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

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 not 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. False: Random Forests can handle both categorical and continuous variables, making them versatile for different types of datasets.

B. True: 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. True: 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. True: 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

3. When growing a decision tree using the ID3 algorithm, which of the following is 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.

A. True: Information gain measures how well a given attribute separates the training examples according to their target classification. It evaluates the effectiveness of an attribute in classifying the training data.

B. True: The attribute with the highest information gain is always selected at each step while growing the tree in the ID3 algorithm. This ensures that the attribute that best reduces uncertainty is chosen.

C. True: Information gain is calculated based on the reduction in entropy after the data is split on an attribute. It measures how much information a feature contributes towards classifying the data.

1 mark for correct option

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

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

d) True: 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

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 a Naïve Bayes classifier with 3 boolean input variables, X_1, X_2, and X_3, and one boolean output, Y. (5 marks)

   1. How many parameters must be estimated to train such a Naïve Bayes classifier? (2.5 marks)
   2. How many parameters would have to be estimated to learn the above classifier if we do not make the Naïve Bayes conditional independence assumption? (2.5 marks)

   Solutions:

   a. For a naive Bayes classifier, we need to estimate parameters:
   $P(Y = 1)$,
   $P(X_1 = 1 \mid Y = 0)$,
   $P(X_2 = 1 \mid Y = 0)$,
   $P(X_3 = 1 \mid Y = 0)$,
   $P(X_1 = 1 \mid Y = 1)$,
   $P(X_2 = 1 \mid Y = 1)$,
   $P(X_3 = 1 \mid Y = 1)$.
   Other probabilities can be obtained with the constraint that the probabilities sum up to 1 (like $P(X_1 = 1 \mid Y = 0) = 1 - P(X_1 = 0 \mid Y = 0)$). So we need to estimate 7 parameters.
   1 mark for correct parameters and 1.5 for correct parameter number

   b. Without the conditional independence assumption, we still need to estimate $P(Y = 1)$. **(0.5 mark)**
   For $Y = 1$, we need to know all the enumerations of $(X_1, X_2, \text{and } X_3)$, i.e., 2^3 of possible $(X_1, X_2, \text{and } X_3)$. **(1 mark)**
   Considering the constraint that the probabilities sum up to 1, we must estimate $2^3 - 1 = 7$ parameters for $Y = 1$. Therefore, the total number of parameters is $1 + 2(2^3 - 1) = 15$.. **(1 mark)**

7. Using the dataset provided below, construct a decision tree to predict whether a person will play tennis or not. The attributes available are:

   - **Outlook** (Sunny, Overcast, Rain)
   - **Temperature** (Hot, Mild, Cool)
   - **Humidity** (High, Normal)
   - **Wind** (Weak, Strong)

   The target variable is whether the person will **Play Tennis** (Yes or No). The dataset is as follows:

   1. Calculate the initial entropy for the target variable Play Tennis. (2 mark)
   2. Calculate the information gain for the attributes: Outlook, Temperature, Humidity, and Wind. Which attribute would be chosen as the root of the decision tree based on the ID3 algorithm? (3 mark).

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**Play Tennis**
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Strong	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

1. Calculate the initial entropy for the target variable Play Tennis.

Initial Entropy of Play Tennis:

$$Entropy(S) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$Entropy(S) = -(\frac{9}{14} \times -0.6374) - (\frac{5}{14} \times -1.4854) \approx 0.918$$

Initial Entropy $= 0.940$
2 mark for correct answer or log

2.

Information Gain for Outlook:

$$Entropy(Sunny) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) \approx 0.971$$

$$Entropy(Overcast) = 0$$

$$Entropy(Rain) = -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right) \approx 0.971$$

$$Entropy(Outlook) = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 \approx 0.693$$

$$Gain(S, Outlook) = 0.940 - 0.693 = 0.247$$

Information Gain for Outlook $= 0.247$
0.5 mark for correct answer or correct log

Information Gain for Temperature:

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

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

$$Entropy(Cool) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) \approx 0.811$$

$$Entropy(Temperature) = \frac{4}{14} \times 1 + \frac{6}{14} \times 0.918 + \frac{4}{14} \times 0.811 \approx 0.911$$

$$Gain(S, Temperature) = 0.940 - 0.911 = 0.029$$

Information Gain for Temperature = 0.029
0.5 mark for correct answer or correct log

Information Gain for Humidity:

$$Entropy(High) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) \approx 0.985$$

$$Entropy(Normal) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) \approx 0.592$$

$$Entropy(Humidity) = \frac{7}{14} \times 0.985 + \frac{7}{14} \times 0.592 \approx 0.789$$

$$Gain(S, Humidity) = 0.940 - 0.789 = 0.151$$

Information Gain for Humidity = 0.151
0.5 mark for correct answer or correct log

Information Gain for Wind:

$$Entropy(Weak) = -\frac{6}{7}\log_2\left(\frac{6}{7}\right) - \frac{1}{7}\log_2\left(\frac{1}{7}\right) \approx 0.592$$

$$Entropy(Strong) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) \approx 0.985$$

$$Entropy(Wind) = \frac{7}{14} \times 0.592 + \frac{7}{14} \times 0.985 \approx 0.789$$

$$Gain(S, Wind) = 0.940 - 0.789 = 0.151$$

Information Gain for Wind = 0.151
0.5 mark for correct answer or correct log

Attribute chosen as root: Outlook (highest information gain = 0.247)
1 mark for correct answer