CSC 422/522: Automated Learning and Data Analysis
Homework 3: Due 3/25 @ 23:55

(This is the numerical part of h/w; should be done with hand/calculator; R part of h/w will be given separately)

Student Name:
Student ID:

Table 1. This data will be needed to answer several problems

1. (Hint) Ignore ID attribute in your computations
2. Total 3 input attributes (temperature, outlook, humidity)
    a. Temperature {hot, cool, mild}
    b. Outlook {sunny, overcast, rainy}
    c. Humidity {high, normal}
3. Output attribute = Paly {yes, no}
4. Total records or objects = 14
5. **Ignore case and spelling mistakes if any**

ID	Temperature	Outlook	Humidity	Y=Play{yes,no}
1	hot	sunny	high	no
2	cool	overcast	normal	no
3	mild	sunny	high	no
4	mild	overcast	high	no
5	hot	sunny	high	yes
6	hot	rainy	high	yes
7	mild	rainy	high	yes
8	cool	rainy	normal	yes
9	cool	rainy	normal	yes
10	cool	sunny	normal	yes
11	mild	rainy	normal	yes
12	mild	sunny	normal	yes
13	mild	rainy	high	yes
14	hot	rainy	normal	yes

Q1. (25 points) Using data given in Table 1 as training data, answer the following questions.

(a) Construct three decision trees using (i) Gini index, (ii) Entropy, and (iii) misclassification error measures for selecting best splits. Show all work and draw the resulting tree (no pruning). **15 points**

(b) Compute following accuracy measures on training data: (i) individual class accuracy, (ii) overall class accuracy measures. **5 points**

(c) For the following data, predict class label for each instance using each tree constructed in (a). **5 points**

ID, Temperature, Outlook, Humidity, Y=play (yes or no)
1,hot,overcast,high,?
2,mild,overcast,high,?
3,cool,sunny,normal,?
4,mild,rainy,normal,?
5,mild,sunny,normal,?

Q2. (12 points) Consider the **decision tree** shown in Figure 2, and the corresponding training (table 2.1) and test (table 2.2). Using this data, answer the following questions **(a)-(c).** (**6** points)

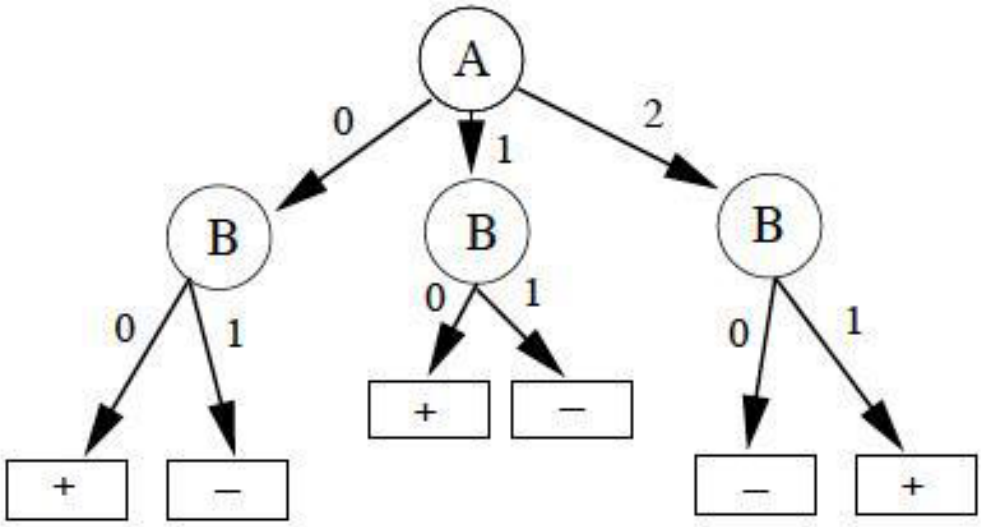


Figure 2: Decision Tree

A	B	Number of (+) instances	Number of (-) instances
0	0	6	2
0	1	2	5
1	0	20	9
1	1	8	31
2	0	2	5
2	1	7	3

Table 2.1 Training data

A	B	Number of (+) instances	Number of (-) instances
0	0	4	1
0	1	3	1
1	0	7	2
1	1	4	14
2	0	6	1
2	1	2	5

Table 2.2 Test data

(a) Estimate the generalization error rate of the tree using both optimistic and pessimistic approach. Use $\Omega = 2$ as the cost of adding a leaf node while calculating the pessimistic error. (**2** points)

(b) Compute the error rate of the tree for the test dataset (**1** point)

(c) Figure 2.2 shows the pruned version of the tree shown in Figure 2.1. Estimate the generalization error of this tree using optimistic and pessimistic approaches (**2** points). Also compute the error rate for the test dataset in table 2.2 (**1** point).

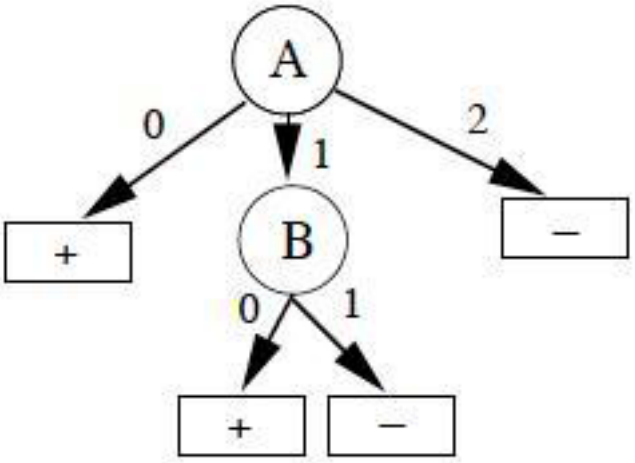


Figure 2.2. Pruned version of the decision tree shown in Figure 2.1

Q3. (12 points) Using the Naïve Bayesian (NB) Classifier and the data given in Table 1 (page 1), answer the following questions (a) – (e):

(a) If there is no information about the weather conditions, predict whether the player is going to play or not? (1 point). Justify your answer (1 point).

(b) Predict the label for the data point X ={Sunny, Mild, Normal}. (1 point). Show your work (1 point).

(c) If the only attribute information you have is that temperature is mild. What will NB classifier prediction (1 point), show your work (1 point).

(d) In addition to temperature is "mild", it is also known that humidity is "high." What will be NB classifier prediction (1 point), show your work (1 point).

(e) If data point X = {Overcast, Mild, High}, using NB classifier, predict if the player is likely to play or not (1 point), show your work (1 point), if there is any problem with the prediction, how do you fix it? (2 points)

Q4. (**10 points**) Rule-based classification. Using **Holt's 1-R** algorithm, answer following sub-questions (a)-(b).

(a) Apply Holt's 1-R on the data given in Table 1 (page 1). Show all work. (8 points)

(**b**) Which is the best attribute (i.e., attribute for which total error is minimum) (2 points)

Q5. **(10 points)** Design a neural network for XOR problem, data is shown in the following table.

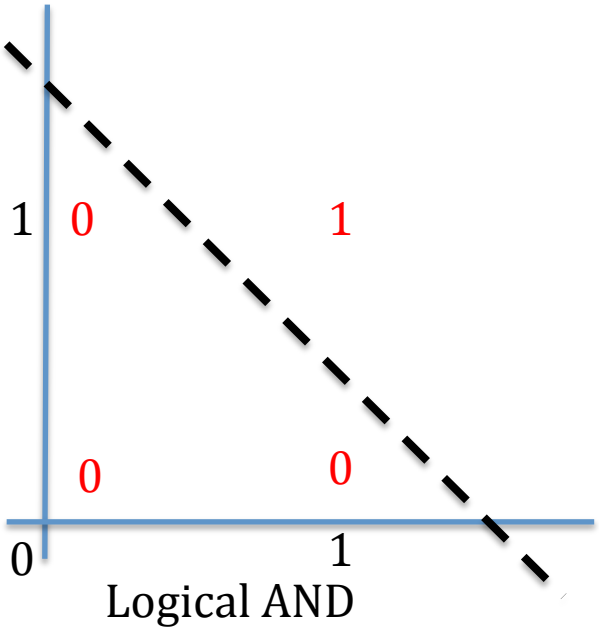
Truth Table for the XOR problem

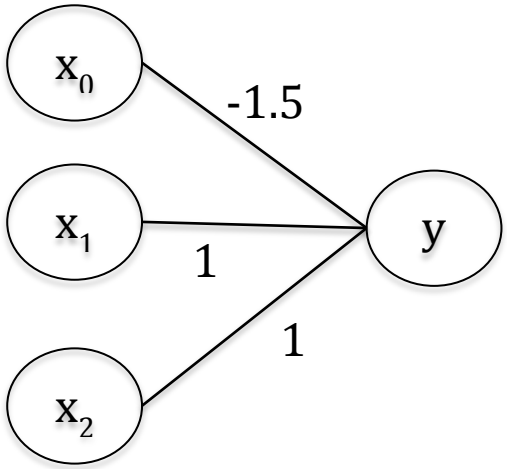
X1	X2	XOR	Class
0	1	1	A
1	0	1	A
0	0	0	B
1	1	0	B

Hint. We know simple perceptron can't solve this problem. However, a two layer perceptron (with one hidden layer). Draw 2-d scatter plot and draw separating planes.

Your objective is to design the network, show the weights and biases, and show resulting line equations of planes. Using your network, predict class for each of the four inputs in the table.

[Below, I am giving an example solution for Logical AND, data is represented the following figure (logical AND values for each input pair is shown in Red color).]





Perceptron Solution

$x_1+x_2-1.5 = 0$

Q6. (11 points) For given dataset below, estimate the error rate of 1-nearest neighbor (1-NN) using leave-one-out cross validation. Consider grids are unit squares, and no need to compute actual distances.