Machine learning

1. R\_ squared is generally considered a better measure of goodness of fit than Rss because it provides a normalized metric that indicates the proportion of variance explained by the model.Rss alone does not account for the number of predictors and can be influenced by the number of data points,whereas R-squared adjusts for that ,making it easier to compare models.
2. Tss (total sum of square): measures the total variance in the response variance. Ess(explained sum of squares): measures the variance explained by the regression model. Rss(residual sum of square): measures the variance that remains unexplained by the model. The relationship is given by TSS=ESS+RSS.
3. Regularization is needed in machine learning to prevent overfitting, which occurs when a model learns noise instead of underlying data pattern. Regularization techniques add a penalty to the loss function, discouraging overly complex models and improving generalization to unseen data.
4. Gini impurity index: The gii measures the impurity or disorder of a dataset.it ranges from 0(pure) to 1(impure). For a node in a decision tree, it is calculated as : Gini = 1-Σpi2 where pi is the probability of class i.
5. Yes, unregularized decision trees are prone to overfitting because they can create overly complex models that perfectly fit the training data, including its noise. This leads to poor generalization to new data.
6. Ensemble technique in machine learning: An ensemble technique combines multiple models to improve overall performance . By aggregating predictions from different models, it reduces the risk of overfitting and increases robustness and accuracy.
7. Bifference between bagging and boosting: Bagging(bootstrap aggregating): Builds multiple models independently using random subsets of data and averages their predictions. It reduces variance. Boosting: Builds model sequentially, which each model corrects the error of its predecessor. It combines weak learners to create a strong learner, reducing data.
8. Out bag error in random forests: is an internal validation measure in random forests. For each tree, it uses data points that were not included in the bootstrap sample to evaluate the model . It provides an unbiased estimate of the models performance.
9. K- fold validation: involves dividing the dataset into k subsets (folds). The model is trained on k-1 folds and validated on the remaining fold. This process is repeated k times, allowing each fold to be used for validation, which helps assess model performance more reliably.
10. Hyper parameter tuning: is the process of optimizing the parameters that are not learned by model during training. It is done to enhance model performance , control overfitting and improve generalization by finding the best configuration of hyperparameters.
11. Issues with large learning rate in gradient descent: a large learning rate can cause gradient descent to overshoot the minimum of loss function, leading to divergence or oscillation instead of convergence. This can result in failure to find the optimal parameters.
12. Logistic regression is inherently a linear classifier, meaning it cannot directly model non-linear relationship. However, you can use polynomial features or kernel methods to enable it to fit non-linear data.
13. d/b btw adaboost and gradient boosting: Adaboost: focuses on adjusting the weights of misclassified instances in each iteration. It combines weak classifiers by focusing more on those that were previously misclassified. Gradient boosting: builds trees sequentially, optimizing the loss function using gradients. Each new model corrects the errors of the previous one , typically producing stronger models.
14. BIAS- VARIANCE TRADE OFF: describes the balance between two sources of error in predictive models.High bias leads to underlifting (oversimplified models), while high variance leads to overfitting (too complex models). The goal is to find a model that minimizes both bias and variance for better generalization.
15. SVM KERNELS: LINEAR KERNEL: Represents the simplest form of kernel , suitable for linearly separable data. It uses the dot product of input features . RBF(Radial basis function) kernel: a non-linear kernel that maps input features into an infinite-dimensional space ,allowing for complex decision boundaries. It is effective for cases where the relationship is not linear. Polynomial kernel: computes the similarity of input features as a polynomial function, allowing for curved decision boundaries. It is useful for data that has polynomial relationship.