

## 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:

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

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)

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

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  $X_{\pi_i}$  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  $|X_{\pi_i}| = L$ :

$$(12) \quad MI_{\hat{P}}(X_i, X_{\pi_i}) = MI_{\hat{P}}(X_i, X_{\pi_i}^1) + \sum_{l=2}^L MI_{\hat{P}}(X_i, X_{\pi_i}^l | \{X_{\pi_i}^1, \dots, X_{\pi_i}^{L-1}\})$$

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

$$(13) \quad \begin{aligned} MI_{\hat{P}}(X_i, X_{\pi_i} \cup \tilde{X}) &= MI_{\hat{P}}(X_i, \tilde{X}) + \sum_{l=2}^{L+1} MI_{\hat{P}}(X_i, X_{\pi_i}^l | \{X_{\pi_i}^1, \dots, X_{\pi_i}^{L-1}\}) \\ &= MI_{\hat{P}}(X_i, \tilde{X}) + MI_{\hat{P}}(X_i, X_{\pi_i}^1) + \sum_{l=3}^{L+1} MI_{\hat{P}}(X_i, X_{\pi_i}^l | \{X_{\pi_i}^1, \dots, X_{\pi_i}^{L-1}\}) \end{aligned}$$

*Proof.*

$$(14) \quad \begin{aligned} MI_{\hat{P}}(X_i, X_{\pi_i}) &< MI_{\hat{P}}(X_i, X_{\pi_i} \cup \tilde{X}) \\ MI_{\hat{P}}(X_i, X_{\pi_i}^1) + \sum_{l=2}^L MI_{\hat{P}}(X_i, X_{\pi_i}^l | \{X_{\pi_i}^1, \dots, X_{\pi_i}^{L-1}\}) &< MI_{\hat{P}}(X_i, \tilde{X}) + \\ &MI_{\hat{P}}(X_i, X_{\pi_i}^1) + \sum_{l=3}^{L+1} MI_{\hat{P}}(X_i, X_{\pi_i}^l | \{X_{\pi_i}^1, \dots, X_{\pi_i}^{L-1}\}) \\ 0 &< MI_{\hat{P}}(X_i, \tilde{X}) \end{aligned}$$

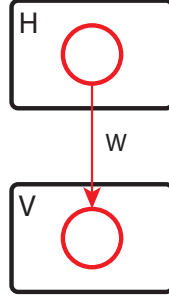
□

## 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**.

We want to find the  $W$  that maximizes the likelihood that the hidden nodes  $H$  generated the data seen by the visible nodes  $V$ . Let us assume we receive a set of values  $D$  for the visible nodes which we represent as the vector  $\mathbf{v}$ . We wish to maximize the likelihood over all data vectors in  $D$  with respect to the weight of each connection  $w_{ij}$ .

In the following derivation (based on the text by Neal Radford [2]), we will be making use of Bayes rule on the probability of the vector  $\mathbf{x}$  of nodes in graph  $X$  which can be viewed as the joint probability over all nodes in the graph, given the values of the visible

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

nodes. Note that we can form the vector of nodes  $x$  by joining the vectors of visible and hidden nodes as  $\mathbf{x} = \langle \mathbf{v}, \mathbf{h} \rangle$ .

$$(15) \quad P(X = \langle \mathbf{v}, \mathbf{h} \rangle) = P(X = \langle \mathbf{v}, \mathbf{h} \rangle | V = \mathbf{v}) P(V = \mathbf{v})$$

Now we maximize the log likelihood of the data with respect to the weights.

$$\begin{aligned}
 (16) \quad \frac{\partial L}{\partial w_{ij}} &= \frac{\partial}{\partial w_{ij}} \log \prod_{\mathbf{v} \in D} P(V = \mathbf{v}) \\
 &= \frac{\partial}{\partial w_{ij}} \sum_{\mathbf{v} \in D} \log P(V = \mathbf{v}) \\
 &= \sum_{\mathbf{v} \in D} \frac{1}{P(V = \mathbf{v})} \frac{\partial}{\partial w_{ij}} P(V = \mathbf{d}) \\
 &= \sum_{\mathbf{v} \in D} \frac{P(X = \mathbf{x} | V = \mathbf{v})}{P(X = \mathbf{x})} \frac{\partial}{\partial w_{ij}} P(V = \mathbf{d}) && \text{bayes rule} \\
 &= \sum_{\mathbf{v} \in D} \frac{P(X = \langle \mathbf{v}, \mathbf{h} \rangle | V = \mathbf{v})}{P(X = \langle \mathbf{v}, \mathbf{h} \rangle)} \sum_{\mathbf{h} \in H} \frac{\partial}{\partial w_{ij}} P(X = \langle \mathbf{v}, \mathbf{h} \rangle) && \text{marginalizing over H} \\
 &= \sum_{\mathbf{v} \in D} \sum_{\mathbf{x}} \frac{P(X = \mathbf{x} | V = \mathbf{v})}{P(X = \mathbf{x})} \frac{\partial}{\partial w_{ij}} P(X = \mathbf{x}) && \text{use given parametrization} \\
 &= \sum_{\mathbf{v} \in D} \sum_{\mathbf{x}} P(X = \mathbf{x} | V = \mathbf{v}) \frac{1}{\sigma(\sum_{j \in \pi_i} w_{ij} X_j)} \frac{\partial}{\partial w_{ij}} \sigma(\sum_{j \in \pi_i} w_{ij} X_j) \\
 &= \sum_{\mathbf{v} \in D} \sum_{\mathbf{x}} P(X = \mathbf{x} | V = \mathbf{v}) X_i X_j \sigma(-\sum_{\pi_i} w_{ij} X_j)
 \end{aligned}$$

From this we can get an update rule for  $w_{ij}$  as:

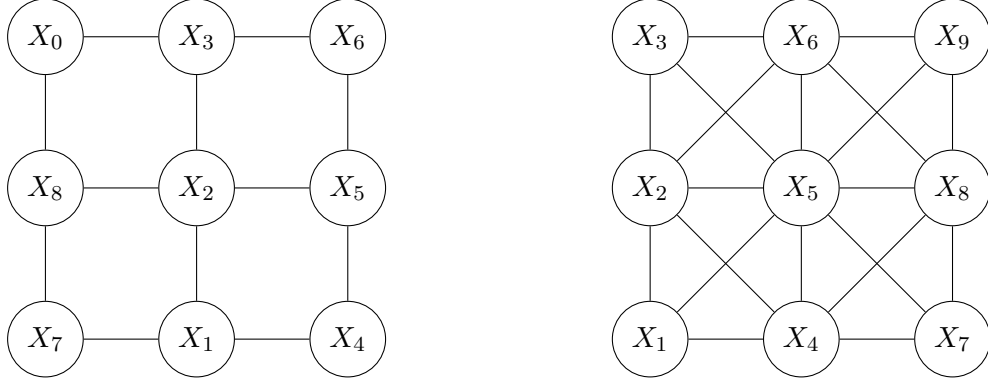


FIGURE 2. Lattice

$$(17) \quad w_{ij}^{new} = w_{ij} + \alpha X_i X_j \sigma \left( -X_i \sum_{\pi_i} w_{ij} X_j \right)$$

We can generate samples for  $P(X = \mathbf{x} | V = \mathbf{v})$  using Gibbs sampling, by fixing the visible nodes to a certain value and sampling for the hidden nodes.

### 3. Q3: MARKOV RANDOM FIELDS

3.1. **(a).** If we parametrize the joint  $p(\mathbf{x}) = \frac{1}{Z} \prod_C \psi_c(x_C)$  the same as before with one parameter per clique the 8-neighbourhood model will be heavily biased due to an overall decrease in the number of parameters with respect to the model complexity. Therefore we need to introduce parameters for the different connections within cliques.

We also retain the same energy computation.

3.2. **(b).** Cons:

- Higher variance due to increased model complexity
- Greater computational cost for training

Pros:

- Capture more complex geometric correlations such as curvatures

### 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.
- [2] Radford M Neal. Connectionist learning of belief networks. *Artificial intelligence*, 56(1):71–113, 1992.