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 23, 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

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

□

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)

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

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.

□

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 $|X_{\pi_i} \cup \tilde{X}| = L + 1$

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

Proof.

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

□

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

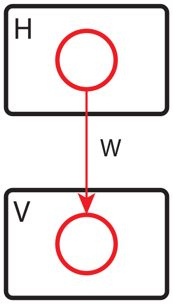

Figure 1. Bayes net with one hidden layer parametrized by weight matrix W.

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

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

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

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

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

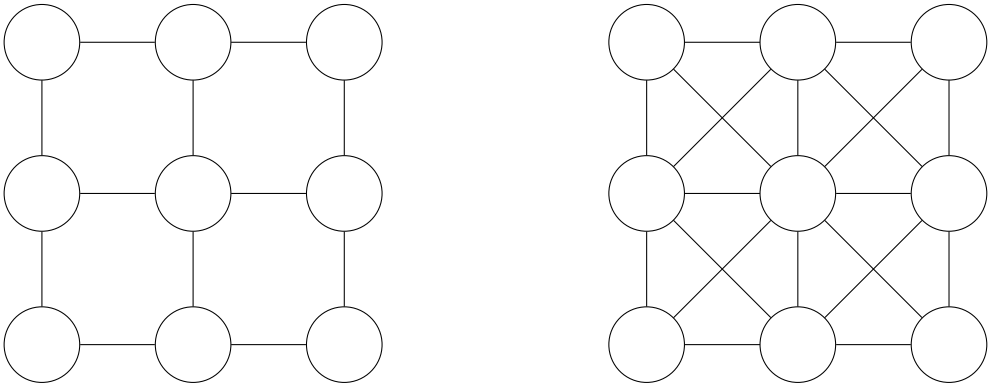

Figure 2. Lattice

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

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. Indeed, we will now have cliques of sizes 1, 2, 3 and 4. It would therefore be convenient to express the markov field instead as a factor graph where each factor incorporates unique cliques of different sizes.

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 and non-linear patterns.

3.3. **(c).** Here we would iteratively sample a value for a given pixel conditioned on all other nodes, and incorporating evidence introduced from the leftmost nodes. Naturally, we would want to perform the sampling in the direction evidence is flowing in from. We denote a random variable on the lattice $X_{ij} \in \tilde{X}$ as the random variable in row i and column j. Nodes $X_{i,j=1}$ will be the nodes receiving observed values $y_{ij} \in \tilde{Y_{ij}}$. We repeat the outer while loop until convergence.

```
input  : $\tilde{X}$, $\tilde{Y}$, $p^{initial}_{X_i|\tilde{X}-X_i}(x_i|\tilde{X}-X_i)$
output: $\tilde{Y}$, $\hat{p}_{X_i|\tilde{X}-X_i}(x_i|\tilde{X}-X_i)$
$X_{ij} \leftarrow$ random
while do
    for $X_{j,i}$ in $cols(\tilde{X})$ do
        if evidence for $X_{i,j}$ then
            $x_{i,j} \leftarrow y_{i,j}$ for any given evidence values
        end
        else
            $x_{i,j} \leftarrow$ draw from$p_{X_i|\tilde{X}-X_i}(x_{ij}|\tilde{x}-x_{ij})$
        end
        $x^{new}_{ij} \leftarrow x_{ij}$
    end
end
return $\tilde{X}$
```

4. Q4: EM Algorithm

4.1. **(a).** The data is a collection D of documents d which are composed of N words. These words can be seen as a collection of independent samples from the distributions corresponding to the document's topic k. Therefore, in order to maximize likelihood of the parameters μ and π we take a likelihood over all documents and words contained in each document.

The resulting maximum likelihood estimators are:

(18)
$$\begin{aligned}\pi_k &= \frac{D_k}{|D|} && \text{The frequency of a subject over the number of documents}\\ \mu_k(i) &= \frac{C_k(i)}{C_k} && \text{The count of occurences of a word of topic } k \text{ over all words of topic } k\end{aligned}$$

4.2. **(b).** Since a document may cover multiple subjects, we have a mixture of distributions where each distribution has some responsibility in explaining the observed value. We call this term $\gamma(\mathbf{z_{kn}})$ where $\mathbf{z}$ is a 1 of k encoding of the distribution that produced word n and $p(z_k = 1) = \pi_k$ for a given instance n.

Step 1: Initialization

Assign random values to the parameters π and μ.

Step 2: Expectation

For all n in sample and for all topics k:

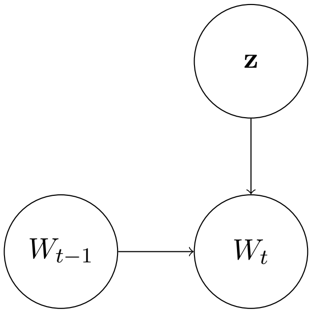

FIGURE 3. Graphical representation of dependence of current word W_t in document on the previous word W_{t-1} and the topic mixture vector $\mathbf{z}$.

$$\gamma(z_{nk}) = \frac{\prod_{i=1}^{M}(\mu_k(i))^{W(i)}}{\sum_{k=1}^{K}\pi_k(\mu_k(i))^{W(i)}} \tag{19}$$

Step 2: Maximization: use maximum likelihood estimators to update π and μ based on responsibilities computed in expectation step.

$$\begin{aligned} \mu_k^{new}(i) &= \frac{\sum_{D_j}\sum_{W_n\in D_j}\gamma(z_{kn})\mu_k(i)}{\sum_{D_j}\sum_{W_n\in D_j}\mu_k(i)} \quad \text{Taken over all words } i \text{ and topics } k \\ \pi_k^{new} &= \frac{\sum_{W_n}^{N}\gamma(z_{kn})}{N} \end{aligned} \tag{20}$$

4.3. **(c).** We can illustrate this scenario with a graphical model in **Fig. 3**. We can have up to $n-1$ possible values for the previous word W_{t-1} and the current word is parametrized by the k distributions of topics over m possible dictionary words so we have a total of $m \times (n-1) \times k$ parameters.

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.