Assignment #4

Wednesday, October 11, 2023 8:24 PM

11413696

Data Mining

Assignment Four **10 points**

Due: Thursday Oct. 19 @ 11:59 PM

Question 1 (6 points). Consider the training examples shown in the Table below for a binary classification problem.

Instance	A1	A2	A3	Class
1	T	T	1.0	Yes
2	T	T	6.0	Yes
3	T	F	4.0	No
4	F	T	7.0	Yes
5	F	T	8.0	Yes
6	F	T	5.0	No
7	F	F	3.0	No
8	F	F	7.0	No
9	F	T	8.0	No

(a) What is the entropy of this collection of training examples with respect to the class attribute?
(b) What are the information gains of *A1* and *A2* relative to these training examples?
(c) For *A3*, which is a continuous attribute, compute the information gain for every possible split.
(d) What is the best split (among *A1*, *A2*, and *A3*) according to the information gain?
(e) What is the best split (between *A1* and *A2*) according to the misclassification error rate?
(f) What is the best split (between *A1* and *A2*) according to the Gini index?

Question 2 (4 points). Consider splitting a parent node P into two child nodes, C1 and C2, using some attribute test condition. The composition of labeled training instances at every node is summarized in the Table below.

	P	C1	C2
Class 0	7	3	4
Class 1	3	0	3

(a) Calculate the Gini index and misclassification error rate of the parent node P.
(b) Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?
(c) Calculate the weighted misclassification rate of the child nodes. Would you consider this attribute test condition if misclassification rate is used as the impurity measure?

Good luck!

1

Entropy at a given node t

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes

2

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Entropy = – 0 log 0 – 1 log 1 = – 0 – 0 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6
Entropy = – (1/6) log_2 (1/6) – (5/6) log_2 (1/6) = 0.65

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6
Entropy = – (2/6) log_2 (2/6) – (4/6) log_2 (4/6) = 0.92

3

Information Gain:

$$Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

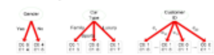
Parent Node, p is split into k partitions (children)
n_i is number of records in child node i

– Choose the split that achieves most reduction (maximizes GAIN)

4

Problem with large number of partitions

- Node impurity measures tend to prefer splits that result in large number of partitions, each being small but pure



– Customer ID has highest information gain because entropy for all the children is zero

5

Measure of Impurity: Classification Error

- Classification error at a node t

$$Error(t) = 1 - \max_i [p_i(t)]$$

– Maximum of $1 - 1/c$ when records are equally distributed among all classes, implying the least interesting situation
– Minimum of 0 when all records belong to one class, implying the most interesting situation

6

Computing Error of a Single Node

$$Error(t) = 1 - \max_i [p_i(t)]$$

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Error = 1 – max (0, 1) = 1 – 1 = 0

C1	1
C2	5

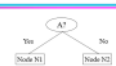
P(C1) = 1/6 P(C2) = 5/6
Error = 1 – max (1/6, 5/6) = 1 – 5/6 = 1/6

C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6
Error = 1 – max (2/6, 4/6) = 1 – 4/6 = 1/3

7

Misclassification Error vs Gini Index



	Parent
C1	7
C2	3
Gini = 0.42	

	N1	N2
C1	3	4
C2	0	3
Gini=0.342		

	N1	N2
C1	3	4
C2	1	2
Gini=0.416		

Misclassification error for all three cases = 0.3 !

8

- Gini Index for a given node t

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

Where $p_i(t)$ is the frequency of class i at node t, and c is the total number of classes

– Maximum of $1 - 1/c$ when records are equally distributed among all classes, implying the least beneficial situation for classification
– Minimum of 0 when all records belong to one class, implying the most beneficial situation for classification

9

- Gini Index for a given node t :

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

– For 2-class problem (p, 1 – p):
  - GINI = 1 – p^2 – $(1 – p)^2$ = 2p (1-p)

C1	0	C1	1	C1	2	C1	3
C2	6	C2	5	C2	4	C2	3
Gini=0.000		Gini=0.278		Gini=0.444		Gini=0.500	

10

Computing Gini Index of a Single Node

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

C1	0
C2	6

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1
Gini = 1 – $P(C1)^2$ – $P(C2)^2$ = 1 – 0 – 1 = 0

C1	1
C2	5

P(C1) = 1/6 P(C2) = 5/6
Gini = 1 – $(1/6)^2$ – $(5/6)^2$ = 0.278

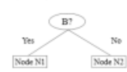
C1	2
C2	4

P(C1) = 2/6 P(C2) = 4/6
Gini = 1 – $(2/6)^2$ – $(4/6)^2$ = 0.444

Binary Attributes: Computing GINI Index

- Splits into two partitions (child nodes)
- Effect of Weighing partitions:
  – Larger and purer partitions are sought

11



	Parent
C1	7
C2	5
Gini = 0.486	

Gini(N1)
= 1 – $(5/6)^2$ – $(1/6)^2$
= 0.278

Gini(N2)
= 1 – $(2/6)^2$ – $(4/6)^2$
= 0.444

	N1	N2
C1	5	2
C2	1	4
Gini=0.361		

Weighted Gini of N1 N2
= 6/12 * 0.278 +
6/12 * 0.444
= 0.361

Gain = 0.486 – 0.361 = 0.125

Question 1

Part A:

$Entropy(t) = -\Sigma\left[P(t) * log2(P(t))\right] = -\left[\left(\frac{4}{9}\right) * log2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * log2\left(\frac{5}{9}\right)\right] = 0.99107605983$

Part B:

"Yes" class labels

$Entropy(A1) = -\Sigma\left[P(t) * log2(P(t))\right] = -\left[\left(\frac{2}{4}\right) * log2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) * log2\left(\frac{2}{4}\right)\right] = 1$

$Entropy(A2) = -\Sigma\left[P(t) * log2(P(t))\right] = -\left[\left(\frac{4}{4}\right) * log2\left(\frac{4}{4}\right) + 0\right] = 0$

"No" class labels

$Entropy(A1) = -\Sigma\left[P(t) * log2(P(t))\right] = -\left[\left(\frac{1}{5}\right) * log2\left(\frac{1}{5}\right) + \left(\frac{4}{5}\right) * log2\left(\frac{4}{5}\right)\right] = 0.72192809488$

$Entropy(A2) = -\Sigma\left[P(t) * log2(P(t))\right] = -\left[\left(\frac{2}{5}\right) * log2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) * log2\left(\frac{3}{5}\right)\right] = 0.97095059445$

$Gainsplit(A1) = Entropy(T) - \Sigma\left(\frac{T_i}{T}\right) * Entropy(i) = 0.99107605983 - \left[\left(\frac{4}{9}\right) * (1) + \left(\frac{5}{9}\right) * (0.72192809488)\right] = 0.14556045156$

$Gainsplit(A2) = Entropy(T) - \Sigma\left(\frac{T_i}{T}\right) * Entropy(i) = 0.99107605983 - \left[\left(\frac{4}{9}\right) * (0) + \left(\frac{5}{9}\right) * (0.97095059445)\right] = 0.45165906291$

Part C: Type equation here.

Sorted Values	1	3	4	5	6	7	8	
Split Positions	0.5	2	3.5	4.5	5.5	6.5	7.5	8.5
	<= | >	<= | >	<= | >	<= | >	<= | >	<= | >	<= | >	<= | >
Yes	0 | 4	1 | 3	1 | 3	1 | 3	1 | 3	2 | 2	3 | 1	4 | 0
No	0 | 5	0 | 5	1 | 4	2 | 3	3 | 2	3 | 2	4 | 1	5 | 0
Gain	$Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{4}{9}) * log2(\frac{4}{9}) + (\frac{5}{9}) * log2(\frac{5}{9})] = 0.99107605983$ $Gainsplit(t) = 0$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{1}{1}) * log2(\frac{1}{1}) + (0) * log2(0)] = 0$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{3}{8}) * log2(\frac{3}{8}) + (\frac{5}{8}) * log2(\frac{5}{8})] = 0.95443400292$ $Weighted\ Average: [(\frac{1}{9}) * (0)] + [(\frac{8}{9}) * (0.95443400292)] = 0.84838578037$ $Gainsplit(t) = 0.99107605983 - (0.84838578037) = 0.14269027946$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{1}{2}) * log2(\frac{1}{2}) + (\frac{1}{2}) * log2(\frac{1}{2})] = 1$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{3}{7}) * log2(\frac{3}{7}) + (\frac{4}{7}) * log2(\frac{4}{7})] = 0.98522813603$ $Weighted\ Average: [(\frac{2}{9}) * (1)] + [(\frac{7}{9}) * (0.98522813603)] = 0.98851077246$ $Gainsplit(t) = 0.99107605983 - (0.98851077246) = 0.00256528737$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{1}{3}) * log2(\frac{1}{3}) + (\frac{2}{3}) * log2(\frac{2}{3})] = 0.91829583405$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{3}{6}) * log2(\frac{3}{6}) + (\frac{3}{6}) * log2(\frac{3}{6})] = 1$ $Weighted\ Average: [(\frac{3}{9}) * (0.91829583405)] + [(\frac{6}{9}) * (1)] = 0.97276527801$ $Gainsplit(t) = 0.99107605983 - (0.97276527801) = 0.01831078182$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{1}{4}) * log2(\frac{1}{4}) + (\frac{3}{4}) * log2(\frac{3}{4})] = 0.81127812445$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{3}{5}) * log2(\frac{3}{5}) + (\frac{2}{5}) * log2(\frac{2}{5})] = 0.97095059445$ $Weighted\ Average: [(\frac{4}{9}) * (0.81127812445)] + [(\frac{5}{9}) * (0.97095059445)] = 0.89998505222$ $Gainsplit(t) = 0.99107605983 - (0.89998505222) = 0.09109100761$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{2}{5}) * log2(\frac{2}{5}) + (\frac{3}{5}) * log2(\frac{3}{5})] = 0.97095059445$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{2}{4}) * log2(\frac{2}{4}) + (\frac{2}{4}) * log2(\frac{2}{4})] = 1$ $Weighted\ Average: [(\frac{5}{9}) * (0.97095059445)] + [(\frac{4}{9}) * (1)] = 0.98386144136$ $Gainsplit(t) = 0.99107605983 - (0.98386144136) = 0.00721461847$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{3}{7}) * log2(\frac{3}{7}) + (\frac{4}{7}) * log2(\frac{4}{7})] = 0.98522813603$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{1}{2}) * log2(\frac{1}{2}) + (\frac{1}{2}) * log2(\frac{1}{2})] = 1$ $Weighted\ Average: [(\frac{7}{9}) * (0.98522813603)] + [(\frac{2}{9}) * (1)] = 0.98851077246$ $Gainsplit(t) = 0.99107605983 - (0.98851077246) = 0.00256528737$	$\leq$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = -[(\frac{4}{9}) * log2(\frac{4}{9}) + (\frac{5}{9}) * log2(\frac{5}{9})] = 0.99107605983$ $>$: $Entropy(t) = -\Sigma[P(t) * log2(P(t))] = 0$ $Weighted\ Average: [(\frac{9}{9}) * (0.99107605983)] + [0] = 0.99107605983$ $Gainsplit(t) = 0.99107605983 - (0.99107605983) = 0$

Part D:

According to the information gain, A2 is the best split between A1, A2, and A3. We prefer to have an information gain value t hat is close to 1. It is hard to tell with A3 because of the large number of partitions. Sure, the nodes are pure but they ar e small measures.

Part E:

According to the misclassification error rate, $Misclassification\ Error(t) = 1 - \max(p_i(t))$, A2 is the best split between A1 and A2 since it has a lower classification error when compared to its counterpart. The lowe r the classification error, the more accurate of a sample set.

$Misclassification\ Error(A1) = 1 - \max(p_i(t)) = 1 - \max\left[\left(\frac{3}{9}\right), \left(\frac{6}{9}\right)\right] = 1 - \left(\frac{6}{9}\right) = 0.3333$

$Misclassification\ Error(A2) = 1 - \max(p_i(t)) = 1 - \max\left[\left(\frac{2}{9}\right), \left(\frac{7}{9}\right)\right] = 1 - \left(\frac{7}{9}\right) = 0.2222$

Part F:

According to the gini index, $Gini(t) = 1 - \sum p_i(t)^2$, A2 is the best split between A1 and A2 since it has a lower gini index when compared to its counterpart. The lower the gini index, the closer a node is to purity.

$$Gini(A1) = 1 - \sum p_i(t)^2 = 1 - \left[\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right] = 0.44444444444$$

$$Gini(A2) = 1 - \sum p_i(t)^2 = 1 - \left[\left(\frac{2}{9}\right)^2 + \left(\frac{7}{9}\right)^2\right] = 0.34567901234$$

Question 2

Part A:

$Gini(t) = 1 - \sum p_i(t)^2,\ Gini(P) = 1 - \sum p_i(t)^2 = 1 - \left[\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2\right] = 0.42$

$Misclassification\ Error\ Rate(t) = 1 - \max(p_i(t)),\ Misclassification\ Error\ Rate(P) = 1 - \max(p_i(t)) = 1 - \max\left[\left(\frac{7}{10}\right), \left(\frac{3}{10}\right)\right] = 1 - \left(\frac{7}{10}\right) = 0.30$

Part B:

$Gini(t) = 1 - \sum p_i(t)^2,\ Gini(C1) = 1 - \sum p_i(t)^2 = 1 - \left[\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right] = 0$

$Gini(t) = 1 - \sum p_i(t)^2,\ Gini(C2) = 1 - \sum p_i(t)^2 = 1 - \left[\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right] = 0.48979591836$

$Weighted\ Gini\ Index: \left[\left(\frac{3}{10}\right) * (0)\right] + \left[\left(\frac{7}{10}\right) * (0.48979591836)\right] = 0.34285714285$

If gini is used as the impurity measure, we would consider this attribute test condition. Consider looking at the gini index of the parent node P before the split. The gini impurity of each child node after the split then calculating the weighted gin i gives a smaller output (which is what we want) in comparison to its parent P. That is the reason we would consider using this attribute test condition.

Part C:

$Misclassification\ Error\ Rate(t) = 1 - \max(p_i(t)),\ Misclassification\ Error\ Rate(C1) = 1 - \max(p_i(t)) = 1 - \max\left[\left(\frac{3}{3}\right), \left(\frac{0}{3}\right)\right] = 1 - \left(\frac{3}{3}\right) = 0$

$Misclassification\ Error\ Rate(t) = 1 - \max(p_i(t)),\ Misclassification\ Error\ Rate(C2) = 1 - \max(p_i(t)) = 1 - \max\left[\left(\frac{4}{7}\right), \left(\frac{3}{7}\right)\right] = 1 - \left(\frac{4}{7}\right) = 0.42857142857$

$Weighted\ Misclassification\ Error\ Rate: \left[\left(\frac{3}{10}\right) * (0)\right] + \left[\left(\frac{7}{10}\right) * (0.42857142857)\right] = 0.29999999999$

If misclassification error rate is used as the impurity measure, we would NOT consider this attribute test condition. Conside r looking at the misclassification error rate of the parent node P before the split. The misclassification impurity of each child node after the split then calculating the weighted misclassification error gives an output equal (there is no drop in the error rate) to its parent P. Therefore, it would bring us no benefit to consider this attribute test condition.