Joint Tech Internship Community Program

ASSIGNMENT - 1

Candidate ID:2024060134

Candidate Name: KAVINNATH RAVICHANDRAN

Submitted by KAVINNATH RAVICHANDRAN

(arkavinnath@gmail.com)

List of Terminologies

Feature

- A **feature** is an individual measurable property or characteristic of a phenomenon being observed. Features are the input variables used by the model to make predictions or classifications.

- **Image Classification:**
  - Features are Pixel values and Color Intensity.

Labels

- Labels are the ground truth or actual values used to train a supervised learning model. They are the answers that the model aims to predict.
- **Characteristics of Labels**:
  - **Known outcomes:**
    - **Labels are known outcomes for training data**: In supervised learning, the model is trained using data where the labels (or outcomes) are already known. This helps the model learn the relationship between the input features and the output labels.
  - **Ground Truth**
    - **Labels represent the ground truth against which the model's predictions are compared**: Labels serve as the benchmark for evaluating the model's performance. The model's predictions are compared against these labels to measure accuracy, precision, recall, and other metrics.

Prediction

- This output is the model's prediction based on the patterns it has learned from the training data. The estimated label is what the

model believes to be the correct outcome for the given input features.

Outliers

- Outliers deviate significantly from the general pattern of the data and may indicate variability, error, or a novel insight.
- Outliers can have a significant impact on statistical measures and model performance, potentially skewing results, and leading to misleading conclusions. It is essential to identify and address outliers to ensure accurate and reliable data analysis.
- **Types of Outliers**:
  - Univariate Outliers
  - Multivariate Outliers

Test Data

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

Training Data

- The subset of the dataset used to train the machine learning model. It includes both the input features and their corresponding labels, enabling the model to learn from the data.
- This data is used to train the machine learning model. The model learns the relationship between the input features (e.g., size, bedrooms, age) and the target variable (e.g., price) from this training data.

Model

- **An ML model is a specific instance of a machine learning algorithm that has been trained on a set of data**: The specific parameters and relationships between variables learned by the

model during training are stored in the model. The model is then used to make predictions on new data.

Validation Data

- It helps assess how well the model generalizes to new data and is used to fine-tune the model's hyperparameters before final evaluation.

Hyperparameter

- **Hyperparameters** are settings defined before training a machine learning model that control the training process and model structure. There are two kinds of parameters in machine learning: **model parameters** and **hyperparameters**. Model parameters are derived from the data and define the internal configuration of the model, such as weights in a neural network. In contrast, hyperparameters are external to the model and must be set manually or through optimization techniques. Proper tuning of hyperparameters is crucial for optimizing model performance and ensuring the model generalizes well to new data.

Epoch

- **An epoch** is one complete pass through the entire training dataset. In neural networks, multiple epochs are typically used to train the model, allowing it to learn and refine the patterns in the data over several iterations.

Loss Function

- **A loss function** is a function that measures how well the model's predictions match the true labels. The goal of training a model is to minimize the loss function, which quantifies the difference between the predicted values and the actual values. By minimizing the loss function, the model's accuracy and performance are improved, leading to better predictions.

Learning Rate

- **The learning rate** is a hyperparameter that controls the extent to which the model's weights are adjusted in response to the estimated error during each update. It determines the step size taken at each iteration when moving toward minimizing the loss function. A well-chosen learning rate helps ensure efficient convergence to the optimal model parameters.

Overfitting

- **Overfitting** occurs when a model learns the training data too well, including its noise and fluctuations, leading to poor performance on new, unseen data. This means the model has become too specialized to the training data and fails to generalize effectively to other data.

Underfitting

- **Underfitting** occurs when a model is too simple to capture the underlying structure of the data. It fails to learn the training data adequately, resulting in poor performance both on the training data and on new, unseen data.

Regularization

- **Regularization** prevents overfitting by adding a penalty for large coefficients, which discourages the model from becoming too complex. For example, Ridge regression applies a penalty proportional to the square of the coefficients, which helps to keep them small and the model more generalizable.

Cross Validation

- **Cross-validation** is a crucial technique in machine learning used to obtain a more reliable evaluation metric, such as an accuracy score. It involves splitting the dataset into multiple subsets or

folds and training the model on different combinations of these subsets, which helps to ensure that the evaluation metric reflects the model's performance across various data splits and reduces the risk of overfitting.

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 can provide additional context that improves the model's ability to make accurate predictions.

Dimensionality Reduction

- **Dimensionality reduction** decreases the number of features in a dataset while retaining essential information. For example, using Principal Component Analysis (PCA) helps simplify the dataset by transforming it into a lower-dimensional space, capturing the most important variance in the data.

Bias

- **Bias** refers to the error introduced by making overly simplistic assumptions in a model. For example, using a linear model to fit a non-linear relationship can lead to high bias, as the model fails to capture the complexity of the data accurately.

Variance

- **Variance** measures a model's sensitivity to small fluctuations in the training data. For example, a highly complex model, such as an overfitted decision tree, may perform well on the training data but poorly on new, unseen data due to its high variance.