**SHALINI Y S**

**22CSR186**

**ASSIGNMENT 1**

|  |  |  |  |
| --- | --- | --- | --- |
| **EMPLOYEE ID** | **HOURS WORKED** | **PROJECTS COMPLETED** | **PERFORMANCE SCORE** |
| 1 | 40 | 3 | 85 |
| 2 | 35 | 2 | 78 |
| 3 | 45 | 4 | 92 |
| 4 | 30 | 2 | 70 |
| |  | | --- | | 5 | | |  | | --- | | 5 | | |  | | --- | | 5 | | |  | | --- | | 5 | |

**Feature:** Individual measurable properties or characteristics used as inputs to the model.

* Eg: the features are hours worked and projects completed.

**Label:** The output variable that the model aims to predict.

* Eg: the label is the performance score.

**Prediction:** The output from the model after it has been trained.

* Eg: hours worked and projects completed will predict a performance score.

**Outlier:** A data point that differs significantly from other observations.

* Eg: if one employee had a performance score of 100 while others are between 70-92, then this score will be considered an outlier.

**Test Data:** The portion of the data used to evaluate the model's performance.

* Eg: employees 4 and 5 can be used as test data to evaluate the model trained on the other employees.

**Training Data:** The portion of the data used to train the model.

* Eg: Employees 1, 2, 3 can be used as training data.

and labels.

* Eg: a linear regression model that predicts performance scores based on hours worked and projects completed.

**Validation Data:** A subset of the training data used to tune hyperparameters and avoid overfitting.

* Eg: using employee 3's data to validate the model trained on employees 1 and 2.

**Hyperparameter:** Parameters whose values are set before the learning process begins.

* Eg: learning rate or the number of epochs.

**Epoch:** One complete pass through the entire training dataset.

* Eg: if we iterate through all the employees' data once during training it counts as one epoch.

**Loss Function:** A method to evaluate how well the model's predictions match the actual data.

* Eg: mean squared error (MSE) could be used to measure the difference between the predicted and actual performance scores.

**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.

* Eg: a learning rate of 0.01 means the model is updated slowly while a learning rate of 1 means the model is updated quickly.

**Overfitting:** When a model learns the training data too well including noise and outliers leading to poor performance on new data.

* Eg: a model that perfectly predicts the training data but performs poorly on the test data.

**Underfitting:** When a model is too simple to capture the underlying pattern in the data leading to poor performance on both training and new data.

* Eg: a model that performs poorly on both the training and test data.

**Regularization:** Techniques used to reduce overfitting by penalizing complex models.

* Eg: adding a penalty to the loss function for large coefficients in a linear regression model.

**Cross-Validation:** A technique to evaluate the model's performance by dividing the data into several subsets and training/testing the model on different combinations of these subsets.

* Eg: 5-fold cross-validation involves splitting the data into 5 parts, training the model on 4 parts and testing on the remaining part, repeating this process 5 times.

**Feature Engineering:** The process of creating new features or modifying existing ones to improve model performance.

* Eg: creating a new feature total effort by multiplying hours worked with projects completed.

**Dimensionality Reduction:** Techniques to reduce the number of features while retaining important information.

* Eg: principal component analysis (PCA) to reduce hours worked and projects completed into a single combined feature.

**Bias:** The error introduced by approximating a real-world problem which may be complex by a much simpler model.

* Eg: a high bias model may consistently predict performance scores far from the actual scores.

**Variance:** The error introduced by the model's sensitivity to small fluctuations in the training set.

* Eg: a high variance model may predict performance scores very accurately for training data but poorly for test data.

**Model:** A mathematical representation of the relationship between features and labels.

* Eg: a linear regression model that predicts performance scores based on hours worked and projects completed.

**Validation Data:** A subset of the training data used to tune hyperparameters and avoid overfitting.

* Eg: using employee 3's data to validate the model trained on employees 1 and 2.

**Hyperparameter:** Parameters whose values are set before the learning process begins.

* Eg: learning rate or the number of epochs.

**Epoch:** One complete pass through the entire training dataset.

* Eg: if we iterate through all the employees' data once during training it counts as one epoch.

**Loss Function:** A method to evaluate how well the model's predictions match the actual data.

* Eg: mean squared error (MSE) could be used to measure the difference between the predicted and actual performance scores.

**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.

* Eg: a learning rate of 0.01 means the model is updated slowly while a learning rate of 1 means the model is updated quickly.

**Overfitting:** When a model learns the training data too well including noise and outliers leading to poor performance on new data.

* Eg: a model that perfectly predicts the training data but performs poorly on the test data.

**Underfitting:** When a model is too simple to capture the underlying pattern in the data leading to poor performance on both training and new data.

* Eg: a model that performs poorly on both the training and test data.

**Regularization:** Techniques used to reduce overfitting by penalizing complex models.

* Eg: adding a penalty to the loss function for large coefficients in a linear regression model.

**Cross-Validation:** A technique to evaluate the model's performance by dividing the data into several subsets and training/testing the model on different combinations of these subsets.

* Eg: 5-fold cross-validation involves splitting the data into 5 parts, training the model on 4 parts and testing on the remaining part, repeating this process 5 times.

**Feature Engineering:** The process of creating new features or modifying existing ones to improve model performance.

* Eg: creating a new feature total effort by multiplying hours worked with projects completed.

**Dimensionality Reduction:** Techniques to reduce the number of features while retaining important information.

* Eg: principal component analysis (PCA) to reduce hours worked and projects completed into a single combined feature.

**Bias:** The error introduced by approximating a real-world problem which may be complex by a much simpler model.

* Eg: a high bias model may consistently predict performance scores far from the actual scores.

**Variance:** The error introduced by the model's sensitivity to small fluctuations in the training set.

* Eg: a high variance model may predict performance scores very accurately for training data but poorly for test data.