COMP2610: INFORMATION THEORY

Week 3 Tutorial

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Australian National University

SOME KEY CONCEPTS:

- Binomial and Bernoulli distribution
- Likelihood function

$$\mathcal{L}(\theta) = p(\mathcal{D} \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

- Maximum likelihood estimate

Maximising $p(\mathcal{D} \mid \theta)$ is same as max log of $\mathcal{L}$

$$\text{ie. } \max \ \log p(\mathcal{D} \mid \theta) = \sum_{i=1}^{N} \log \ p(x_i \mid \theta)$$

- Information content of a random variable outcome

$$h(x) = \log_2 \left(\frac{1}{p(x)} \right) = -\log_2 p(x)$$

- Entropy of a RV

$$\begin{aligned} H(X) &= \mathbb{E}_x(h(x)) \\ &= \sum p(x) \cdot h(x) \\ &= -\sum p(x) \log_2 p(x) \end{aligned}$$

- Conditional entropy

$$H(Y \mid X = x) = -\sum p(y \mid X = x) \log p(y \mid X = x)$$

$$\begin{aligned} H(Y \mid X) &= \sum p(x) H(Y \mid X = x) \\ &= -\sum p(x) \sum p(y \mid x) \log p(y \mid x) \end{aligned}$$

- Joint entropy and chain rule

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log(p(x, y))$$

$$H(X, Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

1. Let X be a random variable with possible outcomes $\{1,2,3\}$. Let the probabilities of the outcomes be

$$p(X=1) = \frac{\theta}{2}$$
$$p(X=2) = \frac{\theta}{2}$$
$$p(X=3) = 1-\theta$$

for some parameter $\theta \in [0,1]$.

Suppose we see N observations of the random variable, $\{x_1,\ldots,x_N\}$. Let n_i denote the number of times that we observe the outcome $X = i$, i.e.

$$n_i = \sum_{k=1}^{N} \begin{cases} 1 & \text{if } x_k = i \\ 0 & \text{else.} \end{cases}$$

(a) Write down the likelihood function of θ given the observations $\{x_1,\ldots,x_N\}$ in terms of n_1,n_2,n_3.

(b) Suppose the observations are

$$\{3,3,1,2,3,2,2,1,3,1\}.$$

Compute the maximum likelihood estimate of θ. (*Hint*: Compute the log-likelihood function, and check when the derivative is zero.)

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

(a) Write down the likelihood function of θ given the observations $\{x_1,\ldots,x_N\}$ in terms of n_1,n_2,n_3.

Solution:

1. (a) Let n_i denote the number of times that we observe outcome $X = i$. The likelihood is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{N} p(X = x_i|\theta) \\ &= \prod_{i:x_i=1} \left(\frac{\theta}{2}\right) \cdot \prod_{i:x_i=2} \left(\frac{\theta}{2}\right) \cdot \prod_{i:x_i=3} (1-\theta) \\ &= \left(\frac{\theta}{2}\right)^{n_1} \cdot \left(\frac{\theta}{2}\right)^{n_2} \cdot (1-\theta)^{n_3} \\ &= \left(\frac{\theta}{2}\right)^{n_1+n_2} \cdot (1-\theta)^{n_3} . \end{aligned}$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

(b) Suppose the observations are

$$\{3,3,1,2,3,2,2,1,3,1\}.$$

Compute the maximum likelihood estimate of θ. (*Hint*: Compute the log-likelihood function, and check when the derivative is zero.)

Solution: (b) The log-likelihood is

$$\mathcal{L}(\theta) = (n_1 + n_2) \cdot \log \frac{\theta}{2} + n_3 \cdot \log(1-\theta).$$

The derivative is

$$\mathcal{L}'(\theta) = \frac{n_1 + n_2}{\ln 2} \cdot \frac{1/2}{\theta/2} + \frac{n_3}{\ln 2} \cdot \frac{-1}{1-\theta}.$$

We have that $n_1 = 3, n_2 = 3, n_3 = 4$. So, we need

$$\frac{6}{\theta} = \frac{4}{1-\theta}$$

for which the solution may be checked to be $\theta = 0.6$. Observe then that we estimate

$$p(X=1) = 0.3 \quad p(X=2) = 0.3$$

$$p(X=3) = 0.4,$$

matching the frequencies of observations of each outcome.

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

2. Consider the following joint distribution over X,Y:

$p(X,Y)$		X = 1	2	3	4
	1	0	0	1/8	1/8
	2	1/8	1/16	1/16	0
Y	3	1/8	1/8	0	0
	4	0	1/16	1/16	1/8

(a) Show that X and Y are not statistically independent. (*Hint*: You need only show that for at least one specific x,y pair, $p(X=x,Y=y)$ not equal to $\mathrm{p}(X=x)p(Y=y)$.)

(b) Compute the following quantities:

(i) $H(X)$

(ii) $H(Y)$

(iii) $H(X|Y)$

(iv) $H(Y|X)$

(v) $H(X,Y)$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

$p(X,Y)$		X			
		1	2	3	4
	1	0	0	1/8	1/8
	2	1/8	1/16	1/16	0
Y	3	1/8	1/8	0	0
	4	0	1/16	1/16	1/8

(a) Show that X and Y are not statistically independent. (*Hint*: You need only show that for at least one specific x,y pair, $p(X = x, Y = y)$ not equal to $p(X = x)p(Y = y)$.)

Solution:

2. (a) We can show that X and Y are not statistically independent by showing that $p(x,y)$ 6= $p(x)p(y)$ for at least one value of x and y. For example: $p(X = 1) = 1/8+1/8 = 1/4$ and $p(Y = 2) = 1/8 + 1/16 + 1/16 = 1/4$. From the given table we see that:

$$p(X = 1, Y = 2) = 1/8 \text{ which is different from } p(X = 1)p(Y = 2) = 1/16.$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

$p(X,Y)$		X			
		1	2	3	4
	1	0	0	1/8	1/8
	2	1/8	1/16	1/16	0
Y	3	1/8	1/8	0	0
	4	0	1/16	1/16	1/8

(b) Compute the following quantities:

(i) $H(X)$

(ii) $H(Y)$

(iii) $H(X|Y)$

(iv) $H(Y|X)$

(v) $H(X,Y)$

Solution:

(b) First, we find the marginal probabilities using the sum rule:

$$\mathbf{p}(X) = \big(P(X=1), P(X=2), P(X=3), P(X=4)\big) = (1/4, 1/4, 1/4, 1/4)$$
$$\mathbf{p}(Y) = \big(P(Y=1), P(Y=2), P(Y=3), P(Y=4)\big) = (1/4, 1/4, 1/4, 1/4).$$

We see that both $p(X)$ and $p(Y)$ are uniform distributions with 4 possible states. Hence:
$H(X) = H(Y) = \log_2 4 = 2$ bits.

To compute the conditional entropy $H(X|Y)$ we need the conditional distributions $p(X|Y)$ which can be computed by using the definition of conditional probability $p(X = x|Y = y) = p(X = x, Y = y)/p(Y = y)$. In other words, we divide the rows of the given table by the corresponding marginal.

$\mathbf{p}(X|Y=1) = (0,0,1/2,1/2)$ $\mathbf{p}(X|Y=2) = (1/2,1/4,1/4,0)$ $\mathbf{p}(X|Y=3) = (1/2,1/2,0,0)$ $\mathbf{p}(X|Y=4) = (0,1/4,1/4,1/2)$.

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

$p(X,Y)$		X			
		1	2	3	4
	1	0	0	1/8	1/8
	2	1/8	1/16	1/16	0
Y	3	1/8	1/8	0	0
	4	0	1/16	1/16	1/8

(b) Compute the following quantities:

(i) $H(X)$

(ii) $H(Y)$

(iii) $H(X|Y)$

(iv) $H(Y|X)$

(v) $H(X,Y)$

Solution:

b) continued

Hence the conditional entropy $H(X|Y)$ is given by:

$$H(X|Y) = \sum_{i=1}^{4} p(Y=i)H(X|Y=i)$$
$$= (1/4)H(0,0,1/2,1/2) + (1/4)H(1/2,1/4,1/4,0)$$
$$+ (1/4)H(1/2,1/2,0,0) + (1/4)H(0,1/4,1/4,1/2)$$
$$= 1/4 \times 1 + 1/4 \times 3/2 + 1/4 \times 1 + 1/4 \times 3/2$$
$$= 5/4 \text{ bits.}$$

Here we note that conditioning has indeed decreased entropy. We can compute the joint entropy by using the chain rule:

$$H(X,Y) = H(X|Y) + H(Y) = 5/4 + 2 = 13/4 \text{ bits.}$$

Additionally, we know that by the chain rule $H(X,Y) = H(Y|X) + H(X)$, hence:

$$H(Y|X) = H(X,Y) - H(X) = 13/4 - 2 = 5/4 \text{ bits.}$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

3. A standard deck of cards contains 4 *suits* — ♥, ♦, ♣, ♠ (“hearts”, “diamonds”, “clubs”, “spades”) — each with 13 *values* — A, $2, 3, 4, 5, 6, 7, 8, 9, 10$, J, Q, K (The A, J, Q, K are called “Ace”, “Jack”, “Queen”, “King”). Each card has a *colour*: hearts and diamonds are coloured red; clubs and spades are black. Cards with values J, Q, K are called *face cards*.

   Each of the 52 cards in a deck is identified by its value v and suit s and denoted vs. For example, 2♥, J♣, and 7♠ are the “two of hearts”, “Jack of clubs”, and “7 of spades”, respectively. The variable c will be used to denote a card’s colour. Let $f = 1$ if a card is a face card and $f = 0$ otherwise.

   A card is drawn at random from a thoroughly shuffled deck. Calculate:

   (a) The information in observing a red King, i.e., $h(c = \text{red}, v = \text{K})$

   (b) The conditional information in observing a King given a face card was drawn, i.e., $h(v = \text{K} | f = 1)$

   (c) The entropies $H(S)$ and $H(V, S)$.

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

(a) The information in observing a red King, i.e., $h(c = \text{red}, v = \text{K})$

(b) The conditional information in observing a King given a face card was drawn, i.e., $h(v = \text{K} | f = 1)$

(c) The entropies $H(S)$ and $H(V,S)$.

Solution:

3. (a) $h(c = \text{red}, v = \text{K}) = \log_2 \frac{1}{P(c=\text{red}, v=\text{K})} = \log_2 \frac{1}{1/26} = 4.7004$ bits.

(b) $h(v = K \mid f = 1) = \log_2 \dfrac{1}{p(v = K \mid f = 1)} = \log_2 \dfrac{1}{1/3} = 1.585\text{bits}$

(c) We have

i. $H(S) = \sum_s p(s) \log_2 \frac{1}{p(s)} = 4 \times \frac{1}{4} \times \log_2 \frac{1}{1/4} = 2$ bits.

ii. $H(V,S) = \sum_{v,s} p(v,s) \log_2 \frac{1}{p(v,s)} = 52 \times \frac{1}{52} \log_2 \frac{1}{1/52} = 5.7$ bits.

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

4. Let X be a random variable taking on a finite number of values. What is the (general) inequality relationship of H(X) and H(Y) if

   a. $Y = 2^X$?

   b. $Y = \cos X$?

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Q4

By chain rule we have $H(X)+H(Y \mid X) = H(Y)+H(X \mid Y)$. If we can find whether $H(Y \mid X)$ & $H(X \mid Y)$ are zero or > 0, then we can find the inequality relationship of $H(X)$ and $H(Y)$.

Solution:

a) Let's first consider $Y = 2^X$. We know this is a 1 to 1 mapping expression, i.e., for every unique value of X, we can find a unique Y.

$$\therefore \text{ if } p(X = x_i) = \theta, \text{ then } p(Y = 2^{x_i}) = \theta.$$

We also have

$$p(Y = 2^{x_i} \mid X = x_i) = 1 \text{ and } p(Y = y_i \mid X = x_i) = 0 \text{ for } y_i \neq 2^{x_i},$$

which can be generalised as

$$p\left(Y = y_i \mid X = x_i\right) = \begin{cases} 1 & y_i = f\left(x_i\right), \\ 0 & \text{otherwise.} \end{cases}$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Solution: 4a) continued

$$p\left(Y=y_i \mid X=x_i\right)= \begin{cases}1 & y_i=f\left(x_i\right), \\ 0 & \text{otherwise.}\end{cases}$$

The conditional entropy formula

$$H(Y \mid X)=-\sum p(x) \sum p(y \mid x) \log p(y \mid x),$$

When $p(Y=y_i \mid X=x_i)=1, \log(p=(Y=y_i \mid X=x_i))=0$

$$p(Y=y_i \mid X=x_i)=0, p(Y=y_i \mid X=x_i) \log(p=(Y=y_i \mid X=x_i))=0.$$

$$\therefore H(Y \mid X)=0.$$

We can further generalise to:

$$\text{if } B=f(A), H(B \mid A)=0. \tag{1}$$

Using (1), we can conclude

$$H(Y \mid X)=0 \text{ since } Y=2^X=f(X)$$
$$\text{Also } H(X \mid Y)=0 \text{ since } X=\log_2(Y)=g(Y).$$

$$\therefore H(X)=H(Y).$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

Solution:

b) $Y = \cos(X)$

We can conclude

$$H(Y \mid X) = 0, \text{ since } Y = \cos(X) = f(X).$$

But we cannot express X as a function of Y, since $Y = \cos(X)$ is a many to 1 mapping expression.

$$\therefore H(X \mid Y) \neq 0$$

$$\therefore H(X) + \cancel{H(Y \mid X)}^{=0} = H(Y) + \cancel{H(X \mid Y)}^{\neq 0}$$

$$\therefore H(X) = H(Y) + H(X \mid Y)$$

$$\therefore H(X) > H(Y).$$

TEQSA PROVIDER ID: PRV12002 (AUSTRALIAN UNIVERSITY) CRICOS PROVIDER CODE: 00120C

THANK YOU

Contact Details:

Office: *B137B, Brian Anderson Building (Ground Floor)*

Email: *yile.zhang@anu.edu.au*

Australian National University