**Introduction to Machine Learning -IITKGP**

**Assignment - 2**

**TYPE OF QUESTION: MCQ/MSQ**

Number of questions: 15 Total mark: 2 * 15 = 30

Data for Q. 1 to 3

The following dataset will be used to learn a decision tree for predicting whether a person is happy (H) or sad (S), based on the color of their shoes, whether they wear a wig, and the number of ears they have.

| Color | Wig | Num. Ears | Emotion (Output) |
|---|---|---|---|
| G | Y | 2 | S |
| G | N | 2 | S |
| G | N | 2 | S |
| B | N | 2 | S |
| B | N | 2 | H |
| R | N | 2 | H |
| R | N | 2 | H |
| R | N | 2 | H |
| R | Y | 3 | H |

Based on the dataset answer the following questions:

1. What is $Entropy\ (Emotion|Wig\ =\ Y)$?
   a. 1
   **b.** 0
   **c.** 0.50
   **d.** 0.20

**Correct answer:** a

**Explanation:**

To calculate the entropy of the target variable (*Emotion*) given the condition *Wig = Y*, we need to compute the distribution of emotions within that subset of the dataset.

Subset of the dataset where *Wig = Y*:

| Color | Wig | Num. Ears | Emotion (Output) |
|---|---|---|---|
| G | Y | 2 | S |
| R | Y | 3 | H |

Within this subset, we have 1 instance of "S" (sad) and 1 instance of "H" (happy). Therefore, the distribution of emotions is equal, with a count of 1 for each class.

To calculate the entropy, we can use the formula: Entropy(X) = - Σ P(x) $\log_2$ P(x)

*Entropy (Emotion | Wig = Y)* = - P(S) $\log_2$ P(S) - P(H) $\log_2$ P(H)

Since P(S) = P(H) = 0.5 (both classes have equal counts), we can substitute these values into the entropy formula:

*Entropy (Emotion | Wig = Y)* = - (0.5) $\log_2$ (0.5) - (0.5) $\log_2$ (0.5)

$$= - (0.5) (-1) - (0.5) (-1) = 1$$

Therefore, $Entropy\ (Emotion|Wig\ =\ Y)$ = 1

---

2. What is $Entropy\ (Emotion|Ears\ =\ 3)$?
   a. 1
   b. 0
   c. 0.50
   d. 0.20

**Correct answer: b**

**Explanation:**

To calculate the entropy of the target variable (Emotion) given the condition *Ears = 3*, we need to compute the distribution of emotions within that subset of the dataset.

Subset of the dataset where *Ears = 3*:

| Color | Wig | Num. Ears | Emotion (Output) |
|---|---|---|---|
| R | Y | 3 | H |

Within this subset, we have 1 instance of "H" (happy) and 0 instances of "S" (sad).

To calculate the entropy, we can use the formula:
Entropy(X) = - Σ P(x) $\log_2$ P(x)

*Entropy (Emotion | Ears=3)* = - P(S) $\log_2$ P(S) - P(H) $\log_2$ P(H)

Since P(S) = 0 and P(H) = 1 (since there are no instances of "S" and 1 instance of "H"), we can substitute these values into the entropy formula:

*Entropy (Emotion | Ears=3)* = - 0 $\log_2$ 0 - 1 $\log_2$ 1 = 0 - 0 = 0

Therefore, $Entropy\ (Emotion|Ears\ =\ 3) = 0$.

---

3. Which attribute should you choose as root of the decision tree?
   a. Color
   b. Wig
   c. Number of ears
   d. Any one of the previous three attributes

**Correct answer: a**

**Explanation:**

To determine the attribute to choose as the root of the decision tree, we need to consider the concept of information gain. Information gain measures the reduction in entropy or impurity achieved by splitting the data based on a specific attribute.

We can calculate the information gain for each attribute by comparing the entropy before and after the split. The attribute with the highest information gain will be chosen as the root of the decision tree.

Let's calculate the information gain for each attribute (Color, Wig, and Num. Ears) based on the given dataset:

Information Gain (Color):

To calculate the information gain for the Color attribute, we need to compute the entropy of the Emotion variable before and after the split based on different colors.
*Entropy (Emotion)* = - (4/9) $\log_2$ (4/9) - (5/9) $\log_2$ (5/9) ≈ 0.991

After the split based on Color, we have the following subsets:

Subset for Color = Green:
*Entropy (Emotion | Color = Green)* = 0 (as all instances are of the same class, "S")

Subset for Color = Blue:
*Entropy (Emotion | Color = Blue)* = -1/2 $\log_2$ (1/2) - 1/2 $\log_2$(1/2) = 1

Subset for Color = Red:
*Entropy (Emotion | Color = Red)* = 0 (as all instances are of the same class, "H")

*Information Gain (Color) = Entropy (Emotion*) - [ (3/9) * 0 + (2/9) * 1 + (4/9) * 0] ≈ 0.7687

Information Gain (Wig):

To calculate the information gain for the Wig attribute, we need to compute the entropy of the Emotion variable before and after the split based on different values of Wig.
*Entropy (Emotion)* = - (4/9) $\log_2$ (4/9) - (5/9) $\log_2$ (5/9) ≈ 0.991

After the split based on Wig, we have the following subsets:

Subset for Wig = Yes:
*Entropy (Emotion | Wig = Yes)* = -1/2 $\log_2$ (1/2) - 1/2 $\log_2$ (1/2) =1

Subset for Wig = No:
*Entropy (Emotion | Wig = No)* = - (4/7) $\log_2$ (4/7) - (3/7) $\log_2$ (3/7) ≈ 0.985

*Information Gain (Wig) = Entropy (Emotion)* - [ (2/9) * 1 + (7/9) * 0.985] ≈ 0.002

Information Gain (Num. Ears):

To calculate the information gain for the Num. Ears attribute, we need to compute the entropy of the Emotion variable before and after the split based on different values of Num. Ears.
*Entropy (Emotion)* = - (4/9) $\log_2$ (4/9) - (5/9) $\log_2$ (5/9) ≈ 0.991

After the split based on Num. Ears, we have the following subsets:

Subset for Num. Ears = 2:
*Entropy (Emotion | Num. Ears = 2)* = - (4/8) $\log_2$ (4/8) - (4/8) $\log_2$ (4/8) ≈ 1

Subset for Num. Ears = 3:
*Entropy (Emotion | Num. Ears = 3)* = 0 (as all instances are of the same class, "H")

*Information Gain (Num. Ears) = Entropy (Emotion)* - [ (8/9) * 1 + (1/9) * 0] ≈ 0.102

Based on the information gain calculations, the attribute with the highest information gain is Color, with an information gain of approximately 0.768. Therefore, Color should be chosen as the root of the decision tree.

---

4. In linear regression, the output is:
   a. Discrete
   b. Categorical
   c. Continuous
   d. May be discrete or continuous

**Correct answer: c**

**Explanation:**
In linear regression, the output variable, also known as the dependent variable or target variable, is continuous. Linear regression is a supervised learning algorithm used to model the relationship between a dependent variable and one or more independent variables.

The goal of linear regression is to find a linear relationship between the independent variables and the continuous output variable. The linear regression model predicts a continuous value as the output based on the input features**.**

---

5. Consider applying linear regression with the hypothesis as $h_\theta(x) = \theta_0 + \theta_1 x$**.**The training data is given in the table.

| **X** | **Y** |
|---|---|
| 6 | 7 |
| 5 | 4 |
| 10 | 9 |
| 3 | 4 |

We define Mean Square Error (MSE), $J_\theta$ =

$$\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

where m is the number of training examples. $h_\theta(x_i)$ is the value of linear regression hypothesis at point, i. If $\theta = [1, 1]$, find $J_\theta$.

a. 0
b. 1
c. 2
d. 0.5

**Correct answer: b**

**Explanation:**

Let's calculate the value of $J_\theta$:

$$\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

We have four training examples, so m = 4.

For θ = [1, 1], the hypothesis function $h_\theta(x)$ becomes $h_\theta(x) = 1 + 1x$.

Substituting the values from the training data into the MSE equation:

$J_\theta$ = 1/(2*4) [ (1 + 1(6) - 7) $^2$ + (1 + 1(5) - 4)$^2$ + (1 + 1(10) - 9)$^2$ + (1 + 1(3) - 4)$^2$ ]

= 1/(8) [ (0)$^2$ + (2)$^2$ + (2)$^2$ + (0)$^2$ ]

= 1/(8) [ 0 + 4 + 4 + 0 ]

= 1/(8) [ 8 ]

= 1

Therefore, the value of $J_\theta$ for θ = [1, 1] is 1.

6. Specify whether the following statement is true or false?

*"The ID3 algorithm is guaranteed to find the optimal decision tree"*

a. True
b. False

**Correct answer**: b

**Explanation:**
The ID3 algorithm uses a greedy strategy to make local decisions at each node based on the information gain or other impurity measures. It recursively builds the decision tree by selecting the attribute that provides the highest information gain or the most significant reduction in impurity at each step. However, this greedy approach does not consider the global optimum for the entire decision tree structure.

Due to the greedy nature of the algorithm, it is possible for ID3 to get stuck in suboptimal solutions or make decisions that do not result in the most accurate or optimal tree. In some cases, the ID3 algorithm may produce a decision tree that is a local optimum but not the global optimum.

---

7**.** Identify whether the following statement is true or false?

*"A classifier trained on less training data is less likely to overfit"*

a. True
b. False

**Correct answer:** b

**Explanation:**
In reality, a classifier trained on less training data is more likely to overfit. Overfitting occurs when a model learns the training data too well, capturing noise or irrelevant patterns that do not generalize to unseen data. When the training dataset is smaller, the model has less exposure to the variety of examples and may struggle to capture the true underlying patterns.

With a limited amount of training data, the model has a higher risk of memorizing specific examples and idiosyncrasies of the training set, resulting in a biased and overfitted model. The lack of diversity in the training data hampers the model's ability to generalize well to new, unseen examples.

To mitigate overfitting, it is generally recommended to have a sufficient amount of diverse training data that represents the underlying distribution of the problem. More data allows the model to learn more robust and generalizable patterns, reducing the likelihood of overfitting.

---

8. Identify whether the following statement is true or false?

"*Overfitting is more likely when the hypothesis space is small*"

a. True
b. False

**Correct answer: b**

**Explanation**: We can see this from the bias-variance trade-off. When hypothesis space is small, it's more biased with less variance. So with a small hypothesis space, it's less likely to find a hypothesis to fit the data very well,i.e., overfit.

---

9. Traditionally, when we have a real-valued input attribute during decision-tree learning, we consider a binary split according to whether the attribute is above or below some threshold. One of your friends suggests that instead we should just have a multiway split with one branch for each of the distinct values of the attribute. From the list below choose the single biggest problem with your friend's suggestion:

a. It is too computationally expensive
b. It would probably result in a decision tree that scores badly on the training set and a test set
c. It would probably result in a decision tree that scores well on the training set but badly on a test set
d. would probably result in a decision tree that scores well on a test set but badly on a training set

**Correct answer:** c

**Explanation**: The single biggest problem with the suggestion of using a multiway split with one branch for each distinct value of a real-valued input attribute is that it would likely result in a decision tree that overfits the training data. By creating a branch for each distinct value, the tree would become more complex, and it would have the potential to fit the training data too closely, capturing noise or irrelevant patterns specific to the training set.

As a consequence of overfitting, the decision tree would likely score well on the training set since it can perfectly match the training examples. However, when evaluated on a test set or unseen data, the tree would struggle to generalize and perform poorly. Overfitting leads to poor performance on new instances, indicating that the model has failed to learn the underlying patterns and instead has become too specialized in the training data.

---

10**.** Which of the following statements about decision trees is/are true?

a. Decision trees can handle both categorical and numerical data.
b. Decision trees are resistant to overfitting.
c. Decision trees are not interpretable.
d. Decision trees are only suitable for binary classification problems.

**Correct answer**: a

**Explanation**: Decision trees can handle both categorical and numerical data as they partition the data based on various conditions during the tree construction process. This allows decision trees to be versatile in handling different types of data.

---

11. Which of the following techniques can be used to handle overfitting in decision trees?
    a. Pruning
    b. Increasing the tree depth
    c. Decreasing the minimum number of samples required to split a node
    d. Adding more features to the dataset

**Correct answers**: a, c

**Explanation**: Overfitting occurs when a decision tree captures noise or irrelevant patterns in the training data, resulting in poor generalization to unseen data. Pruning is a technique used to reduce overfitting by removing unnecessary branches and nodes from the tree.

Decreasing the minimum number of samples required to split a node can also help prevent overfitting by allowing more flexible splits.

---

12. Which of the following is a measure used for selecting the best split in decision trees?
    a. Gini Index
    b. Support Vector Machine
    c. K-Means Clustering
    d. Naive Bayes

**Correct answer**: a

**Explanation**: The Gini Index is a commonly used measure for selecting the best split in decision trees. It quantifies the impurity or dissimilarity of a node's class distribution. The split that minimizes the Gini Index is chosen as the optimal split.

---

13. What is the purpose of the decision tree's root node in machine learning?
    a. It represents the class labels of the training data.
    b. It serves as the starting point for tree traversal during prediction.
    c. It contains the feature values of the training data.
    d. It determines the stopping criterion for tree construction.

**Correct answer**: b

**Explanation:** The root node of a decision tree serves as the starting point for tree traversal during prediction. It represents the first decision based on a feature and directs the flow of the decision tree based on the outcome of that decision. The root node does not contain class labels or feature values but rather determines the initial split based on a selected criterion.

---

14. Which of the following statements about linear regression is true?

a. Linear regression is a supervised learning algorithm used for both regression and classification tasks.
b. Linear regression assumes a linear relationship between the independent and dependent variables.
c. Linear regression is not affected by outliers in the data.
d. Linear regression can handle missing values in the dataset.

**Correct answer**: b

**Explanation**: Linear regression assumes a linear relationship between the independent variables (features) and the dependent variable (target). It seeks to find the best-fitting line to the data.

While linear regression is primarily used for regression tasks, it is not suitable for classification tasks. Outliers can significantly impact linear regression models, and missing values in the dataset require appropriate handling.

---

15. Which of the following techniques can be used to mitigate overfitting in machine learning?

a. Regularization
b. Increasing the model complexity
c. Gathering more training data
d. Feature selection or dimensionality reduction

**Correct answers**: a, c, d

**Explanation**: Regularization techniques, such as L1 or L2 regularization, can help mitigate overfitting by adding a penalty term to the model's objective function, discouraging excessively large parameter values.

Gathering more training data can also reduce overfitting by providing a more representative sample of the underlying data distribution.

Feature selection or dimensionality reduction techniques, such as selecting relevant features or applying techniques like Principal Component Analysis (PCA), can help remove irrelevant or redundant features, reducing the complexity of the model and mitigating overfitting.

---