National University of Computer & Emerging Sciences, Karachi
Fall-2023 FAST School of Computing
Midterm-I
25th September 2023, 10:00 am – 11:00am

Solution

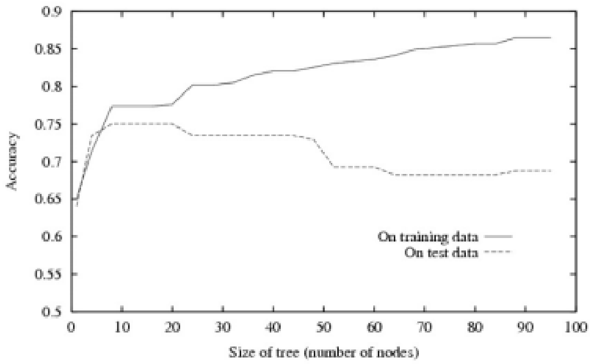
Course Code: CS4101	**Course Name:** Applied Machine Learning	
Instructor Name: Mr. M. Shahzad		
Student Roll No:		**Section:**

Instructions:

- Return the question paper and make sure to keep it inside your answer sheet.
- Read each question completely before answering it. There are **three questions and two pages (front plus back)**.
- In case of any ambiguity, you may make assumption. However, your assumption should not contradict any statement in the question paper.
- Do not write anything on the question paper (except your ID and group).

Total Time: 1 Hour **Max Points**: 15

Question#1: Overfitting: **[5 points, CLO1, 20 mins]**



(a) Consider the training set accuracy and test set accuracy curves plotted above, during decision tree learning, as the number of nodes in the decision tree grows. This decision tree is being used to learn a function f : X → Y , where training and test set examples are drawn independently at random from an underlying distribution P(X), after which the trainer provides a noise-free label Y.
Note error = 1 - accuracy. Please answer each of these **true/false** questions, and **explain/justify** your answer in 1 or 2 sentences.

i. **[1 points]** T or F: Training error at each point on this curve provides an unbiased estimate of true error.
Solution: False. Training error is an optimistically biased estimate of true error, because the hypothesis was chosen based on its fit to the training data.

ii. **[1 point]** T or F: Test error at each point on this curve provides an unbiased estimate of true error
Solution: True. The expected value of test error (taken over different draws of random test sets) is equal to true error.

iii. **[1point]** T of F: Training accuracy minus test accuracy provides an unbiased estimate of the degree of overfitting.
Solution: We defined overfitting as test error minus training error, which is equal to training accuracy minus test accuracy.

iv. **[1point]** T or F: Each time we draw a different test set from P(X) the test accuracy curve may vary from what we see here.
Solution: True. Of course, each random draw from P(X) may vary from another draw.

v. **[1point]** T or F: The variance in test accuracy will increase as we increase the number of test examples
Solution: False. The variance in test accuracy will decrease as we increase the size of the test set

Question 2: [Decision Trees] [5 points, CLO2, 20 mins]

a. **[3 points]** Suppose you are given six training points (listed in Table 1) for a classification problem with two binary attributes X_1, X_2, and three classes $Y \in \{1, 2, 3\}$. We will use a decision tree learner based on information gain.

X_1	X_2	Y
1	1	1
1	1	1
1	1	2
1	0	3
0	0	2
0	0	3

Table 1: Training data for the decision tree learner.

Task: Calculate the information gain for both X_1 and X_2. You can use the approximation $\log_2 3 \approx 19/12$. Report information gains as fractions or as decimals with the precision of three decimal digits. Show your work and circle your final answers for IG(X_1) and IG(X_2).

Solution:

$$\begin{aligned} \mathrm{H}(Y) &= -\sum_{y_i=1}^{n=3} \mathbf{P}(Y = y_i) \log_2 \mathbf{P}(Y = y_i) \\ &= -\sum_{y_i=1}^{n=3} \frac{1}{3} \log_2 \frac{1}{3} = \log_2 3 \approx \frac{19}{12} \end{aligned}$$

For the X_1 split we compute the conditional entropy:

$$\begin{aligned} &= -\left[\frac{2}{6}\left(\frac{0}{2}\log_2\frac{0}{2} + \frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) + \frac{4}{6}\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)\right] \\ &= -\left(-\frac{2}{6} - 1\right) \\ &= \frac{4}{3} \end{aligned}$$

Similarly for the X_2 split we compute the conditional entropy:

$$\begin{aligned} &= -\left[\frac{3}{6}\left(\frac{0}{3}\log_2\frac{0}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) + \frac{3}{6}\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3} + \frac{0}{3}\log_2\frac{0}{3}\right)\right] \\ &\approx -\left(\frac{2}{3} - \frac{19}{12}\right) \\ &= \frac{11}{12} \end{aligned}$$

The final information gain for each split is then:

$$\mathsf{IG}(X_1) = \mathsf{H}(Y) - \mathsf{H}(Y \mid X_1) \approx \frac{19}{12} - \frac{4}{3} = \frac{3}{12} = \frac{1}{4}$$
$$\mathsf{IG}(X_2) = \mathsf{H}(Y) - \mathsf{H}(Y \mid X_2) \approx \frac{19}{12} - \frac{11}{12} = \frac{8}{12} = \frac{2}{3}$$

b. **[2 points]** What is use of ***cost_complexity_pruning_path*** function in scikit-learn's DecisionTreeClassifier? Your answered should be in 3 – 4 lines. Otherwise, answer will not be checked.

Solution:
It gives ***ccp_alphas*** and ***impurities*** arrays with set of values to handle overfitting and makes the decision trees to generalize well to unseen data based.

Question 3: [Naïve Bayes] [5 points, CLO2, 15 mins]

Consider the following **Car Theft** dataset having attributes **Color** , **Type** , **Origin**, and the **subject**, **stolen** can be either *yes* or *no*.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Task: Classify a sample (*Red*, *SUV*, *Domestic*) . Note there is no example of a *Red Domestic SUV* in our data set, so you need to compute this. Use following ***m-estimate*** to estimate $\mathbf{P(a_i|v_j)}$:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

where:

n = the number of training examples for which $v = v_j$
n_c = number of examples for which $v = v_j$ and $a = a_i$
p = a priori estimate for $P(a_i|v_j)$ (assume p=0.5)
m = the equivalent sample size (assume m=3)

Solution:

```
Yes:                       No:
    Red:                       Red:
        n = 5                      n = 5
        n_c= 3                     n_c = 2
        p = .5                     p = .5
        m = 3                      m = 3
    SUV:                       SUV:
        n = 5                      n = 5
        n_c = 1                    n_c = 3
        p = .5                     p = .5
        m = 3                      m = 3
    Domestic:                  Domestic:
        n = 5                      n = 5
        n_c = 2                    n_c = 3
        p = .5                     p = .5
        m = 3                      m =3
```

Looking at $P(Red|Yes)$, we have 5 cases where v_j = Yes , and in 3 of those cases a_i = Red. So for $P(Red|Yes)$, $n = 5$ and $n_c = 3$. Note that all attribute are binary (two possible values). We are assuming no other information so, $p = 1$ / (number-of-attribute-values) = 0.5 for all of our attributes. Our m value is arbitrary, (We will use $m = 3$) but consistent for all attributes. Now we simply apply eqauation (3) using the precomputed values of n , n_c, p, and m.

$$P(Red|Yes) = \frac{3 + 3 * .5}{5 + 3} = .56 \qquad P(Red|No) = \frac{2 + 3 * .5}{5 + 3} = .43$$

$$P(SUV|Yes) = \frac{1 + 3 * .5}{5 + 3} = .31 \qquad P(SUV|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

$$P(Domestic|Yes) = \frac{2 + 3 * .5}{5 + 3} = .43 \qquad P(Domestic|No) = \frac{3 + 3 * .5}{5 + 3} = .56$$

We have $P(Yes) = .5$ and $P(No) = .5$, so we can apply equation (2). For $v = Yes$, we have

```
P(Yes) * P(Red | Yes) * P(SUV | Yes) * P(Domestic|Yes)

    =  .5 * .56 * .31 * .43   = .037
```

and for $v = No$, we have

```
P(No) * P(Red | No) * P(SUV | No) * P (Domestic | No)

    = .5 * .43 * .56 * .56 = .069
```

Since $0.069 > 0.037$, our example gets classified as 'NO'

BEST OF LUCK!