MACHINE learning assignment 1

KARUPPUCHAMY AISHWARYA (22CSR091)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| AGE | INCOME | GENDER | EDUCATION LEVEL | MARITAL STATUS | CREDIT SCORE | PURCHASE |
| 22 | 25000 | F | Bachelor’s | Single | 650 | 0 |
| 45 | 45000 | F | Master’s | Married | 250 | 1 |
| 67 | 35000 | F | High school | Married | 140 | 1 |
| 12 | 30000 | M | PhD | Single | 680 | 0 |
| 45 | 40000 | F | PhD | Single | 700 | 1 |
| 34 | 70000 | M | Midschool | Single | 750 | 0 |
| 89 | 55000 | M | Highschool | Married | 150 | 1 |
| 46 | 25000 | M | Bachelor’s | Married | 280 | 0 |
| 37 | 22000 | F | Master’s | Single | 690 | 0 |
| 39 | 80000 | M | Master’s | Single | 620 | 0 |
| 20 | 30000 | F | Bachelor’s | Married | 700 | 1 |
| 35 | 47000 | M | PhD | Single | 750 | 1 |
| 27 | 32000 | F | Midschool | Married | 710 | 1 |

Terminologies USED

* FEATURE : Feature is an individual measurable property within a recorded data set. features are often called “variables” or “attributes.” In this data set Age, Income, Gender, Education, Marital Status, Credit Score, Purchase.
* LABEL : A label is the thing we're predicting—the y variable in simple linear regression. The target variable we want to predict. Here, "Purchase" is the label, indicating whether a person will purchase the product (1) or not (0).
* PREDICTION : The purpose of prediction in machine learning is to project a probable data set that relates back to the original data. For instance, predicting if Purchase = 1 or Purchase = 0.
* **Outlier**: It is a data point significantly different from other data points in a dataset. For example, if a new row had "Age" 100 and "Income" 150000, it would be an outlier.
* **Test Data**: 20% of the data randomly selected, used to evaluate the model's performance.
* **Training Data**: 80% of the data randomly selected, used to train the model.
* MODEL : A Model defines the relationship between the label (y) and the features (x).
* VALIDATION DATA:  It is a data set of examples used to tune the architecture of a classifier. 10% of the training data, used to tune hyperparameters.
* HYPERPARAMETERS : Parameters like learning rate, regularization strength, or tree depth.
* EPOCH : It refers to the one entire passing of training data through the algorithm.
* LOSS FUNCTION : Binary cross-entropy for classification.
* LEARNING RATE : Controls the step size during optimization.
* OVERRATING : When the model performs well on training data but poorly on test data.
* UNDERFITTING : When the model performs poorly on both training and test data.
* REGULARIZATION : Techniques like L1 (Lasso) or L2 (Ridge) regularization to prevent overfitting
* CROSSVALIDATION : For example, k-fold cross-validation to evaluate model performance.
* FEATURE ENGINEERING : Creating new features, such as interaction terms between "Age" and "Income".
* DIMENSIONALITY REDUCTION : Techniques like PCA to reduce the number of features.
* BIAS: Error due to simplistic model assumptions.
* VARIANCE : Error due to model sensitivity to training data fluctuations.