COMP2610/COMP6261 Tutorial 2 Solutions*

Semester 2, 2018

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

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

(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

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

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

$$
\begin{aligned}
p(X = 1) &= 0.3 \\
p(X = 2) &= 0.3 \\
p(X = 3) &= 0.4,
\end{aligned}
$$

matching the frequencies of observations of each outcome.

*Based in part on solutions by Avraham Ruderman or the 2012 version of the course.

2. (a) We can show that X and Y are not statistically independent by showing that $p(x, y) \neq p(x)p(y)$ 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: $p(X = 1, Y = 2) = 1/8$ which is different from $p(X = 1)p(Y = 2) = 1/16$.

   (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.

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

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

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

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:

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

We see that for this particular example $H(X|Y) = H(Y|X)$, which is not generally the case.

Finally, the mutual information $I(X; Y)$ is given by:

$$I(X; Y) = H(X) - H(X|Y) = 2 - 5/4 = 3/4 \text{ bits.}$$

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

   (b) $h\left(v=\text{K}|f=1\right)=\log_2 \frac{1}{P\left(v=\text{K}|f=1\right)}=\log_2 \frac{1}{1/3}=1.585$ bits.

   (c) We have

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

   (d) Since V and S are independent we have $I\left(V;S\right)=0$ bits.

   (e) Since C is a function of S and by the data processing inequality $I\left(V;C\right)\leq I\left(V;S\right)=0$. However, mutual information must be nonnegative so we must have $I\left(V;C\right)=0$ bits.

4. This is a direct application of Jensen's inequality to the convex function $g(x)=x^2$.

5. (a) We see that $I(X;Y)=0$ as $X \perp\!\!\!\perp Y$.

   (b) To compute $I(X;Y|Z)$ we apply the definition of conditional mutual information:

$$I(X;Y|Z)=H(X|Z)-H(X|Y,Z)$$

   Now, X is fully determined by Y and Z. In other words, given Y and Z there is only one state of X that is possible, i.e it has probability 1. Therefore the entropy $H(X|Y,Z)=0$. We have that:

$$I(X;Y|Z)=H(X|Z)$$

   To determine this value we look at the distribution $p(X|Z)$, which is computed by considering the following possibilities:

X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	2

   Therefore:

$$\mathbf{p}(X|Z=0)=(1,0)$$
$$\mathbf{p}(X|Z=1)=(1/2,1/2)$$
$$\mathbf{p}(X|Z=2)=(0,1)$$

   From this, we obtain: $H(X|Z=0)=0$, $H(X|Z=2)=0$, $H(X|Z=1)=1$ bit. Therefore:

$$I(X;Y|Z)=p(Z=1)H(X|Z=1)=(1/2)(1)=0.5 \text{ bits.}$$

(c) This does not contradict the data-processing inequality (or more specifically the "conditioning on a downstream variable" corollary): the random variables in this example do not form a Markov chain. In fact, Z depends on both X and Y.

6. Gibb's inequality tells us that for any two probability vectors $\mathbf{p} = (p_1, \ldots, p_{|\mathcal{X}|})$ and $\mathbf{q} = (q_1, \ldots, q_{|\mathcal{X}|})$:

$$\sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{q_i} \geq 0$$

with equality if and only if $\mathbf{p} = \mathbf{q}$. If we take $\mathbf{q}$ to be the vector representing the uniform distribution $q_1 = \ldots = q_{|\mathcal{X}|} = \dfrac{1}{|\mathcal{X}|}$, then we get

$$0 \leq \sum_{i=1}^{|\mathcal{X}|} p_i \log \frac{p_i}{\frac{1}{|\mathcal{X}|}} = \sum_{i=1}^{|\mathcal{X}|} p_i \log p_i + \sum_{i=1}^{|\mathcal{X}|} p_i \log|\mathcal{X}| = -H(\mathbf{p}) + \log|\mathcal{X}|$$

with equality if and only if $\mathbf{p}$ is the uniform distribution. Moving $H(\mathbf{p})$ to the other side gives the inequality.