Joint Tech Internship Community Program

TASK - 1

Candidate ID:2024060136

Candidate Name: Kishore S

Submitted by Kishore S

(kishores.22cse@kongu.edu)

List of Terminologies:

Feature

- Features are individual measurable properties or characteristics used as input variables for machine learning models to make predictions. For instance, in image classification, features include pixel values and color intensity. The selection and extraction of relevant features significantly impact the performance and accuracy of machine learning models, enhancing their predictive power by focusing on the most informative aspects of the data.

Labels

Labels are output variables that the model is trained to predict. They are also known as dependent variables, targets, or outputs.

- **Characteristics of Labels:**

- **Known Outcomes:** Labels are known outcomes for training data.
- **Ground Truth:** Labels represent the ground truth against which the model's predictions are compared.
- **Example:** Price

Prediction

- The output of a machine learning model. It is the estimated label produced by the model for given input features.

Outliers

- An outlier is an observation that is substantially different from the other observations.
- Outliers are important because they can change the result of our data analysis.

Types of Outliers:

- Univariate Outliers
- Multivariate Outliers

Test Data

- A subset of the dataset used to assess the performance of a trained machine learning model.
- This data is not utilized during the training process.

Training Data

- The portion of the dataset utilized for training the machine learning model, encompassing both the input features and their associated labels.
- This data serves to instruct the machine learning model, enabling it to understand the correlation between features (such as size, number of bedrooms, and age) and the target variable (price).

Model

- An ML model is a particular instance of a machine learning algorithm that has been trained on a dataset. The model stores the specific parameters and relationships between variables learned during training. It is used to make predictions on new data.

Validation Data

- A subset of the dataset used to offer an unbiased assessment of a model’s performance on the training data while tuning model hyperparameters.

Hyperparameter

- Hyperparameters are crucial in implementing a machine learning model. In machine learning, we typically deal with two types of parameters: model parameters and hyperparameters. Model parameters are derived from the dataset. For example, in a linear regression model, parameters like mmm (slope) and ccc (intercept) are calculated from the data. In contrast, hyperparameters are not derived from the dataset but are set before the training process begins.

Epoch

- An epoch is one complete pass through the entire training dataset. In neural networks, multiple epochs are usually employed to train the model.

Loss Function

- A loss function is a function that evaluates how closely the model's predictions align with the true labels. The objective of training a model is to minimize the value of the loss function.

Learning Rate

- The learning rate is a hyperparameter that dictates the extent to which the model is adjusted in response to the estimated error each time the model weights are updated. It determines the step size at each iteration when moving toward minimizing the loss function.

Overfitting

- Overfitting occurs when a model learns the training data too thoroughly, including the noise, which leads to poor performance on new, unseen data.

Underfitting

- Underfitting happens when a model is too simplistic to capture the underlying structure of the data. It fails to learn the training data adequately and also performs poorly on new, unseen data.

Regularization

- Regularization helps prevent overfitting by introducing a penalty for large coefficients. For instance, Ridge regression adds a penalty proportional to the square of the coefficients.

Cross Validation

- Cross-validation is a crucial technique in machine learning used to obtain a more reliable evaluation metric, such as an accuracy score.

Feature Engineering

- Feature engineering involves creating new features from raw data to enhance model performance.
- For example, extracting the hour of the day from a timestamp.

Dimensionality Reduction

- Dimensionality reduction reduces the number of features while preserving essential information.
- For example, using Principal Component Analysis (PCA) to simplify a dataset.

Bias

- Bias refers to the error that arises from making overly simplistic assumptions in a model.
- For example, using a linear model for a non-linear relationship.

Variance

- Variance measures a model's sensitivity to small changes in the training data.
- For example, a highly complex model like an overfitted decision tree might perform poorly on new data.