ASSIGNMENT - 5

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?

In regression analysis, both R-squared and Residual Sum of Square(RSS) play essential roles in assessing the goodness of fit of a model .Let's delve into each measure and understand their significance

1)R-squared:

➔R-squared represents the proportion of variability in the dependent variable (response) that is explained by the independent variables (predictors) in the regression model.
➔Goodness of fit: The closer the R-squared value is to 1,the better the model fits the data.
➔ R-squared has limitations: It always increases or remains the same when new variables are added to the model,regardless of their significance.It does not account for the number of predictors in the model.

2)Residual Sum of (RSS):

➔RSS measures the remaining error between the regression function and the actual data points after the model has been fitted.
➔Researches often use RSS to compare different models and select the one with lowest RSS.
➔It helps identify how will the model captures the noise in the data.

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.

1)Total Sum of Squares(TSS)

➔TSS represents the total variability in the dependent variable (response) around its overall mean .
➔TSS is the sum of squared differences between the observed dependent variables and the overall mean.

2)Explained Sum of Sqaures(ESS)

➔ ESS quantifies the variability explained by the regression model (how well the model fits the data).
➔ ESS is the sum of squared differences between the predicted values and the mean of the dependent variable

3) Residual Sum of Squares(RSS)

➔ RSS measures the unexplained variability in the dependent variable (the difference between observed and predicted values).
➔ RSS is the sum of squared differences between the actual values and the predicted values.

4)Relationship between TSS,ESS,RSS

➔TSS=ESS+RSS

Total Variability = Explained Variability + Unexplained Variability

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

Regularization is a critical aspect of machine learning models, ensuring they don't succumb to overfitting or underfitting. Essentially, it introduces a penalty term to the loss function, preventing the model from becoming too complex.

Here's why regularization is essential:

1) Overfitting Prevention.
2) Generalization to Unseen Data.
3) Bias-Variance Tradeoff.
4) Common Regularization Techniques;
   ➔Lasso Regularization.
   ➔Ridge Regularization.
   ➔Elastic Net Regularization.

4. What is Gini–impurity index?

➔The Gini Impurity is a crucial concept used in building Decision Trees

➔ The Gini Impurity measures how well a dataset's features should split nodes to form a decision tree.

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

➔ Yes, unregularized decision trees are indeed prone to overfitting.

High Variance: Unregularized decision trees have high variance because they can adapt too well to the training data.
Memorization: Without constraints, decision trees can memorize the training data, including its noise and outliers.

Deep Trees: If allowed to grow without restrictions, decision trees can become deep and have many branches, resulting in a model that fits the training data perfectly but performs poorly on new data.

6.What is an ensemble technique in machine learning?

➔Ensemble learning is a powerful technique in machine learning that combines the predictions of multiple models to improve overall performance. It leverages the collective intelligence of diverse models to enhance accuracy, robustness, and generalization.

➔ They aggregate predictions from different models to make a collective decision.

Examples of Ensemble Techniques:

1)Bagging
2)Boosting

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

Bagging

➔Training data subsets are drawn randomly with replacement from the entire training dataset.
➔Bagging attempts to tackle the over-fitting issue.
➔Every model receives an equal weight.
➔Objective to decrease variance,not bias.
➔Every model is built independently.

Boosting

➔Each new subset contains the components that were misclassified by previous models.
➔Boosting tries to reduce bias.
➔Models are weighted by their performance.
➔Objective to decrease bias,not variance.
➔New models are affected by the performance of the previously developed model.

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

The out-of-bag error is a performance metric specifically used for random forest models.

➔ The OOB error is an estimate of the performance of a random forest on unseen data.

➔It is calculated using the samples that were not included in the training of each individual tree.These samples are called out-of-bag samples.

➔It efficiently estimates model performance during training without additional data splitting.

9. What is K-fold cross-validation?

K-fold cross-validation is a statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem. K-fold cross-validation estimates how well a model will perform on new data. It efficiently uses the available data for both training and validation.

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

➔Hyperparameters are configuration variables that control the learning process of a model. It express important properties of the model, such as its complexity, learning rate, regularization strength, and architecture.

➔ The choice of hyperparameters significantly impacts a model's accuracy, generalization, and other metrics.

➔ Proper tuning helps prevent overfitting.

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

1)Overshooting the Minimum
2)Divergence
3)Slow Convergence
4) Skipping Local Minima
5) Stability Issues

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

No, We can't use Logistic Regression for classification of Non-Linear Data because it assumes a linear relationship between the input features and the output. This means that it cannot capture the complexity and non-linearity of the data.

13. Differentiate between Adaboost and Gradient Boosting.

Adaboost Boosting:
➔ AdaBoost aims to improve the performance of weak learners by combining them into a strong ensemble model.
➔ It focuses on correcting misclassifications made by previous models.
➔AdaBoost uses an exponential loss function.
➔Handles both numerical and categorical features.
➔Robust against overfitting.

Gradient Boosting
➔Gradient Boosting constructs a strong model by iteratively improving weak learners.
➔It minimizes the residual errors made by the previous models.
➔ Gradient Boosting uses a differentiable loss function.
➔Handles non-linear relationships well.
➔Robust and less sensitive to outliers.

14. What is bias-variance trade off in machine learning?
The bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

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

Linear Kernel:

➔It computes the dot product between feature vectors.
➔Suitable for linearly separable data.

RBF (Radial Basis Function) Kernel:

➔It maps data into a higher-dimensional space using the radial basis function.
➔Suitable for non-linearly separable data.

Polynomial Kernel:

➔The polynomial kernel represents the similarity of vectors using polynomial functions of the original variables.
➔It can handle non-linear relationships between features and target.