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:

   R-squared is generally a better measure of the goodness of fit for a regression model than the residual sum of squares (RSS).

   The reason why R-squared is often preferred over RSS as a measure of goodness of fit is due to its standardized nature:

1. **Scalability**: R-squared is scale-invariant, meaning it does not change if the scale of the data changes, whereas RSS is affected by the scale of the dependent variable. This makes R^2 a better choice when comparing models fitted on different scales.
2. **Interpretability**: R-squared has an intuitive interpretation as the proportion of variance explained, which is easier to understand than the sum of squared residuals. An R^2 of 0.75 means that 75% of the variance in the dependent variable is explained by the model, which is a straightforward interpretation.
3. **Benchmarking**: R-squared provides a clear benchmark. An R^2 of 0 indicates that the model explains none of the variability in the response data around its mean, while an R^2 of 1 indicates that the model explains all the variability.
4. **Adjustment for model complexity**: Adjusted R-squared takes into account the number of predictors in the model, which helps in assessing whether the addition of a new predictor really improves the model or is just adding complexity without significantly improving the fit.

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: total sum of squares (TSS) is the sum of squared differences between the observed dependent variables and the overall mean. TSS measures the total variability of a dataset, commonly used in regression analysis and ANOVA.

ESS: **explained sum of squares** (ESS) is the sum of the differences between the *predicted value* and the **mean** of the *dependent variable*. In other words, it describes how well our line fits the data.

RSS: residual sum of squares (RSS, where residual means remaining or unexplained) is the difference between the *observed* and *predicted* values.

Related 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 models.

4. What is Gini–impurity index?

   Answer:

   The Gini impurity index, also known as the Gini index, is a measure of how mixed up or impure a dataset is. It's commonly used in decision tree algorithms, particularly for classification tasks.

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

   Answer:

   Yes, unregularized decision-trees prone to overfitting.

   Unregularized decision-trees techniques unable to add constraints to the learning algorithm and its capacity to overfit.

6. What is an ensemble technique in machine learning?

   Answer:

   Ensemble techniques in machine learning combine multiple models to make predictions or classifications. The goal is to improve the overall performance and accuracy of the predictions by combining the strengths of the individual models.

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

   Answer:

   Bagging and boosting are both machine learning ensemble techniques that improve the accuracy and stability of algorithms. The main difference between the two is how they train and combine base learners.

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

   Answer:

   The out-of-bag (OOB) error is a way to measure the prediction error of a random forest model. It's a useful tool for machine learning professionals and data scientists because it provides an accurate estimate of model performance without the need for cross-validation.

9. What is K-fold cross-validation?

   Answer:

   In K-Fold Cross Validation, we split the dataset into k number of subsets (known as folds) then we perform training on the all the subsets but leave one(k-1) subset for the evaluation of the trained model. In this method, we iterate k times with a different subset reserved for testing purpose each time.

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

    Answer:

    Hyperparameter tuning is the process of finding the best set of hyperparameters for a machine learning model to improve its performance. Hyperparameters are adjustable model arguments that are set before training begins and control the training process.

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 cause a number of issues, including:

- **Overshooting the optimal point**
  A large learning rate can cause the algorithm to take big steps in the direction of the negative gradient, which can lead to overshooting the minimum point. This can result in instability or divergence, and the model may perform poorly.

- **Exploding or oscillating performance**
  A large learning rate can cause the algorithm to perform poorly over the training epochs, and the final performance may be lower.

- **Overfitting**
  A large learning rate can cause the algorithm to overfit the training data, which can lead to poor generalization performance on new data.

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

    Answer:

    Logistic regression can be used to classify non-linear data, but it's not ideal for this purpose:

    Linear assumption
    Logistic regression assumes a linear relationship between the input features and the output, which means it can't capture non-linearity.

    Difficult to find non-linear input variables
    To solve non-linear classification problems with logistic regression, you need to find new input variables that are non-linear functions of the original input variables. This can be difficult, especially in three dimensions or more.

    Nonlinear functional forms
    If you suspect the decision boundary is non-linear, you can try using non-linear functional forms for the logit function. This can be more challenging, but you can use optimization modules to help.

13. Differentiate between Adaboost and Gradient Boosting.

    Answer:

    AdaBoost and Gradient Boosting are both machine learning algorithms that combine multiple weak learners to create a stronger model. However, they differ in several ways, including:

- **How they focus on errors**
  AdaBoost focuses on misclassified instances by adjusting their weights, while Gradient Boosting minimizes a loss function.
- **How they build models**
  AdaBoost builds a sequence of models by adding stumps, while Gradient Boosting builds a sequence of models by adding trees.
- **How they handle outliers**
  Gradient Boosting is more robust to outliers and noise than AdaBoost.
- **How they handle missing values**
  AdaBoost and Gradient Boosting require explicit imputation of missing values, while XGBoost has built-in functionality for this.
- **How they handle multi-class classification**
  AdaBoost and Gradient Boosting require a One-vs-All approach to solve multi-class problems, while XGBoost can handle them natively

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

Answer:

The bias-variance trade-off is a fundamental concept in machine learning that refers to the balance between a model's ability to represent data patterns and its susceptibility to fluctuations in training data

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

Answer:

The most common SVM kernels are linear, good for straight-line data, polynomial, and useful for curves. Radial basis function (RBF), is great for complex patterns. Also, sigmoid can handle different kinds of data changes.