GENERATIVE AI CONSORTIUM SystimaNX

Joint Tech Internship Community Program

ASSIGNMENT - 1

Candidate ID:2024060109

Candidate Name: Pravin S

List of Terminologies:

Feature

ID	Number of Bedrooms	Square Footage	Mileage	Location	Price
1	3	2000	80000	Suburban	$200.00
2	2	1500	92000	Urban	$300.00
3	4	2500	61000	Rural	$400.00

- ❖ Feature are Individual measurable properties or characteristics input variables Machine learning models uses to make predication.

Image Classification:

- ❖ Features are Pixel values and Color Intensity.

Labels

ID	Number of Bedrooms	Square Footage	Location	Price(Label)
1	3	2000	Suburban	$200.00
2	2	1500	Urban	$300.00
3	4	2500	Rural	$400.00

- ❖ Labels are output variable that the model is trained to predict . They are also known as dependent variable, target, or output.
- ❖ 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 evaluate the performance of a trained machine learning model.
- This data is not used during the training process.

Size (sq ft)	**Number of Bedrooms**	**Age**	**Price(Label)**
1500	3	20	$200.00
1600	2	15	$300.00

Training Data

- The subset of the dataset used to train the machine learning model. It includes both the input features and their corresponding labels.

□

Size (sq ft)	Number of Bedrooms	Age	Price(Label)
1500	3	20	$200.00
1600	2	15	$300.00
1700	4	12	$400.00

- This data is used to train the machine learning model. The model learns the relationship between the features (size, bedrooms, age) and the target variable (price).

Model

- ML model is a specific instance of machine learning algorithm that has been trained on a set of data. The specific parameter and relationship between variables learned by the model during training are stored in the model. The model are used to make predications of new model.

Validation Data

- A subset of the dataset used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters.

Hyperparameter

- This is one of the most important things that we have to do while implementing a machine leaning model so in machine learning we usually have two kind of parameters so first kind of parameter is the model parameter and these model parameter are derived from the data set that we have so if we think about a linear regression model we have parameter such a m and c where m is our slope and c is our intercept so m and c values are derived from the dataset and the other set of parameter is called as the hyperparameter.

Epoch

- ❖ One complete pass through the entire training dataset. In neural networks, multiple epochs are typically used to train the model.

Loss Function

- ❖ 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.

Learning Rate

- ❖ The learning rate is a hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated. It determines the step size at each iteration while moving toward a minimum of the loss function.

Overfitting

- ❖ Overfitting happens when a model learns the training data too well, including the noise, causing it to perform poorly on new 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 well enough and also performs poorly on new data.

Regularization

- ❖ Regularization prevents overfitting by adding a penalty for large coefficients. For example Ridge regression adds a penalty proportional to the square of the coefficients.

Cross Validation

- ❖ The cross validation is a very important technique that we use in machine learning in order to get a more reliable evaluation metric such as an accuracy score .

Feature Engineering

- ❖ Feature engineering creates new features from raw data to improve model performance.
- ❖ For example: extracting the hour of the day from a timestamp.

Dimensionality Reduction

- ❖ Dimensionality reduction decreases the number of features while keeping important information.
- ❖ For example:- using Principal Component Analysis (PCA) to simplify a dataset.

Bias

- ❖ Bias is an error 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 fluctuations in the training data.
- For example:- a highly complex model like an overfitted decision tree might perform poorly on new data.