DSAA 5002 - Data Mining and Knowledge Discovery in Data Science

(Fall Semester 2023)

Homework 2

JIANG Zhuoyang

1. (25 marks)Consider the following training data with labels 0 and 1, and three attributes A, B, and C.

$\frac{3}{9}$ $\frac{6}{9}$ $\frac{4}{7}$ $\frac{3}{7}$

id	A	B	C	class	By	Bn	Cy	Cn
1	0.62	yes	yes	0	0		0	
2	3.84	no	no	0		0		0
3	6.61	yes	no	0	0			0
4	6.87	yes	no	0	0			0
5	7.71	no	yes	0		0	0	
6	8.98	no	yes	0		0	0	
7	1.77	yes	no	0	0			0
8	2.02	yes	no	1	1			1
9	2.06	no	yes	1		1	1	
10	2.66	no	yes	1		1	1	
11	3.72	no	yes	1		1	1	
12	4.98	yes	yes	1	1		1	
13	5.73	yes	yes	1	1		1	
14	6.29	yes	yes	1	1		1	
15	9.08	no	no	1		1		1
16	9.45	no	no	1		1		1

(a) (10 marks) Try threshold 2, 5, and 8 for attributes A (that is, use the "A > 2, A < 2", "A > 5, A < 5", and "A > 8, A < 8" respectively). Use the Gini score to determine the best one θ_a among them. Recall

$$Gini(t) := 1 - \sum_{i=1}^{c} [p(i|t)]^2$$

$p(i|t) = \frac{p(i,t)}{p(t)} = \frac{|D_{t,i}|/\cancel{|D|}}{|D_t|/\cancel{|D|}} = \frac{|D_{t,i}|}{|D_t|}$

(b) (15 marks) Use θ_a obtained above, and the Gini score, determine which attributes should firstly be used for developing a decision tree.

Solution:

(a). 1° Sort the value A in increasing order:

D:

id	1	7	8	9	10	11	2	12	13	14	3	4	5	6	15	16
A	0.62	1.77	2.02	2.06	2.66	3.72	3.84	4.78	5.73	6.29	6.61	6.87	7.71	8.98	9.08	9.45
class	0	0	1	1	1	1	0	1	1	1	0	0	0	0	1	1

2° Calculate Information after using θ_a to split the training set D with Gini Score:

① if $\theta_a = 2$

$$Info_{A,2}(D) = \frac{|D_{A<2}|}{|D|} Gini(D_{A<2}) + \frac{|D_{A>2}|}{|D|} Gini(D_{A>2})$$

$$= \frac{2}{16}\cdot(1-1^2-0^2) + \frac{14}{16}(1-(\frac{5}{14})^2-(\frac{9}{14})^2) = 0.4018$$

② if $\theta_a = 5$.

$$Info_{A,5}(D) = \frac{|D_{A<5}|}{|D|} Gini(D_{A<5}) + \frac{|D_{A>5}|}{|D|} Gini(D_{A>5})$$

$$= \frac{8}{16}(1-(\frac{3}{8})^2-(\frac{5}{8})^2) + \frac{8}{16}(1-(\frac{4}{8})^2-(\frac{4}{8})^2) = 0.4844.$$

③ if $\theta_a = 8$.

$$Info_{A,8}(D) = \frac{|D_{A<8}|}{|D|} Gini(D_{A<8}) + \frac{|D_{A>8}|}{|D|} Gini(D_{A>8})$$

$$= \frac{13}{16}(1-(\frac{6}{13})^2-(\frac{7}{13})^2) + \frac{3}{16}(1-(\frac{2}{3})^2-(\frac{1}{3})^2) = 0.4872.$$

3° Because $Info_{A,8} > Info_{A,5} > Info_{A,2}$,

So that, when $\theta_a = 2$, it has the minimum expected information requirement (impurity), it is the best split point $\Rightarrow$

$\theta_a = 2$.

(b). 1° $Info(D) = Gini(D) = 1-(\frac{7}{16})^2-(\frac{9}{16})^2 = 0.4922$

2° ① for attribute A. we have:

不纯度降低最多 ⟺ 信息增益最大.

$Info_{A,2}(D) = 0.4018$.

$Gain_{A_2}(D) = Info(D) - Info_{A,2}(D) = 0.0904$.

② for attribute B, we have:

$$Info_B(D) = \frac{|D_{B=yes}|}{|D_B|}Gini(D_{B=yes}) + \frac{|D_{B=no}|}{|D_B|}Gini(D_{B=no})$$

$$= \frac{8}{16}(1-(\frac{4}{8})^2-(\frac{4}{8})^2) + \frac{8}{16}(1-(\frac{3}{8})^2-(\frac{5}{8})^2) = 0.4844.$$

$Gain_B(D) = Info(D) - Info_B(D) = 0.0078$

③ for attribute C. we have:

$$Info_C(D) = \frac{|D_{C=yes}|}{|D|}Gini(D_{C=yes}) + \frac{|D_{C=no}|}{|D|}Gini(D_{C=no})$$

$$= \frac{9}{16}(1-(\frac{3}{9})^2-(\frac{6}{9})^2) + \frac{7}{16}(1-(\frac{4}{7})^2-(\frac{3}{7})^2) = 0.4643$$

$Gain_C(D) = Info(D) - Info_C(D) = 0.0279$.

3° because $Gain_A(D) > Gain_C(D) > Gain_B(D)$.

so. attribut A should be firstly used for developing a decision tree.

2. (30 marks)The table below is a small part of the Acute Inflammations Data Set.

a1	Temperature of patient (35C-42C)
a2	Occurrence of nausea (yes, no)
a3	Lumbar pain (yes, no)
a4	Urine pushing (continuous need for urination) (yes, no)
a5	Micturition pains (yes, no)
a6	Burning of urethra, itch, swelling of urethra outlet (yes, no)
d1	Decision: Inflammation of urinary bladder (yes, no)
d2	Decision: Nephritis of renal pelvis origin (yes, no)

Here the attributes a1-a6 are observations, and the decisions d1 and d2 are made by a medical expert. The purpose of studying this data set is to predict presumptive diagnosis of two disease of the urinary system, namely, "Inflammation of urinary bladder" and "Nephritis of renal pelvis origin".

$\leftarrow d_1$

$\leftarrow d_2$

a1	a2	a3	a4	a5	a6	d1	d2
37.3	no	yes	no	no	no	no	no
37.4	no	no	yes	no	no	yes	no
37.5	yes	yes	no	no	no	no	no
37.6	no	no	yes	yes	yes	yes	yes
37.7	no	no	yes	no	no	yes	no
37.7	no	no	yes	yes	no	yes	no
37.7	no	no	yes	yes	no	yes	no
37.8	no	yes	no	no	no	no	no
37.9	no	no	yes	yes	yes	yes	no
37.9	no	no	yes	no	no	yes	no
38.0	no	yes	yes	no	yes	no	yes
38.0	no	yes	yes	no	yes	no	yes
38.1	no	yes	yes	no	yes	yes	yes
38.3	no	yes	yes	no	yes	no	yes
38.5	no	yes	yes	no	yes	no	no
38.7	no	yes	yes	no	yes	no	yes
38.9	no	yes	yes	no	yes	yes	yes
39.0	no	yes	yes	no	yes	no	yes
39.4	no	yes	yes	no	yes	no	yes
39.5	no	yes	yes	no	yes	no	yes

(a) (10 marks) Consider the procedures of building a decision tree with Gini score. If we plan only to use the attributes a3 and a5 to predict the decision d2, which attribute should we use first?

(b) (20 marks) Use the naïve Bayes algorithm, the attributes a1 (with the threshold $\theta_1 = 37.95$), a2, and a3 only, to predict the decision d2 for the following data of a new patient. (For simplicity you do NOT need to use the Laplacian correction.)

a1	a2	a3	a4	a5	a6	d1	d2
40.0	yes	no	no	no	no	?	?

Solution:

(a). 1° for d_2, we have:

$$Info(D_{d_2}) = Gini(D_{d_2}) = 1 - (\frac{10}{20})^2 - (\frac{10}{20})^2 = 0.5$$

2° ① for a_3,

$$Info_{a_3}(D_{d_2}) = \frac{|D_{d_2,a_3=y}|}{|D_{d_2}|} Gini(D_{d_2,a_3=y}) + \frac{|D_{d_2,a_3=n}|}{|D_{d_2}|} Gini(D_{d_2,a_3=n})$$

$$= \frac{13}{20}(1 - (\frac{9}{13})^2 - (\frac{4}{13})^2) + \frac{7}{20}(1 - (\frac{1}{7})^2 - (\frac{6}{7})^2) = 0.3626$$

$$Gain_{a_3}(D_{d_2}) = Info(D_{d_2}) - Info_{a_3}(D_{d_2}) = 0.1374$$

② for a_5,

$$Info_{a_5}(D_{d_2}) = \frac{|D_{d_2,a_5=y}|}{|D_{d_2}|} Gini(D_{d_2,a_5=y}) + \frac{|D_{d_2,a_5=n}|}{|D_{d_2}|} Gini(D_{d_2,a_5=n})$$

$$= \frac{4}{20}(1 - (\frac{1}{4})^2 - (\frac{3}{4})^2) + \frac{16}{20}(1 - (\frac{9}{16})^2 - (\frac{7}{16})^2) = 0.4688$$

$$Gain_{a_5}(D_{d_2}) = Info(D_{d_2}) - Info_{a_5}(D_{d_2}) = 0.0312.$$

3° Because $Gain_{a_3}(D_{d_2}) > Gain_{a_5}(D_{d_2})$,
so, attribute a_3 should be used first.

(b) 1° Based on naive bayes assumption.

$$P(d_2 | a_1, a_2, a_3) = \frac{P(a_1, a_2, a_3 | d_2) \cdot P(d_2)}{P(a_1, a_2, a_3)} \Leftrightarrow P(a_1, a_2 a_3, d_2)$$

$$\Leftrightarrow P(a_1 | d_2) \cdot P(a_2 | d_2) \cdot P(a_3 | d_2) \cdot P(d_2)$$

2° We have a new patient, has the syndrome: $a_1 > 37.95$, $a_2 = yes$, $a_3 = no.$, when we train the naive bayes model with the dataset, we will have:

① for a_1: $P(a_1 > 37.95 \mid d_2 = yes) = \frac{9}{10}$

$P(a_1 > 37.95 \mid d_2 = no) = \frac{1}{10}$

② for a_2: $P(a_2 = yes \mid d_2 = yes) = 0$

$P(a_2 = yes \mid d_2 = no) = \frac{1}{10}$

③ for a_3: $P(a_3 = no \mid d_2 = yes) = \frac{9}{10}$

$P(a_3 = no \mid d_2 = no) = \frac{6}{10}$

and $P(d_2 = yes) = \frac{1}{2}$, $P(d_2 = no) = \frac{1}{2}$.

3° So that, we can calculate with naive bayes assumption:

$P(a_1 > 37.5, a_2 = yes, a_3 = no, d_2 = yes)$

$= P(a_1 > 37.5 \mid d_2 = yes) \cdot P(a_2 = yes \mid d_2 = yes) \cdot P(a_3 = no \mid d_2 = yes) \cdot P(d_2 = yes)$

$= 0$

$P(a_1 > 37.5, a_2 = yes, a_3 = no, d_2 = no)$

$= P(a_1 > 37.5 \mid d_2 = no) \cdot P(a_2 = yes \mid d_2 = no) \cdot P(a_3 = no \mid d_2 = no) \cdot P(d_2 = no)$

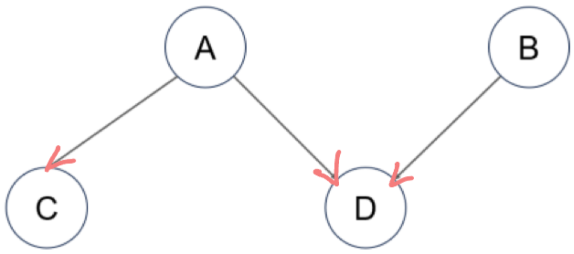
$= \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{6}{10} \cdot \frac{1}{2} = 0.003$

because $P(a_1 > 37.5, a_2 = yes, a_3 = no, d_2 = no) >$

$P(a_1 > 37.5, a_2 = yes, a_3 = no, d_2 = yes)$

so, we predicate the patient's d_2 to be 'no'.

3. (15 marks) There is a BBN below, which comprises four Random Variables(RV). Each RV is a Boolean RV.



$P(D|BC)$

$P(A) = 0.1$ $\quad$ $P(B) = 0.5$ $\quad$ $P(C|A) = 0.7$

$P(C|\neg A) = 0.2$ $\quad$ $P(D|A, B) = 0.9$ $\quad$ $P(D|\neg A, B) = 0.6$

$P(D|A, \neg B) = 0.7$ $\quad$ $P(D|\neg A, \neg B) = 0.3$

(a) (7 marks) What is $P(\neg A, B, \neg C, D)$?

(b) (8 marks) What is $P(A \mid B, C, D)$?

Solution:

(a). $P(\neg A, B, \neg C, D) = P(\neg A) \cdot P(B) \cdot P(\neg C | \neg A) \cdot P(D | \neg A, B)$

$= (1 - P(A)) \cdot P(B) \cdot (1 - P(C|\neg A)) \cdot P(D | \neg A, B)$

$= 0.9 \times 0.5 \times 0.8 \times 0.6 = 0.2160$

(b). $P(A, B, C, D) = P(A) \cdot P(B) \cdot P(C|A) \cdot P(D|A, B)$

$= 0.1 \times 0.5 \times 0.7 \times 0.9 = 0.0315$

$P(B, C, D) = \sum_A P(A, B, C, D)$

$= P(B) \cdot \sum_A P(A) \cdot P(C|A) \cdot P(D|A, B)$

$= P(B) \cdot [P(A) \cdot P(C|A) \cdot P(D|A, B) + P(\neg A) \cdot P(C|\neg A) P(D|\neg A, B)]$

$= 0.5 \times [0.1 \times 0.7 \times 0.9 + 0.9 \times 0.2 \times 0.6] = 0.0855$

$\Rightarrow P(A|B, C, D) = \dfrac{P(A, B, C, D)}{P(B, C, D)} = \dfrac{0.0315}{0.0855} = 0.3682$

4. (30 marks) Consider a simple neural network with a single hidden layer. The input layer consists of three dimensional $\boldsymbol{x} = (x1, x2, x3)^T$. The hidden layer includes two dimensional $\boldsymbol{h} = (h1, h2)$. The output layer includes one scalar o. We ignore bias terms for simplicity.

We use linear rectified (ReLU) as activation function **for the hidden and output layer BOTH**.

$$ReLU(x) = \max(0, x)$$

$$ReLU'(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

Moreover, denote the loss function (also called $error$ in slides) by $\boldsymbol{J}(o, t) = \frac{1}{2}|o - t|^2$ where t is the associated label (target) value for scalar output o.

Denote by W and V weight matrices connecting input and hidden layer, and hidden layer and output respectively. They are **initialized** (i.e., the initial condition before first updating round) as follows:

$W = \begin{bmatrix} 1 & 0 & 1 \\ -3 & -1 & 0 \end{bmatrix}$, $V = [0 \quad 1]$, Moreover, one training sample is $x = (-1,2,-1)^T$, $t = 0$.

Now, try to solve the following parts.

(a) (5 marks) Write out symbolically (thus, no need to plug in the specific values of W and V just yet) the mapping $x \to o$ using ReLU, W, V.

(b) (10 marks) Given the condition $\boldsymbol{x} = (1, 2, 1)^T$, $t = 1$, compute the numerical output value o, clearly show all intermediate steps. You can reuse the results of the previous question.

(c) (15 marks) Compute the gradient of the loss function with respect to the V weights, and evaluate the gradients at specific $\boldsymbol{x} = (1, 2, 1)^T$, $t = 1$.

Solution:

(a). $\vec{x} = (x_1, x_2, x_3)^T$ $W = \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \end{bmatrix}$ $V = [V_{11} \ V_{12}]$

$\vec{h} = (h_1, h_2)^T$

$$\Rightarrow \begin{cases} h_1 = W_{11}x_1 + W_{12}x_2 + W_{13}x_3 \Rightarrow y_1 = ReLU(h_1) \\ h_2 = W_{21}x_1 + W_{22}x_2 + W_{23}x_3 \Rightarrow y_2 = ReLU(h_2) \end{cases}$$

$$z = V_{11} \cdot y_1 + V_{12} \cdot y_2 \Rightarrow o = ReLU(z)$$

so that, $o = ReLU[V_{11} \cdot ReLU(W_{11}x_1 + W_{12}x_2 + W_{13}x_3) + V_{12} \cdot ReLU(W_{21}x_1 + W_{22}x_2 + W_{23}x_3)]$

$\Rightarrow \vec{o} = ReLU[V \cdot ReLU(W\vec{x})]$

(b) $\vec{x} = (1, 2, 1)^T$, $\vec{h} = W\vec{x} = \begin{bmatrix} 1 & 0 & 1 \\ -3 & -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -5 \end{bmatrix}$

$\vec{y} = ReLU(\vec{h}) = \begin{bmatrix} max(0, 2) \\ max(0, -5) \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$

$z = V \cdot \vec{y} = [0 \quad 1] \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 0$

$o = ReLU(z) = max(0, 0) = 0$

(c). $J(o, t) = \frac{1}{2} |o - t|^2$, $\frac{\partial J}{\partial V} = [\frac{\partial J}{\partial V_{11}}, \frac{\partial J}{\partial V_{12}}]$, $z = V \cdot \vec{y}$

$\frac{\partial J}{\partial V_{11}} = \frac{\partial J}{\partial z} \cdot \frac{\partial z}{\partial V_{11}} = \frac{\partial J}{\partial z} \cdot \frac{\partial (V_{11} y_1 + V_{12} y_2)}{\partial V_{11}} = \frac{\partial J}{\partial z} \cdot y_1$

similarly, $\frac{\partial J}{\partial V_{12}} = \frac{\partial J}{\partial z} \cdot y_2$, let $\delta_v = \frac{\partial J}{\partial z}$, we have

$\frac{\partial J}{\partial V_{11}} = \delta_v y_1$. $\frac{\partial J}{\partial V_{12}} = \delta_v y_2$.

$\delta_v = \frac{\partial J}{\partial z} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial z} = \frac{\partial (\frac{1}{2}(o-t)^2)}{\partial o} \cdot \frac{\partial (ReLU(z))}{\partial z}$

$= (o - t) \cdot ReLU'(z) = \begin{cases} (o-t), & z > 0 \\ 0, & z \leq 0. \end{cases}$

$\Rightarrow \frac{\partial J}{\partial V_{11}} = \begin{cases} (o-t) \cdot y_1, & z > 0 \\ 0, & z \leq 0 \end{cases}$; $\frac{\partial J}{\partial V_{12}} = \begin{cases} (o-t) \cdot y_2, & z > 0 \\ 0, & z \leq 0 \end{cases}$

$\frac{\partial J}{\partial V} = [\frac{\partial J}{\partial V_{11}}, \frac{\partial J}{\partial V_{12}}]$ when $\vec{x} = (1, 2, 1)$, based on (b)

we have $z = 0$, so $\frac{\partial J}{\partial V} = [0, 0]$