CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2023)
Quiz - 3 rubric

Date of Examination: 02.11.2023 Duration: 30 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. Which of the following statements about the Perceptron convergence theorem is true? [**1 mark**]
   (A) The Perceptron algorithm always converges in a finite number of steps.
   (B) The Perceptron algorithm converges if and only if the data is linearly separable.
   (C) The Perceptron algorithm converges for any dataset with a fixed learning rate.
   (D) The Perceptron algorithm's convergence is guaranteed for non-linearly separable data.

   (B) The Perceptron algorithm converges if and only if the data is linearly separable.

2. Pruning in decision trees is a technique used for: [**1 mark**]
   1. Reducing model complexity
   2. Reducing overfitting
   3. Introducing more features
   4. Enhancing bias

   (B) Reducing overfitting.

3. Bagging in Random Forest involves: [**1 mark**]
   (A) Training multiple models on the same dataset
   (B) Using the same hyperparameters for all trees
   (C) Assigning equal weights to all features
   (D) Randomly selecting subsets of data and features

   (D) Randomly selecting subsets of data and features.

4. The maximum value of entropy for 2 and 3 classes are - [**1 mark**]
   1. 0.5 and 0.33
   2. 1 and log23
   3. 1 and loge3
   4. 0.5 and 1

For 2 classes,

Entropy = $-p_a \log_2(p_a) - p_b \log_2(p_b)$ is maximum when $p_a = p_b = 1/2$

So max value of entropy = $-(1/2)\log_2(1/2) - (1/2)\log_2(1/2) = \log_2 2 = 1$

For 3 classes,

Entropy = $-p_a \log_2(p_a) - p_b \log_2(p_b) - p_c \log_2(p_c)$ is maximum when $p_a = p_b = p_c = 1/3$

So max value of entropy = $-(1/3)\log_2(1/3) - (1/3)\log_2(1/3) - (1/3)\log_2(1/3) = \log_2 3$

5. In the context of Random Forests, consider the following statements about the 'Out-of-Bag' (OOB) error: **[2 marks]**

    1. It is computed by predicting the class (for classification) or value (for regression) of a data point using only the trees for which it was out-of-bag, and then comparing the predicted value or class to the actual value or class.
    2. On average, around 36.8% of the observations are used to train each tree in a Random Forest.
    3. It provides an unbiased estimate of the test error without the need for a separate validation set.
    4. It is a measure of the variance of the Random Forest model.

    Which of the above statements are TRUE regarding OOB error?

    A. 1 and 2 only

    B. 1, 2, and 3 only

    C. 1 and 3 only

    D. 3 and 4 only

    (C) 1 and 3 only

6. Derive the variance of a random forest, given that the variables are simply i.d. (identically distributed, but not necessarily independent) with positive pairwise correlation ρ. **[3 marks]**

$$
\begin{aligned}
\mathrm{Var}\left(\frac{\sum_{i=1}^{B} X_i}{B}\right) &= \frac{1}{B^2}\sum_{i=1}^{B}\mathrm{Var}(X_i) + \frac{1}{B^2}\sum_{i\neq j}^{B}\mathrm{Cov}(X_i, X_j) \\
&= \frac{\sigma^2}{B} + \frac{B-1}{B}\sigma^2\rho \\
&= \sigma^2\rho + \frac{1-\rho}{B}\sigma^2.
\end{aligned}
$$

Figure 1: 1 mark for each step

7. Using the Perceptron training algorithm, find the decision boundary if it exists for the given data points and validate the decision boundary you get by performing an extra iteration of PTA and showing that no weights are updated. **[3 marks]**

Point	Coordinates	Class
A	(1,2)	+
B	(-1,-2)	-
C	(2,3)	+
D	(3,2)	+
E	(-2,-1)	-

We will solve the following problem using the perception training algorithm, firstly we will assume the initial weights and bias to be all zero (not necessary, to assume all to be zero only).
Initially, $w_1 = 0, w_2 = 0, b = 0$

$Iteration 1:$
$Point A(1, 2):$
$w_1 * x_1 + w_2 * x_2 + b = 0 * 1 + 0 * 2 + 0 = 0$
$error =$ $\Phi(0)$ - 1 = 1 - 1 = 0 (No update required)

Point B (-1,-2):
$w_1 * x_1 + w_2 * x_2 + b = 0 * (-1) + 0 * (-2) + 0 = 0$
$error = -1 - \Phi(0)$ = -1 - 1 = -2
$w_1 = w_1 + error * x_1 = 0 + (-2) * (-1) = 2$
$w_2 = w_2 + error * x_2 = 0 + (-2) * (-2) = 4$
$b = b + error = 0 + (-2) = -2$

$Point C(2, 3):$
$w_1 * x_1 + w_2 * x_2 + b = 2 * 2 + 4 * 3 + (-2) = 14$
$error = 1 - \Phi(14)$ = 1 - 1 = 0 (No update is required)

Similarly, Points D and E decision boundary will not be updated as the error will remain zero.

Iteration 2: (For verification)
Perform another iteration for all 5 points and show that the decision boundary is not getting updated.

Note: Only the solutions showing the calculation of all the points will get full credit. Also, there can be multiple solutions depending on initial weights and bias. As long as the PTA and final decision boundary are correct, they will receive full credit.

Final decision boundary = $2x_1 + 4x_2 - 2$

8. If a decision tree node splits data into two sets: one with 60 instances of class A and 40 instances of class B, and the other with 30 instances of class A and 70 instances of class B, calculate the Gini impurity for the split. **[3 marks]**

Initial data before the split (Z)
Class A - 90
Class B - 110

Left Child(Z1):
Class A – 60
Class B – 40

Gini(Z1) = 1 – $P(A)^2$ – $P(B)^2$ = 1 – $(60/100)^2$ – $(40/100)^2$ = 0.48 (1 mark)

Right Child(Z2):
Class A – 30
Class B – 70
Gini(Z2) = 1 – $P(A)^2$ – $P(B)^2$ = 1 – $(30/100)^2$ – $(70/100)^2$ = 0.42 (1 mark)

Therefore, Gini impurity after the split is
$= |Z1|/|Z| * Gini(Z1) + |Z2|/|Z| * Gini(Z2)$
$= 100/200 * 0.48 + 100/200 * 0.42$
$= 0.45 (1 mark)$