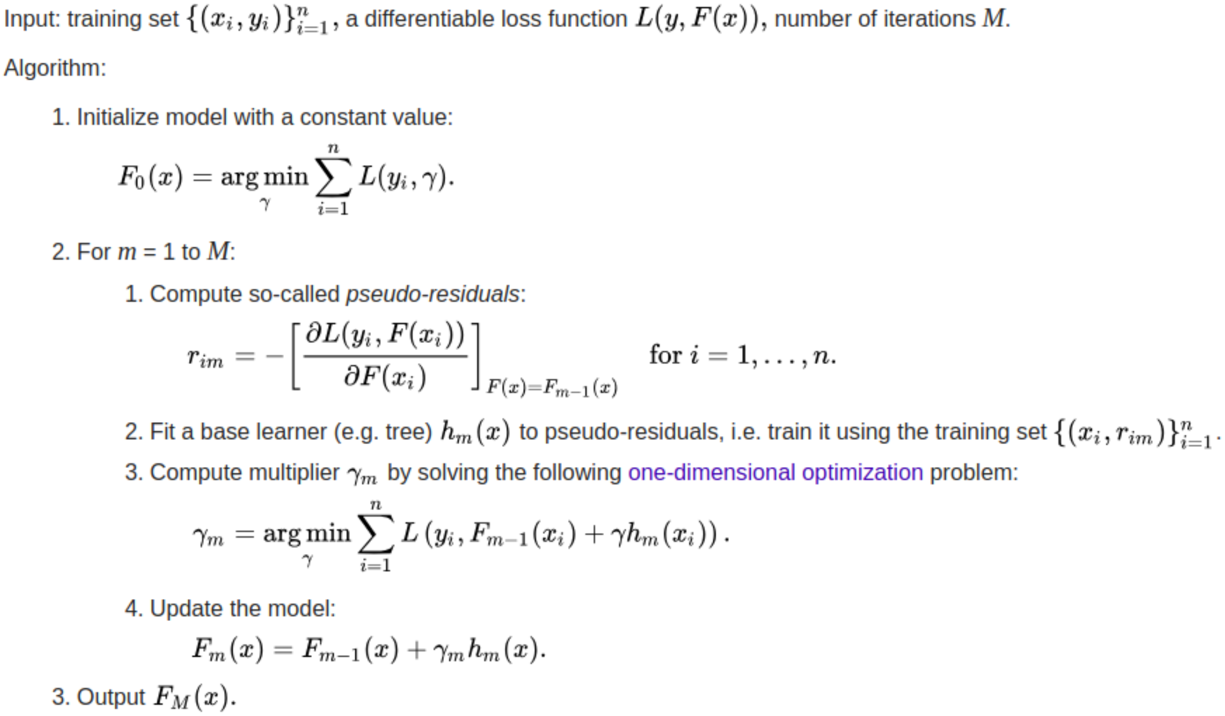
1. What is the optimization equation of GBDT ?
   ANS :



Input: training set $\{(x_i, y_i)\}_{i=1}^{n}$, a differentiable loss function $L(y, F(x))$, number of iterations M.

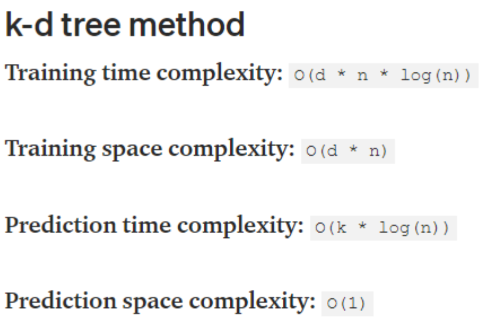
Algorithm:

```
1. Initialize model with a constant value:
       $F_0(x) = \underset{\gamma}{\arg\min} \sum_{i=1}^{n} L(y_i, \gamma).$
2. For m = 1 to M:
    1. Compute so-called pseudo-residuals:
           $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$
    2. Fit a base learner (e.g. tree) $h_m(x)$ to pseudo-residuals, i.e. train it using the training set $\{(x_i, r_{im})\}_{i=1}^{n}$.
    3. Compute multiplier $\gamma_m$ by solving the following one-dimensional optimization problem:
           $\gamma_m = \underset{\gamma}{\arg\min} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right).$
    4. Update the model:
           $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$
3. Output $F_M(x)$.
```

2. Write the formulation of hinge loss ?
   ANS:

$$\min_{\beta \in \mathbb{R}^d} \; F(\beta) := \frac{1}{n} \sum_{i=1}^{n} \left(\max(0, 1 - y_i x_i^T \beta)\right)^2 \; + \lambda \|\beta\|_2^2 \, .$$

3. What is the train time complexity of KNN ?
   ANS:



k-d tree method

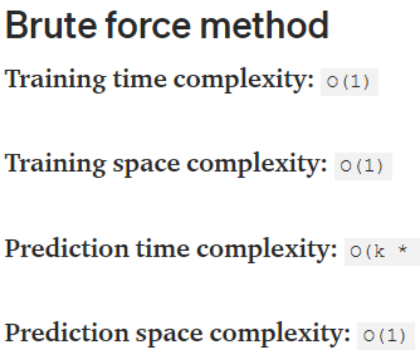
Training time complexity: `O(d * n * log(n))`

Training space complexity: `O(d * n)`

Prediction time complexity: `O(k * log(n))`

Prediction space complexity: `O(1)`

4. What is the Test time complexity of KNN in brute force ?
   ANS: In Brute force there is not training , calculations happen only during prediction time.i.e, We calculate distance to each training point and obtain k-nearest neighbours.



Brute force method

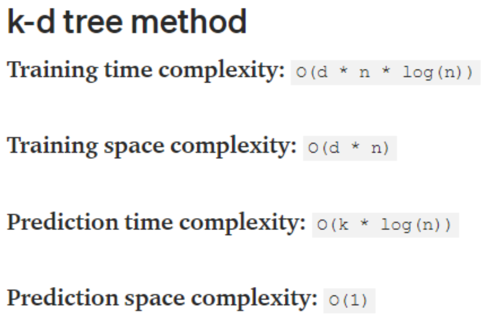
Training time complexity: `O(1)`

Training space complexity: `O(1)`

Prediction time complexity: `O(k * n * d)`

Prediction space complexity: `O(1)`

5. What is the test time complexity of KNN if we use kd-tree ?
   ANS:



k-d tree method

Training time complexity: `O(d * n * log(n))`

Training space complexity: `O(d * n)`

Prediction time complexity: `O(k * log(n))`

Prediction space complexity: `O(1)`

6. How will you regularise the KNN model ?
   ANS: k acts as regularize in KNN model.

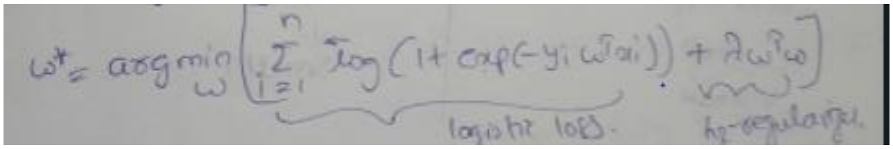
7. Which of these model are preferable when we have low complexity power ?
   a. SVM
   b. KNN
   c. Linear Regressions
   d. Xgboost

   ANS: SVM-O(n2) train time and O(kd) at test time k=support vectors
   KNN-O(nd)
   Linear Regression- O(d)
   We choose linear regression.
8. What is Laplace smoothing ?
   ANS : In naïve bayes if in the test data a category was not present in train data to avoid Problem of zero probability we use Laplace smoothing. We add a small value (alpha) in Numerator and k*alpha in the denominator where k is the number of values that a feature take.
9. How will you regularize your naive bayes model ?
   ANS : alpha in Laplace smoothing will act as regularizer.
   As alpha increases the likelihood probabilities will have uniform distribution i.e. P(0)=0.5 and P(1)=0.5 thus the prediction will be biased towards larger class.
   As alpha decreases even with the small change in the data will affect the probability Thus, it leads to overfitting.

10. Can we solve dimensionality reduction with SGD?
    ANS : we can use SGD.It is because dimensionality reduction can be posed as an optimization problem where we try to reduce the loss function.

    $(||xi - xj||^2 - Xij)^2$ here xi and xj are points in higher dimesion and xij is the distance between the points in lower dimension thus we prefer the neighborhood distance in the lower dimensions with minimal losses.

11. Which of these will be doing more computations GD or SGD ?
    ANS : GD because Gradient descent considers all the points in the training data for update equation.
12. If A is a matrix of size (3,3) and B is a matrix of size (3,4) how many numbers of multiplications that can happen in the operations A*B ?
    ANS : 36
13. What is the optimization equation of Logistic Regression ?
    ANS :



$$w^* = \underset{w}{\arg\min} \left[\underbrace{\sum_{i=1}^{n} \log(1+\exp(-y_i w^T x_i))}_{\text{logistic loss.}} + \underbrace{\lambda w^T w}_{L_2\text{-regularizer.}} \right]$$

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

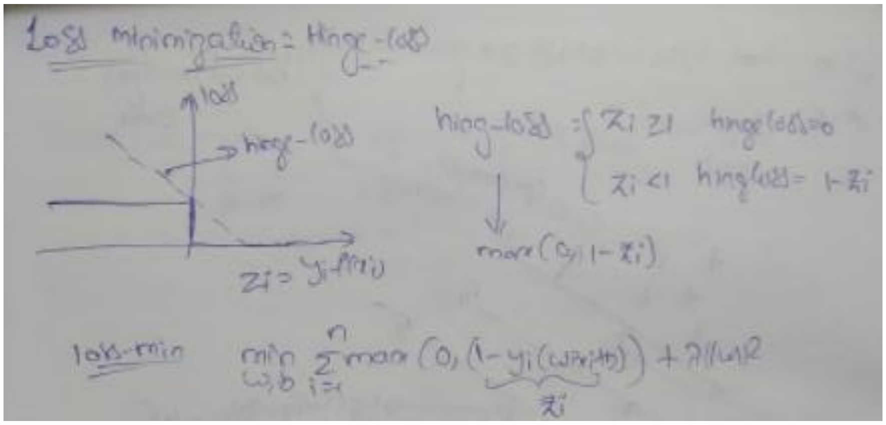
$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

μ_y is the mean of all the values in x_i associated with y.

15. Write the code for proportional sampling.
    ANS :
    - normailise all the values i.e the range will be (0,1)
    - calculate the cummulative sum .
    - Sample one ramdom value from unifrom distribution of (0,1)
    - If the random value <= cummulative sum return the number

16. What are hyperparameters in kernel svm ?
    ANS:
    Depends on the kernel we use
    Example : In RBF kernel we have σ as hyperpameter.

17. What are hyperparameters in SGD with hinge loss ?
    ANS :



Loss minimization = Hinge-loss

loss
hinge-loss
$z_i = y_i f(x_i)$

hing-loss $= \begin{cases} z_i \geq 1 & \text{hinge loss} = 0 \\ z_i < 1 & \text{hing loss} = 1 - z_i \end{cases}$

$\downarrow$

$\max(0, 1 - z_i)$

loss-min $\quad \min_{w,b} \sum_{i=1}^{n} \max\big(0, (\underbrace{1 - y_i(w^T x_i + b)}_{z_i})\big) + \lambda \|w\|^2$

λ is the hyperprameter.

18. Is hinge loss differentiable if not how we will modify it so that you apply SGD ?
    ANS : Hinge loss is not differentiable at 1 as it is not continuous at 1. So, we use squared hinge loss in optimizations. Or we can use smooth approximation for hinge loss as below.

$$H_s(x) = \begin{cases} \frac{1}{2} - x & x \leq 0, \\ \frac{1}{2}(1-x)^2 & 0 < x < 1 \\ 0 & x \geq 1 \end{cases}$$

19. Difference between ADAM vs RMSPROP ?
    ANS :
    RMSPROP:

$$For\ each\ Parameter\ w^j$$

$(j\ subscript\ dropped\ for\ clarity)$

$$\nu_t = \rho \nu_{t-1} + (1 - \rho) * g_t^2$$

$$\Delta \omega_t = -\frac{\eta}{\sqrt{\nu_t + \epsilon}} * g_t$$

$$\omega_{t+1} = \omega_t + \Delta \omega_t$$

$\eta : Initial\ Learning\ rate$

$\nu_t : Exponential\ Average\ of\ squares\ of\ gradients$

$g_t : Gradient\ at\ time\ t\ along\ \omega^j$

ADAM:

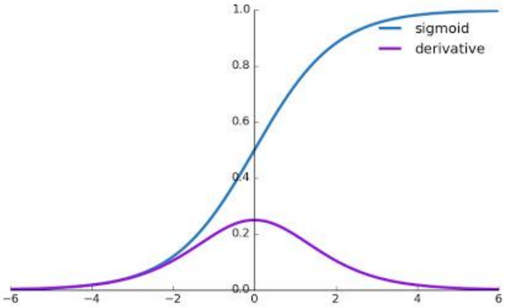
$$
\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\
\hat{v}_t &= \max(\hat{v}_{t-1}, v_t) \\
\theta_{t+1} &= \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} m_t
\end{aligned}
$$

If β1=0 Adam=rmsprop

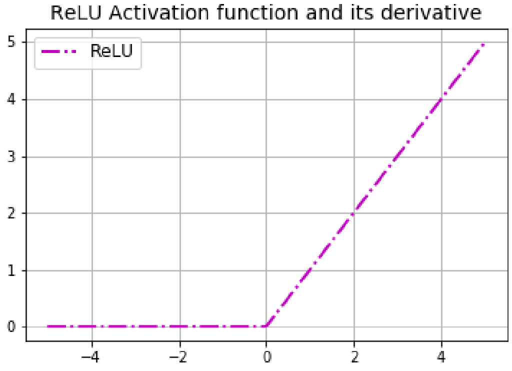
20. What is RMSPROP?
    ANS: Refer question 19

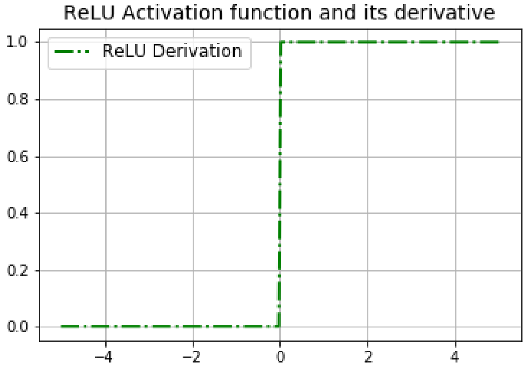
21. What is ADAM ?
    ANS : Refer question 19

22. What is the maximum and minimum values of gradient of the sigmoid function ?
    ANS :



23. What is RELU? Is it differentiable ?
    ANS : It is not differentiable at 0 but can be used as activation function with some hacks.





24. What is F1 score ?

ANS : F1-score is the harmonic mean of precission and recall which balances both the metrics if one of the metrics is bad entire f1-score will be affected.

$$\text{precision} = \frac{\text{tp}}{\text{tp}+\text{fp}}$$

$$\text{recall} = \frac{\text{tp}}{\text{tp}+\text{fn}}$$

$$\text{accuracy} = \frac{\text{tp}+\text{tn}}{\text{tp}+\text{tn}+\text{fp}+\text{fn}}$$

$$\text{F}_1 \text{ score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision}+\text{recall}}$$

25. What is precision and recall ?

ANS :

Precision tells us that out the total positive points how many of them are predicted to be positive.

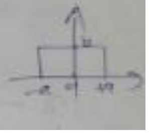
Recall tells us that out of total points that are predicted positive how many points are actually positive.

26. Name a few weight initialization techniques ?

ANS:

- Gaussian distribution
- Uniform initialization

$$w_{ij}^{k} \sim unif\left[\overset{a}{\frac{-1}{\sqrt{fan\text{-}in}}}, \overset{b}{\frac{1}{\sqrt{fanin}}}\right]$$



- Xavier/Glorot

normal:-

(a) $w_{ij}^{k} \sim N(0,\sigma_{ij})$ $\sigma_{ij} = \frac{2}{fanin+fanout}$

(b) uniform:-

$$w_{ij}^{k} \sim U\left[\overset{a}{\frac{-\sqrt{6}}{\sqrt{fanin+fanout}}}, \overset{b}{\frac{\sqrt{6}}{\sqrt{fanin+fanout}}}\right]$$

- He-initialization:

(a) normal: $w_{ij}^{k} \sim N(0,\sigma)$

$$\sigma = \sqrt{\frac{2}{fanin}}$$

(b) uniform: $w_{ij}^{k} \sim U\left[-\sqrt{\frac{6}{fanin}}, +\sqrt{\frac{6}{fanin}}\right]$

27. Which of these will have more numbers of tunable parameters?

    a. 7,7,512) ⇒ flatern ⇒ Dense(512)

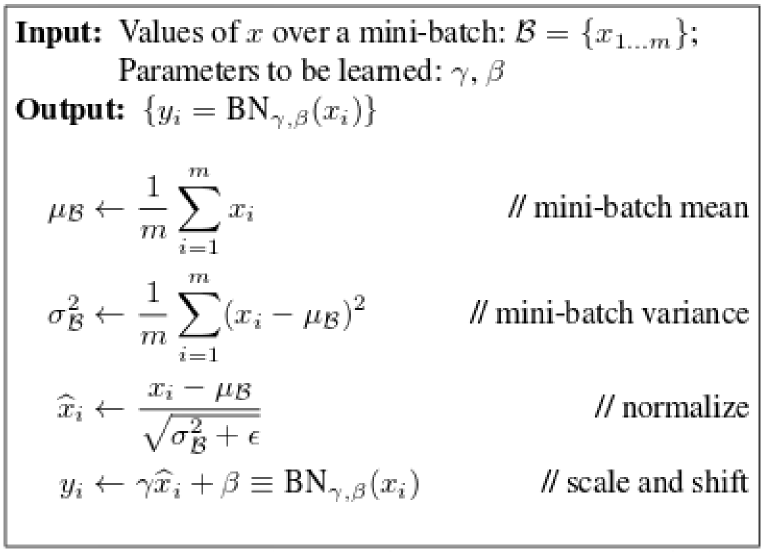
    b. (7,7,512) ⇒ Conv (512,(7,7))

    ANS: both will have paramaters i.e, 7*7*512*512

28. What is overfitting and underfitting ?
    ANS: Overfitting is a condition were the model has high variance even with slight change in data output is differed i.e., model is trained highly on train data. Underfitting is a condition where the model is highly biased and tend to predict the output to be of larger class i.e., is model is not trained well.

29. What do you do if a deep learning model is overfitting?
    ANS :
    - Add dropouts which makes some neurons inactive randomly at each step while training. Dropouts is like Column sampling in random forest.
    - Use Data Augmentation techniques as the model is performing very good on train data increase the complexity of data while training.
    - Decrease the complexity of the model
    - Use regularization techniques on weights.

30. What is the batch Normalization layer ?
    ANS :
    While training a deep neural network the weights may experience covariance shift deep in the network. i.e., the distribution of the weights for each batch has huge difference that leads to degrade model performance. To avoid such problems batch normalization layer is added which normalizes the weights by using scale and shift operations.



```
Input: Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1\dots m}\}$;
       Parameters to be learned: $\gamma$, $\beta$
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

  $\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$                          // mini-batch mean
  $\sigma_\mathcal{B}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_\mathcal{B})^2$        // mini-batch variance
  $\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}}$                    // normalize
  $y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i)$          // scale and shift
```

Algorithm 1: Batch Normalizing Transform, applied to activation x over a mini-batch.

31. Write keras code to add a BN layer in an existing network ?
    ANS: bn=tf.keras.layers.BatchNormalization(axis=-1)(previous layer)
    Preferably use previous layer to be activation layer.
    By default, axis=-1 that mean normalization is performed on different input weights in a batch
32. Number of tunable parameters in the BN layer.
    ANS: β,γ are the tunable parameter in BN layer.

33. What is convolution operation?
    ANS: Convolution operation performs sum of elementwise multiplications of two matrices(generally image pixel matrix) where convolution matrix moves over the input matrix to extract feature matrix which describes the features that are present in Input matrix.

34. Number of parameters in a convolution neural network given in architecture
    ANS: In convolution layer of size k*k with n filters if the previous layer has m channels
    Number of params will be = **k*k*m*n+n**
35. What are the inputs required to calculate the average f1 score ?
    ANS : class labels predicted and actual class labels to calculate precision and recall using
    In F1-score
36. What macro average f1 score for 5 class classification problem.
    ANS: We calculate precision and recall for individual classes and do average of them
    P=P1+P2+P3+P4+P5/5
    R=R1+R2+R3+R4+R5/5
    Now we calculate harmonic mean P and R to get macro-f1-score
    Macro f1-score does not care for class imbalance data.

37. How do you get probabilities for RF classifier outputs?
    ANS: In RF we will be having many base decision trees each predicting a class label.
    For probability of class in RF we calculate total number of DT's predicting the class
    Divided by total number of DT's in RF.

38. Is the Calibration classifier required to get probability values for logistic regression.?
    ANS : No, because logistic regression already gives the exact probability values unlike other
    Algorithms which give probabilistic estimate.

39. How does kernel svm work in test time ?
    ANS : In kernel SVM we calculate the similarity of the points with the others in the train data
    With the help of kernel.

$$F(x_q)=summation(\alpha_i*y_i*kernel(x_i,x_q)+b)_{i=1\text{ to }n}$$

40. What kind of base learners are preferable in random forest classifiers ?
    ANS: In Random forest we prefer decision trees as the base learners because it can be
    trained very deep which can have high variance easily.

41. How does bootstrapping works in RF classification.

ANS : In random forest each base learner is trained on sample of data randomly from total Data such that training many such models will see all the datapoints of train data.

42. Difference between one vs rest and one vs one.

ANS:

The One-vs-Rest strategy splits a multi-class classification into one binary classification problem per class.

The One-vs-One strategy splits a multi-class classification into one binary classification problem per each pair of classes.

43. Which one is better is one vs rest and one vs one?

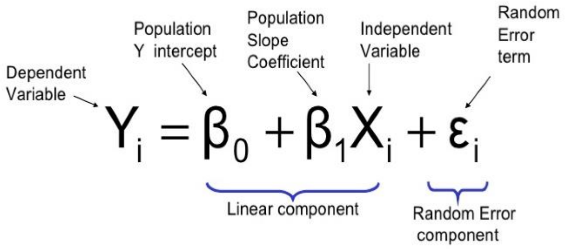
ANS : One vs rest is better because in One vs One we will take pairs of classes in each classification. So, in multiclass there will be many such pairs which will make it difficult to calculate.

44. What will happen if gamma increases in RBF kernel sum.

ANS : As gamma in RBF kernel is increased more of the points are given some similarity as the range of distance for similarity scores is increased. It is similar to K-NN with increased k values.

45. Explain linear regression.

ANS : Linear regression is an algorithm in which it tries to find a hyperplane which tries to fit to the training datapoints and has least sum of squared distances to the plane.



46. What is difference between one hot encoding and a binary bow?

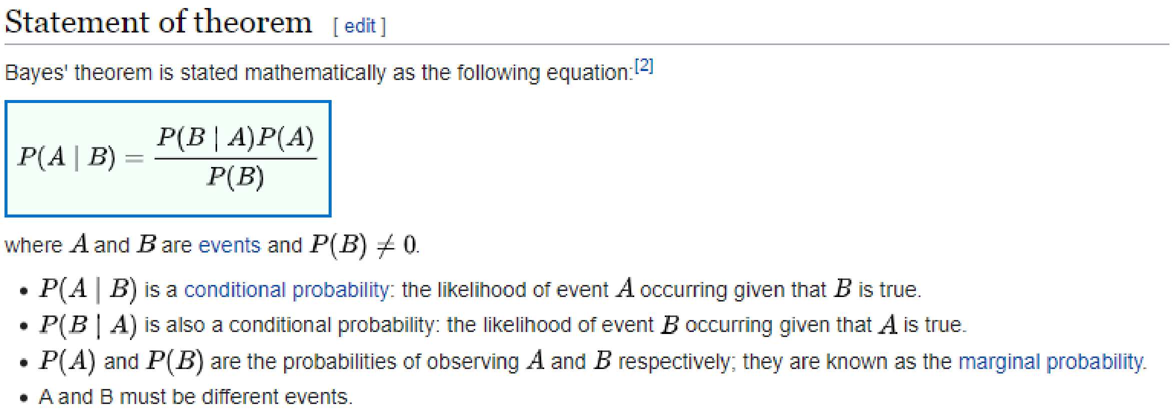
ANS: One hot encoding: In one hot encoding we take the number of times a category or word is repeated in the train data. Whereas in binary BOW we only take whether the category or word is present in train data (0 or 1).

47. Kernel svm and linear svm (SGD classifier with hinge loss). Which has low latency and why.

ANS: Linear SVM is faster because in linear SVM no kernelization is involved. Thus linear tries to operate in lower dim only where as kernel SVM tries to find higher dimensional features which takes time than linear SVM.

48. Explain bayes theorem.

ANS:



Statement of theorem [edit]

Bayes' theorem is stated mathematically as the following equation:[2]

$$\boxed{P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}}$$

where A and B are events and $P(B) \neq 0$.

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true.
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively; they are known as the marginal probability.
- A and B must be different events.

49. How to decrease the test time complexity of a logistic regression model.

ANS: If we get a sparse weight matrix just ignore the features which have the 0 weights.

50. What is the need for sigmoid function in logistic regression?

ANS: In logistic regression Euclidean distances play a major role and the impact of outlier will be high so avoid the problem of outlier we use squashing which can be done by using sigmoid function.

51. Why we need Calibration ?

ANS: After training a model, the model may or may not give the exact probabilities as the model may have many assumptions while training. So, to obtain exact probability values we use calibration.

52. What is MAP ? (mean average precision)

ANS: MAP is calculated in case of multi-class classification problems where Average Precision is calculated for all classes and mean of average precision is calculated.

53. Why do we need gated mechanism in LSTM ?

ANS: In LSTM’s long-term dependencies of input preserved by the mechanism of gates in LSTM which enables model to work on long sequences. Forgot gate enables to change the amount of input to be transferred to another cell and so on.

54. What is stratified sampling ? Explain.

ANS: In stratified sampling each sample gets equally shared class data points after dividing Into smaller groups.

55. How do you compare two distributions ?

ANS: Using Kullback-leiber (KL) divergence.

$$D_{KL}(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} dx$$

56. What will happen to train time of K means of data is very high dimension.
    ANS: As the k means deals with the distance measure with the very high dimensions it may not work properly as curse of dimensionality kicks in.

57. If you have 10mill records with 100dimension each for a clustering task. Which algorithm will you try first and why ?
    ANS: K-Means handles well with large data because it has linear time complexity of O(n) whereas Hierarchical clustering is of order $O(n^2 logn)$

58. What is matrix Factorization? Explain with an Example.
    ANS: Matrix factorization is a class of collaborative filtering algorithms used in recommender systems. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices.

59. Which algorithm will give high time complexity if you have 10million records for a clustering task?
    ANS: Hierarchical clustering gives the high time complexity of order $O(n^2 logn)$

60. Difference between GD and SGD.
    ANS:

    Batch Gradient Descent: Batch Gradient Descent involves calculations over the full training set at each step as a result of which it is very slow on very large training data. Thus, it becomes very computationally expensive to do Batch GD. However, this is great for convex or relatively smooth error manifolds. Also, Batch GD scales well with the number of features.

    Stochastic Gradient Descent: SGD tries to solve the main problem in Batch Gradient descent which is the usage of whole training data to calculate gradients as each step. SGD is stochastic in nature i.e it picks up a “random” instance of training data at each step and then computes the gradient making it much faster as there is much fewer data to manipulate at a single time, unlike Batch GD.

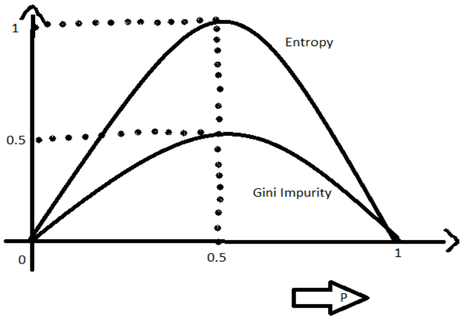
61. Which one will you choose GD or SGD? Why ?
    ANS SGD, refer Q.60

62. Why do we need repetitive training of a model ?
    ANS: If the model that was trained on a data is changing over time then we should retrain our model as the model may not perform well if the nature of data is changed.

63. How do you evaluate the model after productionization ?
    ANS: We should perform A/B testing.
    Ref: https://www.analyticsvidhya.com/blog/2020/10/ab-testing-data-science/

64. What is need for Laplace smoothing in N.B
    ANS: Refer Q.9

65. Explain Gini impurity.
    ANS: Gini Impurity is similar to Entropy which explains the randomness of the data.



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$Gini(E) = 1 - \sum_{j=1}^{c} p_j^2$$

66. Explain entropy?
    ANS:

    Entropy, as it relates to machine learning, is a measure of the randomness in the information being processed. The higher the entropy, the harder it is to draw any conclusions from that information.

    If the Entropy is 0, we can say that the split is done well and if Entropy is 1, we have the Equally distributed classes in the split.

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

68. What is k-fold cross validation ?
    ANS: In training process for hyperparameter tuning we make use of test data but that leads to leakage of data. To use our training data itself for doing hyperparameter effectively we do k-fold cross validation. In which we divide train data in k parts and at each iteration we use k-1 parts for training and other part for hyperparameter thus, we get k iterations enabling no data loss while training.

69. What is need for CV ?
    ANS: While doing hyperparameter tuning we can't use test data as it leads to data leakage So, we use cross validate data for hyperparameter tuning.

70. How do you to CV for a test classification problem using random search.
    ANS: In random search cv we try out all the combinations of hyperparameters from a list of values and try to pick the combination which gives best score.

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

    ANS: Naïve Bayes which can do well on high dimensional data compared to many other Models, this can be used as a benchmark for training.

72. What is AUC?

    ANS: AUC is short form of Area under curve. The Area under a ROC curve is defined as AUC. ROC curve is drawn by plotting FPR vs TPR with each class probability of a datapoint treated as threshold.

    Range of AUC is [0,1] 0-is worst, 1- is best, random model can get AUC of 0.5.

    AUC tells that if two points of different class labels given how likely the model would correctly classify them.

73. Tell me one business case where recall is more important than precision.

    ANS: In Medical treatments we can afford to miss any person in giving vaccine who are tested positive. And we can leave the persons without vaccinating who are tested negative.

74. Tell me one business case where precision is more important.

    ANS: In medical treatments we should not give treatment for wrong person who are tested Negative which can lead to side effects.

    Precision is more important than recall when you would like to have less False Positives in trade off to have more False Negatives. Meaning, getting a False Positive is very costly, and a False Negative is not as much.

    In a zombie apocalypse, of course you would try to accept as many as healthy people you can into your safe zone, but you really don't want to mistakenly pass a zombie into the safe zone. So, if your method causes some of the healthy people mistakenly not to get into the safe zone, then so be it.

    When you let go 100 culprits your recall is low. But if you punish someone, you are sure that you are punishing only a criminal - precision is high.

75. Can we use accuracy for very much imbalance data? If yes/no , why ?

    ANS: No, we can't use accuracy as performance metric if the data is highly imbalanced as accuracy does not care for imbalanced data. It is biased towards larger class

    EX: we have 100 datapoints of 80 +ve and 20 -ve, even if the model predicts all points to be +ve we get an accuracy of 80%

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

    ANS : In micro average F1-score we care for all true positive and true negatives of each Class and take all values into accounts .This deal with imbalanced data well

    In macro average F1-score we don't care for individual classes we do the average of Precision and recall calculated on individual classes. This doesn't account for class Imabalance.

77. Difference between AUC and accuracy ?
    ANS: Accuracy tells with how much percentage the model correctly classifies a point.
    AUC tells that with how much percentage the model correctly classifies two or more points correctly.

78. How do we calculate AUC for a multiclass classification?
    ANS: We need to one vs all approach and calculate AUC for each class.

79. Test the complexity of Kernel sum ?
    ANS:

80. Can we use TSNE for dimensionality reduction i.e., convert the data n to d dimension.
    ANS: No
    The main reason that t-SNE is not used in classification models is that it does not learn a function from the original space to the new (lower) dimensional one. As such, when we would try to use our classifier on new / unseen data we will not be able to map / pre-process these new data according to the previous t-SNE results.
    t-SNE tries to find new distribution of data in lower space such that both the distributions are very similar. This is achieved by KL divergence.

81. What is person correlation coefficient ?
    ANS : person correlation coefficient is a method by which we can obtain the variability of two random variables. Unlike co-variance it gives exact value by which the variables are related to each other.It measures the linear correlation between two variables.
    Its range is -1 to 1.

82. Training time complexity of naive bayes ?
    ANS:
83. Numbers of tunable parameters in maxpooling layer ?
    ANS: 0

84. 100,50) -> Embeddylayer (36) -> output shape ?
    ANS: (100,50,36)

85. Number of tunable parameters in embedding layer (36, vocab size = 75)
    ANS: 75*36 = 2700

86. Relation between KNN and kernel sum ?
    ANS: In KNN we use similarities of points based on distances similarly we use similarities Computed by kernel in kernel SVM.
    In RBF kernel σ is equivalent to K in KNN, i.e., as σ increases we consider points with distances of increased range similarly in KNN if K is increases, we take more points having more distance ranges.

87. Which is faster

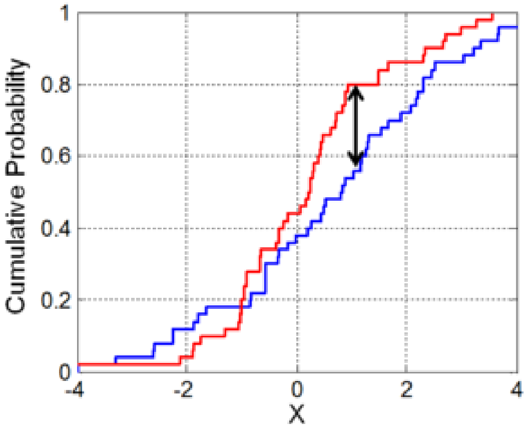
a. SVC(C=1). Fit(x,y)
b. SGD(Log=hinge).fit(x,y)

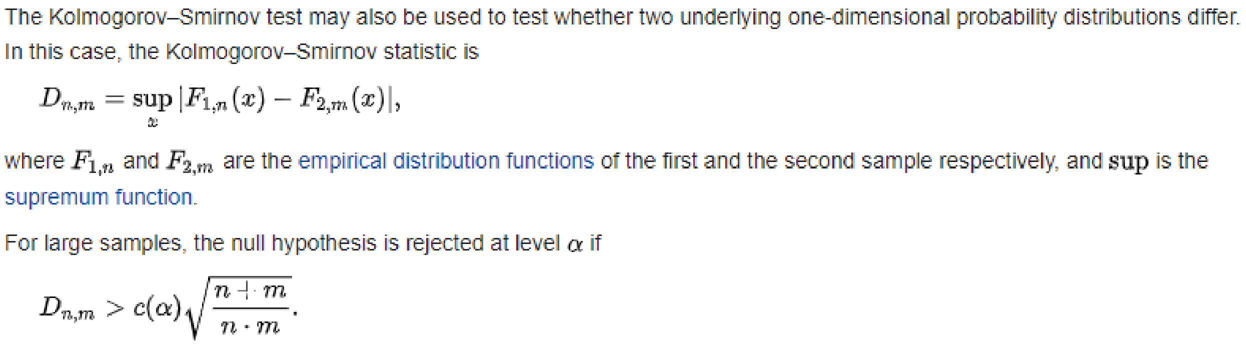
ANS: SGD with hinge loss will be faster because it doesn't have any constraints like SVC with soft margin has

Constraint = $y_i(w^t x_i+b)>=1-\xi_i$

88. Explain about KS test ?

ANS: KS Test tells us that whether two distributions are same or not.





The Kolmogorov–Smirnov test may also be used to test whether two underlying one-dimensional probability distributions differ. In this case, the Kolmogorov–Smirnov statistic is

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|,$$

where $F_{1,n}$ and $F_{2,m}$ are the empirical distribution functions of the first and the second sample respectively, and $\sup$ is the supremum function.

For large samples, the null hypothesis is rejected at level α if

$$D_{n,m} > c(\alpha)\sqrt{\frac{n+m}{n \cdot m}}.$$

89. What is KL divergence ?

ANS : Refer Q.55

90. How QQ plot works ?

ANS: QQ plot is used to know the distribution of unknown data.

The values are sorted, and quantiles is computed.

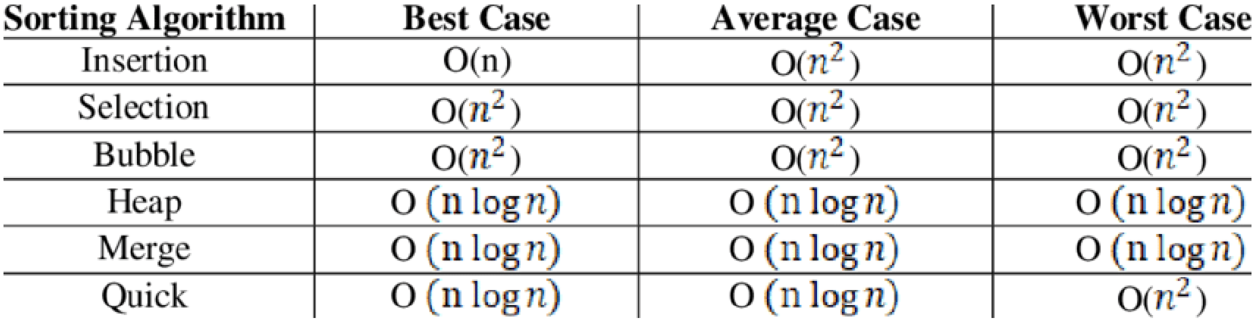
Now we take the values from a known distribution and computed percentiles in the same way.

Now we plot this percentile values if the plot is a straight line, we can say that both the Distributions are same. Thus, we can know the distribution of unknown data

91. What is the need of confidence interval ?
ANS:
Confidence intervals show us the likely range of values of our population mean. When we calculate the mean we just have one estimate of our metric; confidence intervals give us richer data and show the likely values of the true population mean.

92. How do you find the out outliers in the given data set ?
ANS: Using IQR plot
Or by plotting the pdf of the data.

93. Can you name a few sorting algorithms and their complexity ?
ANS:



Sorting Algorithm	Best Case	Average Case	Worst Case
Insertion	O(n)	$O(n^2)$	$O(n^2)$
Selection	$O(n^2)$	$O(n^2)$	$O(n^2)$
Bubble	$O(n^2)$	$O(n^2)$	$O(n^2)$
Heap	$O(n \log n)$	$O(n \log n)$	$O(n \log n)$
Merge	$O(n \log n)$	$O(n \log n)$	$O(n \log n)$
Quick	$O(n \log n)$	$O(n \log n)$	$O(n^2)$

94. What is the time complexity of ”a in list ( )” ?
ANS: O(n)

95. What is the time complexity of “a in set ( ) “?
ANS: O(1) if there is hash table implementation
O(n) worst case.

96. What is percentile ?
ANS: percentile is a value that says what is the value of data present at a specific percentage when the data is sorted.

97. What is IQR ?
ANS: IQR tells what the values are present in 25th and 75th percentile range of data.
IQR=Q3-Q1

98. How do you calculate the length of the string that is available in the data frame column ?
ANS: len(df.iloc[index][column])

99. Can you explain the dict.get() function ?
ANS: It takes a key as argument and returns its value if the key is not present returns None.

100. Is list is hash table ?
ANS: No

101. Is tuple is hash table ?
ANS: No

Array

Each record is placed one next to the other in RAM. So you know that your record is X bytes long, thus the 3rd record starts at position “start of array + 2 record lengths”.

Linked List

Each record is allocated to wherever the memory manager deems most applicable just now. You get the location where it was allocated and save that into the “next” field of the previous record. You only have one value in your data structure … the head with the location of the first item.

Hash table

You make an array that’s longer than you need. You calculate a numeric value from the name using a hash formula. That you then use to define a position inside that array. Into which you store that particular name and its info. Do the same for the rest. If you find that the position is already occupied, recalc a new hash until you get an empty spot.

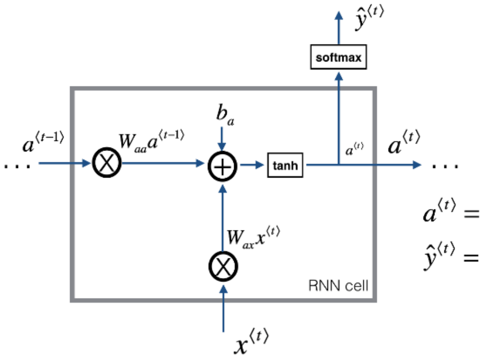
102. What is parameter sharing in deep learning?
ANS: A convolutional neural network learns certain features in images that are useful for classifying the image. Sharing parameters gives the network the ability to look for a given feature everywhere in the image, rather than in just a certain area. This is extremely useful when the object of interest could be anywhere in the image.
Relaxing the parameter sharing allows the network to look for a given feature only in a specific area. For example, if your training data is of faces that are centered, you could end up with a network that looks for eyes, nose, and mouth in the center of the image, a curve towards the top, and shoulders towards the bottom.
It is uncommon to have training data where useful features will usually always be in the same area, so this is not seen often.

103. What will be the alpha value for nonsupport vectors?
ANS: 0 for non-support vectors
>=0 for support vectors

104. What will be the effect of increasing alpha values in multinomial NB ?
ANS: It leads to underfitting

Refer Q.9

105. What is recurrent equation of RNN output function ?


$\hat{y}^{\langle t \rangle}$
softmax
b_a
$a^{\langle t-1 \rangle}$
$W_{aa}a^{\langle t-1 \rangle}$
tanh
$a^{\langle t \rangle}$
$a^{\langle t \rangle}$
$W_{ax}x^{\langle t \rangle}$
RNN cell
$x^{\langle t \rangle}$

$$a^{\langle t \rangle} = \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b_a)$$

$$\hat{y}^{\langle t \rangle} = soft\max(W_{ya}a^{\langle t \rangle} + b_y)$$

ANS:

106. What is the minimum and maximum value of tanh ?

ANS: -1 and +1

107. How many thresholds we need to check for a real valued feature in DT ?

ANS: In DT with real values, we need to treat each value as a threshold

108. How do you compute the feature importance in DT ?

ANS: We take the node which gives high information gain as important or

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

109. How do you compute the feature importance in SVM ?

ANS: By using the weight coefficients.

In Kernel it is not possible because we use the datapoints of higher dimensions for Classification

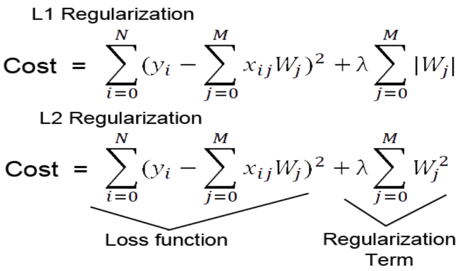
.

110. Prove that L1 will give sparsity in the weight vector ?

ANS: Refer: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/3043/why-l1-regularization-creates-sparsity/3/module-3-foundations-of-natural-language-processing-and-machine-learning

111. What are L1,L2 regularizer ?

ANS: As we can see from the formula of L1 and L2 regularization, L1 regularization adds the penalty term in cost function by adding the absolute value of weight(Wj) parameters, while L2 regularization adds the squared value of weights(Wj) in the cost function.



L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2}_{\text{Loss function}} + \lambda \underbrace{\sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

112. What is elastic net ?

ANS: In elastic net we use both L1 and L2 regularizers

113. What are the assumptions of NB ?

ANS: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

114. What are the assumptions of KNN ?

ANS: It is non-parametric method since it doesn’t make any assumptions of the data.

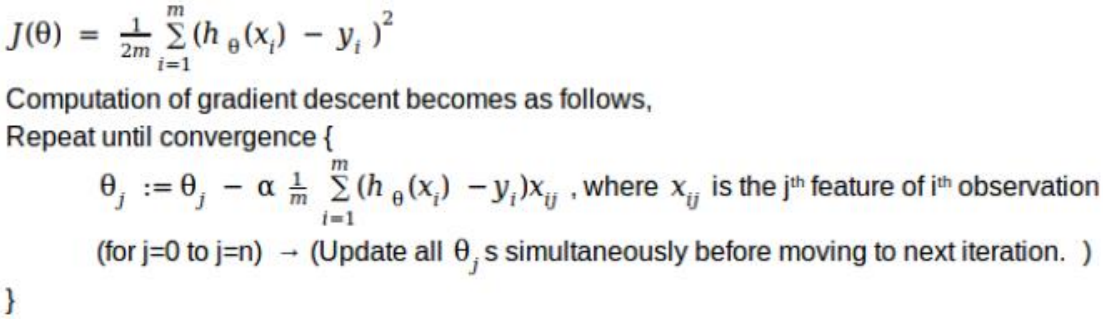
115. What are the assumptions of linear regression ?

ANS: First, logistic regression does not require a linear relationship between the dependent and independent variables. Second, the error terms (residuals) do not need to be normally distributed. Third, homoscedasticity is not required. Finally, the dependent variable in logistic regression is not measured on an interval or ratio scale.

Refer: https://www.statisticssolutions.com/assumptions-of-logistic-regression/?__cf_chl_jschl_tk__=985eb51af8754a7beaabffc1d638f2120071f825-1609051672-0-AdNEoZkMh1myeCM1kge-Cts_hjKBTX7nJx8inMhrGznWgxeOmKAph0E8VoB-5i6hhNyi46a-HkfuoO02haTGhx-E2Q-tDtpUIlSSJWqD_5adYbL8VecJrNim-WCH-cwj73y89h6GBv_Q94527FY-VSs7Iux_aYQSkMEQa1NHFqEd78axZRoW6Vg0nN4JMYoc0vA4Sv4K9FMy5F8E8LbEO-sXnv5TvjRKqTV7uLsIobne8Aq7Kt4uHJM2Yg5pmHfWR7IrL8o9ZY3BHJZFAmom72N3vAxCcFxSezFTv4pjlxpLs-8Mys8C03KWGWnx0wv9y4hxAWo1RrOV0_UVpzWF-iExEfRAAPgOUDGYTuDRJuOh

116. Write the optimization equation of linear regression ?

ANS:



$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

Computation of gradient descent becomes as follows,

Repeat until convergence {

$\theta_j := \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)x_{ij}$, where x_{ij} is the jth feature of ith observation

(for j=0 to j=n) → (Update all θ_j s simultaneously before moving to next iteration.)

}

117. What is time complexity of building KD tree ?

ANS: O(ndlogn)

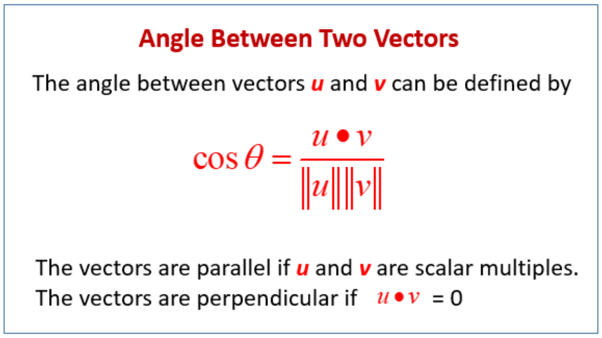
118. What is the time complexity to check if a number is prime or not ?

ANS: O(sqrt(n))

119. Angle between two vectors (2,3,4) (5,7,8).

ANS: cosƟ=1

Ɵ=0º



Angle Between Two Vectors

The angle between vectors ***u*** and ***v*** can be defined by

$$\cos\theta = \frac{u \bullet v}{\|u\|\|v\|}$$

The vectors are parallel if ***u*** and ***v*** are scalar multiples.

The vectors are perpendicular if $u \bullet v$ = 0

120. Angle between the weigh vector of 2x+3y+5=0 and the vector(7,8).

ANS: cosƟ=0.5

121. Distance between (7,9) and the line 7x+4y-120=0.

122. Distance between the lines 4x+5y+15=0, 4x+5y-17=0.

ANS: two lines are parallel we get 15-(-17)=32

123. Which of this hyperplane will classify these two class points?

   a. P: (2,3), (-3,4) N: (-5,7), (-5,-9)
   b. 4x+5y+7=0, -3y+3x+9=0

ANS: by checking the sides of the points w.r.t plane ,No plane will classify correctly

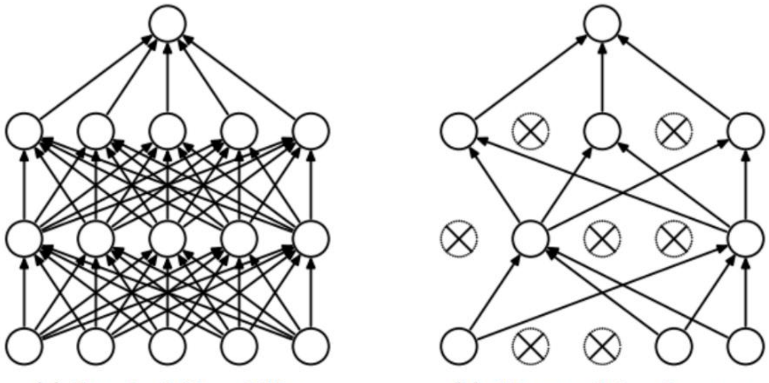
124. Which of the vector pairs perpendicular to each other?

   a. (3,4,5) (-3,-4,5)
   b. (7,4,6) (-4,-7,-12)

ANS: Pair will be perpendicular as angle between them is 90º ,cosƟ=0

125. How dropout works ?

ANS: At each iteration dropouts layer makes a neuron active with a probability of dropout rate. This makes the model to avoid overfitting as only some neurons are at each step .


(a) Standard Neural Net (b) After applying dropout.

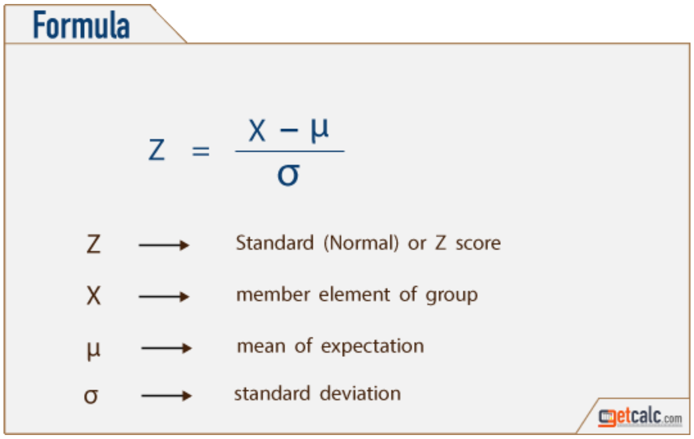
126. Explain the back-propagation mechanism in dropout layers ?
ANS : During test time the weight are multiplied with dropout rate.

127. Explain the loss function used in auto encoders assuming the network accepts images ?
ANS: Mean square error is used as loss function in autoencoder

128. Numbers of tunable parameters in dropout layer ?
ANS : 0
129. When F1 score will be zero? And why ?
ANS: If either precision or recall is 0

130. What is the need of dimensionality reduction?
ANS: Dimensionality reduction refers to techniques for reducing the number of input variables in training data. When dealing with high dimensional data, it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the "essence" of the data

131. What happens if we do not normalize our dataset before performing classification using KNN algorithm.
ANS: **KNN** performance usually requires preprocessing of data to make all variables similarly scaled and centered.
If we don't normalize the data, all the features will be on different scales thus the model doesn't Perform well.
132. What is standard normal variate ?
ANS : Standard normal variate makes the data with 0 centering and with variance 1.



Formula

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

Z → Standard (Normal) or Z score

X → member element of group

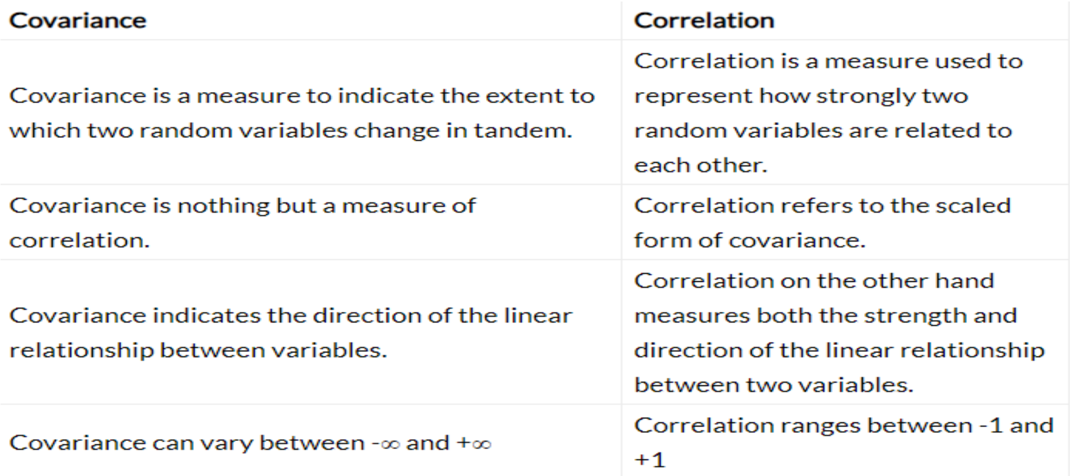
μ → mean of expectation

σ → standard deviation

getcalc.com

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

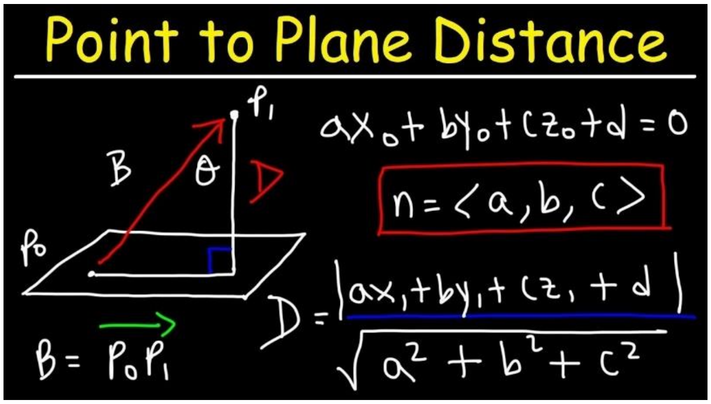
ANS:



Covariance	**Correlation**
Covariance is a measure to indicate the extent to which two random variables change in tandem.	Correlation is a measure used to represent how strongly two random variables are related to each other.
Covariance is nothing but a measure of correlation.	Correlation refers to the scaled form of covariance.
Covariance indicates the direction of the linear relationship between variables.	Correlation on the other hand measures both the strength and direction of the linear relationship between two variables.
Covariance can vary between -∞ and +∞	Correlation ranges between -1 and +1

134. How do we calculate the distance of a point to a plane?

ANS:


Point to Plane Distance
P_1
B
θ
D
P_0
$ax_0 + by_0 + cz_0 + d = 0$
$n = \langle a, b, c \rangle$
$D = \dfrac{|ax_1 + by_1 + cz_1 + d|}{\sqrt{a^2 + b^2 + c^2}}$
$B = \overrightarrow{P_0P_1}$

135. When should we choose PCA over t-sne?
ANS: Refer: https://stats.stackexchange.com/questions/238538/are-there-cases-where-pca-is-more-suitable-than-t-sne

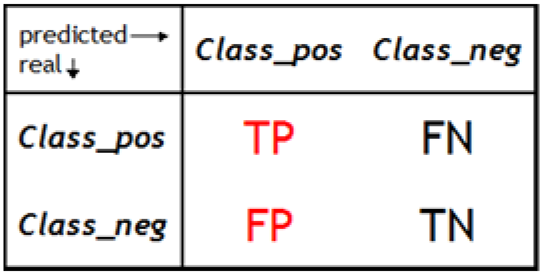
136. How is my model performing if?
    a. Train error and cross validation errors are high.
    b. Train error is low and cross validation error is high.
    c. Both train error and cross validation error are low.

ANS: a. underfitting b. overfitting c. best fit

137. How relevant / irrelevant is time-based splitting of data in terms of weather forecasting ?
ANS: It is required to time-based splitting in weather forecasting as it changes over time.

138. How is weighted knn algorithm better simple knn algorithm.
139. What is the key idea behind using a kdtree?
ANS : It is based on axis parallel lines and it is very useful in performing search queries as

140. What is the relationship between specificity and false positive rate?
ANS:

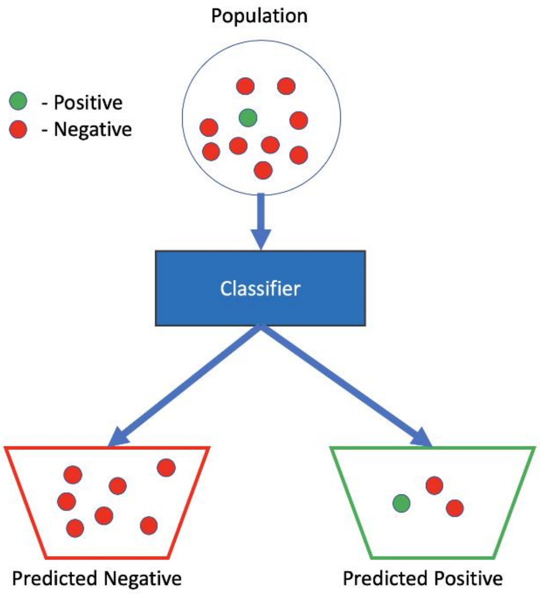


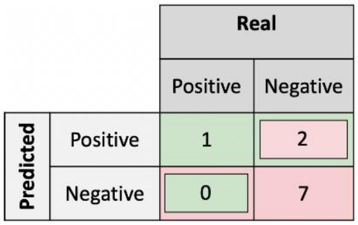
predicted → real ↓	*Class_pos*	*Class_neg*
Class_pos	TP	FN
Class_neg	FP	TN

$$\text{TPR (sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

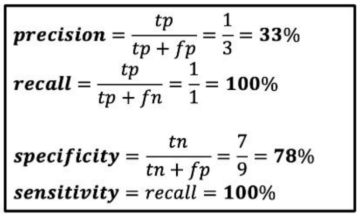
$$\text{FPR (1-specificity)} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

141. What is the relationship between sensitivity,recall,true positive rate and false negative rate?





		Real	
		Positive	Negative
Predicted	Positive	1	2
	Negative	0	7



$$\boldsymbol{precision} = \frac{tp}{tp+fp} = \frac{1}{3} = \mathbf{33\%}$$

$$\boldsymbol{recall} = \frac{tp}{tp+fn} = \frac{1}{1} = \mathbf{100\%}$$

$$\boldsymbol{specificity} = \frac{tn}{tn+fp} = \frac{7}{9} = \mathbf{78\%}$$

$$\boldsymbol{sensitivity} = recall = \mathbf{100\%}$$

FNR=1-sensitivity, sensitivity=recall=TPR

142. What is the alternative to using Euclidean distance in Knn when working with high dimensional data ?

ANS: Cosine distance.

143. 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 ?

144. How does outlies effect the performance of a model and name a few techniques to overcome those effects.

ANS: Random sample consensus (**RANSAC**) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence on the values of the estimates.

Anomaly detection Refer: LOF is one of the anomaly detection algorithm

145. What is reachability distance

ANS: Reachability distance. The k-distance is now used to calculate the reachability distance. This distance measure is simply the maximum of the distance of two points and the k-distance of the second point. Basically, if point a is within the k neighbors of point b, the reach-dist(a,b) will be the k-distance of b.

$$reachdist_k(o \leftarrow o') = \max\{dist_k(o), dist(o,o')\}$$

146. What is the local reachability density ?

ANS: The local reachability density is a measure of the density of k-nearest points around a point which is calculated by taking the inverse of the sum of all of the reachability distances of all the k-nearest neighboring points.

$$\mathrm{lrd}_k(A) := 1 \Big/ \left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|} \right)$$

$$\mathrm{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\mathrm{lrd}_k(B)}{\mathrm{lrd}_k(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \mathrm{lrd}_k(B)}{|N_k(A)| \cdot \mathrm{lrd}_k(A)}$$

147. What is the need of feature selection ?

ANS: If the data is having very dimensional data the model would not perform well.
So, to get the best model we do features selection which gets the useful features of data
And model is trained on selected model to yield better results.

148. What is the need of encoding categorical or ordinal features?

ANS: Machine learning models require all input and output variables to be numeric. This means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model

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

ANS: Bias-variance tradeoff give the flexibility to a model for changing its behavior towards
Model performing for gaining balance between overfitting and underfitting.

150. **Can we use algorithm for real time classification of emails?**

ANS: Yes

151. What does it mean by precision of a model equal to zero is it possible to have precision equal to 0?

ANS: Yes, we can have 0 precision that means model doesn’t give any true positives.

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

ANS: We can’t have FPR=TPR=1 at the same time equal to 1.

153. What does AUC = 0.5 signifies.

ANS: If two points of different classes were given the model would correctly separate them with a probability of 0.5

154. When should we use log loss, AUC score and F1 score?

ANS: If the objective of classification is scoring by probability, it is better to use AUC which averages over all possible thresholds. However, if the objective of classification just needs to

classify between two possible classes and doesn't require how likely each class is predicted by the model, it is more appropriate to rely on F-score using a particular threshold.
We use log loss when we care about probability deviations from the actual probabilities.

155. What performance metric should use to evaluate a model that see a very less no.of positive data points as compared to -ve data points.
ANS: Recall as it considers false negatives.

156. What performance metric does t-sne use to optimize its probabilistic function.
ANS: T-sne uses KL divergence between probabilities of two distributions

157. What happens in Laplace smoothing in my smoothing factor '**α**' is too large.
ANS : Refer Q.9

158. When to use cosine similarity over Euclidean distance.
ANS: when the data is very high dimensional.

159. What is fit, transform and fit transform in terms of BOW,tf-idf,word2vector.
ANS: Fit used to learn the vocabulary and creates a dictionary with key as word and Its words count in the documents as value.
Transform converts the given data to the representations that were learnt during training .
fit transform does the two steps on the same data simultaneously.

160. How do we quantify uncertainty in probability class labels when using KNN model for classifications?
ANS : After choosing k nearest neighbors we do majority voting and get its probabilities.
161. How do we identify whether the distribution of my train and test is similar or not.
ANS: we can use QQ-plot over test and train data
Or we can measure the distribution of classes among train and test data.

162. What does it mean by embedding high dimensional data points to a lower dimension ? what are the advantages and disadvantages of it.
ANS: It means reducing the dimensionality of the data without losing its meaning by preserving the nature like its neighborhood .
By dimensionality reduction it becomes much easy to visualize and understand data and models perform well on low dimensional data. Time and space complexity will also be reduced.

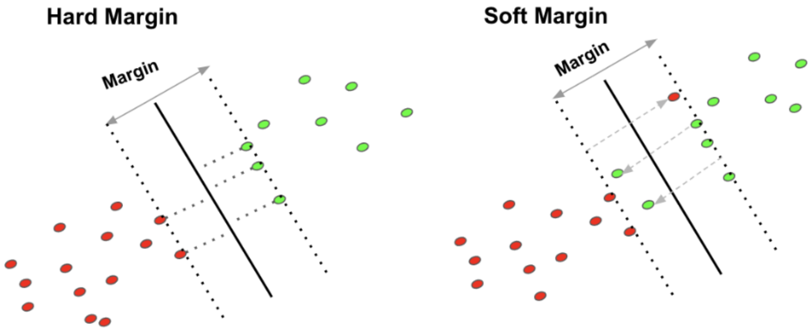
163. What is the crowding problem w.r.t t-sne.
ANS: the area of the two-dimensional map that is available to accommodate moderately distant datapoints will not be nearly large enough compared with the area available to accommodate nearby datapoints

164. What is the need of using log probabilities instead of normal probabilities in naive bayes.

ANS: Since in original formulation of naïve bayes it has lot of product terms. So, it can lead to 0 values easily. To overcome this, we use log of probabilities which converts product terms to sum terms

165. What do you mean by hard margin SVM ?

ANS: In hard margin we only do optimization for increase the margin with the points to be correctly classified.



166. What is kernel function in svm ?

ANS: Kernel function in SVM is used to find the similarities of the datapoints by converting them into higher dimensional data where similarities cannot be calculated easily at lower space.

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

ANS: In SVM we try to find two parallel hyperplanes which classifies the data which gives high distance difference between them.

168. Is svm affected by outliers ?

ANS: Hard margin SVM will be greatly affected by outliers but soft margin SVM resolves this for some extent by fixing a constraint on misclassification rate.

169. What is locality sensitive hashing ?

ANS: Locality sensitive hashing is method to find similar points that are In the same locality or Neighborhood this is done by drawing multiple hyperplanes and calculating the sides of each point w.r.t each plane thus we treat the points which having same sides as similar.

170. What is sigmoid function? What is its range ?

ANS: sigmoid function is a function which has a real value for every real value.
Its range is 0 to 1.

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

ANS: We can use any function that can normalize the data and its derivative exists.
Example: $\frac{Z}{1+|z|}$

172. Why is accuracy not a good measure for classification problem ?
ANS : refer Q.75

173. How to deal with multiclass classification problem using logistic regression ?
ANS: We train multiple classifiers each classifies a specific class by using One vs Rest approach.

174. Can linear regression be used for classification purpose ?
ANS: No, because linear regression predicted values are continuous while classification requires probability score.

We can do workaround on top linear regression to make it for classification.

Refer: https://jinglescode.github.io/2019/05/07/why-linear-regression-is-not-suitable-for-classification/

175. 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

176. When EDA should be performed, before or after splitting data? Why ?
ANS: we should perform EDA before splitting because we need to understand the whole nature of data which the model would work on. If EDA was performed on train and the model works better on train and may work worse on test as the model is trained on data which couldn’t be prepared for future unseen data.

177. How k-means++ is different from k-means clustering ?
ANS: In K-means we randomly initialize the centroids at the start. So, model can vary with Different initializations.

In K-means++ we choose a better of initializing centroids by following some strategies to avoid problems of initialization.

178. Where ensemble techniques might be useful ?
ANS: Ensembles are used to achieve better predictive performance on a predictive modeling problem than a single predictive model. The way this is achieved can be understood as the model reducing the variance component of the prediction error by adding bias (i.e., in the context of the bias-variance trade-off)

179. What is feature forward selection ?
ANS: In feature forward selection we train a model with one feature at a time and choose the best feature which gives the better performance. Now we use the best feature and extra one feature one at a time and get the best pair. This process is repeated until the required most important features are obtained.

180. What is feature backward selection ?

ANS: Feature backward selection is process of selecting important features from that data, where initially we train model on all the features and remove the worse features one by one by until required features are obtained observing the loss or metric.

181. What is type 1 & type 2 error ?

ANS: 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"

182. What is multicollinearity ?

ANS: Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model

183. How is eigenvector different from other general vectors ?

ANS: Eigen vector is a vector is a vector that signifies most of the variance along a feature

184. What is eigenvalue & eigenvectors ?

ANS: Eigen values gives the ammount of variance explained along a vector or feature.

185. What is A/B testing

ANS : A/B testing is used to know the performance of a model after deployment.

EX: If a model is retrained and need to be tested. We allow only some part of the traffic through the retrained model and check the performance while compared to the another model. If the new model gives good results, we redirect entire traffic to our New model.

186. How to split data which has temporal nature.

ANS: We should use time-based splitting.

EX: If we have data of 10 data we use first 9 days data as train and 10th day data as test.

187. What is response encoding of categorical features ?

ANS: It is method for encoding categorical features. We will calculate the probability of a class with a given category out of all the points having the given category.

Response encoding gives the probability for each class.

188. What is the binning of continuous random variables.

ANS: Data binning (also called Discrete binning or bucketing) is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values which fall into a given small interval, a bin, are replaced by a value representative of that interval, often the central value.

189. Regularization parameter in dual form of SVM ?

190. What is the difference between sigmoid and SoftMax ?

ANS: Sigmoid is used if the classes are independent of each other. each class can have any probability value without having the constraint of sum of probabilities across all classes to be one. Thus, we can use sigmoid for multi class classification.

Softmax is used if the chances of predicting one class effects another. The sum of probabilties of each class should sum to 1 .So, we only have one class prediction at a time

191. For a binary classification which among the following cannot be the last layer ?

a. sigmoid(1)
b. sigmoid(2)
c. softmax(1)
d. softmax(2)

ANS: a

192. What is P-value in hypothesis testing ?

ANS: The P value, or calculated probability, is the probability of finding the observed, or more extreme, results when the null hypothesis (H 0) of a study question is true – the definition of 'extreme' depends on how the hypothesis is being tested.

193. How to check if a particular sample follows a distribution or not ?

ANS: Use Q-Q plot or K-S test

194. What is the difference between covariance and correlation ?

ANS: Refer Q.133

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

ANS: If we have similarity matrix, we can use agglomerative clustering.

In high dimensional data time complexity of agglomerative clustering is high so we use k-means then.

196. What is the metric that we use to evaluate unsupervised models?

ANS: Dunn Index

197. What is the difference between model parameters and hyper parameters ?

ANS: Model parameter are used to define the structural and variations of model.

Hyperparameter are used to tune the model performances.

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

ANS: 3m(mn+n+1)

199. What is box cox transform? When can it be used ?

ANS: BOX COX transformation is used to transform the data of unknown distribution to a know Normal distribution.

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

and the two-parameter Box–Cox transformations as

$$y_i^{(\lambda)} = \begin{cases} \dfrac{(y_i + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & \text{if } \lambda_1 \neq 0, \\ \ln(y_i + \lambda_2) & \text{if } \lambda_1 = 0, \end{cases}$$

If $\lambda=0$ then X is log-normal

200. In what format should the data be sent to embedding layer?

ANS: Numeric values to be sent to embedding layers

201. What does trainable = true/false mean in embedding layer ?

ANS: Trainable=True means we train the weights by update equation.
Trainable=False freezes the weights without training.

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

ANS : It tells whether to return last state of output or not.

203. Why are RNN'S and CNN'S called weight shareable layers ?

ANS: A convolutional neural network learns certain features in images that are useful for classifying the image. Sharing parameters gives the network the ability to look for a given feature everywhere in the image, rather than in just a certain area. This is extremely useful when the object of interest could be anywhere in the image.

Relaxing the parameter sharing allows the network to look for a given feature only in a specific area. For example, if your training data is of faces that are centered, you could end up with a network that looks for eyes, nose, and mouth in the center of the image, a curve towards the top, and shoulders towards the bottom.

It is uncommon to have training data where useful features will usually always be in the same area, so this is not seen often.

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

a. Standard scaler
b. Count vectorizer
c. PCA

ANS :refer Q.159

205. Can we use t-sne for transforming test data ? if not why ?
ANS: Refer Q.80

206. Find the sum of diagonals in the numpy array ?
ANS: **numpy.trace(array)**

207. Write the code to get the count of row for each category in the dataframes.
ANS: **df.groupby(columnname)[columnname].nunique()**

208. Difference between categorical cross entropy and binary cross entropy.
ANS: Categorical cross entropy is used for multi-class classification whereas binary cross Entropy is used for two class classification.

209. When you use w2v for test factorization, and we each sentence is having different words how can you forward data into models ?
210. What is tf idf weighted w2v ?
ANS: After obtaining w2v representation of a word we multiply it with a scalar value of tf-idf

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

212. What is the time complexity of SVD decomposition ?
ANS: O(mn **min**(n, m))

213. What is the difference between content based recommendation and collaborative recommendation ?
ANS: Content based RF users metadata and item metadata to get user and item vectors to compute similarities
In collaborative filter we try to find user and items matrix using SVD

214. Why do you think inertia actually works in choosing elbow point in clustering ?
ANS: Inertia: It is the sum of squared distances of samples to their closest cluster center.
If the use elbow method the inertia will abruptly change for the points out of cluster.

215. What is gradient clipping ?
ANS: In training a neural network sometimes we may experience exploding gradients beacause of large weights that exponentially increasing, to avoid this exploding gradient problem we use gradient clipping which limits the weight values with increasing.

216. Which of these layers will be a better option as a last layer in multilabel classification ?
a. Sigmoid
b. Softmax

ANS: Sigmoid because a point may have any number of classes .
In softmax we only get one class that have the max probability.

217. Is there a relation or similarity between LSTM and RESNET ?
ANS : Yes, In LSTM we use forgot gate which enables the to control the amount of data to be forwarded. Similarly in RESNET we use residual connections which makes the data to flow from the residual connection when weights are not useful.

218. What are the values returned by np.histogram()
ANS: It takes a numpy nd array as input, it flattens the array and return histogram values which can be plotted to see the histogram of data.

219. What is PDF, can we calculate PDF for discrete distribution ?
ANS: Yes

220. Can the range of CDF be (0.5 - 1.5).
ANS: No max value of a cdf is 1

221. Number of parameters in the following network :

a. Number of neurons = 4
b. Problem = binary classification
c. no: of FC = 2
d. Neurons in 1st FC = 5
e. Neurons in 2nd FC = 3

ANS: 54

222. How do we interpret alpha in dual form of sum? What is the relation between C and Alpha?
ANS: alpha and C are inversely proportional to each other.

223. How does back propagation work in case of LSTM?

224. What is the difference between supervised and unsupervised models?
ANS: In supervised learning we have the class labels for the datapoints
In unsupervised learning we don’t have any class label data to train the model with.

225. What is the derivative of this fraction 1/(1+e^sinx).
ANS: $(e^{(\sin(x))}\,(\cos(x)-1))/((1+e^{\sin^2(x)}))$

226. What will be the output of a = [1 2 3 10], [4 5 6 11], [7 8 9 12] a[:,:-1]

227. What is the output of this a = [1 5 9],[2 6 10],[3 7 11],[4 8 12] a[:-2,:]

228. What will be the output of

a. a= dict()
b. a[(‘a’,’b’)] = 0
c. a[(a,b)] = 1
d. print(a)

ANS: {(‘a’,’b’)}=1

229. What will be the output of

   a. a = [1 2 3],[4 5 6],[7 8 9]
   b. np.mean(a,axis=1)

ANS: [4,5,6]

230. What will be the output of

   a. a =[3 4 5],[6 7 8],[9 10 11]
   b. b = [1 2 3],[4 5 6],[7 8 9]
   c. np.stack((a,b), axis= 0)

231. What is “local outlier factor”?

ANS: LOF is the anomaly detection technique used in K-NN

232. How RANSAC works?

ANS: We take a sample of data from original data and train the model on sample data. The probability of outlier in sampled data will be reduces. We will remove the outlier by Computing the loss with point. We remove the points which have high loss value from The original data and repeat the same process until the convergence.

233. What are jaccard & Cosine Similarities

234. What are assumptions of Pearson correlation ?

ANS: Linearity and absence of outliers.

235. Differences between Pearson and Spearman correlation?

ANS : The fundamental difference between the two correlation coefficients is that the Pearson coefficient works with a linear relationship between the two variables whereas the Spearman Coefficient works with monotonic relationships as well.

236. What is the train time complexity of DBSCAN?

ANS : O(n^2)

237. Explain the procedure of “prediction in hierarchical clustering”

238. Relation between knn and kernel SVM

ANS : refer Q.86

239. Proof of “convergence of kmeans”

240. What is the optimal value of minpoints for the data (1000, 50)?