Data mining (2022/11/14)

1. (20%) With the following training dataset, use the ~~Naïve Bayes~~ classification algorithm to predict the class label of (A=0, B=1, C=0). Please set the Laplacian smoothing parameter to 1 to prevent the "0-probability problem". Please give all the calculation steps, not just give a random guess for the answer. Please note that the attributes of the dataset include "A", "B"," C" and "Class". Furthermore, "Class" is the class label (the response variable)

$C_1 = Class = P = 5 \Rightarrow P(C_1) = \frac{5}{10}$

$C_2 = Class = N = 5 \Rightarrow P(C_2) = \frac{5}{10}$

$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$

Index	A	B	C	Class
	0	1	0	
1	0	0	0	P
2	1	0	1	N
3	0	1	1	N
4	0	1	1	N
5	0	0	1	P
6	1	0	1	P
7	1	0	1	N
8	1	0	1	N
9	1	1	1	P
10	1	0	1	P

$P(X|C_1) = \frac{2+1}{5+2} \times \frac{1+1}{5+2} \times \frac{1+1}{5+2}$

$= \frac{3}{7} \times \frac{2}{7} \times \frac{2}{7} \times \frac{5}{10}$

$= 0.017$

$P(X|C_2) = \frac{2+1}{5+2} \times \frac{2+1}{5+2} \times \frac{5+1}{5+2}$

$= \frac{3}{7} \times \frac{3}{7} \times \frac{6}{7} \times \frac{5}{10} = 0.079$ $\Rightarrow$ Ans: No

Err → Gini

2. (20%)With the following dataset, answer the following questions.

a2: T (3,3) F (2,2)

Error = 4

$Gini_T = 2 \times \frac{2}{5} \times \frac{3}{5} = 0.48$

$Gini_F = 2 \times \frac{2}{4} \times \frac{2}{4} = 0.5$

$C_1: TC = + = 4,\ \frac{4}{9}$

$C_2: TC = - = 5,\ \frac{5}{9}$

Instance	a1	a2	a3	Target Class
	T=4, F=5	T=5, F=4		+:4, -:5
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a1 Err

Dataset

T (3,1) F (1,4)

Error = 2

$A = \frac{9-2}{9} = 0.778$

a1 $Gini(T) = 2 \times \frac{3}{4} \times \frac{1}{4} = 0.375$

$Gini(F) = 2 \times \frac{1}{5} \times \frac{4}{5} = 0.32$

(a) Using Gini function as attribute selection measure; select the better splitting attribute among attributes a1 and a2 for the root. (5%)

(b) For a3, which is a continuous attribute, compute the Gini value for every possible split. What is the best splitting condition for this attribute? (10%)

(c) According to the results of questions a and b, draw a two-level decision tree and calculate its classification accuracy based on the training dataset. (5%)

An example for splitting a continuous attribute is shown in the following:

a3

$\frac{3}{10}\times\frac{7}{10}\times 2$

Gini = $1-\sum p_i^2 = \sum_{i\neq j} p_i * p_j = 2p_1 * p_2$ (for two-class case)

$= 1 -$

$2\times\frac{2}{5}\times\frac{3}{5} = 0.48$

$2\times\frac{0}{1}\times\frac{1}{1}$

$2\times\frac{1}{5}\times\frac{4}{5} = 0.32$ = 0.4

162 Chapter 4 Classification

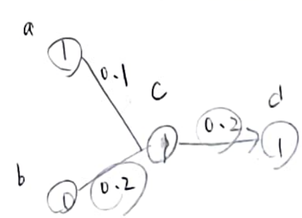
Class	No	No	No	Yes	Yes	Yes	No	No	No	No
Annual Income – Sorted Values	60	70	75	85	90	95	100	120	125	220

Split Positions	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		0.420	

Figure 4.16. Splitting continuous attributes.

0.23 + 0.10 = 0.33

3. (20%) Given a simple MLP with two inputs a and b, one hidden node c and one output node d, please find the updated weight of W_{cd} and W_{bc} after training the network once with the following training example. Assume that the initial values of the weights of w_{ac}, w_{bc}, w_{cd} are 0.1, 0.2, 0.2, respectively, and the learning rate is 0.9.



$W_{cd} = W_{cd} + L\cdot Error_d\ Out_c$

$W_{bc} = W_{bc} + L\cdot Error_c\ Out_b$

The training example:

a	b	d
1	0	1

$Error_c = 0.711\times(1-0.711)\times(1-0.711)$
$= 0.059$

(Notes: 1. Let not consider the biases (i.e., no biases).

2. W_{ac} denotes the weight of the link from node a to node c. Similarly, W_{bc} and W_{cd} denote the weight of the link from node b to node c and the weight of the link from node c to node d, respectively.)

$O_c = \frac{1}{1+e^{-0.9}} = 0.711$

1.407

$Error_d = 0.536\times(1-0.536)\times(1-0.536) = 0.116$

$W_{cd} = 0.2 + 0.9\times 0.536\times 0.116 = 0.2560$

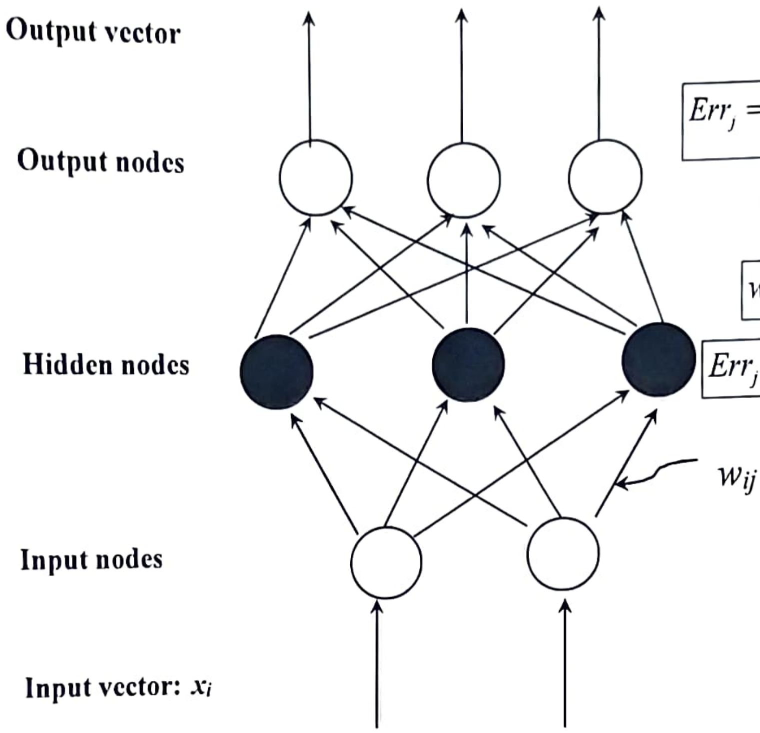
$O_d = \frac{1}{1+e^{-0.1422}} = 0.536$

0.867

$O_b = \frac{1}{1+e^{-0.1422}} = 0.536$

$W_{bc} = 0.2 + 0.9\times 0.059\times 0.536$
$= 0.228$

Appendix :



$$Err_j = O_j(1-O_j)\sum_k Err_k w_{jk}$$

$$\theta_j = \theta_j + (l)Err_j$$

$$w_{ij} = w_{ij} + (l)Err_j O_i$$

$$Err_j = O_j(1-O_j)(T_j - O_j)$$

$$O_j = \frac{1}{1+e^{-I_j}}$$

$$I_j = \sum_i w_{ij} O_i + \theta_j$$

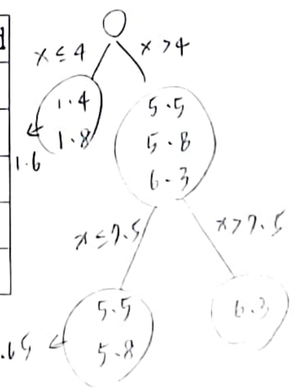
4. (20%) The following dataset contains two columns, X and Y. Please build a regression tree with X as the input variable and Y as the output variable (10%). Then, fill in the predicted value for each X using the regression tree. (10%)

Note: Set the minimum number of samples in a node to three to allow splitting the node in building the regression tree. In other words, if the number of samples in a node is less than or equal to two (≤2), no splitting of this node is allowed.

X	Y	Predicted
1	1.4	1.6
2	1.8	1.6
6	5.5	5.65
7	5.8	5.65
8	6.3	6.3

x = 4 →

x = 7.5 →



5. (20%, Gradient Boosting, with learning rate of 0.5) With the same dataset for problem 4, use gradient boosting to construct a model for predicting Y values. To reduce the required computation, use only one one-level regression tree. That is, use F0 = 4.16 as the first prediction for all Y's; construct a one-level regression tree h1 to predict the residuals. Then, compute the predicted values for all Y's

using the **learning rate of 0.5**

X	Y	pre
1	1.4	1.6
2	1.8	1.6

$Y - \bar{Y} = -2.76$

-2.36

$x = 4 \rightarrow$

$x \leq 4$ $x > 4$