Assignment 1

1) What is Hypothesis in Machine Learning?

In Machine Learning, a hypothesis is a function or a model that map input features to output predictions. It represents the relationship between the input variables and the output.

In linear regression, the hypothesis might be $h(x) = \theta_0 + \theta_1 x$ where θ_0 and θ_1 are parameters to be learned.

2) Explain the impact of taking learning rate? When can you take smaller value of learning rate and in which situation the value can be larger?

The learning rate controls how much the model's parameters are updated during training. It is a hyperparameter that determines the step size at each iteration while moving towards minimizing ~~and~~ the loss function.

Impact:

too high : The model may overshoot the optimal solution, leading to divergence or instability

too low : The model may converge very slowly to get stuck in a local minimum

When to use:

Smaller

→ When data is noisy and complex

→ When model is close to convergence

Larger:

→ When dataset is large and simple

→ During the initial stages of training to speed up convergence

3) What is Overfitting? How to avoid it?

Overfitting occurs when a model learns the training data too well, including noise and outliers, resulting in poor performance on unseen data.

→ Use regularization techniques
→ Perform cross-validation
→ Use drop out in neural networks
→ Simplify the model
→ Increse size of training dataset
→ Use early stopping during training

4) Why the Use of Linear Regression is not preffered for classification task?

Linear Regression predicts the continuous values while classification tasks req. discrete output.

Lineare regression does not provide probabilities or clear decision boundaries for classif.

5) Why do we perform Normalization?

Normalization scales the features to a single values [standard range] to ensure that all feature contribute equally to model

→ Improves convergence speed in gradient-based alg.
→ Prevent features with larger scale from model

6) What is Bias - Variance Trade-off ?

Bias is an error due to overly simplistic assumptions in the model. High bias can cause underfitting.

Variance : error due to the model's sensitivity to small fluctuations in the training set. High variance can cause ~~under~~ fitting

Trade off : Increasing model complexity reduces bias but increase variance and vice versa. the goal is to find the right balance to minimize total errors.

7) What is 'Training set' and 'Test set' in model?

Training set is the portion of dataset used to train the model by adjusting the parameter

Testing set is the portion of dataset used to evaluate the model's performance on unseen data.

8) How to handle missing values in Dataset.

Remove : Drop rows and columns with missing data if the dataset is large enough.

Impute: Replace values with mean, median, mode or use advanced techniques like K.NN

Predict : Use ML models to predict missing values

9) Explain Evalution matrix for classification problems

In classification task, goad is to predic discrete class labels the performance of a model is evaluted using various metrix, depending on the problem and nature of dataset.

1) Confusion Matrix

2) Accuracy = $\frac{TP + TN}{TP + TN + FP + FN}$

3) Precision = $\frac{TP}{TP + FP}$

4) Recall = $\frac{TP}{TP + FN}$

5) F1 score = $2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

10) Evalution matrices for regression Problems

1) MAE = $\frac{1}{n} \Sigma [y_i - \hat{y}_i]$

→ measures absolute differences between predicted and actual values.

2) MSE = $\frac{1}{n} \Sigma [y_i - \hat{y}_i]^2$

→ Panalize larger errors more than MAE

3) RMSE $= \sqrt{MSE}$

∘ Same unit as target variables, help interpret model performance

4) R-squared score

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

→ Measures how well the model explain varias in data. Higher is better

When to use
→ MAE when all errors matter equally
→ MSE /RMSE when larger errors need more panelty
→ R^2 when accessing goodness of fit.