WS

1. Consider a training data-set of n points for a regression problem. Assume that the model is linear. Let w_1 and w_2 be the optimal weight vectors obtained from solving the following optimization problems:

$$w_1 = \arg\min_w \sum_{i=0}^{n} (w^T x_i - y_i)^2$$
$$w_2 = \arg\min_w \sum_{i=0}^{n} (w^T x_i - y_i)^3$$

Choose the most appropriate answer.

(a) w_1 will generalize better than w_2 on the test data-set. ✓

(b) w_2 will generalize better than w_1 on the test data-set.

(c) Both models will show identical performance on the test data-set.

ω_1 will generalize better than ω_2 because, ω_2 is a cubic equation. Cubic equations can give negative answers. If the values are negative, they might even sum upto zero.

2. Consider kernel regression with the kernel function $(x_1^T x_2 + 2)^2$ applied on the following dataset.

$$X = \begin{bmatrix} 1 & 0 & 2 & 0 & 3 & 0 \\ 0 & 1 & 0 & 2 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The optimal weight vector w^* is given by:

$$w^* = \phi(X)[0.1, 2, 3.9, 5, 6, 8]^T$$

where ϕ is the transformation mapping corresponding to the given kernel. What will be the prediction for the data point $[0, 0, 1]^T$?

Given,

Kernel $= K(x_i, x_j) = \phi(x_i)^T \phi(x_j) = (x_1^T x_2 + 2)^2$

$\omega^* = \phi(X)[0.1, 2, 3.9, 5, 6, 8]^T$

$x_{test} = [0, 0, 1]^T$

We know,

Prediction $= \omega^T x$

$\Rightarrow (\omega^*)^T x_{test}$

$\Rightarrow \left(\phi(X)[0.1, 2, 3.9, 5, 6, 8]^T\right)^T [0, 0, 1]^T$

$\Rightarrow [0.1, 2, 3.9, 5, 6, 8] \left(\phi(X)^T \phi([0,0,1])\right)$

$\Rightarrow$ Solving for $\left(\phi(X)^T \phi([0,0,1])\right)$

$\Rightarrow K(X, [0,0,1])$

$\Rightarrow [4, 4, 4, 4, 4, 4]^T$

$\Rightarrow 4[1, 1, 1, 1, 1, 1]^T$

$\Rightarrow [0.1, 2, 3.9, 5, 6, 8] \; 4[1, 1, 1, 1, 1, 1]^T$

$\Rightarrow 4(0.1 + 2 + 3.9 + 5 + 6 + 8)$

$\Rightarrow 4(25) = \underline{\underline{100}}$

3. Consider the following three models for a one-dimensional dataset:

Model 1: $y = w_1 x_1$

Model 2: $y = w_1^2 x_1$

Model 3: $y = w_1^2 x_1 + w_2 x_1$

Select all the correct options. Assume that we have access to sufficiently large data points.

Options:

(a) There may be some datasets for which model 1 performs better than model 2. ✓

(b) There may be some datasets for which model 2 performs better than model 1.

(c) There may be some datasets for which model 3 performs better than model 1.

(d) There may be some datasets for which model 3 performs better than model 2. ✓

(e) Model 1 and model 3 perform eqally well on all datasets. ✓

Model 2 is the worst because it squares w_1, which makes all the values of w_1 Positive.
Hence our predicteons (y_i) will never account for negative values.

Model 1 and Model 3 are similar. square of w_1 in Model 3 is offset by w_2. Model 1 and Model 3 both, can account for negative values.

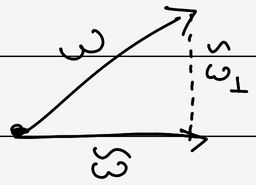
4. Let w be the solution of the linear regression model and $\tilde{w}$ be the projection of w on the linear subspace spanned by the datapoints. Which of the following relationship is true?

(a) training error for $w =$ training error for $\tilde{w}$ ✓

(b) $w = \tilde{w}$ ✓

(c) training error for $w \neq$ training error for $\tilde{w}$

We know that w can be written as,



$w = \tilde{w} + \tilde{w}_\perp$ ——— (1)

as $\tilde{w}$ is the projection of w on the linear subspace spanned by the datapoints.
Here $\tilde{w}$ is the representation of w on the subspace.
$\tilde{w}_\perp$ is the perpendicular component of the component

From (1)

$w = \tilde{w} + \tilde{w}_\perp$

$w^T x_i = (\tilde{w} + \tilde{w}_\perp)^T x_i$

$\Rightarrow \tilde{w}^T x_i + \underbrace{\tilde{w}_\perp^T x_i}$

this will become 0 as $\tilde{w}_\perp$ and x_i are orthogonal to each other.
(x_i lie in the subspace and $\tilde{w}_\perp$ is perpendicular to the subspace itself)

$\Rightarrow w^T x_i = \tilde{w}^T x_i$

From this we can conclude that error for both w and $\tilde{w}$ is exactly the same.

5. Gaussian kernel regression with parameter $\sigma^2 = \frac{1}{2}$ was applied to the following dataset with two features:

Ans = 3

$$X = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} \quad y = [2.1, 1, 2, 1.2]^T$$

The weight vector can be written as $w = \phi(X)\alpha$ where ϕ is the transformation mapping corresponding to the kernel. The vector α is given by $[2.1, -2.1, 3, 0]^T$ which is obtained as $(K)^{-1}y$, where K is the kernel matrix. What will be the prediction for point $[1, 1]^T$?

Solution

→ This solution is from W5 PA Q8. Although the answer is different, the method to solve this question is exactly the same.

The prediction is given by

$$\sum_{i=1}^{n} k(x_i, x_{\text{test}})\alpha_i \tag{1}$$

The kernel function is given by

$$\begin{aligned} k(x_i, x_j) &= \exp\left(\frac{-\|x_i - x_j\|^2}{2(\sigma^2)}\right) \\ &= \exp\left(-\|x_i - x_j\|^2\right)(\because \sigma^2 = 1/2) \end{aligned}$$

Now,

$$\begin{aligned} k(x_1, x_{\text{test}}) &= k([1,0],[1,1]) \\ &= \exp(-(0+1)) \quad = e^{-1} \\ k(x_2, x_{\text{test}}) &= k([0,1],[1,1]) \\ &= \exp(-(1+0)) \quad = e^{-1} \\ k(x_3, x_{\text{test}}) &= k([1,1],[1,1]) \\ &= \exp(-(0+0)) \quad = 1 \\ k(x_4, x_{\text{test}}) &= k([0,0],[1,1]) \\ &= \exp(-(1+1)) \quad = e^{-2} \end{aligned}$$

Putting the values in eq (1), we get

$1.4e^{-1} - 1.4e^{-1} + 1(2) + 0(e^{-2}) = 2$

6. Let X be the data matrix of shape (d, n) and y be the corresponding label vector. A linear regression model of the form $\hat{y} = w^T x_i$ is fit using the squared error on the same dataset. If the solution of w^* to the optimization problem is orthogonal to the subspace spanned by the data points (columns of matrix X), what will be the squared error?

(a) 0
(b) 1
(c) $\|y\|^2$ ✓
(d) Insufficient information to answer

If w is orthogonal to the datapoints,

$\Rightarrow w^T x_i = 0$ —— (1)

We know that error function is given as,

$$\text{Error} = \sum_{i=1}^{n} (w^T x_i - y_i)^2$$

$$\Rightarrow \sum_{i=1}^{n} (-y_i)^2 \quad [\text{From } (1)]$$

$$\Rightarrow \sum_{i=1}^{n} (-1)^2 (y_i)^2$$

$$\Rightarrow \sum_{i=1}^{n} (y_i)^2$$

$$\Rightarrow \|y\|$$

7. Let w^*, w^g and $w^s g$ be the weight vectors obtained using analytical, gradient descent,and stochastic gradient descent approaches,respectively, on the same linear regression model. The following expression holds true for these weight vectors:

$$\|w^g - w^*\| < \|w^{sg} - w^*\|$$

The model obtained by the analytical solution gives a training error of 0.5. Which of the following approaches is more likely to give less training error? Assume that the loss function is a convex function.

(a) Gradient descent ✓
(b) Stochastic gradient descent

w^g has smaller 'weight vector values' compared to w^{sg}.

8. Which of the following can NOT be linear regression model?
Options:

(a) $y = \sum_{i=0}^{m} w_i x_i$
(b) $y = \prod_{i=0}^{m} w_i x_i$ ✓ → If any pair of $w_i x_i$ becomes 0. Whole product becomes 0.
(c) $y = \sum_{i=0}^{m} w_i^{x_i} x_i$ ✓ → Its not even linear! Its exponential
(d) $y = \sum_{i=0}^{m} w_i^2 x_i$ ✓ → w is squared and doesnt span negative subspace.

9. A linear regression model trained on a dataset $X \in \mathbb{R}^{d \times n}$ achieves zero training error for any label vector y. Which of the following options will necessarily hold true? Here I denotes identity matrix of an appropriate size.

(a) $XX^T = I$
(b) $X^T(XX^T)^{-1}X = I$
(c) $(XX^T)^{-1}Xy$ is a vector of all ones.
(d) $(XX^T)^{-1}Xy$ is a vector of all zeros.

Error function is given as,

$$\|X^T w - y\| = \text{error}$$

if error $= 0$,

$\Rightarrow \|X^T w - y\| = 0$

$\Rightarrow (X^T w - y)^T (X^T w - y) = 0$

$\Rightarrow X^T w - y = 0$

We also know that,

$$w = (XX^T)^{-1} Xy$$

Putting w in above equation

$$\Rightarrow \quad X^T(\underbrace{(XX^T)^{-1}Xy}_{\omega}) - y = 0$$

$$\Rightarrow \quad X^T(XX^T)^{-1}Xy = y$$

Dividing both sides by y

$$\Rightarrow \quad X^T(XX^T)^{-1}X = I$$

10. Is the following statement true or false?

    Gradient descent takes more number of iterations to converge to local minima than stochastic gradient descent.

    (a) True

    (b) False ✓

It takes less iterations.

W6

1. Consider a linear regression model that was trained on dataset X of shape (d, n). Which of the following techniques could potentially decrease the loss on the training data (assuming the loss is the square error) ?

   Options :

   (a) Adding a dummy feature in the dataset and learning the intercept W_0 as well. ✓

   (b) Penalizing the model weights with L2 regularization.

   (c) Penalizing the model weights with L1 regularization.

   (d) Training the kernel regression model of degree 2. ✓

b) and c) options refer to ridge and lasso regression respectively.
We know that both of them have more error compared to MSE, because of $\lambda\|w\|$.

2. Which of the following statements is/are true regarding solution of Ridge regression problem ?

   Options :

   (a) If there are multiple w solutions for minimizing mean square error, then w_R will be the one with least norm. ✓

   (b) If there are multiple w solutions for minimizing mean square error, then w_R will be the one with highest norm.

   (c) Prior for w is $N(0, \gamma^2 I)$ and $y_i|x_i \sim N(w^T x_i, \sigma^2)$ ✓

   (d) Prior for w is $N(1, \gamma^2 I)$ and $y_i|x_i \sim N(0, \sigma^2)$

Ridge regression equation is,

$$\hat{\omega}_R = \arg\min \sum_{i=1}^{n} (\omega^T x_i - y_i) + \lambda\|\omega\|$$

The higher the 'weight features' of ω, the higher the norm.
Higher norm gives more error.

3. Consider a ridge regression model with the loss $L(w) = \|X^T w - y\|^2 + \lambda\|w\|^2$ is trained on a given dataset with $\lambda = 0.1, 0, 1, 10, 100$. Which of the value of λ is more likely to underfit the model?

If we choose a higher value of 'λ' we will have to decrease the 'weight features' of 'ω' to maintain the same error, as the ridge regression model penalises high value of 'ω' or 'λ'.

If the 'weight features' of ω are reduced, the model tends to underfit.

$\Rightarrow \lambda = 100$ will underfit the model most.

4. Consider that the three weight vectors w_1, w_2 and w_3 are learned for a six-dimensional dataset using a linear regression regularized linear regression model (Not in any particular order).

$w_1 = [0.5, 0, 0.25, 0, 0, -0.14]$

$w_2 = [0.8, -0.23, 0.45, 0.2, 0.31, -0.54]$

$w_3 = [0.24, -0.03, 0.1, 0.02, 0.09, -0.14]$

Select the most appropriate match for these weight vectors.

(a) $w_1 \rightarrow$ Linear regression, $w_2 \rightarrow$ Ridge regression, $w_3 \rightarrow$ Lasso
(b) $w_1 \rightarrow$ Ridge regression, $w_2 \rightarrow$ Linear regression, $w_3 \rightarrow$ Lasso
(c) $w_1 \rightarrow$ Lasso, $w_2 \rightarrow$ Ridge regression, $w_3 \rightarrow$ Linear regression
(d) $w_1 \rightarrow$ Lasso, $w_2 \rightarrow$ Linear regression, $w_3 \rightarrow$ Ridge regression ✓

ω_1 has the most number of features zero, when compared to ω_2 and ω_3.
$\Rightarrow \omega_1$ is the output of LASSO Regression.

ω_3 has lower 'weight features' when compared to ω_2.
$\Rightarrow \omega_3$ is the output of Ridge Regression.

5. Consider the following dataset:
$X = [-3, 5, 4]$
$y = [-10, 20, 20]$

Assuming a ridge penalty $\lambda = 50$, what will be the value of $\frac{\tilde{w}_{ridge}}{\tilde{w}_{MLE}}$?
Here $\tilde{w}_{ridge}$ and $\tilde{w}_{MLE}$ are the Ridge and MLE estimates of the weight vectors, respectively.

Options:

(a) 2
(b) 1
(c) 0.666
(d) 0.5 ✓
(e) 0.25

$X = [-3, 5, 4]$
$y = [-10, 20, 20]$
$\lambda = 50$

$\omega_{Ridge} = (XX^T + \lambda I)^{-1} Xy$

$\Rightarrow XX^T = [-3, 5, 4][-3, 5, 4]^T = 9 + 25 + 16 = 50$
$\lambda I = 50 \times 1 = 50$
$Xy = [-3, 5, 4][-10, 20, 20]^T = 30 + 100 + 80 = 210$

$\Rightarrow (50 + 50)^{-1} \, 210$
$\Rightarrow \frac{210}{100} = \frac{21}{10}$

$\omega_{MLE} = (XX^T)^{-1} Xy$
$\Rightarrow (50)^{-1} \, 210$
$\Rightarrow \frac{210}{50} = \frac{21}{5}$

$\Rightarrow \omega_{Ridge} / \omega_{MLE} = \frac{21/10}{21/5} = \frac{5}{10} = \underline{\underline{0.5}}$

Direct formula for this question,

$$\frac{w_{Ridge}}{w_{MLE}} = \frac{(XX^T + \lambda I)^{-1} Xy}{(XX^T)^{-1} Xy}$$

$$\Rightarrow \frac{XX^T}{XX^T + \lambda I}$$

6. For a data set with 5000 data points and 500 features, I divide my dataset into training and testing part where I took 25% of my data as test data and rest as training data, I started training model on training data, how many models will be trained during Leave-One-Out cross-validation?
   (a) 500
   (b) 5000
   (c) 1250
   (d) 3750 ✓

Training data = datapoints − test datapoints

$\Rightarrow 5000 - 5000 \times 25\%$

$\Rightarrow \underline{3750}$

- Leave-One-Out Cross Validation: The model is trained using all but one sample in the training set, and the left-out sample is used for validation. This process is repeated for each sample in the dataset. The optimal λ is determined based on the average error across all iterations.

7. In Ridge regression, as the regularization parameter increases, do the regression coefficients decrease?
   (a) True ✓
   (b) False

8. **Statement 1:** The cost function is altered by adding a penalty equivalent to the square of the magnitude of the coefficients
   Statement 2: Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent overfitting which may result from simple linear regression.
   (a) Statement 1 is true and statement 2 is false
   (b) Statement 1 is False and statement 2 is true
   (c) Both statement (1 and 2) is true ✓
   (d) Both Statement (1 and 2) is wrong

Statement 1 :—
$\lambda\|w\|$ is what we use to add penalty to the cost function of simple linear regression.

Statement 2: The penalty term reduces 'weight features' of w to prevent overfitting.

9. Which of the following cross validation versions may not be suitable for very large datasets with hundreds of thousands of samples?
   (a) k-fold cross-validation
   (b) Leave-one-out cross-validation ✓

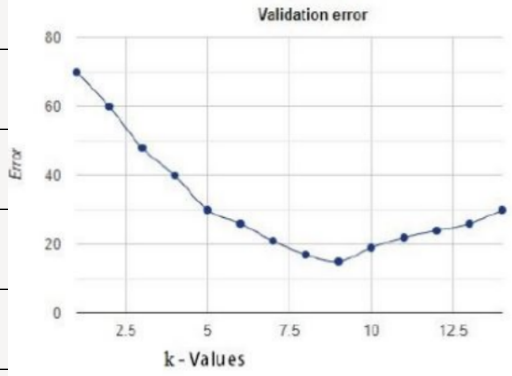
In leave-one-out cross validation, if our training dataset is large, then the number of iterations is also going to be large.

10. When most of the features are redundant, then
   (a) Ridge regression will choose those features that have the highest weight.
   (b) Ridge regression will choose those features that have the least weight. ✓

Redundant features doesn't play any role, Ridge Regression always chooses the least weights.

W 7

1. What would be the best value for k to be used in KNN algorithm based on the graph given below?



Options :

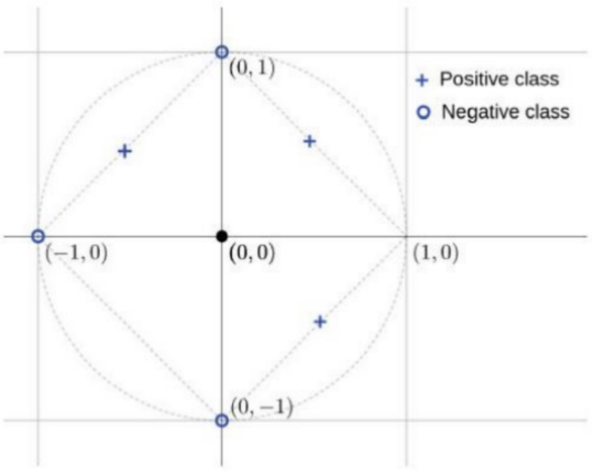
(a) 4
(b) 5
(c) 9 ✓
(d) 12

Validation error is least at 9.

2. Consider a KNN classifier for a binary classification problem with k = 3. A test-point at (0, 0) has to be classified using this model. The training-dataset is as follows:

- The three negative class data-points are fixed and occupy three vertices of the diamond.
- The three positive class data-points can be anywhere on the edges of the diamond except the vertices

Based on the above data, answer the given sub-questions.



(i) If the Manhattan distance metric is used, what is the predicted label of the test-point?
   (a) Positive class
   (b) Negative class
   (c) It could be either of the two classes. An exact decision requires information about how to break ties. ✓

(ii) If the Euclidean distance metric is used, what is the predicted label of the test-point?
   (a) Positive class ✓
   (b) Negative class
   (c) It could be either of the two classes. An exact decision requires information about how to break ties.

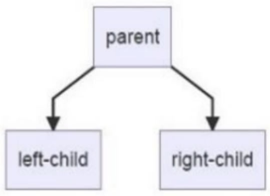
When using manhattan distance, the distance to all the points remains the same.

$|x_1 - x_2| + |y_1 - y_2|$ [Manhattan formula]

When using euclidean distance, the points on edges (+ve points) will be closest to (0,0).

$\sqrt{(x_1-x_2)^2 + (y_2-y_1)^2}$ [Euclidean formula]

3. A decision stump is a decision tree that has exactly one question at the parent node (root) which then splits into two prediction nodes (leaves):



Consider a decision stump for a binary classification problem that has 500 data points at the parent node, out of which 200 data points go into the left child. The number of data points that belong to class 1 in the parent node is 300. The number of data points that belong to class 1 in the left child is 50. The labels are in $\{1, 0\}$.

Note for calculations: Use log_2 for all calculations that involve logarithms. For all questions, enter your answer correct to three decimal places. Use three decimal places even while calculating intermediate quantities.

(i) What is the label assigned to the left child? Enter 1 or 0 . 0

(ii) What is the entropy of the parent?

(iii) What is the entropy of the left child?

(iv) What is the entropy of the right child?

(v) What is the information gain corresponding to the question at the parent node?

i) 200 datapoints went to left child.
50 of those datapoints belong to class 1.
$\Rightarrow$ 150 datapoints belong to class 0.

ii) $p = \frac{300}{500} = \frac{3}{5}$

$\Rightarrow \text{Entropy}(3/5) = -\left(\frac{3}{5}\log_2\left(\frac{3}{5}\right) + \frac{2}{5}\log_2\left(\frac{2}{5}\right)\right)$

$\Rightarrow -(0.6 \times (-0.73) + 0.4 \times (-1.32))$

$\Rightarrow \underline{\underline{0.966}}$

iii) $p = \frac{50}{200} = \frac{1}{4}$

$\Rightarrow \text{Entropy}\left(\frac{1}{4}\right) = -\left(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2\left(\frac{3}{4}\right)\right)$

$\Rightarrow -(0.25 \times (-2) + 0.75(-0.41))$

$\Rightarrow \underline{\underline{0.807}}$

iv) $p = \frac{25\cancel{0}}{30\cancel{0}} = \frac{5}{6}$

$\Rightarrow \text{Entropy}(5/6) = -\left(\frac{5}{6}\log_2\left(\frac{5}{6}\right) + \frac{1}{6}\log_2\left(\frac{1}{6}\right)\right)$

$\Rightarrow -(0.83 \times (-0.263) + 0.166 \times (-2.584))$

$\Rightarrow \underline{\underline{0.647}}$

v) Information gain $= \text{Entropy}(D) - [\gamma\,\text{Entropy}(D_{yes}) + (1-\gamma)\,\text{Entropy}(D_{no})]$

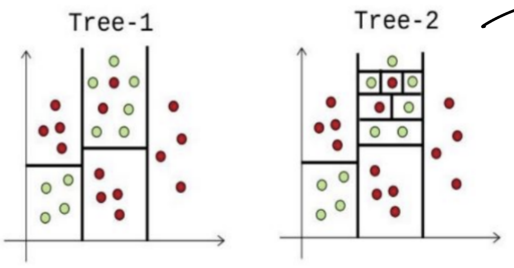
$\Rightarrow 0.966 - \left[\frac{3}{5} \times 0.647 + \frac{2}{5} \times 0.807\right]$

$\Rightarrow 0.966 - [0.6 \times 0.647 + 0.4 \times 0.807]$

$\Rightarrow \underline{\underline{0.255}}$

4. Consider a training dataset for a binary classification problem in R^2. Two decision trees are trained on the same dataset. The decision regions obtained are plotted for both the trees:

This tree is overfitted



Which of these two trees is likely to perform better on test data.

(a) Tree 1 ✓

(b) Tree 2

5. Suppose we have a binary classification dataset with 1000 data points, consisting of 600 points belonging to class 0 and 400 points belonging to class 1. If we use a k-nearest neighbor model with k=900 to predict the class labels of the data points,how many data points will be classified correctly? 600

600 belong to class 0

400 belong to class 1

$$\text{Majority}(\underbrace{500}_{\text{class 1}}, \underbrace{400}_{\text{class 0}}) = \text{label of 500 datapoints} = 0$$

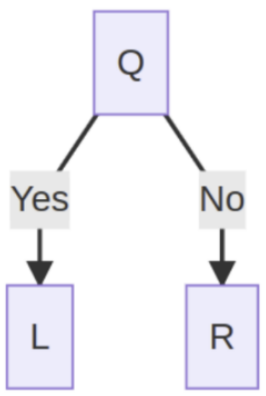
⇒ Even if all class 1 datapoints are counted as nearest neighbours, the answer will still be class 0.

⇒ Our KNN will always predict test point as 0.

In our training dataset, there are 600 class 0 datapoints.

⇒ 600 of them will be predicted correctly.

6. The dataset at node Q has 1000 points. After asking the question Q it splits into two parts, 250 in node L and 750 in node R.



$$\gamma = \frac{250}{1000} = \frac{1}{4} = \underline{\underline{0.25}}$$

What is the value of γ? γ is the cardinality of L divided by cardinality of Q?

7. Consider the following statements:

S1: The impurity of a node is influenced by its ancestors.

S2: The impurity of a node is influenced by its descendants.

   (a) Only S1 is true ✓
   (b) Only S2 is true
   (c) Both S1 and S2 are true
   (d) Neither S1 nor S2 are true

8. Which of the following statements about Decision Trees is true?

   (a) They are only applicable to classification tasks.
   (b) They can handle both categorical and numerical features. ✓
   (c) They are less prone to overfitting compared to other models.
   (d) They are robust to outliers in the training data.