Table of Contents

[Things to learn 2](#_Toc105502197)

[Machine Learning, 2](#_Toc105502198)

[Key Points 7](#_Toc105502199)

[To read 7](#_Toc105502200)

[New things in market- 7](#_Toc105502201)

[Statistics, 7](#_Toc105502202)

[Mathematics, 10](#_Toc105502203)

[Deep learning 10](#_Toc105502204)

[Neural networks 10](#_Toc105502205)

[Interview questions 11](#_Toc105502206)

[Data structures 13](#_Toc105502207)

[Behavioural Interview Topics 13](#_Toc105502208)

[Important links 13](#_Toc105502209)

[Python 13](#_Toc105502210)

[Dunder methods/Magic methods 13](#_Toc105502211)

## Things to learn

1. Deep learning (NN, RNN, LSTM, CNN etc…) with real-time in python packages (NumPy, TensorFlow, Keras, sci-kit learn, torch)
2. Machine learning algorithms
3. Statistics
4. Time-series Forecasting
5. Time-series Forecasting in Deep learning
6. System design
7. Data structures
8. Python API creation
9. OLTP System design
10. OLAP System design
11. NLP
12. Mongo DB
13. Timeseries DB
14. AWS
15. AZURE
16. GCP
17. Docker image creation
18. Kubernetes
19. GO
20. Service mesh
21. Tweedie distribution

JAVA

1. Public <T extends BaseStep> T setLoopWeekDay(int weekDay) { -------how this works <T extends Basestep>

this.LoopWeekDay = weekday

return (T) this; ------------------------how return this works?

}

Gather the interview questions on Data science, Statistics, ML, DL, Python, SQL, Data modelling, System design, DevOps., and Cloud.

## Machine Learning,

1. KNN-
   1. K nearest neighbour is the classification or regression supervised learning algorithm.
   2. When we split the data into train and test and fit the KNeighborsClassifier, we need to give the number of neighbours.
   3. If any new data point is added to the dataset, based on the “Number of neighbours” it will get classified to the required class.
   4. For example, if neighbour = 1, then based on the nearest neighbour (using any distance formula) it will be classified.
   5. If we have neighbour =3, then based on the nearest for most of the data points, it will be classified for the respective class. Let’s consider I have green and red as two classifiers. If a new datapoint comes, if the data point is close to two green points and one red point, then it will be classified as the green point.
   6. We need to trade-off between choosing the neighbours, if we have fewer neighbours training accuracy will be good but not testing accuracy.
   7. KNN can be used for regression also. In regression, the average of the nearest points will be considered as the final value. But for classification, the mode of nearest points is considered.
2. Ridge regression
   1. Ridge regression is popular type of regularized linear regression that includes an L2 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction task.
   2. It penalizes the coefficients by α times. α is the hyperparameter, by default it is 1. Reducing the α and checking the which is the best fitted model.
   3. Ridge regression will not make the coefficients 0. It will reduce the least affected feature coefficient value to avoid overfitting.
   4. It adds “squared magnitude” of coefficient as penalty term to the loss function.

Text

Description automatically generated

1. Lasso regression
   1. Lasso regression is popular type of regularized linear regression that includes an L1 penalty. This has the effect of shrinking the coefficients for those input variables that do not contribute much to the prediction task. And makes the coefficients to 0. Sometimes this works as feature selection which will remove least affected features.
   2. It penalizes the coefficients by α times. α is the hyperparameter, by default it is 1. Reducing the α and checking the which is the best fitted model.
   3. Lasso regression will make the coefficients 0. It will reduce the least affected feature coefficient value to avoid overfitting.
   4. It adds “absolute value of magnitude” of coefficient as penalty term to the loss function.

Text

Description automatically generated

1. Decision Tree
   1. Decision tree algorithm works for both classification and regression dataset.
   2. It works like, if else conditions to the data, like if size of the feature is less than some number then it is 0 else it is 1. Like this it creates multiple if else conditions for each feature. Since decision tree creates the tree structure for the individual feature, scalability/normalization of the feature is not required.
   3. It creates each data point into one leaf. If any datapoint is near to the leaf, then it is classified to the respective leaf type. But it will overfit.
   4. So, pre-pruning and post pruning techniques are used to avoid overfitting the model.
      1. Pre-pruning- Limiting the maximum depth of the tree, maximum number of leaves, minimum number of points in the node.
      2. Post-pruning- Building the tree and collapsing nodes which has little information.
   5. DecisionTreeClassifier from scikit learn library only does the pre-pruning.
   6. The main downside of decision trees is that even with the use of pre-pruning, they tend to overfit and provide poor generalization performance.
   7. In decision tree only single tree will be formed. However, in random forest, multiple trees will be formed which will increase the efficiency of model.
2. Bagging
3. Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce variance within a noisy dataset.
4. In bagging, a random sample of data in a training set is selected with replacement meaning that the individual data points can be chosen more than once.
5. After several data samples are generated, these weak models are then trained independently, and depending on the type of task regression or classification, for example the average or majority of those predictions yield a more accurate estimate.
6. Example, Random forest.
7. Boosting
8. Boosting is the ensemble method to reduce the bias. The main difference between the bagging and boosting is, in bagging weak learners are trained in parallel while in boosting it will train in sequential.
9. This means that a series of models are constructed and with each new model iteration, the weights of the misclassified data in the previous model are increased.
10. This redistribution of weights helps the algorithm identify the parameters that it needs to focus on to improve its performance.
11. Examples, Adaboost, Xgboost, Gradient boost
12. Random Forest
    1. The idea behind random forests is that each tree might do a relatively good job of predicting but will likely be overfit on part of the data. If we build many trees, all of which work well and overfit in different ways, we can reduce the amount of overfitting by averaging their results.
    2. To build the random forest, we need to decide the number of trees to be created. It will be given as n\_estimators in RandomForestRegressor and RandomForestClassifier.
    3. It uses bootstrap splitting and create the trees on split data.
       1. bootstrap sample of the list ['a', 'b', 'c', 'd']. A possible bootstrap sample would be ['b', 'd', 'd', 'c']. Another possible sample would be ['d', 'a', 'd', 'a'].
    4. Instead of looking for the best test for each node, in each node the algorithm randomly selects a subset of the features, and it looks for the best possible test involving one of these features. The number of features that are selected is controlled by the max\_features parameter. This selection of a subset of features is repeated separately in each node, so that each node in a tree can make a decision using a different subset of the features.
    5. To make prediction using the random forest, the algorithm predicts from each tree. For the regression it takes the average of the values, for classification it does the soft voting, based on this it will classify the data point.
    6. Random forest is computationally expensive, but it can run in parallel by choosing the option of n\_jobs.
    7. Higher the number of trees the smoother the decision tree boundaries, but it is computationally expensive.
    8. It also gives importance to all the features unlike decision trees.
    9. Choosing the right hyperparameters like n\_estimators and max\_features is the challenging task.
13. Gradient boosted regression trees
    1. In contrast to the random forest approach, gradient boosting works by building trees in a serial manner, where each tree tries to correct the mistakes of the previous one.
    2. By default, there is no randomization in gradient boosted regression trees; instead, strong pre-pruning is used.
    3. Gradient boosted trees often creates shallow trees depth of one to five, which makes the model smaller and computes faster than random forest.
    4. Apart from the pre-pruning, learning rate also will be passed as hyperparameter. Using which it tries the learn from the previous trees.
    5. This algorithm also gives importance to features like random forest.
    6. The main parameters of gradient boosted tree models are the number of trees, n\_estimators, and the learning\_rate, which controls the degree to which each tree is allowed to correct the mistakes of the previous trees.
14. SVM Completely no clear, need to understand more on this
    1. The SVM learns how important each of the training data points is to represent the decision boundary between the two classes. Typically, only a subset of the training points matters for defining the decision boundary: the ones that lie on the border between the classes. These are called support vectors and give the support vector machine its name.
    2. To make prediction of new datapoint. The distance between the support vectors is measured. A classification decision is made on the distance to the support vectors and importance of the support vectors learned during training.
    3. The important parameters in SVM are ϒ(gamma) and C.
       1. ϒ(gamma) controls the width of the Gaussian kernel (boundary between the division of two classes, in this boundary support vectors will be residing).
       2. C controls is a regularization parameter like the one used in linear models.
    4. As per the book (Introduction to Machine learning in python), SVM works fine for small dataset. For large dataset its recommended to use RandomForest.
    5. In SVM, improving of accuracy has seen on scaling all features in the common scale.
15. Naïve Bayes classifier
    1. Computational is less expensive than linear models.
    2. In scikit, there are three different types. GuassianNB, BernoulliNB, MultinomialNB
       1. GuassianNB- Is used by continuous data.
       2. BernoulliNB- Is used mostly in text data classification.
       3. MultinomialNB- Is used mostly in text data classification.
    3. More to add
16. Adaboost
17. Adaboost is an ensemble learning method which was created to increase the efficiency of the classification. Adaboost uses iterative approach to learn from mistakes of week classifiers and turn them into strong ones.
18. Adaboost can use any classification algorithm, by default it uses decision tree.
19. Below steps are performed in adaboost.
    1. Week classifier is made on the training data based on the weighted samples. Initially all the samples have equal weights.
    2. More weights are assigned to the sample which is classified wrongly.
    3. New model will be created based on the weighted samples.
    4. Step b and step c is performed until the n\_estimators given in the hyperparameters.
    5. Final model will be created based on the weighted average of all models built. And using which prediction will be done to the testing dataset.
20. It has hyperparameter, learning rate (α)
    1. **base\_estimator**: The model to the ensemble, the default is a decision tree.
    2. **n\_estimators**: Number of models to be built.
    3. **learning\_rate**: shrinks the contribution of each classifier by this value.
    4. **random\_state**: The random number seed, so that the same random numbers generated every time.
21. K fold Cross Validation

Calendar

Description automatically generated with medium confidence

1. Xgboost
2. LDA
3. QDA
4. PCA
5. Hierarchy clustering
6. K-means clustering
7. Markov chain
8. Monte Carlo simulations
9. ADDA
10. Ada Match
11. Self-ensemble
12. Model Assessment and selection methods

### Key Points

1. Bagging- Bagging, also known as bootstrap aggregation, is the ensemble learning method that is commonly used to reduce **variance** within a noisy dataset.

Eg, Random forest

1. Boosting- Boosting is an ensemble meta-algorithm for primarily reducing **bias**.

Eg, Adaboost, Xgboost, Gradient boost and Brown boost

1. In bagging, weak learners are trained in parallel, but in boosting, they learn sequentially.
2. Stacking - Improve Predictions.
3. Decision tree doesn’t require standardization/normalization because it create tree like structure for individual feature.

### To read

https://machinelearningmastery.com/dropout-for-regularizing-deep-neural-networks/

### New things in market-

* Transformer
* BERT

## Statistics,

1. Pdf

Answer: probability for a continuous random variable is given by areas under pdf's

1. Cdf

Answer: Integration of pdf is cdf

1. Why p-value should be less than 5%?

Answer: If p-value is less than 5%, it means the data used for the model is greater than 95% of data. Let’s consider if we fit the features to the algorithm if we get one feature as p-value 0.38. This doesn’t contribute to model. Reason being, for that feature only 62% of the data getting used in model. Rest 38% of the data it doesn’t use. So, we remove that feature.

1. Contingency table

Answer: It is the tabular representation of the categorical data. Contingency table shows the frequency distribution of combinations of values of two discrete variable x and y.

Table

Description automatically generated

1. Degrees of freedom

Answer: In the most of formulae we are using n-1 in denominator. The reason for that could be if we use n then there will be slightly biased. But we do n-1 in denominator the bias will be reduced. However, it doesn’t matter much if the dataset is huge. And the n-1 is called as degrees of freedom.

1. Maximum likelihood estimation

Answer: If we have data with mean and standard deviation (sigma), the future data which is similar with current data will have same mean and standard deviation (sigma) of current data. So that it will increase the likelihood of estimation.

1. Central limit theorem

Answer: The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

1. Bootstrap

Answer: A sample taken with replacement from an observed dataset. The sample may not be in normal distribution.

1. Permutation test

Answer: The process of combining two groups of samples and randomly creating the sample from the combined group.

1. Resampling

Answer: Process of taking repeated samples from observed data. Repeated samples taking procedure can be bootstrap or permutation test.

1. Standardization or Normalization

Answer: Subtract each data point with its mean and divide it by standard deviation will make the data as Standardization or Normalization

1. Z distribution

Answer: It is nothing but the Standardization. Z distribution will have 0 mean and 1 SD.

1. T test,

Answer: a method of testing [hypotheses](https://www.merriam-webster.com/dictionary/hypotheses) about the [mean](https://www.britannica.com/science/mean) of a small [sample](https://www.britannica.com/science/sampling-statistics) drawn from a [normally distributed](https://www.britannica.com/topic/normal-distribution) population when the population [standard deviation](https://www.britannica.com/topic/standard-deviation) is unknown.

It is like Z test, but here it is sample standard distribution. However, Z test is population standard distribution. T test is very similar to Z test, but in T test it will have sd greater than 1. T test will have longer tails. Shape of the distribution looks like T shape. T distribution with high degrees of freedom will be exactly same as Z distribution.

1. One tail test

Answer: It will deal with situation where we are interested in testing hypothesis that the population mean is greater or less than the hypothesized value. So only one rejection region.

1. Two tail test

Answer: There will be 2 rejection regions one in each tail. If significance level is 5% then 2.5% in each tail is the rejection ratio.

1. Binomial distribution

Answer: If the outcome of the data is yes/no or True/False. Then it is in binomial distribution. Mean is np and variance are npq. Where q = 1-p

1. Poison distribution

Answer: If the data satisfies below 3 requirements, then it is Poison.

* The random events must take place in a unit of time or space.
* The number of events which might occur in any given unit of time or space must be theoretically unlimited.
* The probability of occurrence in any single unit of time or space is independent of occurrence in any other unit of time or space.

Mean and variance is µ

1. Exponential distribution
2. Chisquare test- It’s the procedure to test the observed count is as expected count which shows the independence. Like for example buy a particular item is independent of gender. In chisquare test, Expected(Ei) should be minimum of 5. If not adjust with adjacent cells and until all the Expected values are greater than 5. If degrees of freedom is 1 then subtracting 0.5 from the chi square formulae will give better results. For higher degrees of freedom we can use the existing formulae.
3. Fisher test- It is also same as Chisquare test to check independence. But it works only smaller dataset.
4. A/B testing – An A/B test is an experiment with two groups to establish which of two treatments, products, procedures, or the like is superior.
5. Kurtosis-

Kurtosis is a measure of whether the data is heavy- tailed or light-tailed relative to the normal distribution. That is data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have small tails or no outliers.

K = μ4/σ4 Where μ4 is the fourth central moment, σ is the standard deviation

Text

Description automatically generated

Where N is number of data points, s is standard deviation.

The kurtosis for a [standard normal distribution](https://www.itl.nist.gov/div898/handbook/eda/section3/eda3661.htm) is three.

1. Stationary poison process
2. Stochastic proximity embedding (SPE)

**Key Points**

1. In most of the formulae, we are using n-1 in denominator. The reason for that could be if we use n then there will be slightly biased. But we do n-1 in denominator the bias will be reduced. However, it doesn’t matter much if the dataset is huge. And the n-1 is called as degrees of freedom

## Mathematics,

1. Where does Fourier transform fit in ML models?
2. Wallet packet decomposition

## Deep learning

### Neural networks

Here is the list of top 10 most popular deep learning algorithms:

* Convolutional Neural Networks (CNNs)
* Long Short Term Memory Networks (LSTMs)
* Recurrent Neural Networks (RNNs)
* Generative Adversarial Networks (GANs)
* Radial Basis Function Networks (RBFNs)
* Multilayer Perceptrons (MLPs)
* Self Organizing Maps (SOMs)
* Deep Belief Networks (DBNs)
* Restricted Boltzmann Machines( RBMs)
* Autoencoders

Good article on neural networks-

<https://www.simplilearn.com/tutorials/deep-learning-tutorial/deep-learning-algorithm>

http://neuralnetworksanddeeplearning.com/chap1.html

#### Activation functions

Sigmoid-

* As a part of standardization, input features will be standardized with centered as 0. But Sigmoid activation function is centred at 0.5. Lowest value is 0 and highest is 1.
* As we use gradient descent for model, derivative of the sigmoid will be close to 0. So learning will be slow. If we have multiple hidden layers then learning rate will be extremely slow.
* It is recommended to use Sigmoid in output for the binary classification

Tanh-

Equation

Text

Description automatically generated

* Tanh looks similar to sigmoid but it is cantered to 0. With lowest value is -1 and highest is 1.
* This works better than sigmoid. But derivative of Tanh also close to 0.

Relu-

* Learning rate is good in Relu because derivative of relu is not 0 when x>0.
* But when x<0 derivative is 0. So some the nearons will become dead. So to overcome this leaky relu has introduced.
* This is most used in industry.

Leaky relu

* Derivative is positive in both x>0 and x<0. So this can overcome the negativity of Relu.
* Mean can be 0.

Elu(Exponential Linear Unit)

Equation

A picture containing chart

Description automatically generated

* As per the documentation, it speeds up the learning and higher the accuracy.
* Mean can be 0.

Softplus

Equation

A picture containing text

Description automatically generated

* Learning rate is good as Relu because derivative of relu is not 0 when x>0.

Softmax

* Activation function which will be used in the final layer for the multiclass classification.

## Interview questions

1. What cases mean is chosen over median?

Answer: If we are solving a regression problem, a new data point has been received to predict the target variable. If the data point is near the group of points, then the mean is better than the median in this case.

1. What cases median is chosen over the mean?

Answer: In the case of data cleaning, on doing the box plot, we can see the distribution of the data. If the data concentration is near the median, then the median can be used to fill the NA’s instead of the mean.

1. What is the p-value?

Answer: p-value is the probability value, is a number determining how likely it is that your data would have occurred by chance.

At 5% significance,

If p value < 0.05 reject the null hypothesis

If the p-value > 0.05 fail to reject the null hypothesis

1. What is hypothesis testing?

Null hypothesis – always predicts no effect or no relationship between the variables.

Alternate hypothesis – states your research prediction of an effect or relationship.

1. Confidence interval?

A confidence interval displays the probability that a parameter will fall between a pair of values around the mean. Confidence intervals measure the degree of uncertainty or certainty in a sampling method. They are most often constructed using confidence levels of 95% or 99%.

1. Statistical power

Statistical power or the power of a hypothesis test is the probability that the test correctly rejects the null hypothesis.

That is the probability of a true positive result. The power is precisely the inverse of the probability of a Type II error.

Power = 1 - Type II Error

Pr(True Positive) = 1 - Pr(False Negative)

1. Assumptions of linear regression?

* Error term and the independent variable are not related.
* The relationship between x and y is linear.
* The observations are independent of each other.
* The distribution of errors is normally distributed and the distribution of y is also normally distributed.

1. Type I and Type II error
   * Type I Error is rejecting the null hypothesis in favour of a false alternative hypothesis.
   * Known as False positive
   * The probability of a Type I error is typically known as Alpha.
   * Type II Error is failing to reject a false null hypothesis in favour of a true alternative hypothesis.
   * Known as False negative
   * The probability of a Type II error is typically known as Beta.
2. What is the difference between One sample T-test and paired T-test?

One sample T-test - is used to compare a single population to a standard value (For example, to determine whether the city average is different from the country average)

Paired T-test - is used to compare a single population before and after experimental intervention or at two different points of time. (For example, measuring student performance before and after being taught the material)

## Data structures

1. Inorder
2. Preorder
3. Postorder

## Behavioural Interview Topics

1. Leadership principles

2. STAR format-(Situation Task Action Result)

## Important links

Statistical power <https://machinelearningmastery.com/statistical-power-and-power-analysis-in-python/>

## Python

### Dunder methods/Magic methods

* Magic methods are special methods in python which start and ends with a double underscore.
* It gets invokes internally from the class on certain activities.
* Let’s consider if we add two variables which are integers with a + operator. Then \_\_add\_\_ gets invoked internally and performs that operation.
* dir() function to see the number of magic methods invoked by the class.
* Dunder methods can be overridden.

Examples of override are in the below link

<https://www.tutorialsteacher.com/python/magic-methods-in-python>

Important links-

Recommendation engine-

<https://uxplanet.org/how-can-we-design-an-intelligent-recommendation-engine-b9bb1db4d050>