Ayiman Bekowei

ASSIGNMENTS 5

Data Science

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

R- squared is commonly used metric to assess goodness of fit because it provides a measure of the overall explanatory power of the model

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

Total Sum of squares

(TSS) represent the total variability in the dependent variable (Y)

It is the sum of the squared differences between each observed Y value and the mean of Y.

Explained sum of squares

(ESS) measures the variability in the dependent variable that is explained by the independent variables in the regression model. Its also the sum of the squared differences between the predicted Y values and the mean Y.

Residual sum of squares

(RSS) measure the unexplained variability in the dependent representing the errors or residuals of model. Its also the sum of the squared differences between the observed Y values and the predicated Y values.

The relationship between TSS, ESS AND RSS can be summarized by this equation.

TSS = ESS + RSS

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

Answer

Regularization is a technique used in machine learning to prevent overfitting and improve the generalization performance of model. And can be applied on preventing overfitting, handling multicollinearity, feature selection etc.

4). What is Gini–impurity index?

Answer

The Gini impurity index is a measure used in decision tree algorithms, particularly in the context of classification problems. It quantifies the impurity or disorder of a set of examples, showing how often a randomly chosen element would be incorrectly classified. Decision trees use impurity measures to make decisions at each node during three building process.

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

Answer

Unregularized decision trees are prone to overfitting. Decision trees have a natural tendency to create complex, deep structures that perfectly fit the training data, capturing even the noise or random fluctuations in data.

Key reason why unregularized trees are prone to overfitting include, high variance, memorization of training data, failure to generalize, pruning, limiting tree depth etc.

6). What is an ensemble technique in machine learning?

Answer

Ensemble techniques in machine learning involve combining multiple individual models to create stronger, more robust predictive model. Furthermore, ensemble methods are widely used across various machine learning tasks and are particular effective in improving predictive performance, generalization and robustness.

Examples of Ensemble techniques include,

Bagging (bootstrap Aggregation), boosting, stacking, voting classifiers.

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

Answers

Bagging is multiple instance of the same learning algorithm are trained on different subsets of training data, each created by sampling with replacement.

While Boosting focuses on training multiple weak learns sequentially with each new model giving more weights to examples that were misclassified by the preceding models.

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

Answer

The out of bag error is a way to estimate the performance of the random forest without the need for a separate set.

9). What is K-fold cross-validation?

Answer

K-cross-validation is popular technique used in machine learning to assess the performance and generalization ability of a model. The basic idea K-fold cross-validation is to partition the dataset into K subsets, called folds. The model is trained and evaluated K times, each time using a different fold as the test set and the remaining data as the training set.

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

Answer

Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the optimal set of hyperparameters for a machine learning model. Hyperparameters for a machine learning model. Hyperparameters are external configurations that are not learned from the training data but are set prior to the training process.

Hyperparameter tuning is an essential step in the machine learning model development process. It helps ensure that the model achieves the best performance, generalization, and robustness, leading to more reliable and accurate predictions on new, unseen data.

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

Answer

Using a large learning rate in gradient descent can lead to several issues, affecting the convergence and performance of the optimization process. Here are some common problems associated with a large learning like divergence, overshooting the minimum, instability, difficulty in finding the optimal solution and slow convergence in the long run.

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

Answer

Logistic regression is a linear model designed for binary classification tasks. It models the probability that a given instance belongs to a particular class using a logistic function. Despite being a linear model, logistic regression can be applied to non-linear data under certain conditions. You can apply logistic regression to non-linear data with feature transformation, interaction terms, kernel trick.

13). Differentiate between Adaboost and Gradient Boosting.

Answer

Adaboost and Gradient boosting are both ensemble learning techniques, but they differ in their approaches to combining weak leaners to form a strong predictive model. Weighting of observation, Adaboost assigns weights to data points, with higher weights given to misclassified points at each iteration. While Gradient boosting uses the residuals (the difference between the true values and predictions) of previous learner as the new target. Each weak learn is trained to predict the residuals, gradually reducing the errors.

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

Answer

Bias-variance tradeoff is a fundamental concept in machine learning that describes the balance between two sources of error that affect the performance of a predictive model: bias and variance. Achieving a good tradeoff between bias and variance is crucial for building models that generalize well to new, unseen data.

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

Answer

Support vector machine (SVM) are powerful supervised learning algorithms used for classification and regression tasks. SVMs use a kernel function to map the input data input data into a higher – dimensional space, allowing the algorithm to find a hyperplane that separates that data into different classes. Here are short descriptions of three common types of kernels used in SVM:

Linear kernel – the linear kernel is the simplest kernel and is used when the relationship between the features and the target variable is approximately linear.

Radial basis function (Kernel) the RBF kernel, also known as Gaussian kernel, transforms the input data into an infinite-dimensional space.