DSAA 5002 - HW2

50015976 Ruiming ZHANG

Q1 [25 Marks]

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

Solution:

1	2	3
1	2	3
4	5	6

Thus we get the candidate 3-itemset C_3.

Q2 [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”.

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 θ_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) Because we only predict the decision d2, we have: $Info(T) = 1 - {\frac{1}{2}}^2 - {\frac{1}{2}}^2 = 0.5$

For attribute a3, we have:

$Info(T_{yes}) = 1 - {\frac{9}{13}}^2 - {\frac{4}{13}}^2 = \frac{72}{169} \approx 0.4260$

$Info(T_{no}) = 1 - {\frac{1}{7}}^2 - {\frac{6}{7}}^2 = \frac{12}{49} \approx 0.2499$

$Info(a3, T) = \frac{13}{20} \times \frac{72}{169} + \frac{7}{20} \times \frac{12}{49} = \frac{33}{91} \approx 0.3626$

$Gain(a3, T) = 0.5 - 0.3626 = 0.1374$

For attribute a5, we have:

$Info(T_{yes}) = 1 - \frac{1}{4}^2 - \frac{3}{4}^2 = \frac{3}{8} = 0.3750$

$Info(T_{no}) = 1 - \frac{9}{16}^2 - \frac{7}{16}^2 = \frac{63}{128} \approx 0.4922$

$Info(a5, T) = \frac{4}{20} \times \frac{3}{8} + \frac{16}{20} \times \frac{63}{128} = \frac{15}{32} \approx 0.4688$

$Gain(a5, T) = 0.5 - 0.4688 = 0.0312$

We can see $Gain(a3, T) > Gain(a5, T)$, so we should use a3 first.

(b) Because we have:

For attribute a1:

$P(a1 > \theta_1 | d2 = yes) = \frac{9}{10} = 0.9$

$P(a1 < \theta_1 | d2 = yes) = \frac{1}{10} = 0.1$

$P(a1 > \theta_1 | d2 = no) = \frac{1}{10} = 0.1$

$P(a1 < \theta_1 | d2 = no) = \frac{9}{10} = 0.9$

For attribute a2:

$P(a2 = yes | d2 = yes) = \frac{0}{10} = 0$

$P(a2 = no | d2 = yes) = \frac{10}{10} = 1$

$P(a2 = yes | d2 = no) = \frac{1}{10} = 0.1$

$P(a2 = no | d2 = no) = \frac{9}{10} = 0.9$

For attribute a3:

$P(a3 = yes | d2 = yes) = \frac{9}{10} = 0.9$

$P(a3 = no | d2 = yes) = \frac{1}{10} = 0.1$

$P(a3 = yes | d2 = no) = \frac{4}{10} = 0.4$

$P(a3 = no | d2 = no) = \frac{6}{10} = 0.6$

Then we have:

$$\begin{aligned} P &= P(a1 > \theta_1, a2 = yes, a3 = no) \\ &= P(a1 > \theta_1) \times P(a2 = yes) \times P(a3 = no) \\ &= 0.00875 \end{aligned}$$

$$\begin{aligned} P_1 &= P(a1 > \theta_1, a2 = yes, a3 = no | d2 = yes) \\ &= P(a1 > \theta_1 | d2 = yes) \times P(a2 = yes | d2 = yes) \times P(a3 = no | d2 = yes) \\ &= 0.9 \times 0 \times 0.1 \\ &= 0 \end{aligned}$$

$$
\begin{aligned}
P_2 &= P(a1 > \theta_1, a2 = yes, a3 = no | d2 = no) \\
&= P(a1 > \theta_1 | d2 = no) \times P(a2 = yes | d2 = no) \times P(a3 = no | d2 = no) \\
&= 0.1 \times 0.1 \times 0.6 \\
&= 0.006
\end{aligned}
$$

Thus:

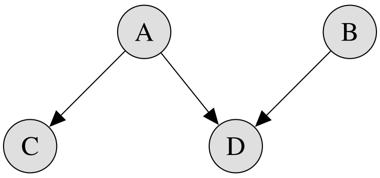
$$
\begin{aligned}
&P(d2 = yes | a1 > \theta_1, a2 = yes, a3 = no) \\
&= \frac{P_1 \times P(d2 = yes)}{P} \\
&= \frac{0 \times 0.5}{0.00875} \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
&P(d2 = no | a1 > \theta_1, a2 = yes, a3 = no) \\
&= \frac{P_2 \times P(d2 = no)}{P} \\
&= \frac{0.006 \times 0.5}{0.00875} \\
&= 0.3429
\end{aligned}
$$

Hence, the decision of d2 is no.

Q3 [15 Marks]

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



$P(A) = 0.1$	$P(B) = 0.5$	$P(C\|A) = 0.7$
$P(C\|\neg A) = 0.2$	$P(D\|A, B) = 0.9$	$P(D\|\neg A, B) = 0.6$
$P(D\|A, \neg B) = 0.7$	$P(D\|\neg A, \neg B) = 0.3$	

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

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

Solution:

(a) According to the BBN above, we have:

$$\begin{aligned}P(\neg A, B, \neg C, D) &= P(B, D|\neg A, \neg C) \times P(\neg A, \neg C)\\ &= P(B, D|\neg A) \times P(\neg C|\neg A) \times P(\neg A)\\ &= P(D|\neg AB) \times P(B) \times [1 - P(C|\neg A)] \times [1 - P(A)]\\ &= 0.6 \times 0.5 \times (1 - 0.2) \times (1 - 0.1)\\ &= 0.216\end{aligned}$$

(b) According to the BBN above, we have:

$$\begin{aligned}P(A|B, C, D) &= \frac{P(A, B, C, D)}{P(B, C, D)}\\ &= \frac{P(B, D|A, C) \times P(A, C)}{P(B, C, D)}\\ &= \frac{P(B, D|A) \times P(A, C)}{P(B, D) \times P(C)}\\ &= \frac{P(D|A, B) \times P(B) \times [P(C|A) \times P(A)]}{[P(B, D, A) + P(B, D, \neg A)] \times P(C)}\\ &= \frac{P(D|A, B) \times P(B) \times [P(C|A) \times P(A)]}{[P(D|A, B) \times P(A, B) + P(D|\neg A, B) \times P(\neg A, B)] \times P(C)}\end{aligned}$$

Except for $P(C)$, all other variables are known. According to Total Probability Formula, we have:

$$\begin{aligned}P(C) &= P(C, A) + P(C, \neg A)\\ &= P(C|A) \times P(A) + P(C|\neg A) \times P(\neg A)\\ &= 0.7 \times 0.1 + 0.2 \times 0.9\\ &= 0.25\end{aligned}$$

Hence:

$$P(A|B, C, D) = \tfrac{0.9 \times 0.5 \times [0.7 \times 0.1]}{[0.9 \times 0.1 \times 0.5 + 0.6 \times 0.9 \times 0.5] \times 0.25} = 0.4$$

Q4 [30 Marks]

Consider a simple neural network with a single hidden layer. The input layer consists of three-dimensional $\mathbf{x} = (x_1, x_2, x_3)^T$. The hidden layer includes two-dimensional $\mathbf{h} = (h_1, h_2)$. 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**.

$$\text{ReLU}(x) = \max(0, x)$$

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

Moreover, denote the loss function (also called error in slides) by $\mathbf{J}(o, t) = \frac{1}{2}|o - t|^2$. 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 the first updating round) as follows:

$$W = \begin{bmatrix} 1 & 0 & 1 \\ -3 & -1 & 0 \end{bmatrix}, V = \begin{bmatrix} 0 & 1 \end{bmatrix}, \tag{1}$$

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 $\mathbf{x} \to o$ using ReLU, W, V.

(b) (10 marks) Given the condition $\mathbf{x} = (1, 2, 1)^T, t = 1$, 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 $\mathbf{x} = (1, 2, 1)^T, t = 1$.

Forward Pass

1. Hidden Layer Calculation:

The output of the hidden layer can be computed as $\mathbf{h} = \text{ReLU}(W \cdot \mathbf{x})$.

Given $W = \begin{pmatrix} 1 & 0 & 1 \\ -3 & -1 & 0 \end{pmatrix}$ and $\mathbf{x} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$,

$$\mathbf{h} = \text{ReLU}\left(\begin{pmatrix} 1 & 0 & 1 \\ -3 & -1 & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}\right) = \text{ReLU}\left(\begin{pmatrix} 2 \\ -5 \end{pmatrix}\right) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$$

2. Output Layer Calculation:

The output o can be computed as $o = \text{ReLU}(V \cdot \mathbf{h})$.

Given $V = \begin{pmatrix} 0 & 1 \end{pmatrix}$ and $\mathbf{h} = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$,

$$o = \text{ReLU}\left(\begin{pmatrix} 0 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 0 \end{pmatrix}\right) = \text{ReLU}(0) = 0$$

So, the numerical output value o would be 0 given the condition $\mathbf{x} = (1, 2, 1)^T$ and $t = 1$.

To compute the gradient of the loss function with respect to the V weights, we first need to define the loss function and then use backpropagation to find the gradients.

Loss Function: The given loss function is

$$J(o, t) = \frac{1}{2}|o - t|^2$$

where o is the predicted output and t is the target value.

Backpropagation for Computing Gradient: We want to find $\frac{\partial J}{\partial V}$.

1. First, compute $\frac{\partial J}{\partial o}$ (the derivative of the loss function with respect to the output):

$$\frac{\partial J}{\partial o} = o - t$$

2. Next, compute $\frac{\partial o}{\partial V}$ (the derivative of the output with respect to the V weights). Since $o = \text{ReLU}(V \cdot h)$, and $h = (2, 0)^T$ (from our earlier calculation), $\frac{\partial o}{\partial V} = h$ if $o > 0$ and $\frac{\partial o}{\partial V} = 0$ otherwise.

3. Finally, use the chain rule to find $\frac{\partial J}{\partial V}$:

$$\frac{\partial J}{\partial V} = \frac{\partial J}{\partial o} \times \frac{\partial o}{\partial V}$$

Evaluate the Gradients: 1. $\frac{\partial J}{\partial o} = o - t = 0 - 1 = -1$ 2. $\frac{\partial o}{\partial V} = (2, 0)^T$ (because $o = 0$ which is not greater than 0) 3. $\frac{\partial J}{\partial V} = -1 \times (2, 0)^T = (-2, 0)$

So, the gradient of the loss function with respect to the V weights is $(-2, 0)$.