MACHINE LEARNING

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?

Ans-> R-squared is a batter measure of goodness of fit model in regression because R-squared is a measure of how well the model explains the variance in the data, while RSS measures the total amount of unexplained variance. Both measures are useful, but R-squared is more commonly used because it is easier to interpret.

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.

Ans-> 1. Total Sum of Squares (TSS):-

- TSS represents the total variability in the dependent variable (Y).
- It is calculated by summing the squared differences between each observed value of Y and the mean of Y.
- TSS measures the total amount of variation in the data, regardless of whether it is explained by the independent variable or not.

2. Explained Sum of Squares (ESS):-

- ESS represents the portion of the variability in Y that is explained by the independent variable (X).
- It is calculated by summing the squared differences between each predicted value of Y (based on the regression line) and the mean of Y.
- ESS measures the amount of variation in Y that is accounted for by the regression model.

3. Residual Sum of Squares (RSS):-

- RSS represents the portion of the variability in Y that is not explained by the independent variable (X).
- It is calculated by summing the squared differences between each observed value of Y and its corresponding predicted value from the regression line.
- RSS measures the amount of variation in Y that is not captured by the regression model.

The relationship between TSS, ESS, and RSS can be expressed by the following equation:

TSS = ESS + RSS

- ➢ This equation states that the total variability in the dependent variable (TSS) can be decomposed into two components: the variability explained by the independent variable (ESS) and the variability not explained by the independent variable (RSS).
- ➢ TSS represents the total variation in the data, ESS represents the variation explained by the regression model, and RSS represents the variation not

explained by the regression model. These metrics provide insights into the goodness of fit and predictive power of the regression model.

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

Ans-> Regularization is a technique used in machine learning to reduce overfitting, which occurs when a model is too closely fit to the training data and does not generalize well to new data. Regularization adds a penalty term to the loss function that is proportional to the size of the model's weights, which encourages the model to find simpler solutions that are less likely to overfit.

There are several reasons why regularization is need in machine learning:

A. **Prevents overfitting:-** Regularization helps to prevent overfitting by penalizing models that have large weights. This encourages the model to find simpler solutions that are less likely to be specific to the training data and more likely to generalize well to new data.
B. **Improves generalization:-** Regularization improves the generalization performance of machine learning models by reducing overfitting. Models that are regularized are less likely to make predictions that are too specific to the training data and more likely to make accurate predictions on new data.
C. **Reduces variance:-** Regularization can help to reduce the variance of machine learning models. Variance is the amount that the predictions of a model can change when the training data is changed. Regularization helps to reduce variance by penalizing models that have large weights, which encourages the model to find simpler solutions that are less sensitive to changes in the training data.
D. **Improves robustness:-** Regularization can help to improve the robustness of machine learning models. Robustness is the ability of a model to make accurate predictions even when the data is noisy or corrupted. Regularization helps to improve robustness by penalizing models that have large weights, which encourages the model to find simpler solutions that are less sensitive to noise and corruption in the data.

4. What is Gini–impurity index?

Ans-> The Gini impurity index is a measure of how impure a set of data is. It is used in decision trees to determine the best way to split the data into subsets. The Gini impurity index is calculated by summing the probability of each class multiplied by the probability of not being in that class. The higher the Gini impurity index, the more impure the data is.

To calculate the Gini impurity index, we first need to calculate the probability of each class. This can be done by dividing the number of data points in each class by the total number of data points. Next, we need to calculate the probability of not being in each class. This can be done by subtracting the probability of each class from 1. Finally, we need to multiply the probability of each class by the probability of not being in that class. The sum of these products is the Gini impurity index.

The Gini impurity index can be used to determine the best way to split the data into subsets. The best split is the one that results in the lowest Gini impurity index. This can be done by finding the feature that has the highest correlation with the class label. The data can then be split into two subsets based on the value of this feature.

The Gini impurity index is a useful tool for decision tree learning. It can help to improve the accuracy of decision trees by finding the best way to split the data into subsets.

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

Ans-> yes, Decision trees are indeed prone to overfitting, especially when they are not regularized. Overfitting occurs when a model learns the specific details of the training data too well and loses its ability to generalize to new, unseen data.

Decision trees make decisions based on a series of binary splits in the feature space. At each split, the tree chooses the feature and threshold that best separates the data into two groups. This process can lead to complex trees with many branches and deep levels, which can capture intricate patterns in the training data. However, these complex trees may not generalize well to new data because they are too specific to the training set.

To mitigate overfitting in decision trees, regularization techniques are employed. Regularization adds a penalty term to the loss function that encourages the tree to be simpler. This can be achieved by limiting the depth of the tree, setting a minimum number of samples required at each leaf node, or using pruning techniques to remove unnecessary branches.

Regularization helps to prevent the decision tree from learning the idiosyncrasies of the training data and promotes generalization to new data. By balancing the complexity of the tree with the regularization penalty, it is possible to find a model that performs well on both the training data and unseen data.

6. What is an ensemble technique in machine learning?

Ans-> In machine learning, an ensemble technique is a type of model that uses a group of individual models, or base learners, to make predictions. The idea behind ensemble techniques is to combine the strengths of multiple models to achieve better predictive performance than any single model could achieve on its own.

There are different types of ensemble techniques, but some of the most commonly used ones include:

A. **Bagging (Bootstrap Aggregating):-** Bagging involves training multiple models on different subsets of the training data. Each model makes predictions independently, and the final prediction is typically obtained by averaging or voting across the predictions of the individual models. Bagging can help reduce variance in the predictions and improve the overall accuracy of the model.
B. **Boosting (Adaptive Boosting):-** Boosting involves training multiple models sequentially, with each subsequent model focused on correcting the errors of the previous models. The final prediction is determined by aggregating the weighted predictions of the individual models, where the weights are assigned based on the accuracy of each model. Boosting can help improve the accuracy of the model, particularly for complex problems.

C. **Stacking (Stacked Generalization):-** Stacking involves training multiple models and combining their predictions using another model, known as a meta-model or blender. The individual models make predictions on the training data, and these predictions, along with the original features, are used to train the meta-model. The meta-model then makes the final prediction. Stacking can help improve the accuracy and robustness of the model by leveraging the strengths of different models.

Ensemble techniques are widely used in machine learning for various tasks, including classification, regression, and anomaly detection. They can help improve the performance of machine learning models by reducing variance, improving accuracy, and handling complex problems.

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

Ans-> Bagging (short for bootstrap aggregating) and boosting are two popular ensemble techniques that leverage multiple models to enhance predictive performance. While both techniques utilize multiple models, they differ in their training procedures and the way they combine the predictions of individual models.

Bagging:-

A. **Training:-** In bagging, multiple models are trained independently on different subsets of the training data. Each model makes predictions on the entire dataset, and the final prediction is typically obtained by averaging the predictions of all the individual models.
B. **Key Idea:**- Bagging reduces variance in the ensemble model's predictions by training models on different data subsets. By leveraging the diversity of these models, bagging helps mitigate overfitting and improves the generalization performance of the ensemble.

Boosting:-

A. **Training:**- In boosting, models are trained sequentially, with each subsequent model focused on correcting the errors of the previous ones. Models are weighted based on their performance, and the final prediction is determined by aggregating the weighted predictions of the individual models.
B. **Key Idea:-** Boosting aims to reduce bias in the ensemble model's predictions by iteratively training models and emphasizing instances that the previous models misclassified. This process helps the ensemble model learn from its mistakes and progressively improve its accuracy.

bagging and boosting are powerful ensemble techniques that offer different approaches to improving predictive performance. Bagging focuses on reducing variance by training models on diverse data subsets, while boosting emphasizes reducing bias by iteratively correcting errors. The choice between bagging and boosting depends on the specific problem and dataset at hand.

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

Ans-> In the context of random forests, out-of-bag (OOB) error refers to the error rate estimated using data that was not used in training a particular decision tree within the forest. Here's how OOB error is calculated:

A. **Bootstrap Sampling:-** When building a random forest, each decision tree is trained on a bootstrapped sample of the training data. This means that a portion of the data is randomly selected with replacement, allowing some data points to appear multiple times in the sample while others are excluded.
B. **Out-of-Bag Data:-** The data points that are not included in the bootstrap sample for a particular tree are called out-of-bag (OOB) data. These OOB data points serve as a holdout set for estimating the error rate of that tree.
C. **OOB Error Calculation:-** For each decision tree in the random forest, the OOB data is passed through the tree to obtain predictions. These predictions are then compared to the true labels of the OOB data points to calculate the error rate.
D. **Averaging:-** The OOB error rates for all the trees in the random forest are averaged to obtain an overall estimate of the generalization error of the forest.

OOB error provides an unbiased estimate of the error rate without the need for a separate validation set. It is particularly useful when the training data is limited or when there is a need to assess the performance of individual trees within the forest.

It's important to note that OOB error is only applicable to random forests and not other ensemble methods like bagging or boosting.

9. What is K-fold cross-validation?

Ans-> In K-fold cross-validation, the training data is divided into K equally sized folds. The model is then trained and evaluated K times, each time using a different fold as the test set and the remaining K-1 folds as the training set.

K-fold cross-validation provides a more robust estimate of a model's performance compared to traditional train-test splits. It reduces the impact of a particular split on the evaluation results and ensures that the model is evaluated on different subsets of the data.

The choice of K depends on the size of the dataset and the computational resources available. Common values for K include 5, 10, or even higher for large datasets.

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

Ans-> In machine learning, hyperparameter tuning refers to the process of finding the optimal set of hyperparameters for a given model. Hyperparameters are the parameters of the model that are not learned during the training process, such as the learning rate, the number of hidden units in a neural network, or the regularization coefficient.

Hyperparameter tuning is important because it can significantly affect the performance of a machine learning model. By finding the optimal set of hyperparameters, we can improve the accuracy, generalization, and robustness of the model.

There are various methods for hyperparameter tuning, including:

A. **Manual tuning:-** This involves manually setting the hyperparameters and evaluating the performance of the model on a validation set.
B. **Grid search:-** This involves systematically searching through a grid of hyperparameter values and selecting the combination that produces the best performance.
C. **Random search:-** This involves randomly sampling hyperparameter values and selecting the combination that produces the best performance.
D. **Bayesian optimization:-** This involves using a Bayesian optimization algorithm to efficiently search for the optimal set of hyperparameters.

The choice of hyperparameter tuning method depends on the size of the search space, the computational resources available, and the desired level of accuracy.

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

Ans-> In Gradient Descent, the learning rate determines the step size taken in the direction of the negative gradient during each iteration of the optimization process. A large learning rate means that the steps taken are larger, while a small learning rate means that the steps are smaller.

Now, let's consider the issues that can arise when using a large learning rate:

A. **Divergence:-** A large learning rate can cause the optimization process to diverge, meaning that the model's parameters move away from the optimal solution instead of converging towards it. This can happen when the learning rate is so large that the steps taken overshoot the optimal point and cause the model to oscillate around it.
B. **Slow Convergence:-** A large learning rate can also lead to slow convergence, meaning that the model takes a long time to reach the optimal solution. This is because a large learning rate can cause the model to skip over important regions of the parameter space and miss the optimal point.
C. **Instability:-** A large learning rate can make the optimization process unstable, meaning that the model's parameters fluctuate wildly during training. This can make it difficult to find a stable solution and can lead to poor performance.
D. **Overfitting**:- A large learning rate can also contribute to overfitting, which occurs when the model learns the specific details of the training data too well and starts to make predictions that are too specific to the training set. This can lead to poor generalization performance on new data.

Therefore, it is important to carefully choose the learning rate when using Gradient Descent to ensure that the optimization process is stable, converges quickly, and finds the optimal solution without overfitting.

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

Ans-> Logistic Regression is a statistical model used for binary classification problems. It is a linear model, which means that it assumes a linear relationship between the independent variables and the dependent variable. In other words, it assumes that the probability of an observation belonging to one class or the other can be represented by a straight line.

However, many real-world datasets are non-linear, meaning that the relationship between the independent and dependent variables is not linear. In such cases, Logistic Regression cannot accurately model the data and will produce poor classification results.

To handle non-linear data, we can use non-linear models such as Decision Trees, Random Forests, or Support Vector Machines. These models are able to capture the complex relationships between the variables and make accurate predictions even for non-linear data.

13. Differentiate between Adaboost and Gradient Boosting.

Ans-> To differentiate between Adaboost and Gradient Boosting-

1. **Adaboost (Adaptive Boosting):-**

   a) Adaboost is an ensemble learning algorithm used for classification and regression tasks.
   b) It works by building a series of weak learners (base classifiers) and combining them to create a strong learner.
   c) In Adaboost, each weak learner is trained on a reweighted version of the training data, where the weights are adjusted based on the performance of the previous weak learner.
   d) The final prediction is made by taking a weighted vote of the predictions from all the weak learners.

2. **Gradient Boosting:-**

   a) Gradient Boosting is also an ensemble learning algorithm used for classification and regression tasks.
   b) It works by building a series of decision trees, where each tree is trained on the residuals (errors) of the previous tree.
   c) The final prediction is made by summing up the predictions from all the individual trees.

Key Differences:-

- **Training Strategy:-** Adaboost focuses on reweighting the training data, while Gradient Boosting focuses on reducing the residuals of the previous tree.
- **Model Structure:-** Adaboost builds a series of weak learners, while Gradient Boosting builds a series of decision trees.
- **Prediction Method:-** Adaboost uses a weighted vote of the weak learners, while Gradient Boosting sums up the predictions from the individual trees.

In general, Gradient Boosting tends to perform better than Adaboost, especially for complex datasets. However, the choice of algorithm depends on the specific problem and dataset at hand.

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

Ans-> In machine learning, the bias-variance tradeoff is a fundamental concept that relates to the generalization performance of models. It arises from the tension between two opposing factors: bias and variance.

- **Bias:-** refers to the systematic error introduced by a model due to inherent assumptions or simplifications. It measures the difference between the expected prediction of the model and the true underlying data-generating process. High bias can lead to underfitting, where the model fails to capture the complexity of the data and makes consistently incorrect predictions.
- **Variance:-** on the other hand, refers to the variability in the predictions of a model due to its sensitivity to training data. It measures the extent to which the model's predictions change with different training sets. High variance can result in overfitting, where the model captures idiosyncrasies of the training data and makes predictions that are too specific to it.

The bias-variance tradeoff arises because reducing bias typically increases variance, and vice versa. As a model becomes more complex, with more parameters or features, it can better fit the training data, reducing bias. However, this increased complexity also makes the model more sensitive to the particular training set, leading to higher variance.

Finding the optimal balance between bias and variance is crucial for achieving good generalization performance. A model with high bias may have low variance but will make consistently incorrect predictions, while a model with high variance may have low bias but will make highly variable predictions. The goal is to find a model that has both low bias and low variance, which can make accurate predictions on unseen data.

To manage the bias-variance tradeoff, several techniques can be employed, such as regularization, early stopping, and model selection. Regularization adds a penalty term to the model's loss function that discourages overly complex models, reducing variance. Early stopping involves terminating the training process before the model fully converges, which helps prevent overfitting. Model selection involves choosing the best model from a set of candidates based on their performance on a validation set.

Understanding and managing the bias-variance tradeoff is essential for developing effective machine learning models that generalize well to new data.

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

Ans-> → **Linear Kernel:-** The linear kernel is the simplest kernel function used in SVM. It computes the dot product between two data points in the input space. Mathematically, it is represented as:

$$K(x_i, x_j) = x_i^T x_j$$

where \(x_i\) and \(x_j\) are two data points.

The linear kernel is efficient to compute and works well when the data is linearly separable. However, it may not be suitable for data that is not linearly separable.

→ **RBF Kernel:-** The RBF (Radial Basis Function) kernel is a non-linear kernel function that measures the similarity between two data points based on their distance in the input space. It is defined as:

$$K(x_i, x_j) = exp(-γ ||x_i - x_j||^2)$$

where \(\gamma\) is a hyperparameter that controls the width of the kernel.

The RBF kernel is more computationally expensive than the linear kernel, but it can handle non-linearly separable data. It is a widely used kernel function in SVM and other kernel-based methods.

→ **Polynomial Kernel:-** The polynomial kernel is another non-linear kernel function that computes the dot product between two data points after raising them to a certain power. It is defined as:

$$K(x_i, x_j) = (x_i^T x_j + 1)^d$$

where \(d\) is the degree of the polynomial.

The polynomial kernel can handle non-linearly separable data, but it is more computationally expensive than the linear and RBF kernels. It is also more prone to overfitting, so the degree of the polynomial must be chosen carefully.

STATISTICS

Q1 to Q10 are MCQs with only one correct answer. Choose the correct option.

1. Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies.

a) Mean b) Actual c) Predicted d) Expected

Ans:- a

2. Chisquare is used to analyse

a) Score b) Rank c) Frequencies d) All of these

Ans:- c

3. What is the mean of a Chi Square distribution with 6 degrees of freedom?

a) 4 b) 12 c) 6 d) 8

Ans:- c

4. Which of these distributions is used for a goodness of fit testing?

a) Normal distribution b) Chisqared distribution

c) Gamma distribution d) Poission distribution

Ans:- b

5. Which of the following distributions is Continuous

a) Binomial Distribution b) Hypergeometric Distribution

c) F Distribution d) Poisson Distribution

Ans:- c

6. A statement made about a population for testing purpose is called?

a) Statistic b) Hypothesis c) Level of Significance d) TestStatistic

Ans:- b

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

a) Null Hypothesis b) Statistical Hypothesis

c) Simple Hypothesis d) Composite Hypothesis

Ans:- a

8. If the Critical region is evenly distributed then the test is referred as?

a) Two tailed b) One tailed c) Three tailed d) Zero tailed

Ans:- a

9. Alternative Hypothesis is also called as?

a) Composite hypothesis b) Research Hypothesis

c) Simple Hypothesis d) Null Hypothesis

Ans:- b

10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by_______

a) np b) n

Ans:- a