ASSIGNMENT – 1

A.TARUN

[22ADR109]

B.TECH[AI-DS]

Let's go through the terminologies and create an example dataset to illustrate each concept in the context of predicting house prices.

Example Dataset: House Prices

Got it! Here’s your data in a table format:

| **ID** | **Location** | **Size (sq ft)** | **Bedrooms** | **Age (years)** | **Price (USD)** |
| --- | --- | --- | --- | --- | --- |
| 1 | Suburb A | 2000 | 3 | 10 | 300,000 |
| 2 | Suburb B | 1500 | 2 | 5 | 250,000 |
| 3 | Suburb A | 2500 | 4 | 8 | 400,000 |
| 4 | Suburb C | 1800 | 3 | 15 | 200,000 |
| 5 | Suburb B | 2200 | 3 | 12 | 350,000 |
| 6 | Suburb A | 2400 | 4 | 3 | 450,000 |

Is there anything specific you’d like to do with this data? For example, analyzing trends or creating visualizations?

Terminologies Explained:

1. Feature : These are the individual measurable properties or characteristics used as input to the model. In the dataset above, the features include "Location," "Size (sq ft)," "Bedrooms," and "Age (years)."

2. Label : The label is the output variable that the model aims to predict. In this dataset, the label is "Price (USD)."

3. Prediction : This is the output generated by the model based on the input features. For example, if we input "Suburb B, 1600 sq ft, 2 bedrooms, 6 years," the model might predict a price of $260,000.

4. Outlier : A data point that deviates significantly from the rest of the data. For example, if a house in Suburb A with similar features as others is priced at $1,000,000, it could be considered an outlier.

5. Test Data : A subset of the dataset used to evaluate the performance of the model. It's not used during the training phase. For instance, rows 2 and 4 could be part of the test data.

6. Training Data : The portion of the dataset used to train the model. It includes features and labels. Rows 1, 3, 5, and 6 could serve as training data.

7. Model : The mathematical representation that maps the features to the label. In this context, a model might be a linear regression equation that predicts house prices.

8. Validation Data : A separate subset of data used to tune the model's hyperparameters. It helps prevent overfitting by evaluating the model during training.

9. Hyperparameter : These are configuration settings used to structure the model, not learned from the data. Examples include the learning rate, the number of trees in a random forest, or the number of layers in a neural network.

10. Epoch : In the context of machine learning, an epoch refers to one complete pass through the entire training dataset. For example, if we have 4 rows of training data and 10 epochs, the model will see the dataset 10 times.

11. Loss Function : A function that measures the error between the predicted value and the actual label. For example, Mean Squared Error (MSE) could be used to calculate the difference between predicted and actual house prices.

12. Learning Rate : A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated.

13. Overfitting : This occurs when the model learns the training data too well, including the noise, and performs poorly on new, unseen data. It might happen if the model starts to memorize the training data.

14. Underfitting : When a model is too simple and cannot capture the underlying trend of the data, it underfits. This leads to poor performance on both the training and test data.

15. Regularization : Techniques used to reduce the risk of overfitting by adding a penalty for larger coefficients. Examples include L1 and L2 regularization.

16. Cross-Validation: A technique to assess the model's performance by dividing the dataset into multiple training and testing sets. For example, in k-fold cross-validation, the dataset is divided into k parts, and the model is trained k times, each time with a different part as the test set.

17. Feature Engineering : The process of selecting, modifying, or creating features to improve model performance. For example, creating a new feature that combines "Size (sq ft)" and "Bedrooms" to represent "Space per Bedroom."

18. Dimensionality Reduction : Techniques used to reduce the number of input variables in a model, such as Principal Component Analysis (PCA), to reduce the "curse of dimensionality."

19. Bias : The error introduced by approximating a real-world problem, which may be complex, by a much simpler model. High bias can cause underfitting.

20. Variance : The error introduced by the model's sensitivity to small fluctuations in the training set. High variance can lead to overfitting.