22F CS-559 A HW 1

Jiangrui Zheng

November 14, 2022

1 Problem 1

1. (1)

Entropy before splitting:

Class	Num	$\mathbf{P_j}$	$-(\mathbf{P_j}^* \log_2(\mathbf{P_j})$
+	4	0.44	0.520
−	5	0.56	0.471
Total Num	9	Total Entropy	0.991

Splits:

Split		**Class**			**+**		**-**	
	Child Nodes	**+**	**-**	**total**	$\mathbf{p_j}$	- $(\mathbf{P_j}^* \log_2(\mathbf{P_j})$	$\mathbf{p_j}$	- $(\mathbf{P_j}^* \log_2(\mathbf{P_j})$
a1	T	3	1	4	0.750	0.311	0.250	0.500
	F	1	4	5	0.200	0.464	0.800	0.258
a2	<=1.5	1	0	1	1.000	0.000	0.000	0.000
	>1.5	3	5	8	0.375	0.531	0.625	0.424
a2	<=3.5	1	1	2	0.500	0.500	0.500	0.500
	>3.5	3	4	7	0.429	0.524	0.571	0.461
a2	<=4.5	2	1	3	0.667	0.390	0.333	0.528
	>4.5	2	4	6	0.333	0.528	0.667	0.390
a2	<=5.5	2	3	5	0.400	0.529	0.600	0.442
	>5.5	2	2	4	0.500	0.500	0.500	0.500
a2	<=6.5	3	3	6	0.500	0.500	0.500	0.500
	>6.5	1	2	3	0.333	0.528	0.667	0.390
a2	<=7.5	4	4	8	0.500	0.500	0.500	0.500
	>7.5	0	1	1	0.000	0.000	1.000	0.000

Information gain:

Result				
SUM - $(P_j * \log_2(P_j)$	**PCT**	**H(T)**	**Entropy**	**Information Gain**
0.811	0.444	0.361	0.762	0.229
0.722	0.556	0.401		
0.000	0.111	0.000	0.848	0.143
0.954	0.889	0.848		
1.000	0.222	0.222	0.989	0.003
0.985	0.778	0.766		
0.918	0.333	0.306	0.918	0.073
0.918	0.667	0.612		
0.971	0.556	0.539	0.984	0.007
1.000	0.444	0.444		
1.000	0.667	0.667	0.973	0.018
0.918	0.333	0.306		
1.000	0.889	0.889	0.889	0.102
0.000	0.111	0.000		

We can see that a1 attribute has largest information gain, so a1 will be chosen as the first splitting for decision tree.

2. If we use "Instance" as another attribute, the result will change a lot. I think this attribute should not be used for a decision in the tree because it is does not contain valid information. It is just a just a marker to count the data like ID or phone number or something else.

2 Problem 2

1. Count matrix:

	Parent
+	35
-	65
Gini	0.455

$$Gini(parent) = 1 - (35/100)^2 - (65/100)^2 = 0.455$$

If first splitting attribute is A:

	T	F
+	20	15
-	30	35
Gini		0.45

$$Gini(T) = 1 - (20/50)^2 - (30/50)^2 = 0.48$$
$$Gini(F) = 1 - (15/50)^2 - (35/50)^2 = 0.42$$
$$Gini(children|A) = 50/100 * 0.48 + 50/100 * 0.42 = 0.45$$

If first splitting attribute is B:

	T	F
+	15	20
-	20	45
Gini		0.448

$$Gini(BT) = 1 - (15/35)^2 - (20/35)^2 = 0.4898$$
$$Gini(BF) = 1 - (20/65)^2 - (45/65)^2 = 0.4260$$
$$Gini(children|B) = 35/100 * 0.4898 + 65/100 * 0.4260 = 0.4484$$

Choose the attribute that minimizes weighted average Gini index of the children. So B would be chosen as the first splitting attribute.

2. By count matrix:

A	+	-
T	0	20
T	20	10
F	15	0
F	0	35

B	+	-
T	0	20
T	15	0
F	20	10
F	0	35

$$COST(A) = -20 + 15 * 100–35 * 10 = 1130$$

$$COST(B) = -15 + 20 * 100–45 * 10 = 1535$$

We can find COST(A) is less than COST(B), so here we choose A as the first splitting attribute.

3 Problem 3

1. Count metrix:

	correct	incorrect
H1	8	2
H2	6	4
H3	9	1

In H1:

D1(1) = D1(2) = ... = D1(10) = 1/10
Err1 = 0.1 * 2 = 0.2
α_1 = 1/2 * log((1-0.2)/0.2) = 0.69

For those classified not correctly: $D2 = 0.1 * e^{0.69*1} = 0.200$
For those classified correctly: $D2 = 0.1 * e^{0.69*(-1)} = 0.050$

In H2:

D1(1) = D1(2) = ... = D1(10) = 1/10
Err1 = 0.1 * 4 = 0.4
α_1 = 1/2 * log((1-0.4)/0.4) = 0.203
For those classified not correctly: $D2 = 0.1 * e^{0.203*1} = 0.123$
For those classified correctly: $D2 = 0.1 * e60.203 * (-1) = 0.082$

In H3:

D1(1) = D1(2) = ... = D1(10) = 1/10
Err1 = 0.1 * 1 = 0.1
α_1 = 1/2 * log((1-0.1)/0.1) = 1.097
For those classified not correctly: $D2 = 0.1 * e^{1.097*1} = 0.300$
For those classified correctly: $D2 = 0.1 * e^{1.097*(-1)} = 0.033$

2. In case H1, there are 2 mis-classified instances(9, 10) which will be re-weighted after the first iteration.
In case H2, instances(1,2,3,8) are mis-classified which will be re-weighted after the first iteration.
In case H3, only instance 9 is mis-classified which will be re-weighted after the first iteration.

4 Problem 4

Part I.

1. By directly observing the chart, we can get 5 nearest points are:
   (4,1), (4,4), (5,1) -
   (6,5), (7,3) +
   which " - " takes majority part.
   So, the test point (triangle) is classified as " - ".

2. First list all the points in a count matrix

	X1	X2	Class
1	1	1	-
2	1	4	-
3	1	5	-
4	1	8	-
5	1	9	-
6	2	2	-
7	3	1	-
8	4	4	-
9	6	2	-
10	3	7	+
11	4	9	+
12	5	6	+
13	6	9	+
14	7	7	+
15	7	8	+
16	8	2	+
17	8	6	+
18	9	5	+
19	9	8	+
20	4	5	

Manhattan distance weighted 3-nearest neighbor(the weight is $1/d^2$):

$$\mathrm{d} = w_m \sum_{m=1}^{D} |x_{im} - x_{jm}|$$

1	2	3	4	5	6	7	8	9	10
0.14	0.25	0.33	0.17	0.14	0.20	0.20	1	0.20	0.33
11	12	13	14	15	16	17	18	19	
0.25	0.50	0.17	0.20	0.17	0.14	0.20	0.20	0.13	

(1)

We can see that 19(+) is the closet one, but other 3 points 16(+), 1(-), 5(-) have the same distance. In this case, we can't tell which class takes majority part since there are 2 "+" and 2 " - ".