AI/ML Internship Program

Assignment 1

K.Vaishnavi

(22CSR228)

For better understanding the ML terminologies, we can see with the sample training dataset.

Here I can take an Employee Salaries:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Employee ID** | **Years of Experience** | **Education Level (1-5)** | **Performance Score (1-10)** | **Age** | **Salary (USD)** |
| 001 | 3 | 4 | 7 | 25 | 50000 |
| 002 | 5 | 5 | 8 | 30 | 60000 |
| 003 | 2 | 3 | 6 | 24 | 45000 |
| 004 | 10 | 5 | 9 | 35 | 120000 |
| 005 | 7 | 4 | 8 | 29 | 70000 |

• **Feature** - An input variable(x) which is used to train a model.

Here employee id, years of experience, education level, performance score and age are the features for the dataset.

• **Label** - An output variable(y) which is used to predict a model.

Here Salary is the label for the following features what we trying to predict.

• **Prediction** – An output generated by a model based on input/features.

Here predicted the salary based on the features like year of experience, age, education level and so on.

• **Model** – It defines the relationship between feature and label. This relationship is derived by trying to fit various readily available algorithms or writing an custom algorithm.

Here by using regression model we predict the salary based on year of experience an education level.

• **Outlier** – The data which is significantly differs from the majority of data.

Here an employee with salary 120000 extremely high when comparing with others, so it would be an outlier.

• **Training Data** - The training set is used to train the model.

Here rows with employees like Id 001,002 and 003 used to train the model on how to predict salary.

• **Validation Data** – The validation set aids in model selection and hyperparameter tuning.

Here rows with employee Id 004 used to tune model parameters and evaluate performance during training.

• **Test Data** - The test set evaluates its performance on unseen data.

Here data from employee Id 005 used to access how well the model predicts salary for unseen employee.

• **Epoch** – The total number of iterations of all the training data in one cycle for training the ML model.

Here one complete iteration through all the employee data to update the model's weights.

• **Loss Function** – It quantifies the difference between the predicted outputs of a ML algorithm and the actual target values.

Here Mean Squared Error (MSE) to measure the difference between the predicted salary and the actual salary.

• **Learning Rate** – It is a tuning parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function.

Here learning rate of 0.01 might be used to adjust the weights in small steps.

• **Hyperparameter** – The parameter whose value is chosen before a learning algorithm is trained.

Here the learning rate or the number of layers in a neural network.

• **Overfitting** – It occurs when the model cannot generalize and fits too closely to the training dataset instead.

Here if the model predicts Salary very accurately for training data but fails to generalize to new employees, it might be overfitting.

• **Underfitting** – It occurs when the machine learning model is not well-tuned to the training set.

Here by using a very basic model that cannot capture the influence of Education Level on Salary might lead to underfitting.

• **Bias** – A model with a high bias error underfits data and makes very simplistic assumptions on it.

If the model is too simplistic and does not capture the relationship between Performance Score and Salary, it might have high bias.

• **Variance** – A model with a high variance error overfits the data and learns too much from it.

If the model performs well on the training data but poorly on new employees due to capturing noise, it has high variance.

• **Regularization** – It is a set of methods for reducing overfitting in a ML models.

Adding a regularization term to penalize large coefficients in a regression model to avoid overfitting.

• **Cross-Validation** – It is a technique used in ML and statistical modeling to assess the performance of a model and to prevent overfitting.

Using 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 for the employee data.

• **Dimensionality Reduction** – It is a process and technique to reduce the number of dimensions or features in a data set.

Using PCA to reduce the number of features if there were many additional features beyond the ones listed.

• **Feature Engineering** – The process of selecting and transforming variables when creating a predictive model using ML.

Creating a new feature like Experience per Year by dividing Years of Experience by Age to better capture experience trend.