ASSIGNMENT-2

1. In logistic regression, what is the logistic function (sigmoid function) and how is it used to compute probabilities?

Ans**: Logistic Function (Sigmoid Function) in Logistic Regression:**

**The logistic function, also known as the sigmoid function, is a mathematical function that maps any real-valued number to a value between 0 and 1. It has an S-shaped curve.**

**In logistic regression, the logistic function is used to model the probability that a given input belongs to a certain class. The output of the logistic function represents the probability of the input belonging to the positive class (class 1).**

1. When constructing a decision tree, what criterion is commonly used to split nodes, and how is it calculated?

Ans: **Criterion for Splitting Nodes in Decision Trees:**

**The commonly used criterion for splitting nodes in decision trees includes measures like Gini impurity and information gain (entropy).**

**Gini impurity measures the probability of incorrectly classifying a randomly chosen element if it was randomly labelled according to the distribution of labels in the node.**

**Information gain, based on entropy, measures the reduction in entropy (or uncertainty) of the target variable after the split.**

1. Explain the concept of entropy and information gain in the context of decision tree construction.

Ans**: Entropy and Information Gain:**

**Entropy is a measure of randomness or uncertainty in a dataset. In decision tree construction, entropy is used to calculate the homogeneity of a sample.**

**Information gain measures the reduction in entropy or uncertainty after a dataset is split based on a feature. It helps in deciding the best feature to split the data on at each node of the tree.**

1. How does the random forest algorithm utilize bagging and feature randomization to improve classification accuracy?

Ans: **Random Forest Algorithm and Bagging:**

**Random Forest utilizes bagging (bootstrap aggregating) by training multiple decision trees on random subsets of the training data and then combining their predictions to reduce overfitting and improve generalization.**

**Feature randomization is used by randomly selecting a subset of features at each split in each decision tree to increase diversity among the trees and prevent overfitting.**

1. What distance metric is typically used in k-nearest neighbours (KNN) classification, and how does it impact the algorithm's performance?

Ans: **Distance Metric in K-Nearest Neighbours (KNN):**

**The Euclidean distance metric is typically used in KNN classification, although other distance metrics such as Manhattan distance, Minkowski distance, etc., can also be used.**

**The choice of distance metric impacts how the algorithm measures similarity between data points, which in turn affects its performance.**

1. Describe the Naïve-Bayes assumption of feature independence and its implications for classification.

Ans: **Naïve-Bayes Assumption of Feature Independence:**

**Naïve-Bayes assumes that the features are conditionally independent given the class label. This means that the presence of a particular feature in a class is independent of the presence of other features.**

**Despite this simplifying assumption, Naïve-Bayes often performs well in practice and is computationally efficient.**

1. In SVMs, what is the role of the kernel function, and what are some commonly used kernel functions?

Ans: **Role of Kernel Function in SVMs:**

**The kernel function in SVMs is used to transform the input data into a higher-dimensional space where it becomes linearly separable.**

**Commonly used kernel functions include linear kernel, polynomial kernel, Gaussian (RBF) kernel, and sigmoid kernel.**

1. Discuss the bias-variance trade-off in the context of model complexity and overfitting.

Ans: **Bias-Variance Tradeoff:**

**The bias-variance tradeoff refers to the balance between bias (error due to overly simplistic assumptions) and variance (error due to sensitivity to fluctuations in the training set) in machine learning models.**

**Increasing model complexity typically reduces bias but increases variance, and vice versa. Overfitting occurs when the model captures noise in the training data instead of the underlying pattern.**

1. How does TensorFlow facilitate the creation and training of neural networks?

Ans: **TensorFlow for Neural Networks:**

**TensorFlow facilitates the creation and training of neural networks by providing a flexible framework for building computational graphs, automatic differentiation for optimizing model parameters, GPU acceleration for faster computations, and high-level APIs like Keras for building and training neural networks more easily.**

1. Explain the concept of cross-validation and its importance in evaluating model performance.

Ans: **Cross-Validation:**

**Cross-validation is a technique used to assess the generalization performance of a predictive model. It involves partitioning the dataset into multiple subsets, training the model on some subsets, and evaluating it on the remaining subset.**

**Cross-validation helps in estimating how well the model will perform on unseen data and reduces the risk of overfitting.**

1. What techniques can be employed to handle overfitting in machine learning models?

Ans: **Techniques for Handling Overfitting:**

**Regularization techniques like L1 and L2 regularization penalize large model coefficients to prevent overfitting.**

**Feature selection to reduce the complexity of the model and focus on relevant features.**

**Early stopping during training to prevent the model from learning noise in the data**.

1. What is the purpose of regularization in machine learning, and how does it work?

Ans**: Regularization in Machine Learning:**

**Regularization is a technique used to prevent overfitting by adding a penalty term to the model's loss function, which discourages overly complex models.**

**Common regularization techniques include L1 regularization (lasso), L2 regularization (ridge), and elastic net regularization.**

1. Describe the role of hyper-parameters in machine learning models and how they are tuned for optimal performance.

Ans: **Hyperparameters in Machine Learning Models:**

**Hyperparameters are parameters that are set prior to training and control the learning process of the model.**

**They are tuned to optimize the model's performance, often through techniques like grid search, random search, or Bayesian optimization.**

1. What are precision and recall, and how do they differ from accuracy in classification evaluation?

Ans: **Precision, Recall, and Accuracy:**

**Precision measures the proportion of true positive predictions among all positive predictions.**

**Recall measures the proportion of true positive predictions among all actual positives.**

**Accuracy measures the proportion of correct predictions among all predictions.**

1. Explain the ROC curve and how it is used to visualize the performance of binary classifiers.

Ans: **ROC Curve:**

**The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier across different threshold settings.**

**It plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold values.**

**The area under the ROC curve (AUC) is used as a summary measure of the classifier's performance, with a higher AUC indicating better performance.**