In all the computations, if you need to round a number, please keep four decimal places, like

$$\frac{1000}{7} = 142.8571, \qquad \frac{0.02}{7} = 0.0029$$

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

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

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

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

class 0: 7 samples (1, 2, 3, 4, 5, 6, 7)
class 1: 9 samples (8, 9, 10, 11, 12, 13, 14, 15, 16)
$Info(T) = 1 - (\frac{7}{16})^2 - (\frac{9}{16})^2 = 0.4922$

(a) ① Threshold 2:

<2 : 2 samples

id		class
1	A = 0.62	0
7	A = 1.77	0

>2 : 14 samples

id	class	id	class
2	0	8	1
3	0	9	1
4	0	10	1
5	0	11	1
6	0	12	1
		13	1
		14	1
		15	1
		16	1

$Info(A<2) = 1 - (\frac{2}{2})^2 - (\frac{0}{2})^2 = 0$
$Info(A>2) = 1 - (\frac{5}{14})^2 - (\frac{9}{14})^2 = 0.4592$

$Info(A, T=2) = \frac{2}{16} \times 0 + \frac{14}{16} \times 0.4592 = 0.4018$

$Gain(A, T=2) = Info(T) - Info(A, T=2) = 0.0904$

② Threshold 5

<5: 8 samples

id	class	id	class
1	0	8	1
2	0	9	1
7	0	10	1
		11	1
		12	1

>5: 8 samples

id	class	id	class
3	0	13	1

$Info(A<5) = 1 - (\frac{3}{8})^2 - (\frac{5}{8})^2 = 0.46875$

$Info(A>5) = 1 - (\frac{4}{8})^2 - (\frac{4}{8})^2 = 0.5$

$Info(A, T=5) = \frac{1}{2} \times 0.46875 + \frac{1}{2} \times 0.5 = 0.484375$

4	0	14	1
5	0	15	1
6	0	16	1

$Gain(A, T=5) = Info(T) - Info(A, T=5) = 0.0078$

③ Threshold 8:

<8: 13 samples

id	class	id	class
1	0	8	1
2	0	9	1
3	0	10	1
4	0	11	1
5	0	12	1
7	0	13	1
		14	1

>8: 3 samples

id	class	id	class
6	0	15	1
		16	1

$Info(A<8) = 1 - (\frac{6}{13})^2 - (\frac{7}{13})^2 = 0.4970$

$Info(A>8) = 1 - (\frac{1}{3})^2 - (\frac{2}{3})^2 = 0.4444$

$Info(A, T=8) = \frac{13}{16} \times 0.4970 + \frac{3}{16} \times 0.4444 = 0.4871375$

$Gain(A, T=8) = Info(T) - Info(A, T=8) = 0.005$

$\therefore \theta_a = 2$

(b) For attribute B.

Yes: 8 samples

id	class	id	class
1	0	8	1
3	0	12	1
4	0	13	1
7	0	14	1

No: 8 samples

id	class	id	class
2	0	9	1
5	0	10	1
6	0	11	1
		15	1
		16	1

$Info(T_{Yes}) = 1 - (\frac{1}{2})^2 - (\frac{1}{2})^2 = 0.5$

$Info(T_{No}) = 1 - (\frac{3}{8})^2 - (\frac{5}{8})^2 = 0.46875$

$Info(A, T) = \frac{1}{2} \times 0.46875 + \frac{1}{2} \times 0.5 = 0.484375$

$Gain(A, T) = Info(T) - Info(A, T) = 0.0078$

For attribute C.

Yes: 9 samples

id	class	id	class
1	0	9	1
5	0	10	1
6	0	11	1
		12	1
		13	1
		14	1

No: 7 samples

id	class	id	class
2	0	8	1
3	0	15	1
4	0	16	1
7	0		

$Info(T_{Yes}) = 1 - (\frac{2}{9})^2 - (\frac{7}{9})^2 = 0.4444$

$Info(T_{No}) = 1 - (\frac{4}{7})^2 - (\frac{3}{7})^2 = 0.4898$

$Info(A, T) = \frac{9}{16} \times 0.4444 + \frac{7}{16} \times 0.4898 = 0.4643$

$Gain(A, T) = Info(T) - Info(A, T) = 0.0279$

$\because 0.0904 > 0.0279 > 0.0078$

$\therefore$ firstly A attribute.

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".

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	?	?

(a) For attribute a_3:

Yes: 13 samples

When a_3 = yes, there are 9 "yes" and 4 "no" for d_2.

calculate the Gini score: $1-(\frac{9}{13})^2-(\frac{4}{13})^2 = 0.4260$

No: 7 samples

When a_3 = no, there is 1 "yes" and 6 "no" for d_2.

calculate the Gini score: $1-(\frac{1}{7})^2-(\frac{6}{7})^2 = 0.2449$

$Info(a_3, T) = \frac{13}{20} \times 0.4260 + \frac{7}{20} \times 0.2449 = 0.3626$

For attribute a_5:

Yes: 4 samples

When a_5 = yes, there is 1 "yes" and 3 "no" for d_2.

calculate the Gini score: $1-(\frac{1}{4})^2-(\frac{3}{4})^2 = 0.375$

No: 16 samples

When a_5 = no, there is 9 "yes" and 7 "no" for d_2.

calculate the Gini score: $1-(\frac{9}{16})^2-(\frac{7}{16})^2 = 0.4922$

$Info(a_5, T) = \frac{4}{20} \times 0.375 + \frac{16}{20} \times 0.4922 = 0.4688$

$\because Info(a_3, T) < Info(a_5, T)$

$\therefore Info(T) - Info(a_3, T) > Info(T) - Info(a_5, T)$

$Gain(a_3, T) > Gain(a_5, T)$

$\therefore$ we use a_3 first.

(b) $P(d_2 = yes) = \frac{10}{20} = 0.5$

$P(d_2 = no) = \frac{10}{20} = 0.5$

For attribute a_1: $P(a_1 < \theta_1 \mid yes) = \frac{1}{10} = 0.1$ $\quad P(a_1 < \theta_1 \mid no) = 0.9$

$P(a_1 > \theta_1 \mid yes) = 0.9$ $\quad P(a_1 > \theta_1 \mid no) = 0.1$

For attribute a_2: $P(a_2=yes|yes)=0$ $P(a_2=yes|no)=0.1$

$P(a_2=no|yes)=1$ $P(a_2=no|no)=0.9$

For attribute a_3: $P(a_3=yes|yes)=0.9$ $P(a_3=yes|no)=0.4$

$P(a_3=no|yes)=0.1$ $P(a_3=no|no)=0.6$

$P(a_1>\theta_1, a_2=yes, a_3=no|yes)=P(a_1>\theta_1|yes)\times P(a_2=yes|yes)\times P(a_3=no|yes)=0.9\times 0\times 0.1=0$

(a_1, a_2, a_3 are conditionally independently)

$P(a_1>\theta_1, a_2=yes, a_3=no|no)=P(a_1>\theta_1|no)\times P(a_2=yes|no)\times P(a_3=no|no)=0.1\times 0.1\times 0.6=0.006$

$P(yes|a_1>\theta_1, a_2=yes, a_3=no)=\frac{P(a_1>\theta_1, a_2=yes, a_3=no|yes)\,P(d_2=yes)}{P(a_1>\theta_1, a_2=yes, a_3=no)}=0$

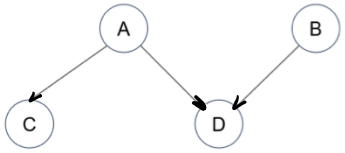
$P(no|a_1>\theta_1, a_2=yes, a_3=no)=\frac{P(a_1>\theta_1, a_2=yes, a_3=no|no)\,P(d_2=no)}{P(a_1>\theta_1, a_2=yes, a_3=no)}=\frac{0.003}{P(a_1>\theta_1, a_2=yes, a_3=no)}>0$

Since $P(no|a_1>\theta_1, a_2=yes, a_3=no) > P(yes|a_1>\theta_1, a_2=yes, a_3=no)$

we predict d_2 is 'no' for the following data for a new patient.

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

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



$P(\bar{A}, B, \bar{C}, D)=P(\bar{C}|\bar{A}, B, D)\cdot P(\bar{A}, B, D)$

$=P(\bar{C}|\bar{A})\cdot P(\bar{A}, B, D)$

$=[1-P(C|\bar{A})]\cdot P(D|\bar{A}, B)\cdot P(\bar{A}, B)$

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

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

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

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

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

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

Since A node is conditionally independent of its non-descendants if its parents are know.

∴ A and B, C and D. B and C are independent.

(a) $P(\bar{A}, B, \bar{C}, D)=P(D|\bar{A}, B, \bar{C})\cdot P(\bar{C}|\bar{A}, B)\cdot P(B|\bar{A})\cdot P(\bar{A})$

$=P(D|\bar{A}, B)\cdot P(\bar{C}|\bar{A})\cdot P(B)\cdot P(\bar{A})$

$=P(D|\bar{A}, B)\cdot[1-P(C|\bar{A})]\cdot P(B)\cdot[1-P(A)]$

$=0.6\times(1-0.2)\times 0.5\times(1-0.1)$

$=0.216$

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

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

$=0.9\times 0.7\times 0.5=0.315$

$P(B, C, D|\bar{A})=P(D|\bar{A}, B, C)\cdot P(C|\bar{A}, B)\cdot P(B|\bar{A})$

$=P(D|\bar{A}, B)\cdot P(C|\bar{A})\cdot P(B)$

$=0.6\times 0.2\times 0.5=0.06$

$P(B, C, D)=P(B, C, D|A)\cdot P(A)+P(B, C, D|\bar{A})\cdot P(\bar{A})$

$=0.0315+0.054$

$=0.0855$

$P(A|B, C, D)=\frac{P(B, C, D|A)\cdot P(A)}{P(B, C, D)}=\frac{0.0315}{0.0855}=0.3684$

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 $\boldsymbol{x} = (1, 2, 1)^T$, $t = 1$.

(a) $h = ReLU(Wx) = \max(0, Wx)$

$o = ReLU(Vh)$

where: $x = (x_1, x_2, x_3)^T$ is the input vector

$h = (h_1, h_2)$ is the hidden layer vector

o is the output scalar

W is the weight matrix connecting the input and hidden layer

$V = [V_1 \; V_2]$ is the weight matrix connecting hidden layer and output

(b) To compute the numerical output value o.

Step 1: computing the hidden layer

$$h = ReLU(Wx) = ReLU\left(\begin{bmatrix} 1 & 0 & 1 \\ -3 & -1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}\right) = ReLU([2, -5]^T) = [2, 0]^T$$

Step 2: computing the output

$$o = ReLU(Vh) = ReLU([0 \; 1] \cdot [2, 0]^T) = ReLU(0) = 0$$

Therefore, the numerical output value o is 0

(c) step 3: compute the loss function

$$J(o, t) = \frac{1}{2}|o - t|^2 = \frac{1}{2}|0 - 1|^2 = 0.5$$

step 4: compute the partial derivatives of the loss function with the respect to the elements of V.

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

$$\frac{\partial J}{\partial V_i} = \frac{\partial J}{\partial o} \cdot \frac{\partial o}{\partial net} \cdot \frac{\partial net}{\partial V_i} \quad i = 1, 2$$

$$= \begin{cases} 0, & net \leq 0 \\ (o - t) \cdot 1 \cdot h_i, & net > 0. \end{cases}$$

for $x = (1, 2, 1)^T$, $t = 1$, $net = 0$ $\quad \frac{\partial J}{\partial V_1} = 0$. $\frac{\partial J}{\partial V_2} = 0$

$\therefore \frac{\partial J}{\partial V} = [0 \; 0]$