Joint Tech Internship Community Program

ASSISGNMENT – 1

SUBMITTED BY

SARAVANAN S

(sarasakthi1995@gmail.com)

Feature

- An input variable used to make predictions.
- Example: In a dataset predicting house prices, features could include the size of the house, number of bedrooms, and location.

Label

- The output variable that the model is trying to predict.
- Example: In the house price prediction dataset, the house price is the label.

Prediction

- The estimated value or category generated by a model.
- Example: If a model predicts a house price, a prediction could be $300,000.

Outlier

- A data point that differs significantly from other data points.
- Example: In a dataset of house prices, a house priced at $1,000,000 in a neighbourhood where most houses are $200,000 is an outlier.

Test Data

- Data used to evaluate the performance of a trained model.
- Example: After training a model on house prices, you use test data to see how well it predicts prices on new, unseen houses.

Training Data

- Data used to train a machine learning model.
- Example: In a house price prediction model, the training data includes many houses' features and their prices.

Model

- The output generated after training an algorithm on a dataset.
- Example: A model trained to predict house prices based on features like size and location.

Validation Data

- Data used to tune hyperparameters and evaluate a model during training.
- Example: A separate set of house price data used to adjust the model before final testing.

Hyperparameter

- A parameter set before training that controls the training process.
- Example: The learning rate that determines how much the model's weights are updated with each step.

Epoch

- One complete pass through the training data.
- Example: Training a house price model for 10 epochs means the model sees the entire training data 10 times.

Loss Function

- A measure of how well the model's predictions match the actual values.
- Example: Mean Squared Error (MSE) is a common loss function that calculates the average squared difference between predicted and actual house prices.

Learning Rate

- A hyperparameter that controls how much the model's weights are updated during training.
- Example: A small learning rate means the model's weights are updated slightly at each step, leading to slower but potentially more accurate training.

Overfitting

- When a model learns the training data too well, including noise.
- Example: A house price model that perfectly predicts the training data prices but performs poorly on new data.

Underfitting

- When a model is too simple to capture the underlying patterns in the data.
- Example: A house price model that performs poorly on both training and new data because it doesn't capture the complexity of the data.

Regularization

- A technique to prevent overfitting by adding a penalty for large coefficients.
- Example: L1 (Lasso) regularization adds a penalty to the absolute values of the coefficients in a house price model.

Cross-Validation

- A technique to evaluate how well a model generalizes by partitioning the data into subsets.
- Example: Using 10-fold cross-validation to train and validate a house price model on different subsets of data.

Feature Engineering

- Creating new features from raw data to improve model performance.
- Example: Combining the number of bedrooms and bathrooms into a single feature in a house price model.

Dimensionality Reduction

- Reducing the number of features while preserving information.
- Example: Using Principal Component Analysis (PCA) to reduce a large set of house features to a smaller set of principal components.

Bias

- Error introduced by approximating a complex problem with a simplified model.
- Example: A house price model with high bias might predict all house prices as the average price, missing the variations.

Variance

- Error introduced by the model's sensitivity to small fluctuations in the training data.
- Example: A house price model with high variance might predict training data prices very well but fail on new data.