**MACHINE LEARNING ASSIGNMENT**

|  |  |  |  |
| --- | --- | --- | --- |
| Number of bed room | Square footage | location | Price |
| 3 | 2000 | Urban | $500,000 |
| 2 | 1500 | Suburban | $300,000 |
| 4 | 2500 | Urban | $600,000 |
| 3 | 1800 | Rural | $350,000 |
| 5 | 3000 | Urban | $800,000 |
| 2 | 1200 | Suburban | $250,000 |
| 4 | 2200 | Rural | $400,000 |

**List of Terminologies:**

• Feature

Feature is the input used to feed the ML model. For example in the above sample dataset the first three columns are the features(no. of bedroom, squareFootage , location).

• Label

It is the output for the above inputs. Here, price is the label.

• Prediction

Based on the features and label the model is trained and when we give new input the output will be predicted.

• Outlier

These are the points that differ from the other points and cause error in the pattern . It should be removed. Some algorithms are less sensitive to these outliers like decision tree. Consider in the above example one of the price is just $100 .It will be wrong . And that point is outlier.

• Test Data

Some part of the dataset is taken and used as test data .ie, once the model is trained it is tested with this dataset.

• Training Data

Some part of the dataset is for the test data and the remaining data is the training data. It is used to build the model ,find patterns between feature and label.

Here ,consider the first three rows are for test data and the remaining rows are training data.

• Model

It is a program that can find patterns and make decisions. Models are trained with dataset and algorithms. Here, the model will predict the price of the house based on the features given.

• Validation Data

It is similar to test data but the validation data is used during training . In case of test data ,it is used to test the final model. Example: A separate set of house data used to validate the model's performance during training.

• Hyperparameter

It’s like a configuration that is external to the model . It cannot be predicted from the data.

• Epoch

One complete pass through the entire training dataset during the training process.

Example : here the dataset has 7 rows ,one epoch means the has processed all the 7 rows once.

• Loss Function

A function that measures the difference between the predicted values and the actual values. The goal of training is to minimize this loss.

Example :Mean squared error (MSE) used in regression tasks.

• Learning Rate

It is a hyperparameter. It controls the step size during the optimization process. It determines how much the model weights are updated in response to the estimated error.

• Overfitting

Here the model learns the training data well but performs poorly for new or unseen data .

Example :if the model is trained using the above data well but when a new unseen data is given and the predicted value is wrong .

• Underfitting

Here a model is too simple to capture the underlying patterns in the data, resulting in poor performance on both training and test data.

Example : A house price model that cannot capture the relationship between features and price due to its simplicity.

• Regularization

It is a techniques used to reduce overfitting by adding a penalty to the loss function for large coefficients.

• Cross-Validation

It a technique for assessing the performance of a model by partitioning the data into multiple subsets and training/testing the model on different combinations.

Example: K-fold cross-validation where the data is divided into k subsets, and the model is trained and tested k times.

• Feature Engineering

It is the process of improving the model performance

Example : Creating a new feature called "price per square foot" from the house price and square footage.

• Dimensionality Reduction

It is a technique used to reduce the number of input variables in a dataset by combining or removing features.

• Bias

Bias is also known as algorithm or AI bias, is when a model's predictions are systematically prejudiced. This can happen when the model's assumptions about the data are incorrect, or when human involvement in the data collection process introduces bias.

Example : A linear model predicting house prices when the actual relationship is nonlinear.

• Variance

 It's a measure of how much the model's predictions vary, or how sensitive it is to small fluctuations in the data. Ie, when trained on different subsets of training data. Variance is not a measure of overall accuracy.

Example : A complex model that predicts well on training data but poorly on test data due to its sensitivity to training data noise.