Joint Tech Internship Community Program

Assignment 1

SUBMITTED BY

PARTHASARATHI S

(parthasarathi262004@gmail.com)

Example Dataset: Car price prediction

Make	Year	Mileage	Engine size	No. of doors	Price
Toyota	2018	30,000	2.0	4	5,00,000
Honda	2017	40,000	2.2	4	4,85,000
Ford	2019	20,000	2.5	2	5,20,000
BMW	2020	10,000	3.5	4	8,50,000

Features:

Features are individual measurable properties of the data. These are the input variables. The machine learning model used to make prediction using this input variables.

Ex: Make, year, Mileage, Engine size, number of doors are the input variables. This is called Features.

Label:

Label are the output variables that the model is trained to predict.

Ex: Price is the output variable. This is called Label.

Prediction:

The model can used to make prediction on new data.

Outliers:

- Outliers are data points that significantly differ from the majority of the data in a dataset. Outliers does not make the pattern.
- Outliers are handled using data visualization.
- Decision trees are less sensitive to outliers.

Ex: Car mileage around 10,000 to 1,00,000. a car mileage at above 1,00,000 might be considered at outlier.

Training Data:

- Training dataset is used for training purpose and this thing is called train model.
- It has attributes used for training machine learning algorithm to prepare model. From that they find relationship.

Test Data:

- Testing dataset is used for testing purpose and this thing is called evaluate model.

- Testing error is occurred by accessing the model by providing the unknown data to the model.

Model:

Model is mathematical representation of algorithm that is trained using data to predict label.

Validation data:

- After training, validation data gives which model is best.
- Choose the best hyperparameters.

Hyperparameter:

- Hyperparameters are configuration setting defined before training that control the learning process.
- It is manually specified.

Epoch:

When we have input the entire dataset once. Entire dataset has passed through the network, the network has seen every single training example once.

Loss Function:

- Measures the difference between the predicted output of a model and actual target values.
- It quantifies a model prediction error and guides the optimization process.

Learning Rate:

It controls how fast or slow a model learn. A high learning rate can make the model miss the best solution, a low learning rate can make learning very slow.

Overfitting:

- Model learns the training data too well.
- It performs well on training data but poorly on unseen data.

Ex: Memorizing the prices in the table

Underfitting:

It performs poorly on both training data and new data because it fails to learn and generalize from the data effectively.
Ex: Predicting the same price for all cars.

Regularization:

- A technique used to reduce errors by fitting the function appropriately on the given training set and avoiding overfitting.
- Lasso regularization- L1 regularization
- Ridge regularization- L2 regularization
- Elastic net regularization- L1 and L2 regularization

Ex: L2 regularization to penalize large coefficients

Cross-Validation:

- It access the performance of machine learning model by dividing the dataset into multiple subsets.
- Minimize training time.
- Minimize running time.
- Maximum accuracy.

Feature Engineering:

It perform select, transform or create new features from raw data to improve the performance of machine learning models.
Ex: Combining mileage and year to create “age” feature.

Dimensionality Reduction:

- It used to reduce the number of features in a dataset while preserving much information as possible.
- This process helps in simplifying models and reduce the risk of overfitting.

Ex: Using PCA to combine Engine Size and Number of Doors into a single feature

Bias:

- It refers to error introduced by approximating a real world problems.
- It can lead to errors in prediction or estimations.
- High bias -training performance is low. It causes underfitting.

Ex: Consistently predicting lower prices than actual

Variance:

- High variance- Validation performance is low.
- It causes overfitting.

Ex: Predicting significantly different prices for similar cars based on minor changes in the data