# MACHINE LEARNING

Q1 to Q15 are subjective answer type questions, Answer them briefly.

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

Answer 1 In multiple regression models, R-square is sensible to the number of independent variables you include in the regression model, is a kind of inflation in the power of explanation of the model because independent of real contribution in the value of R-square, the R-square value increases as you add more independent variables.

That's the formula to calculate:

Adjusted R squared = 1 - [(1-R squared) * [(N - 1)] / (N - 1 - p)]

Where N is the total sample size and p is the number of predictors.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

Answer : TSS or Total sum of squares gives the total variation in Y. We can see that it is very similar to the variance of Y. While the variance is the average of the squared sums of difference between actual values and data points, TSS is the total of the squared sums.
$Tss=\sum(yi-\bar{y})2$
The ESS is then: where. is the value estimated by the regression line . In some cases (see below): total sum of squares (TSS) = explained sum of squares (ESS) + residual sum of squares (RSS).

RSS: RSS is the residual (error) term we have been talking about so far and TSS is the total sum of squares.

3. What is the need of regularization in machine learning?

Answer : Regularization in machine learning serves as a method to forestall a model from overfitting. Overfitting transpires when a model not only discerns the inherent pattern within the training data but also incorporates the noise, potentially leading to subpar performance on fresh, unobserved data.

4. What is Gini–impurity index?

Answer : More precisely, the Gini Impurity of a dataset is a number between 0-0.5, which indicates the likelihood of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset.

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Answer : Complexity: Decision trees become overly complex, fitting training data perfectly but struggling to generalize to new data. Memorizing Noise: It can focus too much on specific data points or noise in the training data, hindering generalization.

6. What is an ensemble technique in machine learning?

Answer : Ensemble learning is a machine learning technique that enhances accuracy and resilience in forecasting by merging predictions from multiple models. It aims to mitigate errors or biases that may exist in individual models by leveraging the collective intelligence of the ensemble.

7. What is the difference between Bagging and Boosting techniques?

Answer :Bagging is a learning approach that aids in enhancing the performance, execution, and precision of machine learning algorithms. Boosting is an approach that iteratively modifies the weight of observation based on the last classification. It is the easiest method of merging predictions that belong to the same type

Bagging and boosting are ensemble learning techniques. Bagging (Bootstrap Aggregating) reduces variance by averaging multiple models, while boosting reduces bias by combining weak learners sequentially to form a strong learner.

8. What is out-of-bag error in random forests?

Answer :Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging). Bagging uses subsampling with replacement to create training samples for the model to learn from.

9. What is K-fold cross-validation?

Answer : By employing K-fold cross-validation, with features like test_index and train_index, we can mitigate overfitting and understand how the model generalizes to unseen data. Furthermore, we will examine the role of neural networks in classification tasks, highlighting their application in subsamples and their ability to learn complex patterns. Join us on this journey to optimize your machine learning models and enhance their performance.

10. What is hyper parameter tuning in machine learning and why it is done?

Answer : Hyperparameters directly control model structure, function, and performance. Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success.

11. What issues can occur if we have a large learning rate in Gradient Descent?

Answer : If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge. Overfitting: Gradient descent can overfit the training data if the model is too complex or the learning rate is too high

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Answer : Logistic regression is simple and easy to implement, but it also has some drawbacks. One of them is that it assumes a linear relationship between the input features and the output. This means that it cannot capture the complexity and non-linearity of the data.

13. Differentiate between Adaboost and Gradient Boosting.

Answer : Overall gradient boosting is more robust to outliers and noise since it equally considers all training instances when optimizing the loss function. AdaBoost is faster but more impacted by dirty data since it fixates on hard examples.
AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

14. What is bias-variance trade off in machine learning?

Answer : BAIS: Bais represents the error introduced by a model's assumptions and simplifications about the data.
A model with high bais tends to be too simplistic, making it unable t
O capture the true underlying patterns in the data. It often leads to underfitting, meaning the model performs poorly both on the training data and new, unseen data because it oversimplifies the problem.

Variance : variance on the other hand, refers to the model's sensitivity to the fluctuations or noise in the training data.
A model with high variance is too complex and flexible, essentially "memorizing" the noise in the traning data trainer than generalizing well to new, unseen data.
High variance leads to overfitting, where the model performs execellently on the training data but poorly on new data because it hasn't learned the true patterns but instead has adapted to the specific examples in the training set.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Answer : Linear Kernel

A linear kernel is a type of kernel function used in machine learning, including in SVMs (Support Vector Machines). It is the simplest and most commonly used kernel function, and it defines the dot product between the input vectors in the original feature space.

Polynomial Kernel

A particular kind of kernel function utilised in machine learning, such as in SVMs, is a polynomial kernel (Support Vector Machines). It is a nonlinear kernel function that employs polynomial functions to transfer the input data into a higher-dimensional feature space.

Gaussian (RBF) Kernel

The Gaussian kernel, also known as the radial basis function (RBF) kernel, is a popular kernel function used in machine learning, particularly in SVMs (Support Vector Machines). It is a nonlinear kernel function that maps the input data into a higher-dimensional feature space using a Gaussian function.

FLIP ROBO