Flip Robo Machine LearningAssignment-5

Adeola Olabode

July 2024

1 Introduction

Question 1

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

Answer:

R-squared: measures the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It provides a sense of how well the model fits the data, with a value between 0 and 1.

Residual Sum of Squares (RSS): RSS measures the total deviation of the response values from the fit to the response values. It represents the sum of the squares of the residuals, which are the differences between observed and predicted values.

Which is better? R-squared is generally considered a better measure of goodness of fit because it is normalized between 0 and 1, making it easier to interpret. It provides a clear sense of how well the independent variables explain the variation in the dependent variable. RSS, on the other hand, can vary widely based on the scale of the data, making it less intuitive to understand.

Question 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): TSS is the total variance in the dependent variable. It measures the total deviation of the observed values from the mean of the observed values.

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Explained Sum of Squares (ESS): ESS measures the amount of variance that is explained by the model. It is the sum of the squared deviations of the predicted values from the mean of the observed values.

$$ESS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Residual Sum of Squares (RSS): RSS measures the variance that is not explained by the model. It is the sum of the squared deviations of the observed values from the predicted values.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Equation relating TSS, ESS, and RSS:

$$TSS = ESS + RSS$$

Question 3

. What is the need of regularization in machine learning?

Answer:

Regularization is needed in machine learning to prevent overfitting, which occurs when a model learns not only the underlying pattern but also the noise in the training data. Regularization techniques add a penalty to the loss function for large coefficients, discouraging the model from fitting the noise. This leads to better generalization to unseen data.

Question 4

. What is the Gini–impurity index?

Answer:

The Gini–impurity index is a measure used in decision trees to determine the best feature to split the data. It quantifies the probability of a randomly chosen element being incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. A lower Gini index indicates a purer node.

Question 5

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

Answer:

Yes, unregularized decision trees are prone to overfitting. This is because they can grow very deep, creating a highly complex model that captures noise in the training data. Without regularization methods such as pruning or setting a maximum depth, the tree can fit the training data perfectly but fail to generalize to new, unseen data.

Question 6

. What is an ensemble technique in machine learning?

1.1 Answer:

An ensemble technique in machine learning combines multiple models to produce a single improved predictive model. The idea is that by aggregating the predictions of several models, the ensemble can often perform better than any individual model. Common ensemble techniques include bagging, boosting, and stacking.

Question 7

. What is the difference between Bagging and Boosting techniques?

Answer:

Bagging (Bootstrap Aggregating): It builds multiple models (typically of the same type) using different subsets of the training data created by random sampling with replacement. The final prediction is made by averaging (regression) or voting (classification) the predictions of all models. It aims to reduce variance and prevent overfitting.

Boosting: It builds models sequentially, each trying to correct the errors of the previous one. Each model is trained with a weighted dataset where more weight is given to previously misclassified instances. The final model is a weighted sum of all models. It aims to reduce bias and variance.

Question 8

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

Answer:

Out-of-bag (OOB) error is an estimate of the prediction error of a random forest model. It is computed using the data points that were not included in the bootstrap sample for each tree (approximately one-third of the data). These OOB samples are used to test the corresponding tree, providing an unbiased estimate of the model's error without the need for a separate validation set.

Question 9

. What is K-fold cross-validation?

Answer:

K-fold cross-validation is a technique for evaluating the performance of a machine learning model. The data is divided into K equally sized subsets. The model is trained on K-1 subsets and tested on the remaining subset. This process is repeated K times, with each subset used exactly once as the test set. The final performance metric is the average of the K test results, providing a more robust estimate of the model's performance.

Question 10

. What is hyperparameter tuning in machine learning and why is it done?

Answer:

Hyperparameter tuning involves selecting the optimal set of hyperparameters for a machine learning model. Hyperparameters are settings that govern the training process, such as learning rate, number of trees in a random forest, or regularization strength. Proper tuning is crucial because it can significantly affect the model's performance and its ability to generalize to unseen data.

Question 11

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

1.2 Answer:

A large learning rate in Gradient Descent can cause the algorithm to overshoot the optimal solution, leading to divergence rather than convergence. This means the algorithm may fail to find the minimum of the loss function and can oscillate around or move away from the optimal value, resulting in poor model performance.

Question 12

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

Answer:

Logistic Regression is inherently a linear classifier, meaning it can only separate data that is linearly separable. For non-linear data, Logistic Regression cannot capture the complex relationships between features and the target variable. However, by using techniques such as the kernel trick or polynomial feature transformation, we can extend Logistic Regression to handle non-linear data.

Question 13

. Differentiate between Adaboost and Gradient Boosting.

Answer:

Adaboost (Adaptive Boosting): Adaboost adjusts the weights of incorrectly classified instances so that subsequent classifiers focus more on difficult cases. It combines weak learners into a strong learner by emphasizing the errors of the previous models. Adaboost primarily uses decision stumps as weak learners.

Gradient Boosting: Gradient Boosting builds models sequentially, each new model correcting the errors of the previous ones by fitting to the residuals. It optimizes a loss function using gradient descent. Gradient Boosting can use any differentiable loss function and various base learners, not limited to decision stumps.

Question 14

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

Answer:

The bias-variance trade-off is a fundamental concept that describes the trade-off between the error introduced by bias (assumptions made by the model) and the error introduced by variance (sensitivity to fluctuations in the training data). High bias can cause underfitting, while high variance can cause overfitting. The goal is to find a balance that minimizes the total error.

Question 15

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

Answer:

Linear Kernel: The linear kernel is the simplest kernel function. It is used when the data is linearly separable. The decision boundary is a straight line (or hyperplane in higher dimensions).

$$K(x, x') = x \cdot x'$$

RBF (Radial Basis Function) Kernel: The RBF kernel is a popular choice for non-linear data. It measures the distance between two points in an infinite-dimensional space and can handle complex boundaries.

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$$

Polynomial Kernel: The polynomial kernel represents the similarity of vectors in a feature space over polynomials of the original variables. It can capture interactions of features up to the specified degree.

$$K(x, x') = (x \cdot x' + c)^d$$