Machine Learning Assignment 5 – Answer

Question # 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:

Since R-squared offers a normalized measure of how well the model fits the data and is relative and normalized, making it easier to interpret and compare across different models or datasets, it is generally regarded as a better measure of goodness of fit than the Residual Sum of Squares (RSS).

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:

TSS, ESS and RSS are used to evaluate the fit of a regression model. Following is the equation that relates them to each other.

$$TSS = ESS + RSS$$

Question # 3:

What is the need of regularization in machine learning?

Answer:

In machine learning, regularization plays a crucial role in guaranteeing generalization and enhancing model performance. In machine learning, regularization plays a crucial role in enhancing the model's stability, handling multicollinearity, preventing overfitting, reducing model complexity, and enhancing generalization to fresh data.

Question # 4:

What is Gini–impurity index?

Answer:

The "impurity" or "heterogeneity" of a dataset is measured using the Gini impurity index, a statistic that decision tree algorithms use, especially in classification tasks. It aids in figuring out how well a dataset is divided according to a certain characteristic or attribute.

Question # 5:

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

Answer:

Deep trees with large variation, extensive branching, and improper pruning make unregularized decision trees susceptible to overfitting.

Question # 6:

What is an ensemble technique in machine learning?

Answer:

When using ensemble techniques in machine learning, several models are combined to get better overall performance than when using individual models. The theory behind this is that an ensemble of models, rather than a single model working alone, can frequently generate forecasts that are more reliable and accurate.

Question # 7:

What is the difference between Bagging and Boosting techniques?

Answer:

While they both use ensemble methods to enhance the performance of machine learning models, boosting and bagging (also known as bootstrap aggregating) differ in their methods and features.

Question # 8:

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

Answer:

The out-of-bag (OOB) error in Random Forests is a technique for assessing the model's performance using data that was not part of the bootstrap sample used to train each individual tree. With the use of this method, one can determine the Random Forest model's error rate without needing a different validation set.

Question # 9:

What is K-fold cross-validation?

Answer:

A statistical method called K-fold cross-validation is used to assess how well a machine learning model performs and to make sure it generalizes to new data. It is a reliable technique that lessens problems like overfitting and evaluates a model's performance using independent data.

Question # 10:

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

Answer:

The process of choosing the ideal collection of hyperparameters for a machine learning model to maximize its performance is known as hyperparameter tuning. Hyperparameters are predetermined and govern the training process and model architecture, in contrast to model parameters, which are learnt during training.

Question # 11:

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

Answer:

Gradient descent using a high learning rate can cause several problems that negatively impact the convergence and performance of the model, including divergence, oscillation, and missing the optimal solution.

Question # 12:

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

Answer:

The main application of logistic regression is in binary classification, where there is a linear relationship between the features and the result. When it comes to categorizing non-linear data, it is limited. Since logistic regression is intrinsically restricted to linear decision limits, it is less useful for categorizing non-linear data directly.

Question # 13:

Differentiate between AdaBoost and Gradient Boosting.

Answer:

Both Gradient Boosting and AdaBoost are effective ensemble techniques, each with certain advantages and disadvantages. The particulars of the issue and the intended results will determine which option is best. One distinction is that while Gradient Boosting has high accuracy and is adaptable to different loss functions, AdaBoost is more effective with weak learners and is simpler and faster.

Question # 14:

What is bias-variance trade off in machine learning?

Answer:

A key idea in statistical modelling and machine learning, the bias-variance trade-off characterizes the equilibrium between two types of error that impact a predictive model's performance.

Question # 15:

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

Answer:

Kernels are used by Support Vector Machines (SVMs) to convert data into a higher-dimensional space where a separating hyperplane may be simpler to locate.

- RBF Kernel: Adaptable and capable of handling complex decision boundaries, ideal for non-linearly separable data.
- Polynomial Kernel: Effective for applications where feature interactions are crucial, this kernel captures polynomial correlations between features.

The SVM can adapt to different kinds of patterns in the data since each kernel changes the input in a unique way.

_____-