Important Question and Answer

1. What is the optimization equation of GBDT?

Ans :- At the end of stage ‘k’ :-

1. Write the formulation of hinge loss?

Ans:- Linear – SVM [ hinge – loss + reg if asked is given below:-

L[w, b] = \* + C\* ( + b)]

1. What is the train time complexity of KNN ?

Ans:- if we use bruteforce then train time complexity is O(d \* n \* n)

And

If we use kd – tree then train time complexity is O(d \* n \* log(n)).

1. Which of these model are preferable when we have low complexity power ?

a. SVM

b. KNN

c. Linear Regressions

d. XGboost

Ans :- Training time for SVM is

Training time for KNN is O(d \* n \* n)

Training time for Linear Regression is O(n \* d)

Training time for XGboost is O[ n\*d \* M]

If we are solving regression problem then Linear regression is perfect choice.

1. What is Laplace smoothing ?

Ans :- Laplace smoothing is a smoothing technique that helps tackle the

problem of zero probability in the Naïve Bayes machine learning

algorithm.

1. How will you regularise your naive bayes model ?

Ans:- we can regularize the naïve bayes model with changing laplace

smoothing value .

if the is reduce then model undergoes overfitting and if the

value increase then model undergoes underfitting.

1. Can we solve dimensionality reduction with SGD?

Ans :- yes we can solve dimensionality reduction problem with SGD.

<https://medium.com/analytics-vidhya/dimensionality-reduction-by-stochastic-gradient-descent-f617ebde3c1b>

1. What is the optimization equation of Logistic Regression?

Ans : - first one log loss:-

w , b = argmin

Where = σ[ + b]

Second one is :-

w , b = argmin [1 + exp[- + b)]

1. How will you calculate the P(x/y=0) in case of gaussian naive baiyes ?

Ans : - first we separate the whole dataset into 0 and 1 class label . we

assume the numerical feature to be Gaussian naïve baiyes so , we

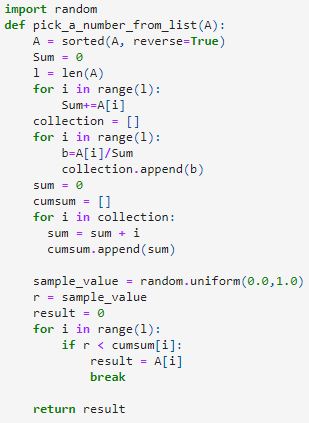
plot the pdf curve for 0 class for particular feature the we

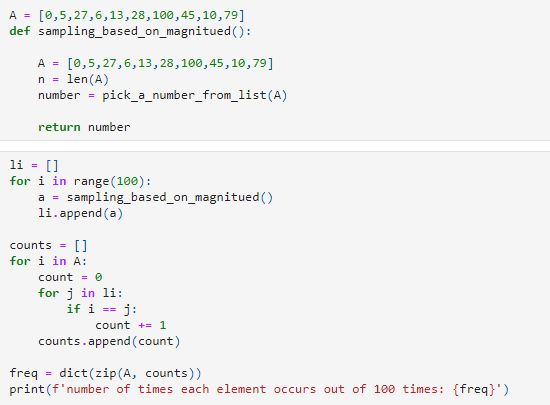
calculate likehood for P(x/y=0) where y is give class label and x

is events.

10. Write the code for proportional sampling.

Ans :-





11. What are hyperparameters in kernel svm ?

Ans: - if we are using rbf kernel then we have 2 hyperparameter i.e C

from soft form of svm and σ.

12. What are hyperparameters in SGD with hinge loss ?

Ans :- C attach with hinge loss. Additionally there are penalty, l1\_ratio.

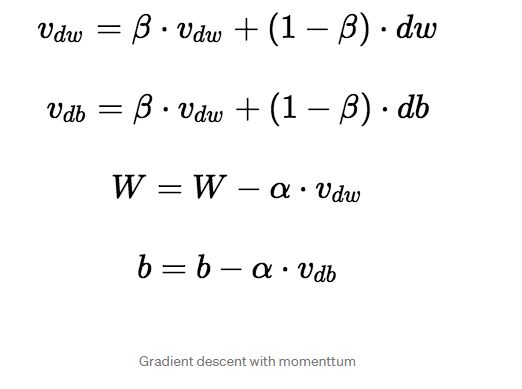
13. Difference between ADAM vs RMSPROP ?

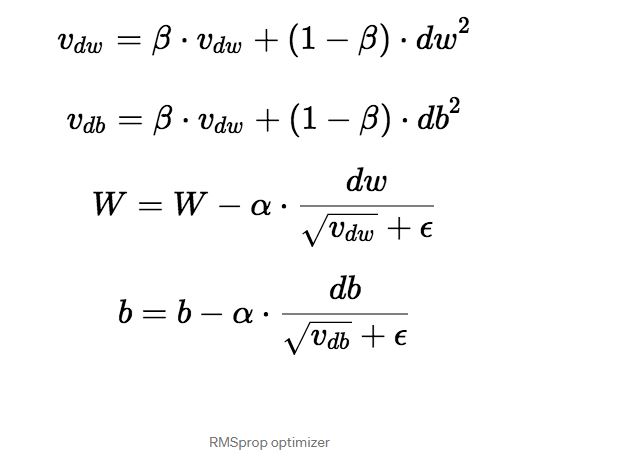
Ans :- RMSProp uses the second moment by with a decay rate to speed

up from AdaGrad. Adam uses both first and second moments, and

is generally the best choice.

RMSProp optimizer explained below:-





14. What is ADAM ?

Ans:- Adam is an optimization algorithm that can be used instead of the

classical stochastic gradient descent procedure to update network

weights iterative based in training data.

15. What is RELU? Is it differentiable ?

Ans :- it is not differential at 0 .

16. What is F1 score ?

Ans:- f1 score deals with both precision and recall. It is defined as the

harmonic mean of precision and recall and measures how low is

the false positives and low false negatives. f1 score 1 means all

predicted Y labels are the same as the true Y labels and 0 means

the model is a total failure. So apart from accuracy, it also deals

with how to labels are classified.

17. Name a few weight initialization techniques?

Ans:- they are :-

1. Xavier / Glovot initialization
2. He initialization

18. What is the batch Normalization layer?

Ans:- Batch-Normalization (BN) is an algorithmic method which makes

the training of Deep Neural Networks (DNN) faster and more

stable. It consists of normalizing activation vectors from hidden

layers using the first and the second statistical moments (mean and

variance) of the current batch.

19. Number of tunable parameters in the BN layer.

Ans:- While μ and σ² parameters are estimated from the input data, γ and

β are trainable. Thus, they can be leveraged by the

backpropagation algorithm to optimize the network.

20. What is convolution operation?

Ans:- In mathematics (in particular, functional analysis), convolution is

a mathematical operation on two functions (f and g) that produces

a third function ({\displaystyle f\*g}f\*g) that expresses how the

shape of one is modified by the other. The term convolution refers

to both the result function and to the process of computing it. It is

defined as the integral of the product of the two functions after one

is reversed and shifted. The integral is evaluated for all values of

shift, producing the convolution function.

21. Number of parameters in a convolution neural network given in architecture.

Ans:- filter size , number of kernel, stride and padding.

22. How do you get probabilities for RF classifier outputs.

Ans: - <https://www.linkedin.com/pulse/decoding-probabilities-random-forest-sanchit-tiwari/>

23. Is the Calibration classifier required to get probability values for logistic regression.?

Ans :- No, since sigmoid function is already a probabilistic interpretation

whose values lies between 0 to 1 . And also Logistic regression is

directly computing log loss for the model.

24.What kind of base learners are preferable in random forest classifiers ?

Ans:- The base learner which has high variance. Sometime dataset are

more sensitive in such a way that if we insert some new points

that would slightly change the model. So the base learner like

decision tree where depth is by default kept large. We train the

model until we find leaf node.

25. Difference between one vs rest and one vs one.

Ans: - Binary classification models like logistic regression and SVM do

not support multi-class classification natively and require meta-

strategies.

The One-vs-Rest strategy splits a multi-class classification into

one binary classification problem per class. For ‘n’ class we have

train ‘n’ model corresponding to each classes.

The One-vs-One strategy splits a multi-class classification into

one binary classification problem per each pair of classes. For ‘n’

classes we have to train (n \* (n – 1)) / 2.

So number of commputaion is more in one vs one than one vs

rest.

26. Kernal svm and linear svm ( SGD classifier with hinge loss). Which

has low latency and why.

Ans:- For training:

The sklearn SVM is computationally expensive compared to

sklearn SGD classifier with loss='hinge'. Hence we use SGD

classifier which is faster. This is good only for linear SVM. If we

are using 'rbf' kernel, then SGD is not suitable.

For testing:

Kernal svm :- O(k \* d)

Linear svm :- O(d)

So for low latency we would prefer Linear svm.

27. Why we need Calibration ?

Ans:- when we computing log loss , we need actual probabilities not the

approximate one. Example in Naïve bayes , we are just getting

approximate probalities . So calibration help us fix this problem.

We do not need calibration in logistic regression because it has

inbuilt probability function which is sigmoid.

28. What is MAP ? (mean average precision)

Ans:- https://www.youtube.com/watch?v=QdWidmgLwbw

29. Why do we need gated mechanism in LSTM ?

Ans:- LSTM ’s and GRU’s were created as the solution to short-term

memory. They have internal mechanisms called gates that can

regulate the flow of information. These gates can learn which data

in a sequence is important to keep or throw away. By doing that, it

can pass relevant information down the long chain of sequences to

make predictions. Almost all state of the art results based on

recurrent neural networks are achieved with these two networks.

When you read the review, your brain subconsciously only

remembers important keywords. You pick up words like “amazing”

and “perfectly balanced breakfast”. You don’t care much for words

like “this”, “gave“, “all”, “should”, etc. If a friend asks you the next

day what the review said, you probably wouldn’t remember it word for

word. You might remember the main points though like “will definitely

be buying again”. If you’re a lot like me, the other words will fade

away from memory.

And that is essentially what an LSTM or GRU does. It can learn to keep

only relevant information to make predictions, and forget non relevant

data. In this case, the words you remembered made you judge that it was

good.

30. What is stratified sampling ? Explain.

Ans:- In statistics, stratified sampling is a method of sampling from a

population which can be partitioned into subpopulations.

Stratified sampling example:-

In statistical surveys, when subpopulations within an overall

population vary, it could be advantageous to sample each

subpopulation (stratum) independently. Stratification is the

process of dividing members of the population into homogeneous

subgroups before sampling.

31. What will happen to train time of K means of data is very high dimension.

Ans:- since the training time complexity of K – means is O(n \* k \* d \* i).

Where n is number of data points , is number of cluster, d is dimension and i is number of iteration.

If we assume k and i to be small , then time complexity depend on n \* d,

In case of high dimensionality , it would affect the training time drastically.

32.If you have 10mill records with 100dimension each for a clustering task. Which algorithm will you try first and why ?

Ans:- need to answer.

33. Which algorithm will give high time complexity if you have 10million records for a clustering task.

Ans:- need to answer.

34. Difference between GD and SGD.

Ans: - Gradient Descent is an iterative method to solve the optimization

problem. There is no concept of "epoch" or "batch" in classical

gradient decent. The key of gradient decent are

1. Update the weights by the gradient direction.
2. The gradient is calculated precisely from all the data points.

Stochastic Gradient Descent can be explained as: quick and dirty

way to "approximate gradient" from one single data point. If we

relax on this "one single data point" to "a subset of data", then the

concepts of batch and epoch come.

35.Why do we need repetitive training of a model ?

Ans:- Once a model is trained and you get new data which can be used

for training, you can load the previous model and train onto it. For

example, you can save your model as a .pickle file and load it and

train further onto it when new data is available. Do note that for the

model to predict correctly, the new training data should have a

similar distribution as the past data.

Predictions tend to degrade based on the dataset you are using. For

example, if you are trying to train using twitter data and you have

collected data regarding a product which is widely tweeted that day.

But if you use use tweets after some days when that product is not

even discussed, it might be biased. The frequency will be dependent

on dataset and there is no specific time to state as such. If you

observe that your new incoming data is deviating vastly, then it is a

good practise to retrain the model.

36.How do you evaluate the model after productionization ?

OR

What are the common techniques, model stability tests, model performance tests, metrics after deployment? How to decide when to replace current model with newer one?

so to be more clear lets consider the problem of loan default prediction. Let's say I have trained and tested off-line multiple classifiers and ensembled them. Then I gave this model to production.

But because people change, data and many other factors change as well. And performance of our model eventually will decrease. So then it needs to be replaced with the new, better model.

Ans:- What you should consider more often in production scenario is revenue for your model , and A/B test is a must .

As in your case , you can exactly measure how much money can your model for loan default prediction bring to you , or how much loss can your model can save for you .

Besides , you can check if the distribution of your prediction is consistent with that of ground truth concerning accuracy and stability for your model

.

37. How to do multi-class classification with random forest?

Ans:- since we use high variance model in random forest like decision tree with high depth until we find leaf node. We solve it by one vs rest. We train k model for k class label. For each model using mojorty vote we can come to conclusion .

38. Assume We have very high dimension data. Which model will you try and which model will be better in a classification problem.

Ans :- for high dimension , we cannot use KNN model because it would create curse of dimensionality.

For naïve bayes, we have order of (d \* c) where d is dimension and c number of class label.

For logistic regression , we have order of d at run time.

For svm , we have oder of k \* d where k is number of support vector and d is dimension.

39.What is AUC?

Ans:-AUC - ROC curve is a performance measurement for the

classification problems at various threshold settings. ROC is a

probability curve and AUC represents the degree or measure

of separability. It tells how much the model is capable of

distinguishing between classes. Higher the AUC, the better the

model is at predicting 0s as 0s and 1s as 1s. By analogy, the

Higher the AUC, the better the model is at distinguishing

between patients with the disease and no disease.

40. Difference between micro average F1 and macro average F1 for a 3 class classification.

Ans:- need to answer.

41.Difference between AUC and accuracy ?

Ans:- AUC (based on ROC) and overall accuracy seems not the same

concept.

Overall accuracy is based on one specific cutpoint, while ROC tries

all of the cutpoint and plots the sensitivity and specificity. So when

we compare the overall accuracy, we are comparing the accuracy

based on some cutpoint. The overall accuracy varies from different

cutpoint.

42. What is pearson correlation coefficient ?

Ans:- In statistics, the Pearson correlation coefficient (PCC, pronounced

/ˈpɪərsən/), also referred to as Pearson's r, the Pearson product-

moment correlation coefficient (PPMCC), or the bivariate

correlation,[1] is a measure of linear correlation between two sets of

data.

43. Training time complexity of naive bayes ?

Ans:- O(n\*d\*c)

44.How to calculate the number of parameters in the CNN?

Ans:- <https://medium.com/@iamvarman/how-to-calculate-the-number-of-parameters-in-the-cnn-5bd55364d7ca>

45. Number of tunable parameters in embedding layer (36, vocab size = 75)

Ans:- 75\*36 I think so.

46. Which is faster

a. SVC(C=1). Fit(x,y)

b. SGD(Log=hinge).fit(x,y)

Ans:- need to answer

47.Explain about KS test ?

Ans:- In statistics, the Kolmogorov–Smirnov test (K–S test or KS test) is

a nonparametric test of the equality of continuous (or

discontinuous, see Section 2.2), one-dimensional probability

distributions that can be used to compare a sample with a reference

probability distribution (one-sample K–S test).

48.What is KL divergence ?

Ans:- In mathematical statistics, the Kullback–Leibler divergence,

{\displaystyle D\_{\text{KL}}}{\displaystyle D\_{\text{KL}}}

(also called relative entropy), is a measure of how one probability

distribution is different from a second, reference probability

distribution.

49.What is the need of confidence interval ?

Ans:- confidence interval tells you more than just the possible range

around the estimate. It also tells you about how stable the estimate

is. A stable estimate is one that would be close to the same value if

the survey were repeated.

50.Can you name a few sorting algorithms and their complexity ?

Ans:- Merge sort –

Best, average and worst case time complexity: nlogn which is

independent of distribution of data.

Bubble sort and Insertion sort –

Average and worst case time complexity: n^2

Best case time complexity: n when array is already sorted.

Worst case: when the array is reverse sorted.

51.What is percentile ?

Ans:- In statistics, a percentile (or a centile) is a score below which a given

percentage of scores in its frequency distribution falls (exclusive

definition) or a score at or below which a given percentage falls

(inclusive definition).

52.What is IQR ?

Ans:- To find the interquartile range (IQR), ​first find the median (middle

value) of the lower and upper half of the data. These values are

quartile 1 (Q1) and quartile 3 (Q3). The IQR is the difference

between Q3 and Q1.

53. Can you explain the dict.get() function ?

Ans:- The get() method returns the value of the item with the specified key.

54.What is parameter sharing in deep learning?

Ans:- To reiterate parameter sharing occurs when a feature map is

generated from the result of the convolution between a filter and

input data from a unit within a plane in the conv layer. All units

within this layer plane share the same weights; hence it is called

weight/parameter sharing.

55. what is sample with replacement and without replacement?

Ans:- <https://www.statisticshowto.com/sampling-with-replacement-without/>

56. how to find feature importance in decision tree?

Ans:- in decision tree , to compute feat. Importance. We see where we

have used in decision tree. Then we take weight sum of sort. The

overall reduction in entropy or gini impurity because of this feature

at various level in decision tree.

Detail :- <https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3#:~:text=Feature%20importance%20is%20calculated%20as,the%20more%20important%20the%20feature>.

Another:- <https://sefiks.com/2020/04/06/feature-importance-in-decision-trees/>

57.What are the assumption of NB ?

Ans:- the assumption is as follows:-

Each feature is conditionally independent of every other feature

for j != i , given the category.

58.What are the assumptions of KNN ?

Ans: - The KNN algorithm assumes that similar things exist in close

proximity. In other words, similar things are near to each other.

59.What are the assumptions of linear regression ?

Ans:-

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.
2. The residue is Gaussian distributed with N(0,)
3. For each value of x i.e and , € will be independent and identically distributed.
4. Homoscedasticity: The residuals have constant variance at every level of x.

60.What is time complexity of building KD tree ?

Ans:- the time to build decision tree and KD tree would be the same . in

decision tree we first the list and then create binary tree. The same

way we do in KD tree . so the time complexity for Building the KD

tree would be O(n \* log(n) \* d ) where d is dimension. So basically

term n \* log(n) is time complexity for sorting the list using merge

sort concept.

61.What is the time complexity to check if a number is prime or not ?

Ans: - <https://softwareengineering.stackexchange.com/questions/197374/what-is-the-time-complexity-of-the-algorithm-to-check-if-a-number-is-prime#:~:text=On%20the%20interval%20%5B0%2Cn,ln(n)%20prime%20numbers.&text=The%20maximum%20execution%20time%20of,of%20two%20large%20prime%20numbers>.

62. Explain the back propagation mechanism in dropout layers ?

Ans:- Yes, the neurons are considered zero during backpropagation as well. Otherwise dropout wouldn't do anything! Remember that forward propagation during training is only used to set up the network for backpropagation, where the network is actually modified (as well as for tracking training error and such).

In general, it's important to account for anything that you're doing in the forward step in the backward step as well – otherwise you're computing a gradient of a different function than you're evaluating.

The way it's implemented in Caffe, for example, is (as can be verified from the source):

In forward propagation, inputs are set to zero with probability p, and otherwise scaled up by .

In backward propagation, gradients for the same dropped units are zeroed out; other gradients are scaled up by the same .

63.Explain the loss function used in auto encoders assuming the network accepts images ?

Ans:- need to answer.

64. Numbers of tunable parameters in dropout layer ?

Ans:- need to answer.

65.What happens if we do not normalize our dataset before performing classification using KNN algorithm.

Ans:- We do not want our algorithm to be affected by the magnitude of feature variables. The algorithm should not be biased towards variables with higher magnitude. To overcome this problem, we can bring down all the variables to the same scale.

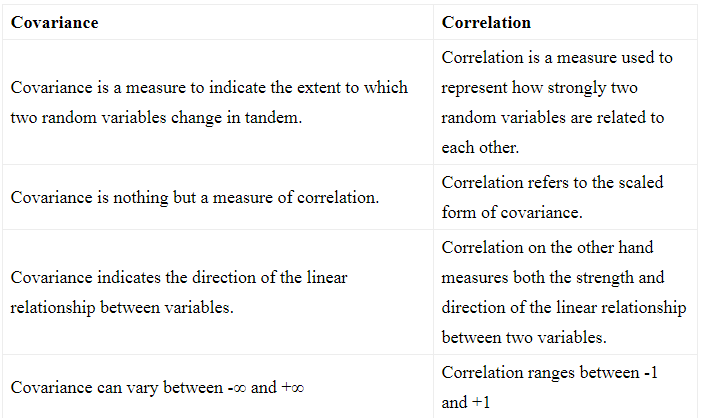
Fordetail:- <https://medium.com/analytics-vidhya/why-is-scaling-required-in-knn-and-k-means-8129e4d88ed7>

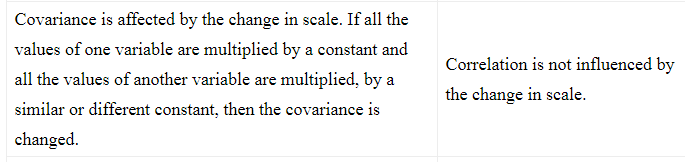
66.What is standard normal variate ?

Ans:- if random variable are normally distributed , mean of variable is at origin , and standard deviation of variable is 1, then it is called standard normal variate.

So if for any random variable if the mean is 0 and std-dev is 1. Then is normally distributed.

67.What is the significance of covariance and correlation and in what cases can we not use correlation



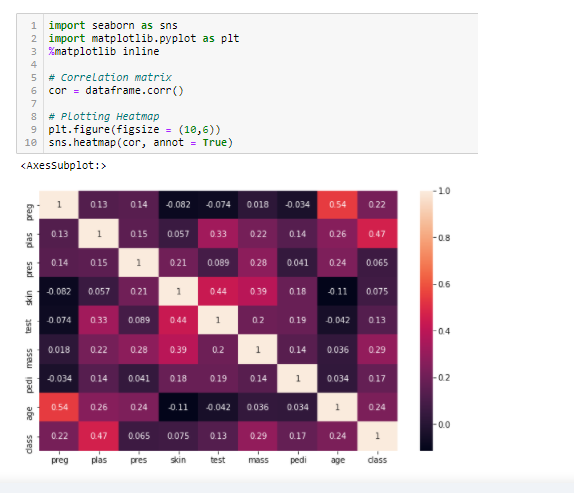


Co-relation does good work for linear relation. What if two variable follows the monotonicity regardless of linear pattern , then co-relation would not work well.

**Correlation Coefficient**

Correlation is a measure of the linear relationship of 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that the good variables are highly correlated with the target. Furthermore, variables should be correlated with the target but should be uncorrelated among themselves.

If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only really needs one of them, as the second one does not add additional information. We will use the Pearson Correlation here.



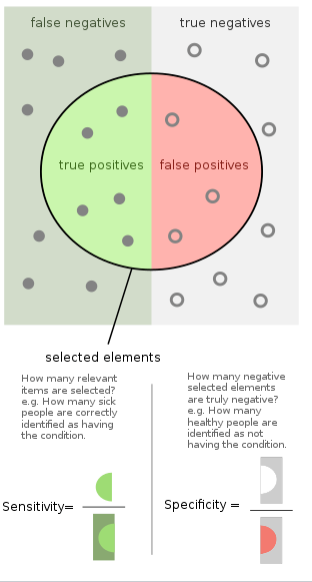
We need to set an absolute value, say 0.5 as the threshold for selecting the variables. If we find that the predictor variables are correlated among themselves, we can drop the variable which has a lower correlation coefficient value with the target variable. We can also compute multiple correlation coefficients to check whether more than two variables are correlated to each other. This phenomenon is known as multicollinearity.

68.When should we choose PCA over t-sne.

Ans:- when we want to visualize the data , t-sne is faborable, but for dimensionality reduction PCA works better.

69.What is the relationship between specificity and false positive rate

Ans:-



70.What are the challenges with time based splitting? How to check whether the train / test split will work or not for given distribution of data ?

Ans:- need to answer

71.How does outlies effect the performance of a mode.

Ans:- since KNN use the distance measure based model, outlier has impact on model.

Hyperplane is drastically change in linear regression due to outlier.

Logistic regression has less impact of outlier since it squash the outlier to one.

Outlier also impact the naïve bayes model because for few points we have to calculate likelihood.

Outlier has impact on decision tree .

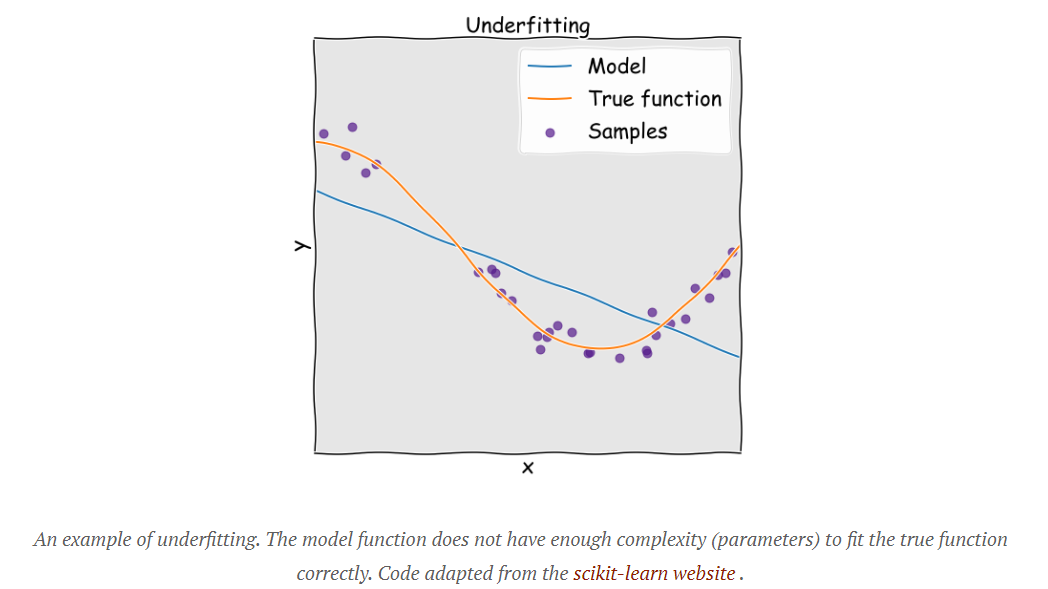
72.What is the need of encoding categorical or ordinal features

Ans:- need to answer.

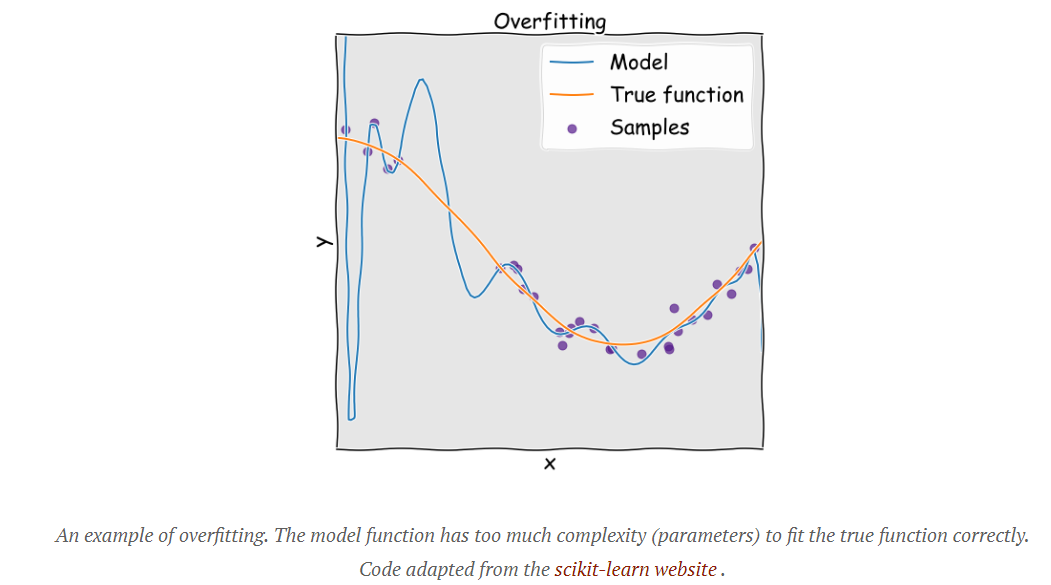
73.What is the intuition behind bias-variance tradeoff ?

Ans:- If we have an underfitted model, this means that we do not have enough parameters to capture the trends in the underlying system. Imagine for example that we have data that is parabolic in nature, but we try to fit this with a linear function, with just one parameter. Because the function does not have the required complexity to fit the data (two parameters), we end up with a poor predictor. In this case the model will have high bias. This means that we will get consistent answers, but consistently wrong answers.

Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data.



If we have overfitted, this means that we have too many parameters to be justified by the actual underlying data and therefore build an overly complex model. Again imagine that the true system is a parabola, but we used a higher order polynomial to fit to it. Because we have natural noise in the data used to fit (deviations from the perfect parabola), the overly complex model treats these fluctuations and noise as if they were intrinsic properties of the system and attempts to fit to them. The result is a model that has high variance. This means that we will not get consistent predictions of future results. For a striking and devastating example of the dangers of overfitting, see this excellent article which includes a section on the Fukushima disaster.



74.What does it mean by FPR = TPR = 1 of a model.

Ans:- this means model is biased toward positive points or toward one complete class label.

75.When should we use log loss .

Ans:- Log-loss is an appropriate performance measure when you're model output is the probability of a binary outcome.

The log-loss measure considers confidence of the prediction when assessing how to penalize incorrect classification. For instance consider two predictions of an outcome P(Y=1|X), where the predictions are 0.51 and 0.99 respectively. In the former case the model is only slightly confident of the class prediction (assuming a 0.5 cutoff), while in the latter it is extremely confident. Since in our case both are wrong, the penalty will be more harsh for the more confident (but incorrect) prediction by employing a log-loss penalty.

76.What performance metric does t-sne use to optimize its probabilistic function.

Ans:- need to answer.

77.When to use cosine similarity over euclidean distance.

Ans:- when dataset is of high dimension, we should use cosine similarity over euclidean distance.

78.What is fit, transform and fit transform in terms of BOW,tf-idf,word2vector

Ans:- need to answer.

79. How do we quantify uncertainty in probability class labels when using KNN model for classifications.

Ans:- need to answer.

80. How do we identify whether the distribution of my train and test is similar or not.

Ans:- we we first label train one class label and test data as another class label. In this way it will be classification problem. Then we can apply any of classification model like KNN , if model is predicting well in the given data , the distribution of test is differ from train.

81. What do you mean by hard margin SVM?

Ans:- hard margin SVM means for a given point if + b is greater than 1 or -1 .

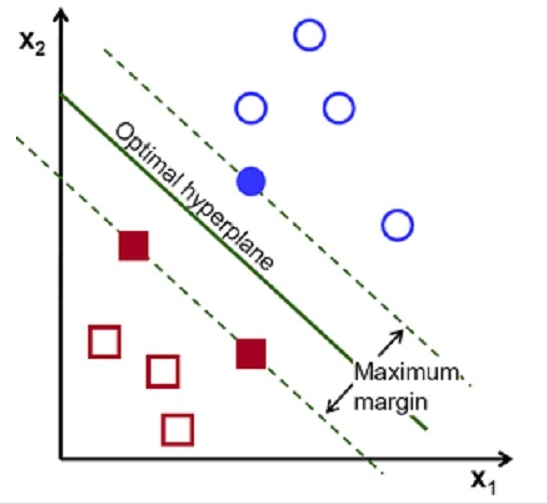
Or while training model , support vector are constructed in such a way that if + b) is greater than 1 then points ae correctly classify assuming that there is no points lies between the two support vector.

82.What is kernel function in svm ?

Ans:- In machine learning, a “kernel” is usually used to refer to the kernel trick, a method of using a linear classifier to solve a non-linear problem. It entails transforming linearly inseparable data like to linearly separable ones. The kernel function is what is applied on each data instance to map the original non-linear observations into a higher-dimensional space in which they become separable.

83.Why do we call an svm a maximum margin classifier ?

Ans:-



SVM is a type of classifier which classifies positive and negative examples, here blue and red data points

As shown in the image, the largest margin is found in order to avoid overfitting ie,.. the optimal hyperplane is at the maximum distance from the positive and negative examples(Equal distant from the boundary lines).

To satisfy this constraint, and also to classify the data points accurately, the margin is maximised, that is why this is called the large margin classifier.

84.Is svm affected by outliers ?

Ans:- Despite its popularity, SVM has a serious drawback, that is sensitivity to outliers in training samples. The penalty on misclassification is defined by a convex loss called the hinge loss, and the unboundedness of the convex loss causes the sensitivity to outliers.

But this condition is only true outlier are in opposite sides or else it does not get affected.

85.What is locality sensitive hashing ?

Ans:- locality-sensitive hashing (LSH) is an algorithmic technique that hashes similar input items into the same "buckets" with high probability.

86. Instead of sigmoid function can we use any other function in LR?

Ans:- yes , but it should satisfy two condition i.e it should give probabilistic interpretation and it should squash the outlier effect to minimum which sigmoid does well.

87. Can linear regression be used for classification purpose ?

Ans:- There are two things that explain why Linear Regression is not suitable for classification. The first one is that Linear Regression deals with continuous values whereas classification problems mandate discrete values. The second problem is regarding the shift in threshold value when new data points are added.

88. What is the use of ROC curve ?

Ans:- ROC curves are frequently used to show in a graphical way the connection/trade-off between clinical sensitivity and specificity for every possible cut-off for a test or a combination of tests. In addition the area under the ROC curve gives an idea about the benefit of using the test(s) in question.

As the area under an ROC curve is a measure of the usefulness of a test in general, where a greater area means a more useful test, the areas under ROC curves are used to compare the usefulness of tests.

For more details:- <https://acutecaretesting.org/en/articles/roc-curves-what-are-they-and-how-are-they-used#:~:text=ROC%20curves%20are%20frequently%20used,test(s)%20in%20question>.

89. When EDA should be performed, before or after splitting data? Why ?

Ans:- We should clearly understand the data before splitting. Clearly understanding the data helps in knowing which attributes to consider for building the model

Another reason to do EDA before model building is, as you have mentioned we have treat the missing, outliers in the data which would highly influence the model.

EDA includes pre-processing of data like missing values,outliers,scaling changing variables,etc after that only do spiltting.

90. What is type 1 & type 2 error ?

Ans:- In statistical hypothesis testing, a type I error is the rejection of a true null hypothesis (also known as a "false positive" finding or conclusion; example: "an innocent person is convicted"), while a type II error is the non-rejection of a false null hypothesis (also known as a "false negative" finding or conclusion; example: "a guilty person is not convicted")

91. What is multicollinearity ?

Ans:- Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.

92. How is eigenvector different from other general vectors ?

Ans:- An eigenvector is a vector whose direction remains unchanged when a linear transformation is applied to it. This unique, deterministic relation is exactly the reason that those vectors are called 'eigenvectors' (Eigen means 'specific' in German)

<https://www.visiondummy.com/2014/03/eigenvalues-eigenvectors/#:~:text=Eigenvectors%20(red)%20do%20not%20change,Other%20vectors%20(yellow)%20do.&text=the%20square%20matrix-,.,'specific'%20in%20German)>.

93.What is eigenvalue & eigenvectors ?

Ans:- In linear algebra, an eigenvector (/ˈaɪɡənˌvɛktər/) or characteristic vector of a linear transformation is a nonzero vector that changes at most by a scalar factor when that linear transformation is applied to it. The corresponding eigenvalue, often denoted by {\displaystyle \lambda }\lambda ,[1] is the factor by which the eigenvector is scaled.

Geometrically, an eigenvector, corresponding to a real nonzero eigenvalue, points in a direction in which it is stretched by the transformation and the eigenvalue is the factor by which it is stretched. If the eigenvalue is negative, the direction is reversed.[2] Loosely speaking, in a multidimensional vector space, the eigenvector is not rotated.

94.What is response encoding of categorical features ?

Ans:- later

95. What is the binning of continuous random variables.

Ans:-need to answer

96.Regularization parameter in dual form of SVM ?

Ans:- As C increase , number of support vector increase which leads to underfit and as C decrease , number of suppot vector decrease which leads to ovefit .

97. On what basis would you choose agglomerative clustering over k means clustering and vice versa ?

Ans:- I would say hierarchical clustering is usually preferable, as it is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.

With k-Means clustering, you need to have a sense ahead-of-time what your desired number of clusters is (this is the 'k' value). Also, k-means will often give unintuitive results if (a) your data is not well-separated into sphere-like clusters, (b) you pick a 'k' not well-suited to the shape of your data, i.e. you pick a value too high or too low, or (c) you have weird initial values for your cluster centroids (one strategy is to run a bunch of k-means algorithms with random starting centroids and take some common clustering result as the final result).

In contrast, hierarchical clustering has fewer assumptions about the distribution of your data - the only requirement (which k-means also shares) is that a distance can be calculated each pair of data points. Hierarchical clustering typically 'joins' nearby points into a cluster, and then successively adds nearby points to the nearest group. You end up with a 'dendrogram', or a sort of connectivity plot. You can use that plot to decide after the fact of how many clusters your data has, by cutting the dendrogram at different heights. Of course, if you need to pre-decide how many clusters you want (based on some sort of business need) you can do that too. Hierarchical clustering can be more computationally expensive but usually produces more intuitive results.

98. What is the difference between model parameters and hyper parameters

Ans:- model parameters are estimated from data automatically and model hyperparameters are set manually and are used in processes to help estimate model parameters

99. Number of parameters in LSTM is 4m(m+n+1). How many number of parameters do we have in GRU ?

Ans:- need to answer

100. What is box cox transform?

Ans:- A Box Cox transformation is a transformation of a non-normal dependent variables into a normal shape.

Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests.

101. what is stemming?

Ans:- Stemming is the process of producing morphological variants of a root/base word. Stemming programs are commonly referred to as stemming algorithms or stemmers. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”.

There are different types of stemming available . few are :-

**Porter’s Stemmer algorithm :-**

It is one of the most popular stemming methods proposed in 1980. It is based on the idea that the suffixes in the English language are made up of a combination of smaller and simpler suffixes. This stemmer is known for its speed and simplicity. The main applications of Porter Stemmer include data mining and Information retrieval. However, its applications are only limited to English words. Also, the group of stems is mapped on to the same stem and the output stem is not necessarily a meaningful word. The algorithms are fairly lengthy in nature and are known to be the oldest stemmer.

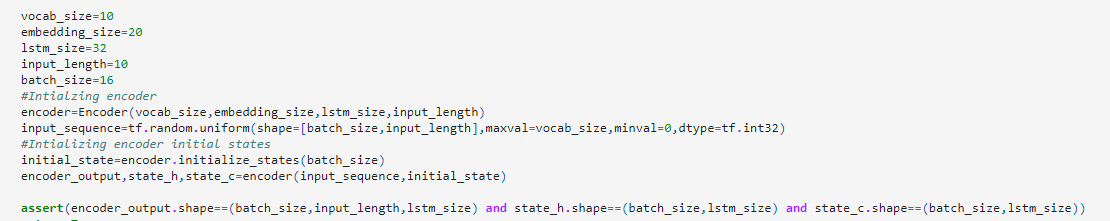
Example: EED -> EE means “if the word has at least one vowel and consonant plus EED ending, change the ending to EE” as ‘agreed’ becomes ‘agree’.

**Snowball Stemmer:-**

When compared to the Porter Stemmer, the Snowball Stemmer can map non-English words too. Since it supports other languages the Snowball Stemmers can be called a multi-lingual stemmer. The Snowball stemmers are also imported from the nltk package. This stemmer is based on a programming language called ‘Snowball’ that processes small strings and is the most widely used stemmer. The Snowball stemmer is way more aggressive than Porter Stemmer and is also referred to as Porter2 Stemmer. Because of the improvements added when compared to the Porter Stemmer, the Snowball stemmer is having greater computational speed.

102. What happens when we set return sequence = true in LSTM ?

Ans:-



103. What happens during the fit and transform of following modules ?

a. Standard scaler

b. Count vectorizer

c. PCA

d. tf-idf

Ans:-

1. Mean and variance
2. All unique words in a series with most most frequent
3. Eigen value and Eigen vector
4. Idf values

104. Can we use t-sne for transforming test data ? if not why ?

Ans:- no , because it does not have any function which can help transform test data. It uses probabilistic approach to embed data point.

105. Difference between categorical cross entropy and binary cross entropy.

Ans:- need to answer

106. How to you use weighted distance in content based recommendation ?

Ams:-need to answer

107. What is the time complexity of SVD decomposition ?

Ans:- need to answer

108. What is the difference between content based recommendation and collaborative recommendation ?

Ans:-

**Content based filtering** –

The point of content-based filtering system is to know the content of both user and item. Usually it constructs and then compare user-profile and item-profile using the content of shared attribute space. For example, for a movie, you represent it with the movie stars in it and the genres (using a binary coding for example).

For user profile, you can do the same thing based on the users likes some movie stars/genres etc

**Collaborative filtering –**

Collaborative algorithm uses “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 behavior and preferences over the items are used to recommend items to the new users. In this case, features of the items are not known.

109. Which of these layers will be a better option as a last layer in multilabel classification ?

a. Sigmoid

b. Softmax

Ans:- a

110. What is PDF, can we calculate PDF for discrete distribution ?

Ans:- no.

111. How do we interpret alpha in dual form of sum? What is the relation between C and Alpha?

Ans:- need to answer

112. What is the parameter sigma in svm?

Ans:- The gamma parameter in the RBF kernel determines the reach of a single training instance. If the value of Gamma is low, then every training instance will have a far reach. Conversely, high values of gamma mean that training instances will have a close reach. So, with a high value of gamma, the SVM decision boundary will simply be dependent on just the points that are closest to the decision boundary, effectively ignoring points that are farther away. In comparison, a low value of gamma will result in a decision boundary that will consider points that are further from it. As a result, high values of gamma typically produce highly flexed decision boundaries, and low values of gamma often results in a decision boundary that is more linear.