Ensemble- Bagging Interview Questions

We believe that you have learned both theoritical and practical knowledge on Naive Bayes classification algorithm through your assignment.

So let's test your knowledge here. This will help you to be prepared for interviews too!

Best with Quest ¶

1. What do you mean by Random Forest Algorithm? Can Random Forest Algorithm be used both for Continuous and Categorical Target Variables?

Random forest is an ensemble machine learning technique that averages several decision trees on different parts of the same training set, with the objective of overcoming the overfitting problem of the individual decision trees.

In other words, a random forest algorithm is used for both classification and regression problem statements that operate by constructing a lot of decision trees at training time.

Yes, Random Forest can be used for both continuous and categorical target (dependent) variables.

In a random forest i.e, the combination of decision trees, the classification model refers to the categorical dependent variable, and the regression model refers to the numeric or continuous dependent variable.

2. Explain the working of the Random Forest Algorithm.

The steps that are included while performing the random forest algorithm are as follows:

Step-1: Pick K random records from the dataset having a total of N records.

Step-2: Build and train a decision tree model on these K records.

Step-3: Choose the number of trees you want in your algorithm and repeat steps 1 and 2.

Step-4: In the case of a regression problem, for an unseen data point, each tree in the forest predicts a value for output. The final value can be calculated by taking the mean or average of all the values predicted by all the trees in the forest.

and, in the case of a classification problem, each tree in the forest predicts the class to which the new data point belongs. Finally, the new data point is assigned to the class that has the maximum votes among them i.e, wins the majority vote.

3. Why do we prefer a Forest (collection of Trees) rather than a single Tree?

While building a machine learning model, our aim is to generalize the model properly for giving predictions on unseen data.

The problem of overfitting takes place when we have a flexible model. A flexible model is having high variance because the learned parameters like the structure of the decision tree, etc will vary with the training data. On the contrary, an inflexible model is said to have a high bias as it makes assumptions about the training data and an inflexible model may not have the capacity to fit even the training data and in both situations, the model has high variance, and high bias implies the model is not able to generalize new and unseen data points properly.

So, we have to build a model carefully by keeping the bias-variance tradeoff in mind.

The main reason for the overfitting of the decision tree due to not put the limit on the maximum depth of the tree is because it has unlimited flexibility, which means it keeps growing unless, for every single observation, there is one leaf node present.

Moreover, instead of limiting the depth of the tree which results in reduced variance and an increase in bias, we can combine many decision trees that eventually convert into a forest, known as a single ensemble model (known as the random forest).

4. What do you mean by Bootstrap Sample? Why does the Random Forest algorithm not require split sampling methods?

It is basically random with the replacement sampling method.

For Example, Suppose we have a box of lottery tickets in which there are 100 unique numbers from 0 to 99. We want to select a random sample of tickets from the box. If we put the ticket back in the box, it may be selected more than once. Therefore, in this process, we are picking the samples randomly from the box with replacement.

Random Forest does not require a split sampling method to assess the accuracy of the model.

This is because it performs internal testing on 2/3rd of the available training data that is used to grow each tree and the remaining one-third portion of training data is always used to calculate out-of-bag error to compute the model performance.

5. What is Out-of-Bag Error? How to determine the overall OOB score for the classification problem statements in a Random Forest Algorithm?

Out-of-Bag is equivalent to validation or test data. In random forests, there is no need for a separate testing dataset to validate the result. It is calculated internally, during the algorithm run, in the following manner –

As the forest is built on training data, each tree is tested on 1/3rd of the samples (36.8%) that are not used in building that tree (similar to the validation data set).

This is known as the out-of-bag error estimate which in short is an internal error estimate of a random forest as it is being constructed.

For each tree, by using the leftover (36.8%) data, compute the misclassification rate, which is known as out of bag (OOB) error rate. Finally, we aggregate all the errors from all trees and we will determine the overall OOB error rate for the classification.

For Example, If we grow 300 trees then on average a record will be OOB for about 37*3 =111 trees.

6. What does random refer to in 'Random Forest'?

'Random' in Random Forest refers to mainly two processes –

Random observations to grow each tree.
Random variables selected for splitting at each node.
Random Record Selection: Each tree in the forest is trained on roughly 2/3rd of the total training data (exactly 63.2%) and here the data points are drawn at random with replacement from the original training dataset. This sample will act as the training set for growing the tree.

Random Variable Selection: Some independent variables(predictors) say, m are selected at random out of all the predictor variables, and the best split on this m is used to split the node.

NOTE:

By default, m is taken as the square root of the total number of predictors for classification whereas m is the total number of all predictors divided by 3 for regression problems.
The value of m remains constant during the algorithm run i.e, forest growing.

7. List down the features of Bagged Trees. What are the Limitations of Bagging Trees?

The main features of Bagged Trees are as follows:

1. Reduces variance by averaging the ensemble's results.

2. The resulting model uses the entire feature space when considering node splits.

3. It allows the trees to grow without pruning, reducing the tree-depth sizes which result in high variance but lower bias, which can help improve the prediction power.

The major limitation of bagging trees is that it uses the entire feature space when creating splits in the trees.

Suppose from all the variables within the feature space, some are indicating certain predictions, so there is a risk of having a forest of correlated trees, which actually increases bias and reduces variance. So, our objective is not achieved due to these issues.

8. List down the factors on which the forest error rate depends upon.

The forest error rate in Random forest depends on the following two factors:

1. How correlated the two trees in the forest are i.e,

The correlation between any two different trees in the forest. Increasing the correlation increases the forest error rate.

2. How strong each individual tree in the forest is i.e,

The strength of each individual tree in the forest. In a forest, a tree having a low error rate is considered a strong classifier. Increasing the strength of the individual trees eventually leads to a decrement in the forest error rate.

Moreover, reducing the value of mtry i.e, the number of random variables used in each tree reduces both the correlation and the strength. Increasing it increases both. So, in between, there exists an “optimal” range of mtry which is usually quite a wide range.

Using the OOB error rate, a value of mtry can quickly be found in the range. This parameter is only adjustable from which random forests are somewhat sensitive.

9. List down the advantages and disadvantages of the Random Forest Algorithm.

Advantages:

Random Forest is unbiased as we train multiple decision trees and each tree is trained on a subset of the same training data.
It is very stable since if we introduce the new data points in the dataset, then it does not affect much as the new data point impacts one tree, and is pretty hard to impact all the trees.
Also, it works well when you have both categorical and numerical features in the problem statement.
It performs very well, with missing values in the dataset.

Disadvantages:

Complexity is the major disadvantage of this algorithm. More computational resources are required and also results in a large number of decision trees combined together.
Due to their complexity, training time is more compared to other algorithms.

10. What is the difference between Extra Tree and Random Forest Algorithms. How are they different from descision tree?

ExtraTreesClassifier is like a brother of RandomForest but with 2 important differences.

We are building multiple decision trees. For building multiple trees, we need multiple datasets. Best practice is that we don't train the decision trees on the complete dataset but we train only on fraction of data (around 80%) for each tree. In a random forest, we draw observations with replacement. So we can have repetition of observations in a random forest. In an ExtraTreesClassifier, we are drawing observations without replacement, so we will not have repetition of observations like in random forest.

The split is the process of converting a non-homogeneous parent node into 2 homogeneous child node (best possible). In RandomForest, it select the best split to convert the parent into the two most homogeneous child nodes. In an ExtraTreesClassifier, it selects a random split to divide the parent node into two random child nodes.

Let’s look at some ensemble methods ordered from high to low variance, ending in ExtraTreesClassifier.

1. Decision Tree (High Variance)

A single decision tree is usually overfits the data it is learning from because it learn from only one pathway of decisions. Predictions from a single decision tree usually don’t make accurate predictions on new data.

2. Random Forest (Medium Variance)

Random forest models reduce the risk of overfitting by introducing randomness by:

building multiple trees (n_estimators)

drawing observations with replacement (i.e., a bootstrapped sample)
splitting nodes on the best split among a random subset of the features selected at every node. Split is process to convert non-homogeneous parent node into 2 homogeneous child node(best possible).

3. Extra Trees (Low Variance)

Extra Trees is like a Random Forest, in that it builds multiple trees and splits nodes using random subsets of features, but with two key differences: it does not bootstrap observations (meaning it samples without replacement), and nodes are split on random splits, not best splits. So in summary, ExtraTrees:

builds multiple trees with bootstrap = False by default, which means it samples without replacement
nodes are split based on random splits among a random subset of the features selected at every node
In Extra Trees, randomness doesn't come from bootstrapping the data, but rather comes from the random splits of all observations. ExtraTrees is named for (Extremely Randomized Trees).

11. What is the difference between Random Search and Grid Search Hyperparameter tuning methods and which one is better?

In Grid Search, we try every combination of a preset list of values of the hyper-parameters and evaluate the model for each combination. The pattern followed here is similar to the grid, where all the values are placed in the form of a matrix. Each set of parameters is taken into consideration and the accuracy is noted. Once all the combinations are evaluated, the model with the set of parameters which give the top accuracy is considered to be the best.4
One of the major drawbacks of grid search is that when it comes to dimensionality, it suffers when the number of hyperparameters grows exponentially.

Random search is a technique where random combinations of the hyperparameters are used to find the best solution for the built model. It tries random combinations of a range of values. To optimise with random search, the function is evaluated at some number of random configurations in the parameter space.The chances of finding the optimal parameter are comparatively higher in random search because of the random search pattern where the model might end up being trained on the optimised parameters without any aliasing.

Random search works best for lower dimensional data since the time taken to find the right set is less with less number of iterations. Random search is the best parameter search technique when there are less number of dimensions.
Hence random search has been shown to find equal or better values than grid search within fewer function evaluations for certain types of problems.

12. What are the benefits of ensemble model? Explain bagging.

There are two major benefits of Ensemble models:

- Better prediction
- More stable model

The aggregate opinion of a multiple models is less noisy than other models. In finance, we called it “Diversification” a mixed portfolio of many stocks will be much less variable than just one of the stocks alone. This is also why your models will be better with ensemble of models rather than individual. One of the caution with ensemble models are over fitting although bagging takes care of it largely.

Bagging stands for bootstrap aggregating. With bagging you uniformly sample with replacement from the data in order to make a bunch of different subsets. Then train a learner on each subset of the data. And combine every learning by simply taking the average of all the individual learner’s outputs.

Bagging is considered to be parallel because each model is built independently (subsets are sampled uniformly and with replacement). Bagging is used to decrease variance (decrease an overfit model). It can not improve the overall predictive force of the model because it is using the same data but it can be used on complex models to smooth output and lead to a better out of sample prediction. Random forest is an example of bagging.

Now Rest with this Quest :)