10601 Machine Learning HW2

Mingyang Song

(1) Did you *receive* any help whatsoever from anyone in solving this assignment? No.

(2) Did you *give* any help whatsoever to anyone in solving this assignment? No.

Q1.

In the famous smart phone game Fruity Ninjas, random fruits appear on your screen and you are required to slice them. There is a steady stream of these random fruits. The only fruits that appear are Watermelons and Oranges. The game randomly generates a watermelon with a prob 0.7 and an orange with a prob 0.3.

(a) How surprised are you (in bits) to observe

- A Watermelon [1 pt]

I[Observe a watermelon]=log $\frac{1}{p(watermelon)}$ =log $\frac{1}{0.7}$ =0.5146bits

- An Orange [1 pt]

I[Observe an orange]=log $\frac{1}{p(orange)}$ =log $\frac{1}{p(0.3)}$ =1.7370bits

(b) What is the average information content of a game during which fruit shows up on the screen 100 times?[2 Pts]

E[100times game]=100*(p(watermelon)log $\frac{1}{p(watermelon)}$ +p(orange)log $\frac{1}{p(orange)}$)=88.129bits

Q2.

On the roll of two six-sided fair dice,

(a) Calculate the distribution of the sum (S) of the total. [2 pts]

Distribution List.

S	2	3	4	5	6	7	8	9	10	11	12
P	1/36	1/18	1/12	1/9	5/36	1/6	5/36	1/9	1/12	1/18	1/36

(b) How surprised are you (in bits) to observe S=3, S=12, S=6, S=8. [.5+.5+.5+.5 pts]

I[S=3]=log $\frac{1}{p(3)}$ =4.1699bits

I[S=12]=log $\frac{1}{p(12)}$ =5.1699bits

I[S=6]=log $\frac{1}{p(6)}$ =2.8480bits

I[S=8]=log $\frac{1}{p(8)}$ =2.8480bits

(c) Calculate the entropy of S. [2 pts]

H(S)= $-\sum_{S} P(s_i)\log P(s_i)$ =-

(1/36*log1/36+1/18*log1/18+1/12*log1/12+1/9*log1/9+5/36*log5/36+1/6*log1/6+5/36*log5/36+1/9*log1/9+1/12*log1/12+1/18*log1/18+1/36*log1/36)=3.2744bits

(d) Lets say you throw the die one at a time, and the first die shows 5. What is the entropy of S after this observation. Did the uncertainty in S change in the process of observing the outcome of the first die? If so, calculate the change (in bits), and whether it was positive or negative? [3+1 pts]

Because the two die results are independent, so that if the first observation is decided, it will not affect the second. Therefore, H(S|first die=5)= $-\sum_{1}^{6} p(x_i)\log p(x_i)$ =-6*(1/6*log1/6)=2.5850bits.

However, if the observing is changing, the uncertainty in S(that is entropy of S) will change.

H(change)=H(S|first ide=5)-H(S)=2.5850-3.2744=-0.6894bits.

It's negative and means that when we know one die the information for total sum will decrease because this one die will have some information contribution to total sum.

Q3.

An expedition of Pittsburgh Mountaineers heads up Mount Everest to study the effects of altitude on human physiology. Mountaineer Deep volunteers to make a series of recordings of his Blood Oxygen Level(O) and Pulse Rate(P) numbers once he reaches Death Zone(> 26,000ft). Back in Pittsburgh, Deep analyses the data and finds that his readings fluctuate somewhat. Deep decides to treat this uncertainty in readings as a probability distribution. He groups the O and P readings into CRITICAL, ABNORMAL and NORMAL categories and creates the following probability table.

	P=CRITICAL	P=ABNORMAL	P=NORMAL
O=CRITICAL	0.05	0.025	0.0
O=ABNORMAL	0.025	0.4	0.2
O=NORMAL	0.0	0.2	0.1

(a) Calculate the marginal distributions Pr(P), Pr(O). [2+2 pts] Hint: Use P(X = x) = $\sum_Y P(X = x, Y = y)$

Pr(P): P(P=CRITICAL)= $\sum_O P(P = CRITICAL, O = o)$ =0.05+0.025+0.0=0.075

P(P=ABNORMAL)= $\sum_O P(P = ABNORMAL, O = o)$ =0.025+0.4+0.2=0.625

P(P=NORMAL)= $\sum_O P(X = NORMAL, O = o)$ =0.0+0.2+0.1=0.3

$$\Pr(\mathrm{P}) = \begin{cases} 0.075, CRITICAL \\ 0.625, ABNORMAL \\ 0.3, NORMAL \end{cases}$$

Pr(O):P(O=CRITICAL)= $\sum_P P(O = CRITICAL, P = p)$ =0.05+0.025+0.0=0.075

P(O=ABNORMAL)= $\sum_P P(O = ABNORMAL, P = p)$ =0.025+0.4+0.2=0.625

P(O=NORMAL)= $\sum_P P(O = NORMAL, P = p)$ =0.0+0.2+0.1=0.3

$$\Pr(\mathrm{O}) = \begin{cases} 0.075, CRITICAL \\ 0.625, ABNORMAL \\ 0.3, NORMAL \end{cases}$$

(b) Calculate H(P), H(O). [2+2 pts]

H(P)= $-\sum_P p(p_i)\log p(p_i)$ =1.2252bits

H(O)= $-\sum_O p(o_i)\log p(o_i)$ =1.2252bits

(c) Calculate H(P|O), H(O|P). [5+5 pts]

H(P|O=CRITICAL)=H(0.05/0.075,0.025/0.075)=H(0.667, 0.0333)=0.9180bits

H(P|O=ABNORMAL)=H(0.025/0.625,0.4/0.625,0.2/0.625)=1.1239bits

H(P|O=NORMAL)=H(0.2/0.3,0.1/0.3)=0.9183bits

H(P|O)=p(O=CRITICAL)H(P|O=CRITICAL)+p(O=ABNORMAL)H(P|O=ABNORMAL)+p(O=NORMAL)H(P|O=NORMAL)=0.075*0.9180+0.625*1.1239+0.3*0.9183=1.0468bits

H(O|P)=p(P=CRITICAL)H(O|P=CRITICAL)+p(P=ABNORMAL)H(O|P=ABNORMAL)+p(P=NORMAL)H(O|P=NO

RMAL)=1.0468bits

Q4.

Prove that for any 3 discrete random variables A,X,Y the following properties hold,

(a) H(X)-H(X|Y) = H(Y)-H(Y|X). [5 pts]

(b) H(A,Y|X) = H(Y|X) + H(A|Y,X). [5 pts]

Explain intuitively in one sentence each what the above properties mean. [2 pts]

(a) H(X)-H(X|Y) = H(Y)-H(Y|X).

H(X)+H(Y|X)= $-\sum_{i=1}^{N} p(x_i)\log p(x_i) - \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i)$

$= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(x_i) - \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i)$

$= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log\left[p(x_i)\,p(y_j \mid x_i)\right]$

$p(xi)p(y_j|x_i)=p(x_iy_j)=p(y_j)p(x_i|y_j)$, then

H(X)+H(Y|X)= $-\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log\left[p(y_j)\,p(x_i \mid y_j)\right]$

=H(Y)+H(X|Y)

Therefore, H(X)-H(X|Y)=H(Y)-H(Y|H)

The meaning for this property means the order of two events does not matter the change between information of one event and information of one event when knowing the other.

(b) H(A,Y|X) = H(Y|X) + H(A|Y,X). [5 pts]

H(A,Y|X)= $-\sum_{k=1}^{L}\sum_{j=1}^{M}\sum_{i=1}^{N} p(a_k y_j x_i)\log p(a_k y_j \mid x_i)$

H(Y|X)+H(A|Y,X)= $-\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i) - \sum_{k=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{M} p(a_k x_i y_j)\log p(a_k \mid y_j x_i)$

$= -\sum_{k=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{M} p(a_k x_i y_j)\log p(y_j \mid x_i) - \sum_{k=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{M} p(a_k x_i y_j)\log p(a_k \mid y_j x_i)$

$$= -\sum_{k=1}^{L}\sum_{i=1}^{N}\sum_{j=1}^{M} p(a_k x_i y_j)\log\left[p(y_j \mid x_i)\,p(a_k \mid y_j x_i)\right]$$

$$= -\sum_{k=1}^{L}\sum_{j=1}^{M}\sum_{i=1}^{N} p(a_k y_j x_i)\log p(a_k y_j \mid x_i)$$

Therefore, H(A,Y|X) = H(Y|X) + H(A|Y,X).

The meaning for this property is that when knowing the first event, the information of joint other two events equals to sum of information of second event when knowing the first event and information of third event when knowing the first two events.

Q5.

Prove that entropy is always non-negative. [4 pts]

Based on the defination, $H(X) = -\sum_{x} P(x)\log P(x)$

Because $0 \le P(x) \le 1$ then $\log P(x) \le 0$. So that $H(X) \ge 0$.

Q6.

Prove that for any two variables X,Y

(a) H(X,Y) ≤ H(X) + H(Y) (slide 14) [7 pts]

(b) If X,Y are independent i.e. P(X, Y) = P(X)P(Y), then H(X, Y) = H(X) + H(Y). [7 pts]

Hint: Use the fact that I(X,Y) ≥ 0.

(a) $H(X,Y) = -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(x_i y_j)$

$$= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log\left[p(x_i)\,p(y_j \mid x_i)\right]$$

$$= -\sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(x_i) - \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i)$$

$$= -\sum_{i=1}^{N} p(x_i)\log p(x_i) - \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i)$$

=H(X)+H(Y|X)

Therefore, the question transfer to prove that $H(X)+H(Y|X) \leq H(X) + H(Y)$. As we talked in the class, that $H(Y|X) \leq H(Y)$ because if we know X we will have some information about Y so that the entropy of Y is bigger than that of Y when knowing X.

Therefore $H(X,Y) \leq H(X) + H(Y)$.

(b)According to above derivation , we know that

$$H(X,Y)= -\sum_{i=1}^{N} p(x_i)\log p(x_i) - \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i)$$

Because Y and X are independent and P(X,Y)=P(X)P(Y), P(Y|X)=P(Y). Then,

$$-\sum_{i=1}^{N} p(x_i)\log p(x_i) - \sum_{i=1}^{N}\sum_{j=1}^{M} p(x_i y_j)\log p(y_j \mid x_i)$$

$$= -\sum_{i=1}^{N} p(x_i)\log p(x_i) - \sum_{i=1}^{N} p(x_i)\sum_{j=1}^{M} p(y_j)\log p(y_j)$$

$$= -\sum_{i=1}^{N} p(x_i)\log p(x_i) - \sum_{j=1}^{M} p(y_j)\log p(y_j)$$ ($\sum_{i=1}^{N} p(x_i)$ does not affect Y)

=H(X)+H(Y)

Q7.

Let X and Y be discrete random variables which are identically distributed (so H(X) = H(Y)) but not necessarily independent. Define

$$r = 1 - \frac{H(Y \mid X)}{H(X)}$$

(a) Show $r = \frac{I(X;Y)}{H(X)}$ **[2 pts]**

$$r = 1 - \frac{H(Y \mid X)}{H(X)} = 1 - \frac{\sum_{X}\sum_{Y} p(x_i y_j)\log p(y_j \mid x_i)}{\sum_{X} p(x_i)\log p(x_i)} = \frac{\sum_{X} p(x_i)\log p(x_i) - \sum_{X}\sum_{Y} p(x_i y_j)\log p(y_j \mid x_i)}{\sum_{X} p(x_i)\log p(x_i)}$$

$$= \frac{\sum_{X} p(y_j)\log p(y_j) - \sum_{X}\sum_{Y} p(x_i y_j)\log p(y_j \mid x_i)}{\sum_{X} p(x_i)\log p(x_i)}$$ (because H(X) = H(Y))

$$= \frac{\sum_X \sum_Y p(x_i y_j) \log p(y_j) - \sum_X \sum_Y p(x_i y_j) \log p(y_j \mid x_i)}{\sum_X p(x_i) \log p(x_i)} = \frac{\sum_X \sum_Y p(x_i y_j) \log \frac{p(y_j)}{p(y_j \mid x_i)}}{\sum_X p(x_i) \log p(x_i)}$$

$$= \frac{\sum_X \sum_Y p(x_i y_j) \log \frac{p(y_j \mid x_i)}{p(y_j)}}{H(X)} = \frac{\sum_X \sum_Y p(x_i y_j) \log \frac{p(y_j \mid x_i) p(x_i)}{p(y_j) p(x_i)}}{H(X)} = \frac{\sum_X \sum_Y p(x_i y_j) \log \frac{p(x_i y_j)}{p(y_j) p(x_i)}}{H(X)}$$

$= \frac{I(X;Y)}{H(X)}$ (As $I(X;Y) = \sum_X \sum_Y p(x_i y_j) \log \frac{p(x_i y_j)}{p(y_j) p(x_i)}$)

(b) Show $0 \le r \le 1$ [2 pts]

$$r = 1 - \frac{H(Y \mid X)}{H(X)}$$

Because H(Y) ≥ H(Y|X) and H(Y)=H(X), then $0 \le \frac{H(Y \mid X)}{H(X)} \le 1$.

Therefore $0 \le r \le 1$

(c) Under what conditions is r = 0? [2 pts]

When $\frac{H(Y \mid X)}{H(X)} = 1$, r=0. That means H(Y|X)=H(X)=H(Y) so that X Y are completely independent.

(d) Under what conditions is r = 1? [2 pts]

When $\frac{H(Y \mid X)}{H(X)} = 0$, r=1. That means H(Y|X)=0 so that there must be one 1 and other are all 0 for p(x,y) distribution, because p(y|x)=p(x,y)/p(x).

Q8.

Let H(p) denote the entropy of a coin flip with Pr(head) = p. Derive the maximum value of H(p), and the value of p that attains this maximum. [4 pts]

H(p)=Pr(head)*log1/Pr(head)+[1-Pr(head)]*log1/[1-Pr(head)]=p*log1/p+(1-p)*log1/(1-p)

Let H(p)=H

So that the above equation is H=-(p*logp+(1-p)*log(1-p))

For calculate convenience, we can set the log base e(it will not affect the result).

H=-(p*lnp+(1-p)*ln(1-p))

$$H' = -(1*\ln p + p*\frac{1}{p} + (-1)\ln(1-p) + (1-p)*(-1)\frac{1}{1-p})$$

$$= -(\ln p - \ln(1-p))$$

$$=\ln(1-p)-\ln p$$

We set H' =0 if we want to get maximum, so that p=1/2=0.5.

Put p=0.5 back into the first equation, so that $H_{MAX}(p)$=-(0.5*log0.5+0.5*log0.5)=1

When p=0.5, we can get the maximum H(p)=1.

Q9.

Give an example of joint distribution between 3 binary random variables X,Y,Z such that all the following properties hold,

(a) I(X;Z) = 0 [6 pts]

(b) I(Y;Z) = 0 [6 pts]

(c) 0<I(X,Y;Z) = H(Z) [6 pts]

After writing down the joint distribution table, calculate I(X;Z) , I(Y;Z), I(X,Y;Z) and H(Z).

To let I(X;Z) and I(Y;Z) be 0, we should keep X,Z and Y, Z independent with each other. And Z only has dependency with (X,Y). And because I(X,Y;Z) = H(Z), so there must be the situation like 2 keys, that we have any one key we do not get any info, and we have 2 keys we have all info.

Y \ X	1	0	P_Y
1	0.5	0	0.5
0	0	0.5	0.5
P_X	0.5	0.5	

Z \ X	1	0	P_Z
1	0.25	0.25	0.5
0	0.25	0.25	0.5
P_X	0.5	0.5	

Y \ Z	1	0	P_Y
1	0.25	0.25	0.5
0	0.25	0.25	0.5
P_Z	0.5	0.5	

Z \ X,Y	1,1	1,0	0,1	0,0	P_Z
1	0.5	0	0	0	0.5
0	0	0	0	0.5	0.5
P_{XY}	0.5	0	0	0.5	

$$I(X;Z)=\sum_{x,z} P(x,z)\log\frac{P(x,z)}{P(x)P(z)}=0$$

$$I(Y;Z)=\sum_{y,z} P(y,z)\log\frac{P(y,z)}{P(y)P(z)}=0$$

$$I(X,Y;Z)=\sum_{x,y,z} P(x,y,z)\log\frac{P(x,y,z)}{P(x,y)P(z)}=2*0.5\log\frac{0.5}{0.5*0.5}=\log 2=1\text{bits}>0$$

H(Z)=H(0.5,0.5)=1bits

Q10.

The probabilities of the six possible outcomes of tossing a (not necessarily fair) die is given by P_{true}=(p1,p2,p3,p4,p5,p6). Two people, A and B, each suggested a different probability model, P_A and P_B for the die:

$$P_{true} = (.08, .55, .15, .12, .05, .05)$$

$$P_A = (.07, .14, .24, .24, .05, .26)$$

$$P_B = (.25, .13, .21, .03, .11, .27)$$

(a) Calculate the cross-entropies $CH(P_{true},P_A)$, $CH(P_{true},P_B)$ and $CH(P_{true},P_{true})$. [5 pts]

$CH(P_{true},P_A)=H(P_{true})+D_{KL}(P_{true}||P_A)$

For discrete distribution $D_{KL}(P_{true}||P_A)=\sum_X P_{true}(x)\log\frac{P_{true}(X)}{P_A(X)}$, So that

$CH(P_{true},P_A)=\sum_X P_{true}(x)\log\frac{1}{P_{true}(X)}+\sum_X P_{true}(x)\log\frac{P_{true}(X)}{P_A(X)}=-\sum_X P_{true}(x)\log P_A(X)$

=-(0.08*log0.07+0.55*log0.14+0.15*log0.24+0.12*log0.24+0.05*log0.05+0.05*log0.26)

=0.3069+1.5601+0.3088+0.2471+0.2161+0.0972=2.7362bits

$CH(P_{true},P_B)=-\sum_X P_{true}(x)\log P_B(X)$ =0.16+1.6189+0.3377+0.6071+0.1592+0.0972=2.9801bits

$CH(P_{true},P_{true})=-\sum_X P_{true}(x)\log P_{true}(X)$ =H(P_{true})=H(.08, .55, .15, .12, .05, .05)=1.9757bits

(b) If we choose cross-entropy as the measure, which model (A or B) is better ? [1 Pts]

As the fewer the cross entropy is, the better the model is against the real probability distribution P_{true}. Therefore, we know that model A is better because $CH(P_{true},P_A)<CH(P_{true},P_B)$.

(c) Suggest an alternative measure of goodness, and answer question (b) again based on it. [2 Pts]

We can easily use $\frac{\sum(P_A-P_{True})^2}{6}$ which is like the deviation of the difference of P_A and P_{True} as the model of measurement.

$\frac{\sum(P_A-P_{True})^2}{6}$ =(0.0001+0.1681+0.0081+0.0144+0+0.0441)/6=0.03913

$\frac{\sum(P_B-P_{True})^2}{6}$ =(0.0289+0.1764+0.0036+0.0081+0.0036+0.0484)/6=0.04483

As a result, A is better because it has a less deviation.