Machine Learning interview question with answers

1. **What is ROC curve and what does it represent?**

Receiver Operating Characteristic curve (or ROC curve) is a fundamental tool for diagnostic test evaluation and is a plot of the true positive rate (Sensitivity) against the false positive rate (Specificity) for the different possible cut-off points of a diagnostic test. Y neri arancqi vra True Positive rate I arjeqna 0-1 mijakayqum isk x eri arancqna vra false positive I rate@

True positive I rate@ aysinqn vor@ menq positive enq gushake irakanum el true a, isk false positive aysinqn menq positive enq gushake bayc irakanum false e exe et kaxumna cuyc talis. Tanumenq y = x uxix@ u dranic verec ROC curve na linum u inchqan mec lini et makeres@ etqan mer test seti vra predictionner@ lavnen.Ev karevor past chen kara erkusne mecanan ete True Positive I rate @ mecanuma False Positive in@ poqranuma ev hakarak@, bayc bnakanabar mez arachin case na petq.

Tesnelu hamar curve gnam es linkov - <https://www.edureka.co/blog/interview-questions/machine-learning-interview-questions/>

1. **Is it better to have too many false positives or too many false negatives? Explain.**

It depends on the question as well as on the domain for which we are trying to solve the problem. If you’re using Machine Learning in the domain of medical testing, then a false negative is very risky, since the report will not show any health problem when a person is actually unwell. Similarly, if Machine Learning is used in spam detection, then a false positive is very risky because the algorithm may classify an important email as spam.

1. Inchia anhnar ogtagorcel ordinary least squares approach@ erb vor feature neri qanak@ aveli meca qan te training example neri.

<https://medium.com/@jennifer.zzz/more-features-than-data-points-in-linear-regression-5bcabba6883e>

1. What is the difference between R squared and adjusted r squared?

Both R2 and the adjusted R2 give you an idea of how many data points fall within the line of the regression equation. However, there is one main difference between R2 and the adjusted R2: R2 assumes that every single variable explains the variation in the dependent variable. The adjusted R2 tells you the percentage of variation explained by only the independent variables that actually affect the dependent variable

The adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. Suppose you compare a five-predictor model with a higher R-squared to a one-predictor model. Does the five predictor model have a higher R-squared because it’s better? Or is the R-squared higher because it has more predictors? Simply compare the adjusted R-squared values to find out! The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it’s usually not.  It is always lower than the R-squared.

1. What is the difference between validation and test set.

Himnakan tarberutyun@ ayn e vor ete unenq model u menak bajanelelnq train test split ov vor evaluate anenq test seti vra menq karanq patahakan lav score stananq test set I vra bayc hnaravora vor konkret test seti vra enq overfit are u ete urish test set lini karoxa lriv urish vat score stananq.Dra hamar baci test set ic pahumenq validation set @ u ogtagorcum orinak cross validation texnikan vor haskananq ardyoq mer model@ lava generalize anum validation seteri vra u ete lava anum arden verchum evaluate enq anum test seti vra inch vor arumov vstah linelov vor sax test seteri vra overfit chenq ani.

1. How long do we usually grow a tree and how to avoid overfitting after growing a deep tree. Difference between weekest link pruning and cost complexity pruning.

We know that in order to represent the data with a decision tree first thing to do is to grow a tree from to root to it’s leaf where during each split to nodes we use some information theory and information gain( measuring the entrpopy and gini index). Or we grow a tree using a top-down greedy approach until we reach some criterion for example if in each decision region there are only threshold or less elements left which we can correctly classify or if our loss( points – average squared) is changing very little after next split. Using either criterion mentioned above we end up with very deep trees which are pron to overfitting.

To avoid overfiting we need to prun the tree either using weekest link prunng or cost complexity pruning. In weekest link pruning menq sksumenq amenanerqevic u amen anqam node enq ktrum very barcranalov u tenc m anqam anelov stanumenq m hat tarber tree vordex mer m rd tree n stacvuma vor uni mi hat leaf node. Heto cross validation anelov stanumenq vorna ameaeffektiv tree n .

Cost complexity pruning I jamanak mer loss function I decision tree I avelacnumenq alpha anqam modul T andam@ vordex alphan hyperparametra isk modul T n mer To mec deep tree I subtree I erkarutyunna. Aysinqn sa nmana nran vor regularization parameter enq avelacnum vor mer carer@ shat deep chlinen. Parza ete alpha n anverchutyun vercnenq uremn petqa null treen vercnenq isk ete alpha n zro ya mer full To tree n enq vercnelu. Fun fact is that for each alpha the solution is in trees we got from other pruning technique.

Now let’s understand how to use cross-validation here.

1. Split the training points into 10 folds.

2. For k = 1, . . . , 10, using every fold except the kth:

Construct a sequence of trees T1, . . . , Tm for a range of values of and find the prediction for each region in each one and u ete m = 9 vekalumenq 9 hat car orinak alphan = 0.1 ov u tenum et 9ic vor na amenavoch xor@ car@ vordex error@ der shat chi poxvum( aysinqn ete mi hat node hanem u uje error@ shat mecana uremn kvercnem alphan = 0.1 I depqum henc et car@ arachi foldic)

Ardyunqum amen foldi hamar stacanq lav tree er u iranc voroshumeri average nenq vekalum alpha = 0.1 I depqum u heto anum 3rd qayl@ u heto veradarnumenq skzbic nuyn baner@ anumenq bayc alphan urish vercnelov u tenc gtnumenq lavaguyn alphan vori depqum error@ amenapoqrna

3. Select the parameter that minimizes the average test error.

1. Why is the mean of highly correlated quantities higher than the mean of low correlated quantities.

When the elements of each sample in this process are positively correlated, when one value is high the others tend to be high, too. Their mean will then be high. When one value is low the others tend to be low, too. Their mean will then be low. Thus, the means tend either to be high or low.

When elements of each sample are not correlated, the amount by which some elements are high is often balanced (or "canceled out") by other low elements. Overall the mean tends to be very close to the average of the population from which the samples are drawn--and rarely much greater or much less than that.

1. What are the advantages and disadvantages of stochastic gradient descent over gradient descent.

## Advantages of Stochastic Gradient Descent

1. It is easier to fit into memory due to a single training sample being processed by the network
2. It is computationally fast as only one sample is processed at a time
3. For larger datasets it can converge faster as it causes updates to the parameters more frequently
4. Due to frequent updates the steps taken towards the minima of the loss function have oscillations which can help getting out of local minimums of the loss function (in case the computed position turns out to be the local minimum)

## Disadvantages of Stochastic Gradient Descent

1. Due to frequent updates the steps taken towards the minima are very noisy. This can often lead the gradient descent into other directions.
2. Also, due to noisy steps it may take longer to achieve convergence to the minima of the loss function
3. Frequent updates are computationally expensive due to using all resources for processing one training sample at a time

## Advantages of Batch Gradient Descent

1. Less oscillations and noisy steps taken towards the global minima of the loss function due to updating the parameters by computing the average of all the training samples rather than the value of a single sample
2. It can benefit from the vectorization which increases the speed of processing all training samples together
3. It produces a more stable gradient descent convergence and stable error gradient than stochastic gradient descent
4. It is computationally efficient as all computer resources are not being used to process a single sample rather are being used for all training samples

## Disadvantages of Batch Gradient Descent

1. Sometimes a stable error gradient can lead to a local minima and unlike stochastic gradient descent no noisy steps are there to help get out of the local minima
2. The entire training set can be too large to process in the memory due to which additional memory might be needed
3. Depending on computer resources it can take too long for processing all the training samples as a batch

This ensures the following advantages of both stochastic and batch gradient descent are used due to which Mini Batch Gradient Descent is most commonly used in practice.

1. Easily fits in the memory
2. It is computationally efficient
3. Benefit from vectorization
4. If stuck in local minimums, some noisy steps can lead the way out of them
5. Average of the training samples produces stable error gradients and convergence
6. What is the main differnce in ridge and lasson and which is used when?

in ridge regression regression we make squared penalty and coefficients chen kara darnan 0, aysinqn menq ridge kogtagorcenq en jamanak erb mer useless featurener@ qichen isk ete mer useless featurener@ shaten menq kogtagorcenq lasso regression vortev lasson vat featureneri coefficient@ karuma sarqi 0 moduli hashvin.Lambdan gtnumenq cross validationov u ogtagorcumenq overfitting ic avoid anelu hamar.

Elastic Net@ ogtagorcuma Lasso u Ridge regression penalty n irar het u inq@ aveli lava correlacvac featureneri het vortev elastici oqnutyamb karumenq rad anenq dranq.

1. What are the main disadvantages of K means.
2. The globally optimal result may not be achieved, although the E–M procedure is guaranteed to improve the result in each step, there is no assurance that it will lead to the global best solution. For example, if we use a different random seed in our simple procedure, the particular starting guesses lead to poor results: Aysinqn kaxvac mer random initialize arac centroidneric menq karanq stananq lriv tarber resultner. Dra hamar asenq Sckit Learn@ ogtagorcuma 10 hat random initialization.
3. Another common challenge with k-means is that you must tell it how many clusters you expect: it cannot learn the number of clusters from the data. For example, if we ask the algorithm to identify six clusters, it will happily proceed and find the best six clusters:
4. The fundamental model assumptions of *k*-means (points will be closer to their own cluster center than to others) means that the algorithm will often be ineffective if the clusters have complicated geometries.In particular, the boundaries between *k*-means clusters will always be linear, which means that it will fail for more complicated boundaries.
5. Because each iteration of k-means must access every point in the dataset, the algorithm can be relatively slow as the number of samples grows. You might wonder if this requirement to use all data at each iteration can be relaxed; for example, you might just use a subset of the data to update the cluster centers at each step.
6. What is the difference between Kmeans and Hierarchical Clustering.

* Hierarchical clustering can’t handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. O(𝑛2n2).
* In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
* K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
* K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

1. When we need to use DBscan or t-SNE ?

DBSCAN (Density-based spatial clustering of applications with noise) is a density based clustering algorithm, which ahs the feature of detecting outliers.

t-SNE (T-distributed Stochastic Neighbor Embedding) is a non-linear density based clustring technique. It is mainly used to reduce the dimension of data and have easier way to visualize it.

**t-SNE** helps make the cluster more accurate because it converts data into a 2-dimension space where dots are in a circular shape. **t-SNE** puts similar cases together, handling non-linearities of data very well. After using the algorithm on several data sets, I believe that in some cases it creates something like circular shapes like islands, where these cases are similar.