithm can save them?**

**Answer:** You might have started hopping through the list of ML algorithms in your mind. But, wait! Such questions are asked to test your machine learning fundamentals.

This is not a machine learning problem. This is a route optimization problem. A machine learning problem consist of three things:

1. There exist a pattern.
2. You cannot solve it mathematically (even by writing exponential equations).
3. You have data on it.

Always look for these three factors to decide if machine learning is a tool to solve a particular problem.

**Q9.** **You came to know that your model is suffering from low bias and high variance. Which algorithm should you use to tackle it? Why?**

**Answer:**  Low bias occurs when the model’s predicted values are near to actual values. In other words, the model becomes flexible enough to mimic the training data distribution. While it sounds like great achievement, but not to forget, a flexible model has no generalization capabilities. It means, when this model is tested on an unseen data, it gives disappointing results.

In such situations, we can use bagging algorithm (like random forest) to tackle high variance problem. Bagging algorithms divides a data set into subsets made with repeated randomized sampling. Then, these samples are used to generate  a set of models using a single learning algorithm. Later, the model predictions are combined using voting (classification) or averaging (regression).

Also, to combat high variance, we can:

1. Use regularization technique, where higher model coefficients get penalized, hence lowering model complexity.
2. Use top n features from variable importance chart. May be, with all the variable in the data set, the algorithm is having difficulty in finding the meaningful signal.

**Q10.** **You are given a data set. The data set contains many variables, some of which are highly correlated and you know about it. Your manager has asked you to run PCA. Would you remove correlated variables first? Why?**

**Answer:** Chances are, you might be tempted to say No, but that would be incorrect. Discarding correlated variables have a substantial effect on PCA because, in presence of correlated variables, the variance explained by a particular component gets inflated.

For example: You have 3 variables in a data set, of which 2 are correlated. If you run PCA on this data set, the first principal component would exhibit twice the variance than it would exhibit with uncorrelated variables. Also, adding correlated variables lets PCA put more importance on those variable, which is misleading.

**Q11. After spending several hours, you are now anxious to build a high accuracy model. As a result, you build 5 GBM models, thinking a boosting algorithm would do the magic. Unfortunately, neither of models could perform better than benchmark score. Finally, you decided to combine those models. Though, ensembled models are known to return high accuracy, but you are unfortunate. Where did you miss?**

**Answer:** As we know, ensemble learners are based on the idea of combining weak learners to create strong learners. But, these learners provide superior result when the combined models are uncorrelated. Since, we have used 5 GBM models and got no accuracy improvement, suggests that the models are correlated. The problem with correlated models is, all the models provide same information.

For example: If model 1 has classified User1122 as 1, there are high chances model 2 and model 3 would have done the same, even if its actual value is 0. Therefore, ensemble learners are built on the premise of combining weak uncorrelated models to obtain better predictions.

**Q12.** **How is kNN different from kmeans clustering?**

**Answer:** Don’t get mislead by ‘k’ in their names. You should know that the fundamental difference between both these algorithms is, kmeans is unsupervised in nature and kNN is supervised in nature. kmeans is a clustering algorithm. kNN is a classification (or regression) algorithm.

kmeans algorithm partitions a data set into clusters such that a cluster formed is homogeneous and the points in each cluster are close to each other. The algorithm tries to maintain enough separability between these clusters. Due to unsupervised nature, the clusters have no labels.

kNN algorithm tries to classify an unlabeled observation based on its k (can be any number ) surrounding neighbors. It is also known as lazy learner because it involves minimal training of model. Hence, it doesn’t use training data to make generalization on unseen data set.

**Q13.** **How is True Positive Rate and Recall related? Write the equation.**

**Answer:** True Positive Rate = Recall. Yes, they are equal having the formula (TP/TP + FN).

Know more: [Evaluation Metrics](https://www.analyticsvidhya.com/blog/2016/02/7-important-model-evaluation-error-metrics/)

**Q14. You have built a multiple regression model. Your model R² isn’t as good as you wanted. For improvement, your remove the intercept term, your model R² becomes 0.8 from 0.3. Is it possible? How?**

**Answer:** Yes, it is possible. We need to understand the significance of intercept term in a regression model. The intercept term shows model prediction without any independent variable i.e. mean prediction. The formula of R² = 1 – ∑(y – y´)²/∑(y – ymean)² where y´ is predicted value.

When intercept term is present, R² value evaluates your model wrt. to the mean model. In absence of intercept term (ymean), the model can make no such evaluation, with large denominator, ∑(y - y´)²/∑(y)² equation’s value becomes smaller than actual, resulting in higher R².

**Q15.** **After analyzing the model, your manager has informed that your regression model is suffering from multicollinearity. How would you check if he’s true? Without losing any information, can you still build a better model?**

**Answer:** To check multicollinearity, we can create a correlation matrix to identify & remove variables having correlation above 75% (deciding a threshold is subjective). In addition, we can use calculate VIF (variance inflation factor) to check the presence of multicollinearity. VIF value <= 4 suggests no multicollinearity whereas a value of >= 10 implies serious multicollinearity. Also, we can use tolerance as an indicator of multicollinearity.

But, removing correlated variables might lead to loss of information. In order to retain those variables, we can use penalized regression models like ridge or lasso regression. Also, we can add some random noise in correlated variable so that the variables become different from each other. But, adding noise might affect the prediction accuracy, hence this approach should be carefully used.

Know more: [Regression](https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/)

**Q16. When is Ridge regression favorable over Lasso regression?**

**Answer:** You can quote ISLR’s authors Hastie, Tibshirani who asserted that, in presence of few variables with medium / large sized effect, use lasso regression. In presence of many variables with small / medium sized effect, use ridge regression.

Conceptually, we can say, lasso regression (L1) does both variable selection and parameter shrinkage, whereas Ridge regression only does parameter shrinkage and end up including all the coefficients in the model. In presence of correlated variables, ridge regression might be the preferred choice. Also, ridge regression works best in situations where the least square estimates have higher variance. Therefore, it depends on our model objective.

Know more: [Ridge and Lasso Regression](https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-regression-python/)

**Q17.** **Rise in global average temperature led to decrease in number of pirates around the world. Does that mean that decrease in number of pirates caused the climate change?**

**Answer:** After reading this question, you should have understood that this is a classic case of “causation and correlation”. No, we can’t conclude that decrease in number of pirates caused the climate change because there might be other factors (lurking or confounding variables) influencing this phenomenon.

Therefore, there might be a correlation between global average temperature and number of pirates, but based on this information we can’t say that pirated died because of rise in global average temperature.

Know more: [Causation and Correlation](https://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/)

**Q18.** **While working on a data set, how do you select important variables? Explain your methods.**

**Answer:** Following are the methods of variable selection you can use:

1. Remove the correlated variables prior to selecting important variables
2. Use linear regression and select variables based on p values
3. Use Forward Selection, Backward Selection, Stepwise Selection
4. Use Random Forest, Xgboost and plot variable importance chart
5. Use Lasso Regression
6. Measure information gain for the available set of features and select top n features accordingly.

**Q19.** **What is the difference between covariance and correlation?**

**Answer:** Correlation is the standardized form of covariance.

Covariances are difficult to compare. For example: if we calculate the covariances of salary ($) and age (years), we’ll get different covariances which can’t be compared because of having unequal scales. To combat such situation, we calculate correlation to get a value between -1 and 1, irrespective of their respective scale.

**Q20. Is it possible capture the correlation between continuous and categorical variable? If yes, how?**

Answer: Yes, we can use ANCOVA (analysis of covariance) technique to capture association between continuous and categorical variables.

**Q21. Both being tree based algorithm, how is random forest different from Gradient boosting algorithm (GBM)?**

**Answer:** The fundamental difference is, random forest uses bagging technique to make predictions. GBM uses boosting techniques to make predictions.

In bagging technique, a data set is divided into n samples using randomized sampling. Then, using a single learning algorithm a model is build on all samples. Later, the resultant predictions are combined using voting or averaging. Bagging is done is parallel. In boosting, after the first round of predictions, the algorithm weighs misclassified predictions higher, such that they can be corrected in the succeeding round. This sequential process of giving higher weights to misclassified predictions continue until a stopping criterion is reached.

Random forest improves model accuracy by reducing variance (mainly). The trees grown are uncorrelated to maximize the decrease in variance. On the other hand, GBM improves accuracy my reducing both bias and variance in a model.

Know more: [Tree based modeling](https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/)

**Q22. Running a binary classification tree algorithm is the easy part. Do you know how does a tree splitting takes place i.e. how does the tree decide which variable to split at the root node and succeeding nodes?**

**Answer:** A classification trees makes decision based on Gini Index and Node Entropy. In simple words, the tree algorithm find the best possible feature which can divide the data set into purest possible children nodes.

Gini index says, if we select two items from a population at random then they must be of same class and probability for this is 1 if population is pure. We can calculate Gini as following:

1. Calculate Gini for sub-nodes, using formula sum of square of probability for success and failure (p^2+q^2).
2. Calculate Gini for split using weighted Gini score of each node of that split

Entropy is the measure of impurity as given by (for binary class):

Entropy, Decision Tree

Here p and q is probability of success and failure respectively in that node. Entropy is zero when a node is homogeneous. It is maximum when a both the classes are present in a node at 50% – 50%.  Lower entropy is desirable.

**Q23.** **You’ve built a random forest model with 10000 trees. You got delighted after getting training error as 0.00. But, the validation error is 34.23. What is going on? Haven’t you trained your model perfectly?**

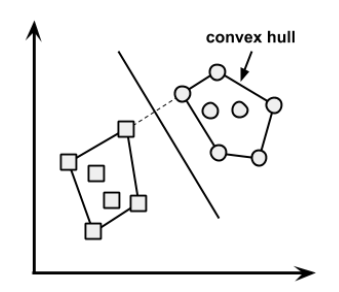
**Answer:** The model has overfitted. Training error 0.00 means the classifier has mimiced the training data patterns to an extent, that they are not available in the unseen data. Hence, when this classifier was run on unseen sample, it couldn’t find those patterns and returned prediction with higher error. In random forest, it happens when we use larger number of trees than necessary. Hence, to avoid these situation, we should tune number of trees using cross validation.

**Q24.** **You’ve got a data set to work having p (no. of variable) > n (no. of observation). Why is OLS as bad option to work with? Which techniques would be best to use? Why?**

**Answer:** In such high dimensional data sets, we can’t use classical regression techniques, since their assumptions tend to fail. When p > n, we can no longer calculate a unique least square coefficient estimate, the variances become infinite, so OLS cannot be used at all.

To combat this situation, we can use penalized regression methods like lasso, LARS, ridge which can shrink the coefficients to reduce variance. Precisely, ridge regression works best in situations where the least square estimates have higher variance.

Among other methods include subset regression, forward stepwise regression.

**Q25. What is convex hull ? (Hint: Think SVM)**

**Answer:** In case of linearly separable data, convex hull represents the outer boundaries of the two group of data points. Once convex hull is created, we get maximum margin hyperplane (MMH) as a perpendicular bisector between two convex hulls. MMH is the line which attempts to create greatest separation between two groups.

**Q26.** **We know that one hot encoding increasing the dimensionality of a data set. But, label encoding doesn’t. How ?**

**Answer:** Don’t get baffled at this question. It’s a simple question asking the difference between the two.

Using one hot encoding, the dimensionality (a.k.a features) in a data set get increased because it creates a new variable for each level present in categorical variables. For example: let’s say we have a variable ‘color’. The variable has 3 levels namely Red, Blue and Green. One hot encoding ‘color’ variable will generate three new variables as Color.Red, Color.Blue and Color.Green containing 0 and 1 value.

In label encoding, the levels of a categorical variables gets encoded as 0 and 1, so no new variable is created. Label encoding is majorly used for binary variables.

**Q27. What cross validation technique would you use on time series data set? Is it k-fold or LOOCV?**

**Answer:** Neither.

In time series problem, k fold can be troublesome because there might be some pattern in year 4 or 5 which is not in year 3. Resampling the data set will separate these trends, and we might end up validation on past years, which is incorrect. Instead, we can use forward chaining strategy with 5 fold as shown below:

* 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]

where 1,2,3,4,5,6 represents “year”.

**Q28.** **You are given a data set consisting of variables having more than 30% missing values? Let’s say, out of 50 variables, 8 variables have missing values higher than 30%. How will you deal with them?**

**Answer:** We can deal with them in the following ways:

1. Assign a unique category to missing values, who knows the missing values might decipher some trend
2. We can remove them blatantly.
3. Or, we can sensibly check their distribution with the target variable, and if found any pattern we’ll keep those missing values and assign them a new category while removing others.

**29.** **‘People who bought this, also bought…’ recommendations seen on amazon is a result of which algorithm?**

**Answer:** The basic idea for this kind of recommendation engine comes from collaborative filtering.

Collaborative Filtering algorithm considers “User Behavior” for recommending items. They exploit behavior of other users and items in terms of transaction history, ratings, selection and purchase information. Other users behaviour and preferences over the items are used to recommend items to the new users. In this case, features of the items are not known.

Know more: [Recommender System](https://www.analyticsvidhya.com/blog/2015/10/recommendation-engines/)

**Q30.** **What do you understand by Type I vs Type II error ?**

**Answer:** Type I error is committed when the null hypothesis is true and we reject it, also known as a ‘False Positive’. Type II error is committed when the null hypothesis is false and we accept it, also known as ‘False Negative’.

In the context of confusion matrix, we can say Type I error occurs when we classify a value as positive (1) when it is actually negative (0). Type II error occurs when we classify a value as negative (0) when it is actually positive(1).

**Q31**. **You are working on a classification problem. For validation purposes, you’ve randomly sampled the training data set into train and validation. You are confident that your model will work incredibly well on unseen data since your validation accuracy is high. However, you get shocked after getting poor test accuracy. What went wrong?**

**Answer:** In case of classification problem, we should always use stratified sampling instead of random sampling. A random sampling doesn’t takes into consideration the proportion of target classes. On the contrary, stratified sampling helps to maintain the distribution of target variable in the resultant distributed samples also.

**Q32.** **You have been asked to evaluate a regression model based on R², adjusted R² and tolerance. What will be your criteria?**

**Answer:** Tolerance (1 / VIF) is used as an indicator of multicollinearity. It is an indicator of percent of variance in a predictor which cannot be accounted by other predictors. Large values of tolerance is desirable.

We will consider adjusted R² as opposed to R² to evaluate model fit because R² increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R² would only increase if an additional variable improves the accuracy of model, otherwise stays same. It is difficult to commit a general threshold value for adjusted R² because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R² and still provide fairly good predictions, as compared to a stock market data where lower adjusted R² implies that model is not good.

**Q33.** **In k-means or kNN, we use euclidean distance to calculate the distance between nearest neighbors. Why not manhattan distance ?**

**Answer:** We don’t use manhattan distance because it calculates distance horizontally or vertically only. It has dimension restrictions. On the other hand, euclidean metric can be used in any space to calculate distance. Since, the data points can be present in any dimension, euclidean distance is a more viable option.

Example: Think of a chess board, the movement made by a bishop or a rook is calculated by manhattan distance because of their respective vertical & horizontal movements.

**Q34.** **Explain machine learning to me like a 5 year old.**

**Answer:** It’s simple. It’s just like how babies learn to walk. Every time they fall down, they learn (unconsciously) & realize that their legs should be straight and not in a bend position. The next time they fall down, they feel pain. They cry. But, they learn ‘not to stand like that again’. In order to avoid that pain, they try harder. To succeed, they even seek support from the door or wall or anything near them, which helps them stand firm.

This is how a machine works & develops intuition from its environment.

*Note: The interview is only trying to test if have the ability of explain complex concepts in simple terms.*

**Q35.** **I know that a linear regression model is generally evaluated using Adjusted R² or F value. How would you evaluate a logistic regression model?**

**Answer:** We can use the following methods:

1. Since logistic regression is used to predict probabilities, we can use AUC-ROC curve along with confusion matrix to determine its performance.
2. Also, the analogous metric of adjusted R² in logistic regression is AIC. AIC is the measure of fit which penalizes model for the number of model coefficients. Therefore, we always prefer model with minimum AIC value.
3. Null Deviance indicates the response predicted by a model with nothing but an intercept. Lower the value, better the model. Residual deviance indicates the response predicted by a model on adding independent variables. Lower the value, better the model.

Know more: [Logistic Regression](https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/)

**Q36.** **Considering the long list of machine learning algorithm, given a data set, how do you decide which one to use?**

**Answer:** You should say, the choice of machine learning algorithm solely depends of the type of data. If you are given a data set which is exhibits linearity, then linear regression would be the best algorithm to use. If you given to work on images, audios, then neural network would help you to build a robust model.

If the data comprises of non linear interactions, then a boosting or bagging algorithm should be the choice. If the business requirement is to build a model which can be deployed, then we’ll use regression or a decision tree model (easy to interpret and explain) instead of black box algorithms like SVM, GBM etc.

In short, there is no one master algorithm for all situations. We must be scrupulous enough to understand which algorithm to use.

**Q37.** **Do you suggest that treating a categorical variable as continuous variable would result in a better predictive model?**

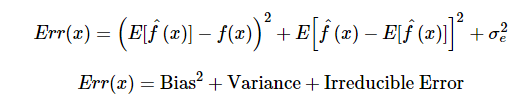
**Answer:** For better predictions, categorical variable can be considered as a continuous variable only when the variable is ordinal in nature.

**Q38.** **When does regularization becomes necessary in Machine Learning?**

**Answer:** Regularization becomes necessary when the model begins to ovefit / underfit. This technique introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term. This helps to reduce model complexity so that the model can become better at predicting (generalizing).

**Q39.** **What do you understand by Bias Variance trade off?**

**Answer:**  The error emerging from any model can be broken down into three components mathematically. Following are these component :

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/07/error-of-a-model.png)

**Bias error** is useful to quantify how much on an average are the predicted values different from the actual value. A high bias error means we have a under-performing model which keeps on missing important trends. **Variance**on the other side quantifies how are the prediction made on same observation different from each other. A high variance model will over-fit on your training population and perform badly on any observation beyond training.

**Q40.** **OLS is to linear regression. Maximum likelihood is to logistic regression. Explain the statement.**

**Answer:** OLS and Maximum likelihood are the methods used by the respective regression methods to approximate the unknown parameter (coefficient) value. In simple words,

Ordinary least square(OLS) is a method used in linear regression which approximates the parameters resulting in minimum distance between actual and predicted values. Maximum Likelihood helps in choosing the the values of parameters which maximizes the likelihood that the parameters are most likely to produce observed data.