Assignment 2.1

University of San Diego

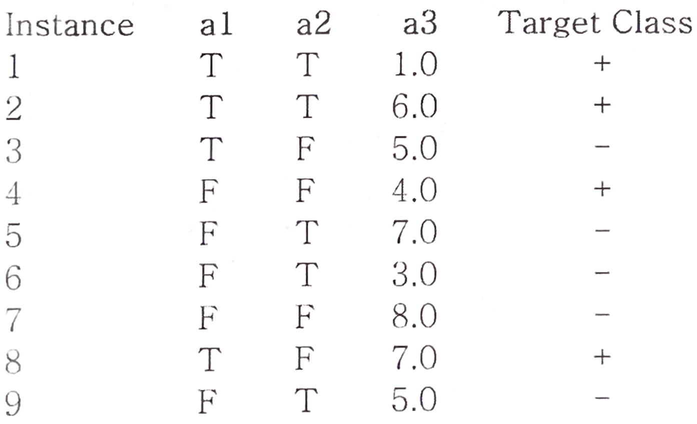
ADS 502

Dingyi Duan

Introduction to Data Mining: Exercises 3.11 – Page 186: Question #3

3. Consider the training examples shown in Table 3.6 for a binary classification problem.

Table 3.6. Data set for Exercise 3.



Instance	a1	a2	a3	Target Class
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	−

a. **What is the entropy of this collection of training examples with respect to the class attribute?**

$P(\text{positive}) = \frac{4}{9}$.

$P(\text{negative}) = \frac{5}{9}$

Entropy for positive class

$$= -\left(\frac{4}{9}\log_2\left(\frac{4}{9}\right) + \frac{5}{9}\log_2\left(\frac{5}{9}\right)\right)$$

$$= \boxed{0,991}$$

b. What are the information gains of a1 and a2 relative to these training examples?

$\boxed{a_1}$	+	-
T	3	1
F	1	4

entropy for a1 : $\frac{4}{9}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right)-\frac{1}{4}\log_2\left(\frac{1}{4}\right)\right]+\frac{5}{9}\left[-\frac{1}{5}\log_2\left(\frac{1}{5}\right)-\frac{4}{5}\log_2\left(\frac{4}{5}\right)\right]$

$= 0.762$

Information gain : $0.991-0.762$

$= \boxed{0.229}$

$\boxed{a_2}$	+	-
T	2	3
F	2	2

entropy for a2 : $\frac{5}{9}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]+\frac{4}{9}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)\right]$

$= 0.984$

Information gain : $0.991-0.984$

$= \boxed{0.007}$

DP

c. **For a3, which is a continuous attribute, compute the information gain for every possible split.**

a3: 1 3 4 5 6 7 8.

split pos.	0.5		2		3.5		4.5		5.5		6.5		7.5		8.5	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>.		
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
-	0	5	0	5	1	4	1	4	3	2	3	2.	4	1	5	0.

Split 1:

$$\text{Entropy} = -[(\frac{4}{9}\log_2(\frac{4}{9}) + \frac{5}{9}\log_2(\frac{5}{9}))]$$

$$= 0.991$$

$$\text{Infor. gain} = 0.991 - 0.991 = \boxed{0}$$

Split 2:

$$<= \text{Entropy} = -[1 \cdot \log_2(1) + 0 \cdot \log_2(0)] = 0.$$

$$> \text{Entropy} = -[\frac{3}{8} \cdot \log_2\frac{3}{8} + \frac{5}{8} \cdot \log_2\frac{5}{8}] = 0.954$$

~~Infor. gain=~~

$$\text{Info. gain} = 0.991 - (\frac{1}{9} \cdot 0 + \frac{8}{9} \cdot 0.954) = \boxed{0.143}$$

Split 3:

$$<= \text{Entropy} = -\left[\frac{1}{2}\cdot\log_2\frac{1}{2} + \frac{1}{2}\cdot\log_2\frac{1}{2}\right] = 1$$

$$> \text{Entropy} = -\left(\frac{3}{7}\cdot\log_2\frac{3}{7} + \frac{4}{7}\cdot\log_2\frac{4}{7}\right) = 0{,}985.$$

$$\text{Info gain} = 0{,}991 - \left(\frac{2}{9}\cdot 1 + \frac{7}{9}\cdot 0{,}985\right) =$$ [redacted] $$\boxed{0{,}00249}$$

Split 4:

$$<= \text{Entropy} = -\left(\frac{2}{3}\cdot\log_2\frac{2}{3} + \frac{1}{3}\cdot\log_2\frac{1}{3}\right) = 0{,}918$$

$$> \text{Entropy} = -\left(\frac{2}{6}\cdot\log_2\frac{2}{6} + \frac{4}{6}\cdot\log_2\frac{4}{6}\right) = 0{,}918.$$

$$\text{Info gain} = 0{,}991 - \left(\frac{3}{9}\cdot 0{,}918 + \frac{6}{9}\cdot 0{,}918\right) = \boxed{0{,}0727}$$

Split 5:

$$<= \text{Entropy} = -\left(\frac{2}{5}\cdot\log_2\frac{2}{5} + \frac{3}{5}\cdot\log_2\frac{3}{5}\right) = 0{,}971$$

$$> \text{Entropy} = -\left(\frac{2}{4}\cdot\log_2\frac{2}{4} + \frac{2}{4}\cdot\log_2\frac{2}{4}\right) = 1$$

$$\text{Info. gain} = 0.991 - \left(\frac{5}{9}\cdot 0.971 + \frac{4}{9}\cdot 1\right) = \boxed{0.00714}$$

Split 6 :

$<=$ Entropy $= -\left(\frac{3}{6} \cdot \log_2 \frac{3}{6} + \frac{3}{6} \cdot \log_2 \frac{3}{6}\right) = 1$

$>$ Entropy $= -\left(\frac{1}{3} \cdot \log_2 \frac{1}{3} + \frac{2}{3} \cdot \log_2 \frac{2}{3}\right) = 0{,}918.$

Info. gain $= 0{,}991 - \left(\frac{6}{9} \cdot 1 + \frac{3}{9} \cdot 0{,}918\right) = \boxed{0{,}0182}$

Split 7 :

$<=$ Entropy $= -\left(\frac{4}{8} \cdot \log_2 \frac{4}{8} + \frac{4}{8} \cdot \log_2 \frac{4}{8}\right) = 1$

$>$ Entropy $= -\left(0 \cdot \log_2 0 + 1 \cdot \log_2 1\right) = 0.$

Info. gain $= 0.991 - \left(\frac{8}{9} \cdot 1 + \frac{1}{9} \cdot 0\right) = \boxed{0.102}$

Split 8 :

$<=$ Entropy $= -\left(\frac{4}{9} \cdot \log_2 \frac{4}{9} + \frac{5}{9} \cdot \log_2 \frac{5}{9}\right) = 0{,}992.$

$>$ Entropy $= -\left(0 \cdot \log_2 0 + 0 \cdot \log_2 0\right) = 0.$

Info. gain $= \boxed{0.}$

DP

d. **What is the best split (among a1, a2 and a3) according to the information gain?**

$\boxed{a_1,}$ due to its higher gain of 0,229

e. What is the best split (between a1 and a2) according to the misclassification error rate?

Classification Error Rate :

$a_1 = 1 - (\frac{7}{9}, \frac{2}{9}) = 1 - \frac{7}{9} = \frac{2}{9}$. ✓

$a_2 = 1 - (\frac{5}{9}, \frac{4}{9}) = 1 - \frac{5}{9} = \frac{4}{9}$.

$\boxed{a_1}$ is the better split for its lower MER of $\frac{2}{9} \approx 0,223$

DP

f. **What is the best split (between a1 and a2) according to the Gini index?**

Gini Index :

$$a_1 = 1 - \left[\left(\frac{7}{9}\right)^2 + \left(\frac{2}{9}\right)^2 \right] = 0,346 . \quad ✓$$

$$a_2 = 1 - \left[\left(\frac{4}{9}\right)^2 + \left(\frac{5}{9}\right)^2 \right] = 0,494$$

$\boxed{a_1}$ is the best split for its lower GI.

of 0,346 .

10