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:

$$H(X) = H(X|Y) + I(X;Y) \tag{1}$$

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:

Date: March 20, 2017.

(2)

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

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

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

$$D_{KL} = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{3}$$

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

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

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 $I(X;Y) \geq 0$.

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

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

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

$\square$

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.

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

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

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 $H(X|Y) = H(X,Y) - H(X)$.

(9)

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

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

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(X|Y) \tag{10}$$

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

$\square$

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_{π_i}.

(11)

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

1.6. **(f).** If graphs G_1 and G_2 are identical except for one extra arc in G_2 we consider the node with an extra incoming arc in G_2 whose set of parents is now $x_{\pi_i \cup \tilde{x}}$ Using the equation for the likelihood of a graph derived in Eq. 11 we see that the second term depends only on the entropy of each node and not on the parents these quantities will not change. The mutual information term only depends on the node being considered and its set of parents X_{π_i} therefore all other terms in the sum remain equal and we can only consider the node X_i in G_1 and G_2 where the additional arc was inserted.

We use an expansion of mutual information between a node and its parents derived in [1] where $|X_{\pi_i}| = L$:

$$MI_{\hat{P}}(X_i, X_{\pi_i}) = MI_{\hat{P}}(X_i, X^1_{\pi_i}) + \sum_{l=2}^{L} MI_{\hat{P}}(X_i, X^l_{\pi_i} | \{X^1_{\pi_i}, ..., X^{L-1}_{\pi_i}\}) \tag{12}$$

Using the same expansion we can compute the mutual information at the node with extra parent $\tilde{X}$ and so $|X_{\pi_i} \cup \tilde{X}| = L + 1$

(13)

$$\begin{aligned} MI_{\hat{P}}(X_i, X_{\pi_i} \cup \tilde{X}) &= MI_{\hat{P}}(X_i, \tilde{X}) + \sum_{l=2}^{L+1} MI_{\hat{P}}(X_i, X^l_{\pi_i} | \{X^1_{\pi_i}, ..., X^{L-1}_{\pi_i}\}) \\ &= MI_{\hat{P}}(X_i, \tilde{X}) + MI_{\hat{P}}(X_i, X^1_{\pi_i}) + \sum_{l=3}^{L+1} MI_{\hat{P}}(X_i, X^l_{\pi_i} | \{X^1_{\pi_i}, ..., X^{L-1}_{\pi_i}\}) \end{aligned}$$

Proof.

(14)

$$\begin{aligned} MI_{\hat{p}}(X_i, X_{\pi_i}) &< MI_{\hat{p}}(X_i, X_{\pi_i \cup \tilde{X}}) \\ MI_{\hat{P}}(X_i, X^1_{\pi_i}) + \sum_{l=2}^{L} MI_{\hat{P}}(X_i, X^l_{\pi_i} | \{X^1_{\pi_i}, ..., X^{L-1}_{\pi_i}\}) &< MI_{\hat{P}}(X_i, \tilde{X}) + \\ & MI_{\hat{P}}(X_i, X^1_{\pi_i}) + \sum_{l=3}^{L+1} MI_{\hat{P}}(X_i, X^l_{\pi_i} | \{X^1_{\pi_i}, ..., X^{L-1}_{\pi_i}\}) \\ 0 &< MI_{\hat{p}}(X_i, \tilde{X}) \end{aligned}$$

□

2. Q2: Sigmoid Bayes nets

Since by construction, bayes nets are DAGs and nodes are separated into two layers, we assume the architecture illustrated in **Fig 1.**

References

[1] Luis M de Campos. A scoring function for learning bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct):2149–2187, 2006.

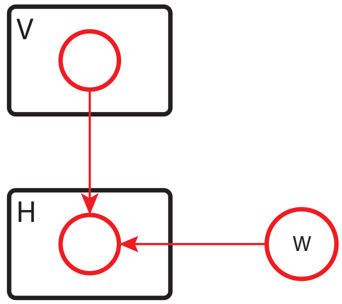

FIGURE 1. Bayes net with one hidden layer parametrized by weight matrix W.