ASSIGNMENT-1

SHRIYA SURAVARAPU
241003

SECTION-A

1. True 2. True 3. True 4. True 5. True

6. False 7. False 8. True 9. True 10. True

SECTION-B

1.

Model	Loss function	Regularizer
SVM	Hinge loss	$L_2 \|\|w^2\|\|$
LASSO	Mean Squared Error (MSE) (regression)	$L_1 \|\|w_i\|\|$
RIDGE	Mean Squared Error (regression)	$L_2 \|\|w^2\|\|$

3. (a) Gradient descent requires the loss function to be differentiable so that the slope can be computed.

∴ The standard loss functions like MSE can be optimized.

(b) Newton's method requires the second order derivatives,

∴ twice differentiable functions like MSE can be optimized.

SECTION-C

1. Underfitting occurs when the model is too simple and does not have enough capacity to capture the complexity in the data.

2. This happens due to underfitting, from high bias. The model cannot learn the structure of the data.

3. Bagging trains multiple models on different samples of the data, and each model makes different predictions & they have high variance. So we average these out to make the overall model's prediction more stable with less variance.

4. Boosting works iteratively adding new models to correct the errors made by previous models. ∴ This reduces bias, though it can increase variance because all the models would be very corelated.

SECTION - B

1. Let a leaf region R_m contain the target values y_i $\forall i = \{1, \ldots, n\}$ for n data points
We want to find the optimal prediction c_m that minimizes the Squared loss.

$$\text{Squared loss} \quad L(c_m) = \sum_{i \in R_m} (y_i - c_m)^2$$

$$\text{Minimum} \Rightarrow \text{derivative} = 0 \Rightarrow \frac{\partial L}{\partial c_m} = \sum_{i \in R_m} 2(y_i - c_m)(-1) = 0$$

$$\Rightarrow \sum_{i \in R_m} y_i - \sum_{i \in R_m} c_m = 0$$

$$\therefore c_m \text{ is const for all points in } R_m \Rightarrow \sum_{i \in R_m} y_i = n c_m$$

$$\Rightarrow c_m = \frac{1}{N_m} \sum_{i \in R_m} y_i$$

∴ Optimal prediction is the mean of the target values in the leaf region

2. Gini impurity $G = 1 - \sum_{k=1}^{3} p_k^2$ p_k = probability for each k

Minimum G ⇒ node is homogeneous (pure distribution)

Max G ⇒ node is perfectly mixed (equal dist)

Min → $(1,0,0) \Rightarrow G = 1 - (1^2 + 0^2 + 0^2) = 0$

Max → $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \Rightarrow G = 1 - \left(\frac{1}{3}^2 + \frac{1}{3}^2 + \frac{1}{3}^2\right) = \frac{2}{3}$

3. Decision trees are "myopic" because they use greedy search algorithm at each node, they look only one step ahead and choose the split that yields maximum gain in the short term.

4. (i) Simplifying the tree by removing the branches & nodes that do not fit well

(ii) Stop the tree construction process before it creates overly specific branches

SECTION F

1. No, because the purpose of testing is to evaluate the model's ability to generalize unseen data. So if we test on training data, we get a low test error, which is not real, and is not reliable to measure the model's true performance.

It would make the results highly optimistic.

2.

Bagging	Boosting
→ Models are trained independently & parallel to each other.	→ Models are trained sequentially and each new model is trained to correct the errors of the previous.
→ Uses random bootstrap samples of the training data for each model.	→ Uses the full dataset; adjusts (weights) predictions based on previous model's performance.
→ Main goal is to reduce variance. It averages out the errors of all models.	→ Reduce bias. Creates a single model sequentially, by weighted sums.