MACHINE LEARNING INTERVIEW QUESTIONS

1. What is Machine Learning?

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

Speech Recognition *Face Recognition* *Anti Virus* *Weather Prediction*

2. What is Supervised Learning?

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

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

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

3. What is Unsupervised learning?

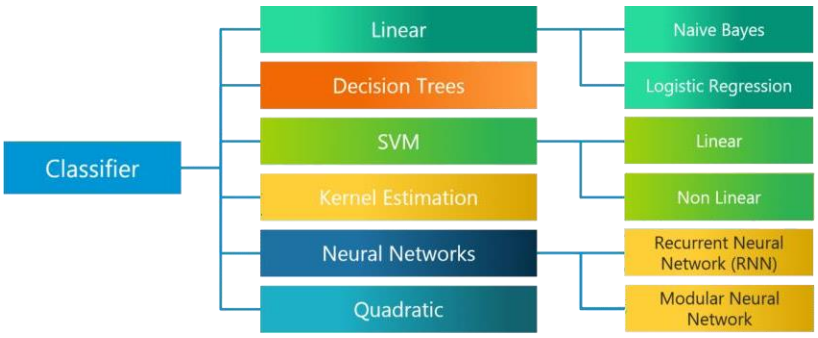
Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labelled responses.

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

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

4. What are the various classification algorithms?

The diagram lists the most important ***classification algorithms***.



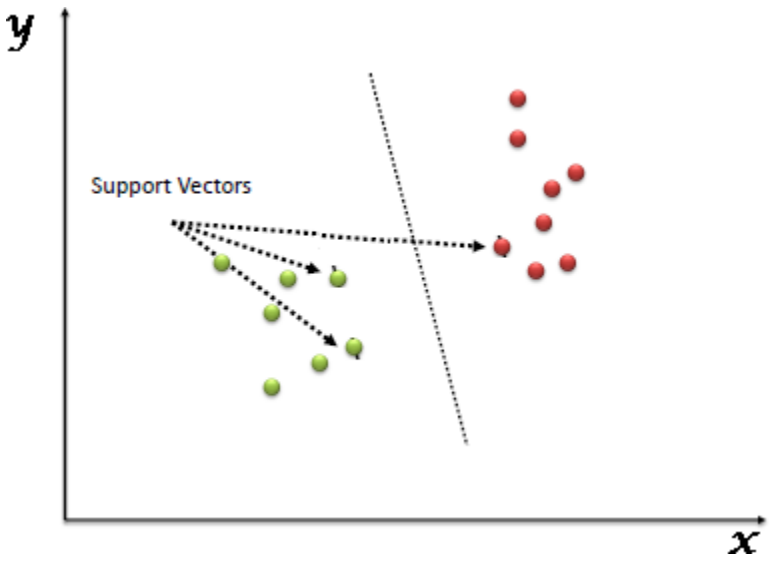
5. What is 'Naive' in a Naive Bayes?

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

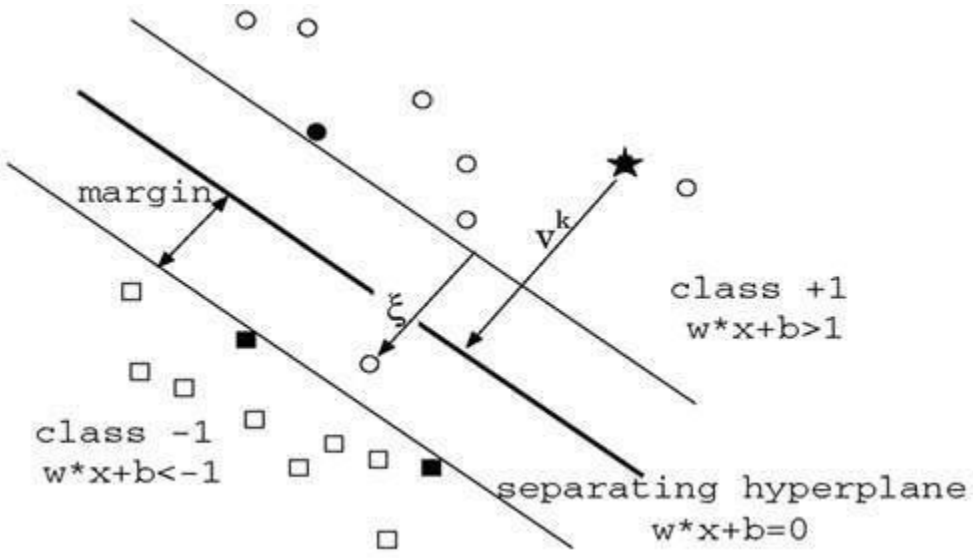
The Algorithm is 'naive' because it makes assumptions that may or may not turn out to be correct.

6. Explain SVM algorithm in detail.

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



7. What are the support vectors in SVM?



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

margin.

8. What are the different kernels in SVM?

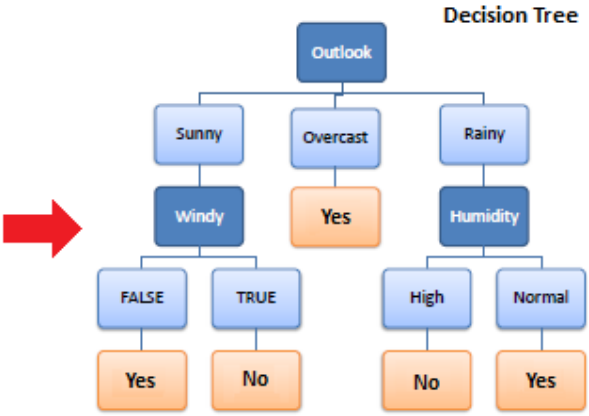
There are four types of kernels in SVM.

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

9. Explain Decision Tree algorithm in detail.

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

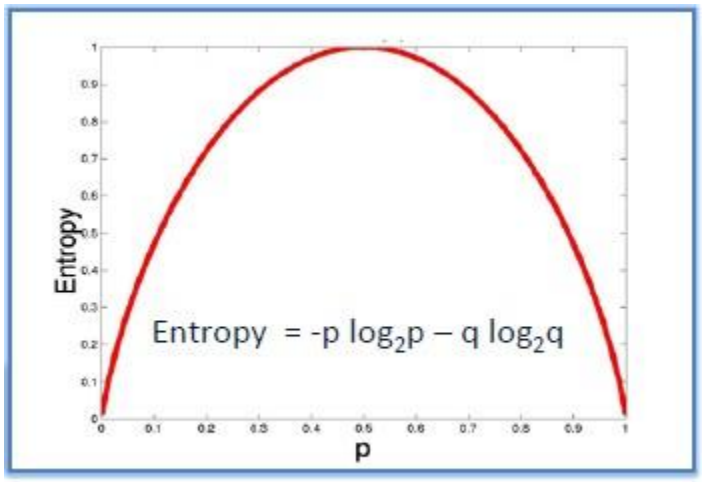
Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



10. What are Entropy and Information gain in Decision tree algorithm?

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

A decision tree is built top-down from a root node and involve partitioning of data into homogenious subsets.



$$Entropy = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$$

ID3 uses enteropy to check the homogeneity of a sample. If the sample is completely homogenious then entropy is zero and if the sample is an equally divided it has entropy of one.

11. Information Gain

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

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

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

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

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

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

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

= 0.940 – 0.693 = 0.247

12. What is pruning in Decision Tree?

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

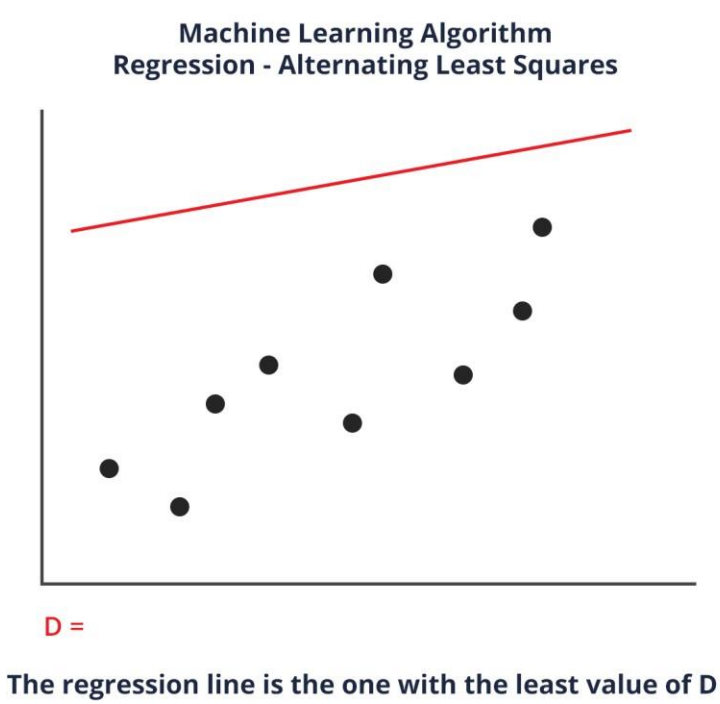
13. What is logistic regression? State an example when you have used logistic regression recently.

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

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

14. What is Linear Regression?

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



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

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

16. What are Recommender Systems?

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

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

17 What is Collaborative filtering?

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

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

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

An example of collaborative filtering can be to predict the rating of a particular user based on his/her ratings for other movies and others' ratings for all movies. This concept is widely used in recommending movies in IMDB, Netflix & BookMyShow, product recommenders in e-commerce sites like Amazon, eBay & Flipkart, YouTube video recommendations and game recommendations in Xbox.

Q18. How can outlier values be treated?

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

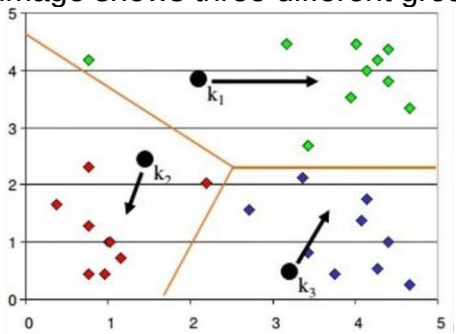
All extreme values are not outlier values. The most common ways to treat outlier values

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

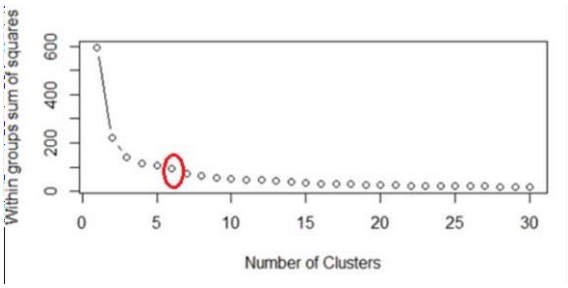
19. How will you define the number of clusters in a clustering algorithm?

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

For example, the following image shows three different groups.



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



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

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

20. What is Ensemble Learning?

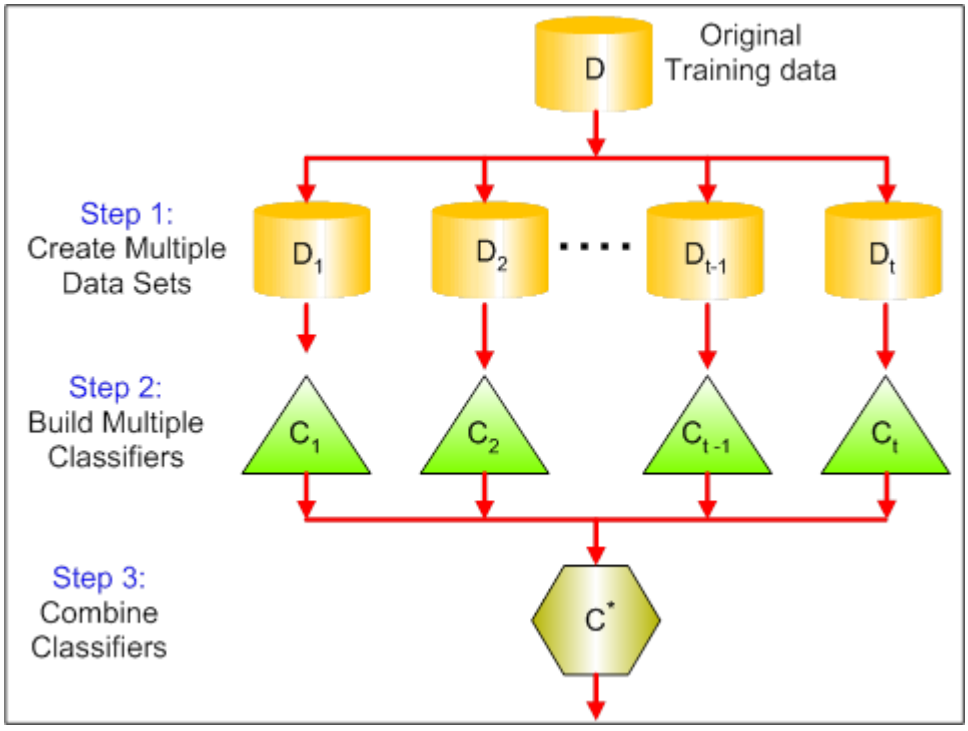
Ensemble Learning is basically combining a diverse set of learners(Individual models) together to improvise on the stability and predictive power of the model.

21. Describe in brief any type of Ensemble Learning?

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

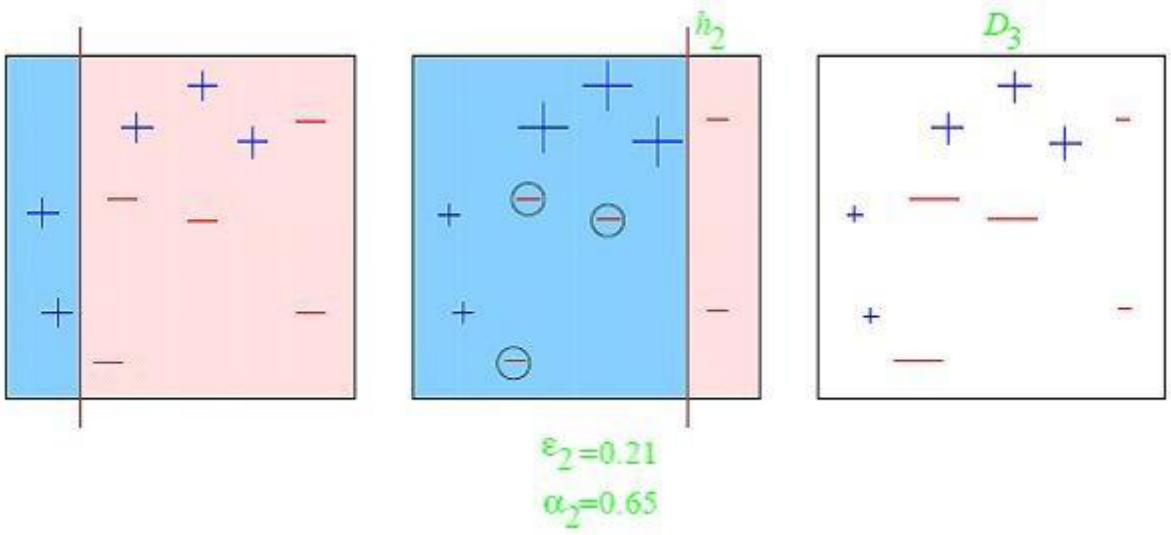
Bagging

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



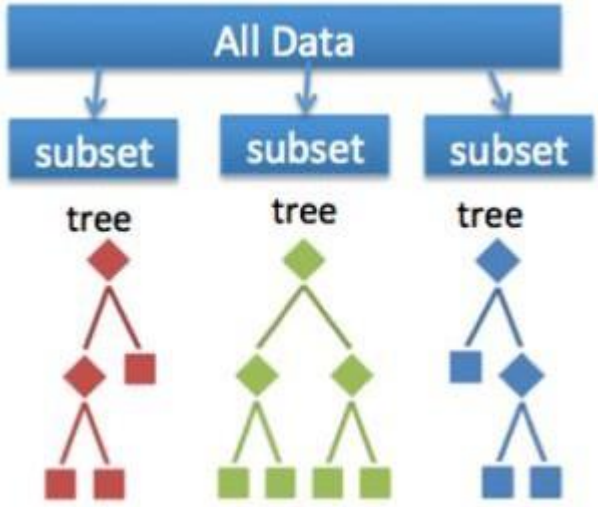
Boosting

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



22. What is a Random Forest? How does it work?

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



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

23. How Do You Work Towards a Random Forest?

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

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

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

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

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

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

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

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

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

25. How Regularly Must an Algorithm be Updated?

You will want to update an algorithm when:

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

26. What Are the Drawbacks of the Linear Model?

Some drawbacks of the linear model are:

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

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

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

28. 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?

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 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.

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

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.

30. How is kNN different from kmeans clustering?

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 unlabelled 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.

31. When is Ridge regression favourable over Lasso regression?

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.

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

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.

33. 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?

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.

34. You have been asked to evaluate a regression model based on R^2, adjusted R^2 and tolerance. What will be your criteria?

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 R^2 as opposed to R^2 to evaluate model fit because R^2 increases irrespective of improvement in prediction accuracy as we add more variables. But, adjusted R^2 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 R^2 because it varies between data sets. For example: a gene mutation data set might result in lower adjusted R^2 and still provide fairly good predictions, as compared to a stock market data where lower adjusted R^2 implies that model is not good.

35. I know that a linear regression model is generally evaluated using Adjusted R^2 or F value. How would you evaluate a logistic regression model?

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 R^2 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

36.When does regularization becomes necessary in Machine Learning?

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).

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

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

38.Name a few *Machine Learning libraries* for various purposes.

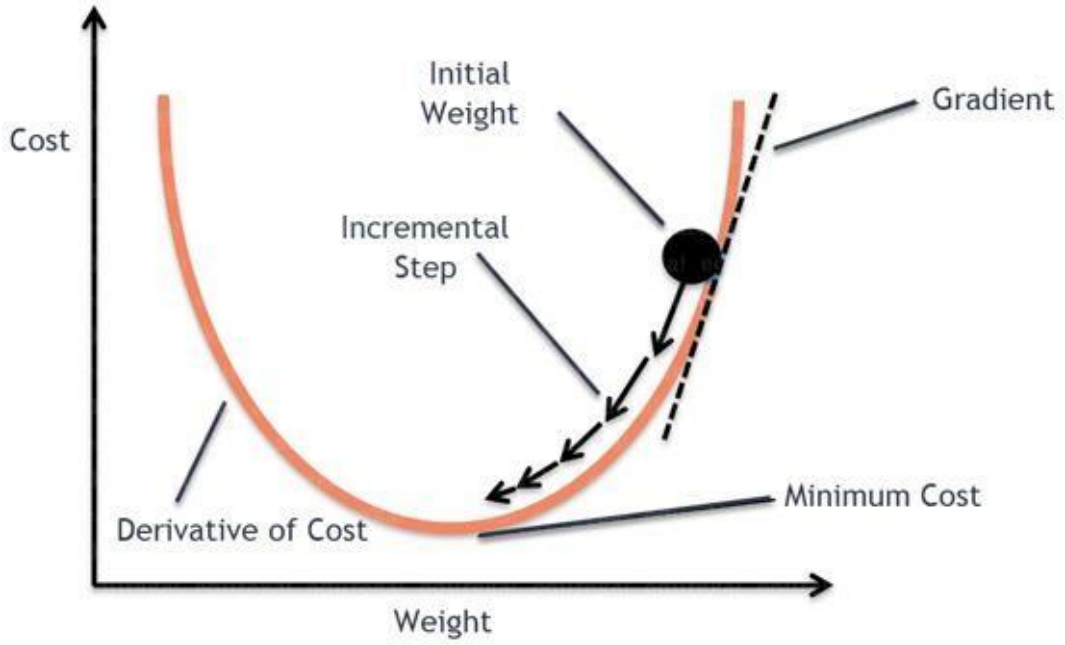
Purpose	Libraries
Scientific Computation	Numpy
Tabular Data	Pandas
Data Modelling & Preprocessing	Scikit Learn
Time-Series Analysis	Statsmodels
Text processing	Regular Expressions, NLTK
Deep Learning	Tensorflow, Pytorch

38. Explain Gradient Descent.

To Understand Gradient Descent, Let's understand what is a ***Gradient*** first.

A **gradient** measures how much the output of a function changes if you change the inputs a little bit. It simply measures the change in all weights with regard to the change in error. You can also think of a gradient as the slope of a function.

Gradient Descent can be thought of climbing down to the bottom of a valley, instead of climbing up a hill. This is because it is a minimization algorithm that minimizes a given function (**Activation Function**).



39. What is DBSCAN Clustering ?

DBSCAN is a **clustering** method that is used in machine learning to separate **clusters** of high density from **clusters** of low density

40.Various Methods to find out Optimal K Clusters:

A. Pair Plots
B. Elbow Method
C. Silhoutte Coefficient
D. Dendogram
E. Box and