

## COMP 652: ASSIGNMENT 2

CARLOS G. OLIVER (ID: 260425853)

### 1. Q1: PROPERTIES OF ENTROPY AND MUTUAL INFORMATION, AND BAYES NET CONSTRUCTION

1.1. **(a).** Prove that  $H(X) \geq H(X|Y)$ , with equality achieved when  $X$  and  $Y$  are independent.

*Proof.* We begin with the following relation:

$$(1) \quad H(X) = H(X|Y) + I(X; Y)$$

Where  $I(X; Y)$  is the mutual information between the two random variables  $X$  and  $Y$ . This quantity represents the amount of information that can be obtained about one random variable, knowing the other. We can arrive at the above equation from the formal definition of mutual information:

(2)

$$\begin{aligned}
I(X; Y) &= \sum_{x,y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right) \\
&= \sum_{x,y} p(x, y) \left[ \log \left( \frac{p(x, y)}{p(y)} \right) - \log p(x) \right] \\
&= \sum_{x,y} p(x, y) \log \left( \frac{p(x, y)}{p(y)} \right) - \sum_{x,y} p(x, y) \log p(x) \\
&= \sum_{x,y} p(y) p(x|y) \log p(x|y) - \sum_{x,y} p(x, y) \log p(x) \quad \text{using: } p(x, y) = p(x|y)p(y) = p(y|x)p(x) \\
&= \sum_y p(y) \sum_x p(x|y) - \sum_x \log p(x) \sum_y p(x, y) \quad \text{breaking up summations} \\
&= - \sum_y p(y) H(X|Y = y) - \sum_x \log p(x) \sum_y p(x, y) \quad \text{by definition of entropy} \\
&= -H(X|Y) - \sum_x p(x) \log p(x) \quad \text{marginal probability} \\
&= -H(X|Y) - H(X) \\
&= H(X) - H(X|Y)
\end{aligned}$$

In order to prove the original statement, it suffices to show that  $I(X; Y) \geq 0$  with equality when  $X \perp Y$ .

We need a few definitions in order to show this. First, we define the KL-divergence between two probability distributions  $P$  and  $Q$  as:

$$(3) \quad D_{KL} = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

Which is related to the mutual information of two random variables  $X$  and  $Y$  as:

$$(4) \quad I(X; Y) = D_{KL}(P(X, Y) || P(X)P(Y)) = \sum_{x,y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$

We will use Jensen's inequality which applies to the expected value of convex functions of random variables, such that if  $f(x)$  is a convex function, then  $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$ . By letting the negative logarithm be the convex function, we can show that  $I(X; Y) \geq 0$ .

$$\begin{aligned}
(5) \quad -\sum_{x,y} p(x,y) \log\left(\frac{p(x)p(y)}{p(x,y)}\right) &\geq -\log\left(\sum_{x,y} p(x,y) \frac{p(x)p(y)}{p(x,y)}\right) \\
&\geq -\log\left(\sum_{x,y} p(x)p(y)\right) \\
&\geq -\log\left(\sum_x p(x) \sum_y p(y)\right) \\
&= 0 \qquad \text{probabilities sum to 1}
\end{aligned}$$

It is easy to see that for the case where  $X \perp\!\!\!\perp Y$  we have  $p(x,y) = p(x)p(y)$  so

$$(6) \quad \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \sum_{x,y} p(x,y) \log\left(\frac{p(x)p(y)}{p(x)p(y)}\right) = 0$$

□

1.2. **(b).** Given the relation between KL divergence and mutual information in Equation 4 we showed in Equation 5 that this quantity is  $D_{KL} \geq 0$ .

The KL divergence is not a symmetric quantity. Let  $P(x) = 1$  for all values of  $x$  and let  $Q(x) = 0$  for all values of  $x$ .

$$(7) \quad D_{KL}(P, Q) = \sum_x 0 \log \frac{0}{1} = 0$$

$$(8) \quad D_{KL}(Q, P) = \sum_x 1 \log \frac{1}{0} = \infty$$

1.3. **(c).** Show that  $I(X; Y) = H(X) + H(Y) - H(X, Y)$ , also known as the chain rule for conditional entropy.

*Proof.* We first show that  $H(X|Y) = H(X, Y) - H(Y)$ .

(9)

$$\begin{aligned}
H(X|Y) &= \sum_{x,y} p(x,y) \log \left( \frac{p(y)}{p(x,y)} \right) \\
&= \sum_{x,y} p(x,y) \left[ \log p(y) - \log p(x,y) \right] \\
&= - \sum_{x,y} p(x,y) \log p(x,y) + \sum_{x,y} p(x,y) \log p(y) \\
&= H(X,Y) + \sum_{x,y} p(x,y) \log p(y) && \text{definition of entropy} \\
&= -H(X,Y) - H(X) && \text{marginalize out x as before}
\end{aligned}$$

From this we obtain the expression  $H(X,Y) = H(X) - H(X|Y)$  and substitute it into the statement to prove.

(10)

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

Which we proved from the definition of  $I(X;Y)$  in Equation 2 above.

□

1.4. **(d).** Shown in Equation 5

1.5. **(e).** Let  $\mathcal{X}_i$  represent the possible values of the random variable  $x_i$  and  $\mathcal{X}_{\pi_i}$  represent the possible values of the set of parents  $x_{\pi_i}$ .

(11)

$$\begin{aligned}
L(G|D) &= \prod_{j=1}^m p(\mathbf{x}_j|G) && \text{by definition of likelihood} \\
&= \prod_{j=1}^m \prod_{i=1}^n p(\mathbf{x}_{j,i}|G) && \text{by graph factorization} \\
&= \sum_{j=1}^m \sum_{i=1}^n \log p(x_{j,i}|G) \\
&= \sum_{j=1}^m \sum_{i=1}^n \log p(x_{j,i}|x_{\pi_i}) && \text{sum over } m \text{ produces empirical distribution} \\
&= \sum_{i=1}^n \sum_{j=1}^m \log p(x_{j,i}|x_{\pi_i}) \\
&= \sum_{i=1}^n \left[ \sum_{\mathcal{X}_i} \sum_{\mathcal{X}_{\pi_i}} N(x_i, x_{\pi_i}) \log(\hat{p}(x_i|x_{\pi_i})) \right] \\
&= m \sum_{i=1}^n \left[ \sum_{\mathcal{X}_i} \sum_{\mathcal{X}_{\pi_i}} \frac{N(x_i, x_{\pi_i})}{m} \log(\hat{p}(x_i|x_{\pi_i})) \right] && \text{multiply by } \frac{m}{m} \\
&= m \sum_{i=1}^n \left[ \sum_{\mathcal{X}_i} \sum_{\mathcal{X}_{\pi_i}} \frac{N(x_i, x_{\pi_i})}{m} \log \frac{\hat{p}(x_i, x_{\pi_i})}{\hat{p}(x_{\pi_i})} \right] && \text{definition of empirical dist.} \\
&= m \sum_{i=1}^n \left[ \sum_{\mathcal{X}_i} \sum_{\mathcal{X}_{\pi_i}} \frac{N(x_i, x_{\pi_i})}{m} \log \frac{\hat{p}(x_i, x_{\pi_i})}{\hat{p}(x_{\pi_i})} \frac{\hat{p}(x_i)}{\hat{p}(x_i)} \right] \\
&= m \sum_{i=1}^n \left[ \sum_{\mathcal{X}_i} \sum_{\mathcal{X}_{\pi_i}} \hat{p}(x_i, x_{\pi_i}) \log \frac{\hat{p}(x_i, x_{\pi_i})}{\hat{p}(x_{\pi_i})} \frac{\hat{p}(x_i)}{\hat{p}(x_i)} \right] && \text{frequencies of joint divided by } m \text{ is joint.} \\
&= m \sum_{i=1}^n \left[ \sum_{\mathcal{X}_i} \sum_{\mathcal{X}_{\pi_i}} \hat{p}(x_i, x_{\pi_i}) \left\{ \log \frac{\hat{p}(x_i, x_{\pi_i})}{\hat{p}(x_{\pi_i})\hat{p}(x_i)} + \log \hat{p}(x_i) \right\} \right] && \text{break up log of products} \\
&= m \sum_{i=1}^n MI_{\hat{p}}(x_i, x_{\pi_i}) - m \sum_{i=1}^n H_{\hat{p}}(x_i) && \text{by definition of MI and entropy}
\end{aligned}$$