CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2024)
Quiz - 2 Set B

Date of Examination: 24.09.2024 Duration: 45 mins Total Marks: 15 marks

---

Instructions –

- Attempt all questions.
- MCQs have a single correct option.
- State any assumptions you have made clearly.
- Standard institute plagiarism policy holds.
- No evaluation without suitable justification.
- 0 marks if the option or justification of MCQs is incorrect.

---

1. Consider the Perceptron algorithm applied to a binary classification task. Which of the following statements are correct? (Select all that apply) (1 mark)

   (A) The Perceptron algorithm is guaranteed to converge for any dataset, as long as the learning rate is appropriately chosen.

   (B) The weights are updated in the Perceptron algorithm based on the dot product of the input vector and the error term, even when the sample is correctly classified.

   (C) The bias in the Perceptron is updated when a sample is misclassified, but the magnitude of the update is not dependent on the input values.

   (D) If the Perceptron converges, it will find a decision boundary that minimizes the number of misclassifications on the training data.

   A. False

   Reason: The Perceptron algorithm is only guaranteed to converge if the data is linearly separable. The algorithm does not require a learning rate, and convergence depends solely on the separability of the data, not on a learning rate.

   B. False

   Reason: The weights are updated only when there is a misclassification. If the sample is classified correctly, no update to the weights occurs. The update rule is:

   $$\mathbf{W} = \mathbf{w} + \text{error}_i \cdot x_i$$

   C. True

   Reason: The bias is updated when a sample is misclassified, using the rule:

   $$B = b + \text{error}_i$$

   The magnitude of the bias update is independent of the input values, and it only depends on the error term.

   D. True

   Reason: If the Perceptron converges (for linearly separable data), it finds a decision boundary that minimizes the number of misclassifications, reducing them to zero on the training data. However, it does not necessarily produce the most optimal decision boundary, as there may be multiple valid solutions.
   1 mark for correct option and correct reason

2. Which of the following statements regarding Random Forests and ensemble methods are TRUE? (1 mark)

   (A) Random Forests can handle both categorical and continuous variables, allowing for greater flexibility in modeling various types of data.

   (B) In a Random Forest, each tree is built in a sequential manner, where each tree depends on the results of the previous tree.

(C) The primary advantage of using Random Forests over a single Decision Tree is the significant increase in bias while keeping variance the same.

(D) Random Forests utilize an averaging scheme for classification tasks, where the final prediction is the average of the predictions from all individual trees.

A. True: Random Forests can handle both categorical and continuous variables, making them versatile for different types of datasets.

B. False: In a Random Forest, each tree is built independently and in parallel; there is no dependency between the trees, which distinguishes them from boosting methods, where models are built sequentially.

C. False: The primary advantage of using Random Forests is a reduction in variance without significantly increasing bias. They typically improve generalization compared to a single Decision Tree.

D. False: For classification tasks, Random Forests utilize a majority voting scheme, where the final prediction is based on the mode of predictions from all individual trees, not an average.
1 mark for correct option and correct reason

3. When growing a decision tree using the ID3 algorithm, which of the following is not true about the role of information gain? (1 mark)

   (a) Information gain measures how well a given attribute separates the training examples according to their target classification.

   (b) The attribute with the highest information gain is always selected at each step while growing the tree.

   (c) Information gain is based on the reduction in entropy after the data is split on an attribute.

   (d) Information gain ensures that the tree will never overfit the training data.

   Solution: (d) while the ID3 algorithm uses information gain to select the best attribute at each step, it does not guarantee that the tree will never overfit the training data. Overfitting can still occur if the tree becomes too complex, fitting noise or specific details in the training data rather than generalizing to unseen examples.
   1 mark for correct option and correct reason

4. Given the Perceptron Convergence Theorem, what can you say about the margin of a classifier and how it affects the convergence of the Perceptron algorithm? Which of the following statements is true? (1 mark)

   (a) A small margin is more desirable because it leads to faster convergence of the Perceptron algorithm.

   (b) A large margin is more desirable because it leads to faster convergence of the Perceptron algorithm.

   (c) A small margin is more desirable because it leads to more updates, improving accuracy.

   (d) The margin of the classifier does not affect the convergence of the Perceptron algorithm.

   a) False: A small margin means the classifier is less confident, leading to more updates. It does not necessarily result in faster convergence.

   b) True: A large margin between the classes leads to fewer mistakes and quicker convergence of the Perceptron algorithm. The algorithm requires fewer updates when the margin is large, as the data points are further from the decision boundary. A small margin would result in more updates and slower convergence.

   c) False: More updates due to a small margin do not necessarily improve accuracy and can lead to overfitting.

   d) False: The margin directly affects convergence. The Perceptron Convergence Theorem states that the number of updates depends on the margin size.
   1 mark for correct option and correct reason

5. Suppose that $X_1, ..., X_m$ are categorical input attributes and Y is the categorical output attribute. Suppose we plan to learn a decision tree without pruning using the standard algorithm. Which of the following is true? (1 marks)

   (a) If X_i and Y are independent in the distribution that generated this dataset, then X_i will not appear in the decision tree.

   (b) If $G(Y|X_i) = 0$ according to the values of entropy and conditional entropy computed from the data, then X_i will not appear in the decision tree.

   (c) The maximum depth of the decision tree must be less than $m + 1$.

(d) Suppose the data has R records. The maximum depth of the decision tree must be less than $1 + \log_2 R$.

A: False (because the attribute may become relevant further down the tree when the records are restricted to some value of another attribute) (e.g. XOR)

B: False for same reason

C: True because the attributes are categorical and can each be split only once

D: False because the tree may be unbalanced

1 mark for correct option and correct reason

6. Consider the following binary classification problem where the class label $Y \in \{0, 1\}$, and each training example X consists of two binary attributes $X_1, X_2 \in \{0, 1\}$. In this problem, assume that the attributes X_1 and X_2 are conditionally independent given Y. Furthermore, the class priors are given by $P(Y = 0) = P(Y = 1) = 0.5$. When $Y = 0$, there's a 70% chance that $X_1 = 0$ and a 30% chance that $X_1 = 1$. Similarly, if $Y = 1$, the probability shifts, giving $X_1 = 0$ only 20% of the time, while $X_1 = 1$ occurs 80% of the time. For X_2, when $Y = 0$, there's a 90% chance that $X_2 = 0$, but only a 10% chance for $X_2 = 1$. When $Y = 1$, both values of X_2 are equally probable.

   The expected error rate is the probability that a classifier provides an incorrect prediction for an observation: if Y is the true label, let $\hat{Y}(X_1, X_2)$ be the predicted class label, then the expected error rate is:

$$P_D(Y = 1 - \hat{Y}(X_1, X_2)) = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P(X_1, X_2, Y = 1 - \hat{Y}(X_1, X_2))$$

   (a) Write down the Naïve Bayes prediction for all the 4 possible configurations of X_1, X_2. (3 marks)

   (b) Compute the expected error rate of this Naïve Bayes classifier, which predicts Y given both of the attributes X_1, X_2. Assume that the classifier is learned with infinite training data. (2 marks)

Rubrics:

a) 2 marks for computation of $P(X_1, X_2, Y = 0)$ and $P(X_1, X_2, Y = 1)$ and 1 for calculating $\hat{Y}(X_1, X_2)$.

b) 1 mark for identifying incorrect predictions and 1 for the final answer.

$P(X_1\|Y)$	$X_1 = 0$	$X_1 = 1$
$Y = 0$	0.7	0.3
$Y = 1$	0.2	0.8

$P(X_2\|Y)$	$X_2 = 0$	$X_2 = 1$
$Y = 0$	0.9	0.1
$Y = 1$	0.5	0.5

Solution:

$$P(X_1, X_2, Y = 0) = P(X_1, X_2 \mid Y = 0) \times P(Y = 0)$$

(Conditional independence)

$$P(X_1, X_2, Y = 0) = P(X_1 \mid Y = 0) \times P(X_2 \mid Y = 0) \times P(Y = 0)$$

$$P(X_1, X_2, Y = 1) = P(X_1 \mid Y = 1) \times P(X_2 \mid Y = 1) \times P(Y = 1)$$

X_1	X_2	$P(X_1, X_2, Y = 0)$	$P(X_1, X_2, Y = 1)$
0	0	$0.7 \times 0.9 \times 0.5 = 0.315$	$0.2 \times 0.5 \times 0.5 = 0.05$
0	1	$0.7 \times 0.1 \times 0.5 = 0.035$	$0.2 \times 0.5 \times 0.5 = 0.05$
1	0	$0.3 \times 0.9 \times 0.5 = 0.135$	$0.8 \times 0.5 \times 0.5 = 0.2$
1	1	$0.3 \times 0.1 \times 0.5 = 0.015$	$0.8 \times 0.5 \times 0.5 = 0.2$

Prediction: $\hat{Y}(X_1, X_2)$

(a) $0.315 > 0.05$, so $\hat{Y}(X_1, X_2) = 0$

(b) $0.035 < 0.05$, so $\hat{Y}(X_1, X_2) = 1$

(c) $0.135 < 0.2$, so $\hat{Y}(X_1, X_2) = 1$

(d) $0.015 < 0.2$, so $\hat{Y}(X_1, X_2) = 1$

2. **Expected error rate** $\rightarrow P_D(Y) = \sum_{X_1=0}^{1} \sum_{X_2=0}^{1} P_D(X_1, X_2, Y)$

where $Y = 1 - \hat{Y}(X_1, X_2)$

To compute error rate we need to sum over all possible values of X_1 and X_2, and for each of them we compute the probability that Y is opposite of the predicted label.

For $X_1 = 0$, $X_2 = 0$ & $\hat{Y}(X_1, X_2) = 0$
$Y = 1$
$P(X_1 = 0, X_2 = 0, Y = 1) = 0.05 \quad -(i)$

For $X_1 = 0$, $X_2 = 1$ & $\hat{Y}(X_1, X_2) = 1$
$Y = 0$
$P(X_1 = 0, X_2 = 1, Y = 0) = 0.035 \quad -(ii)$

For $X_1 = 1$, $X_2 = 0$ & $\hat{Y}(X_1, X_2) = 1 \implies Y = 0$
$P(X_1 = 1, X_2 = 0, Y = 0) = 0.135 \quad -(iii)$

For $X_1 = 1$, $X_2 = 1$ & $\hat{Y}(X_1, X_2) = 1 \implies Y = 0$
$P(X_1 = 1, X_2 = 1, Y = 0) = 0.015 \quad -(iv)$

Summing $(i), (ii), (iii), (iv)$ we get
Expected error term $= 0.235$

7. You are in the mood to play tennis. However, you are unsure if your friend would be willing to play with you. You want to construct a decision tree that you will use to determine whether your friend is in the mood to play tennis or not. You know that the willingness of your friend to play tennis is dependent on three binary attributes:

- Weather (which takes values Sunny/Rainy)
- Worked Out? (which takes values Yes/No)
- Injured? (which takes values Yes/No)

Here is the information for the last 8 times you tried contacting your friend to play tennis:

Weather	Worked Out?	Injured?	Played?
Sunny	Yes	Yes	No
Sunny	No	Yes	Yes
Sunny	No	No	Yes
Sunny	Yes	Yes	No
Rainy	Yes	Yes	Yes
Rainy	No	Yes	No
Rainy	Yes	No	Yes
Rainy	No	Yes	No

1. What is the initial entropy in the Played variable in the training dataset? (2 mark)
2. At the root of the tree, what is the mutual information offered about the label by each of the three attributes? Which attribute would ID3 place at the root? (3 marks)

1.

Initial counts:

- Played Yes: 4
- Played No: 4

Initial Entropy of Played:

$$Entropy(S) = -\frac{4}{8}\log_2\left(\frac{4}{8}\right) - \frac{4}{8}\log_2\left(\frac{4}{8}\right)$$

$$Entropy(S) = -\frac{4}{8} \times 1 - \frac{4}{8} \times 1 = 1$$

Initial Entropy = 1
1 mark for each correct step

2.

Information Gain for Weather:

- Sunny: 4 instances, 2 Yes, 2 No

$$Entropy(Sunny) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

- Rainy: 4 instances, 2 Yes, 2 No

$$Entropy(Rainy) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$$

$$Entropy(Weather) = \frac{4}{8} \times 1 + \frac{4}{8} \times 1 = 1$$

$$Gain(S, Weather) = 1 - 1 = 0$$

Information Gain for Weather = 0
0.75 mark for correct answer

Information Gain for Worked Out?:

- Yes: 4 instances, 2 Yes, 2 No

$$Entropy(Yes) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) \approx 1$$

- No: 4 instances, 2 Yes, 2 No

$$Entropy(Yes) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) \approx 1$$

$$Entropy(WorkedOut?) = \frac{4}{8} \times 1 + \frac{4}{8} \times 1 = 1$$

$$Gain(S, WorkedOut?) = 1 - 1 = 0$$

Information Gain for Worked Out? = 0
0.75 mark for correct answer

Information Gain for Injured?:

- Yes: 6 instances, 2 Yes, 4 No

$$Entropy(Yes) = -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right) \approx 0.918$$

- No: 2 instances, 2 Yes, 0 No

$$Entropy(No) = -\frac{2}{2}\log_2\left(\frac{2}{2}\right) = 0$$

$$Entropy(Injured?) = \frac{6}{8} \times 0.918 + \frac{2}{8} \times 0 \approx 0.688$$

$$Gain(S, Injured?) = 1 - 0.688 = 0.312$$

Information Gain for Injured? $= 0.312$
0.75 mark for correct answer

Attribute chosen as root: Injured? (highest information gain $= 0.312$)
0.75 mark for correct answer