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?
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.
3. What is the need of regularization in machine learning?
4. What is Gini_impurity index?
5. Are unregularized decision-trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?
7. What is the difference between Bagging and Boosting techniques?
8. What is out-of-bag error in random forests?
9. What is K-fold cross-validation?
10. What is hyper parameter tuning in machine learning and why it is done?
11. What issues can occur if we have a large learning rate in Gradient Descent?
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
13. Differentiate between Adaboost and Gradient Boosting.
14. What is bias-variance trade off in machine learning?
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

MACHINE LEARNING(ASSIGNMENT-5)

Q1. 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?

Ans. Both metrics used to refine our assessment of regression models.

R-squared, also known as the coefficient of determination, measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. It gives result in terms of ranges from 0 to 1, with a higher value indicating a better fit. **R-squared is useful for comparing different models or for determining the proportion of the variability in the dependent.** But it has a limitation of being biased.

Whereas, **Residual Sum of Squares (RSS)** measures the total amount of unexplained variance in the dependent variable that remains after the model has been fit. It is the sum of the squared differences between the actual and predicted values of the dependent variable. A lower value of RSS indicates a better fit. **RSS is useful for evaluating the accuracy of the predictions of the model.**

In general, both measures are important and should be considered together when evaluating the goodness of fit of a model. However, R-squared is often considered to be a better measure of goodness of fit than RSS because it provides a single number that summarizes the proportion of variance in the dependent variable that is explained by the model, which is more interpretable and easier to compare across models.

R-squared provides a high-level summary of the explanatory power of your model, while RSS gives you a detailed view of the model's prediction accuracy.

Q2. 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.

Ans. **TSS**: It finds the squared difference between each variable and the mean. It measures how much variation there is in the observed data.

ESS: The ESS is the portion of the total variation that measures how well the regression equation explains the relationship between X and Y. It is the sum of squared deviations of the predicted values from the mean of the data.

RSS: Residual Sum of Squares (RSS) is a statistical technique that measures the amount of variance in a data set that is not explained by a regression model. It is also known as the Sum of Squared Errors (SSE). It is the sum of squared deviations of the observed values from the predicted values.

In regression, the total sum of squares (TSS) can be broken down into two components: the explained sum of squares (ESS) and the residual sum of squares (RSS).

TSS = ESS + RSS

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

Ans: Regularization is a type of techniques used in machine learning to prevent overfitting. Overfitting occurs when a model is too complex and noisy.

Q4. What is Gini–impurity index?

Ans: Gini-impurity is a measure used in decision tree algorithms to quantify a dataset's impurity level or disorder that's how is impure our dataset is. It's used in decision tree algorithms to determine the best way to split data for classification.

It is calculated by subtracting the sum of the squared probabilities of each class from one. It ranges from zero to one, with higher values indicating purer nodes and lower values indicating less pure nodes.

Q5. Are unregularized decision-trees prone to overfitting? If yes, why?

Ans: Yes, unregularized decision trees are prone to overfitting. Overfitting occurs when a model learns to memorize the training data rather than generalize patterns that can be applied to unseen data. As the tree becomes deeper, it becomes more susceptible to overfitting.

Q6. **What is an ensemble technique in machine learning?**

Ans: An ensemble technique in machine learning is a method that combines the predictions of multiple individual models to produce a more accurate and robust prediction than any single model does alone.

There are many ensemble technique such as,
-Bagging
-Boosting
-Random Forest

Q7. What is the difference between Bagging and Boosting techniques?

Ans:

i) **Bagging**- it method combines predictions that belongs to the same type, where as **Boosting** combines the predictions that belongs to different types.
ii) **Bagging-** decrease the variance, where as Boosting decrease the bias.
iii) **Bagging-** base classifier trained parallelly, in Boosting base classifier trained subsequently.
iv) **Bagging-** the models created independently, whereas Boosting models creation is depends on previous one.

Q8. What is out-of-bag error in random forests?

Ans: In Random Forest, the out-of-bag (OOB) error is an estimate of the model's prediction error on unseen data. It's calculated by evaluating the predictions of each individual decision tree in the forest on the instances that were not used to train that particular tree.

Q9. What is K-fold cross-validation?

Ans: K-fold cross-validation is a technique used to assess the performance of a machine learning model. It involves dividing the dataset into k equal-sized subsets or folds, using k-1 folds for training the

model and the remaining fold for testing. This process is repeated k times, each time using a different fold as the test set and the remaining folds as the training set.

Q10. What is hyper parameter tuning in machine learning and why it is done?

Ans: Hyperparameter tuning in machine learning refers to the process of selecting the optimal values for the hyperparameters of a machine learning model. They control the behaviour of the learning algorithm and can have a significant impact on the performance of the model.
Hyperparameter is done to improve performance, avoiding over/under fitting and generalisation.
Hyperparameter tuning is typically performed using techniques like grid search, random search, etc.

Q11. What issues can occur if we have a large learning rate in Gradient Descent?

Ans: Having a large learning rate in Gradient Descent can lead to several issues, including:

i) **Overshooting the Minimum**
ii) **Divergence**
iii) **Instability**
iv) **Difficulty in Convergence**
v) **Poor Generalization**
vi) **Failure to train**

Q12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
Ans: No, logistic regression is not used to classify non-linear data because it has a linear decision surface.

Q13. Differentiate between Adaboost and Gradient Boosting.
Ans: AdaBoost and Gradient Boosting are both machine learning algorithms. AdaBoost is a supervised learning algorithm that combines weak learners into a strong learner. Gradient Boosting is a machine learning technique that uses pseudo-residuals as targets instead of traditional residuals.

AdaBoost can be used in classification and regression tasks. Gradient Boosting can be used for classification, regression, ranking, and survival analysis.

Gradient Boosting is highly versatile and has unmatched accuracy and performance for tabular supervised learning tasks.

Q14. What is bias-variance trade off in machine learning?

The bias-variance trade-off is a fundamental concept in machine learning that describes the balance between the bias of a model and its variance. It relates to the generalization ability of a model to unseen data.
Bias: Bias refers to the error introduced by approximating a real-world problem with a simplified model. A high bias means the model makes strong assumptions about the form of the underlying data distribution, potentially leading to underfitting. Models with high bias may overlook relevant patterns in the training data.
Variance: Variance refers to the model's sensitivity to small fluctuations or noise in the training data. A high variance means the model is overly sensitive to the training data and captures noise along with

the underlying patterns, potentially leading to overfitting. Models with high variance may perform well on the training data but generalize poorly to unseen data.

Q15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Ans:

Linear kernel-The most basic and common kernel function. Linear kernels are best used when data is linearly separable, meaning it can be separated by a single line. Linear and polynomial kernels are usually less time consuming and less accurate than RBF or Gaussian kernels.

Polynomial kernel-A more generalized representation of the linear kernel. Polynomial kernels are best used when data has nonlinear patterns or interactions between features.

RBF kernel-Also known as a Gaussian radial basis function, this is one of the most popular and widely used kernel functions in SVMs. RBF kernels are best used when data has complex and nonlinear patterns or clusters