Assignment 2 – Comp 652

Sumana Basu, McGill University, 260727568 23/03/2017

1. Properties of entropy and mutual information, and Bayes net construction

(a)

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$
$$H(X|Y) = -\sum_{x \in X, y \in Y} p(x,y) \log p(x|y)$$

$$\begin{aligned} H(X) - H(X|Y) &= -\sum_{x \in X} p(x) \log p(x) + \sum_{x \in X, y \in Y} p(x,y) \log p(x|y) \\ &= -\sum_{x \in X} p(x,y) \log p(x) + \sum_{x \in X, y \in Y} p(x,y) \log p(x|y) \\ &= -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x)}{p(x|y)} \\ &= -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x)p(y)}{p(x,y)} \end{aligned}$$

When X and Y are not independent :-

Using Jensen's inequality,

$$\begin{aligned} H(X) - H(X|Y) &\geq -\log \sum_{x \in X, y \in Y} p(x,y) \frac{p(x)p(y)}{p(x,y)} \\ &\geq -\log \sum_{x \in X, y \in Y} p(x)p(y) \\ &\geq -\log \sum_{x \in X} p(x) \sum_{y \in Y} p(y) \\ &\geq -\log 1 \\ &\geq 0 \end{aligned}$$

When X and Y are independent :-

$$\begin{aligned} H(X) - H(X|Y) &= -\sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x)p(y)}{p(x)p(y)} \\ &= -\sum_{x \in X, y \in Y} p(x,y) \log 1 \\ &= 0 \end{aligned}$$

Hence, $H(X) \geq H(X|Y)$, with equality achieved when X and Y are independent

(b)

$$\begin{aligned} D_{KL}(P,Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &= -\sum_x P(x) \log \frac{Q(x)}{P(x)} \\ &\geq \log \sum_x P(x) \frac{Q(x)}{P(x)} \\ &\geq \log \sum_x Q(x) \\ &\geq 0 \end{aligned}$$

Example of P and Q for which $D_{KL}(P,Q) \neq D_{KL}(Q,P)$:

P = Multivariate Normal Distribution $\mathcal{N}(\mu_0, \Sigma_0)$

Q = Multivariate Normal Distribution $\mathcal{N}(\mu_1, \Sigma_1)$

(c)

$$\begin{aligned} I(X;Y) &= D_{KL}(P(X,Y)||P(X)P(Y) \\ &= \sum_{x \in X, y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= \sum_{x \in X, y \in Y} P(x,y)[\log P(x,y) - \log P(x) - \log P(y)] \\ &= \sum_{x \in X, y \in Y} P(x,y) \log P(x,y) - \sum_{x \in X, y \in Y} P(x,y) \log P(x) - \sum_{x \in X, y \in Y} P(x,y) \log P(y) \\ &= -H(X,Y) + H(X) + H(Y) \end{aligned}$$

(d) By definition of mutual information,

$$\begin{aligned} MI(Y,Z) &= \sum_{y \in Y, z \in Z} p(y,z) \log \left(\frac{p(y,z)}{p(y)p(z)} \right) \\ &= -\sum_{y \in Y, z \in Z} p(y,z) \log \left(\frac{p(y)p(z)}{p(y,z)} \right) \end{aligned}$$

If Y and Z are not independent:-

By Jensen's inequality,

$$\begin{aligned} MI(Y,Z) &\geq -\log \sum_{y \in Y, z \in Z} p(y,z) \frac{p(y)p(z)}{p(y,z)} \\ &\geq -\log \sum_{y \in Y, z \in Z} p(y)p(z) \\ &\geq -\log \sum_{y \in Y} p(y) \sum_{z \in Z} p(z) \\ &\geq -\log 1 \\ &\geq 0 \end{aligned}$$

If Y and Z are independent:-

$$\begin{aligned} MI(Y,Z) &= -\sum_{y\in Y, z\in Z} p(y,z)\log\left(\frac{p(y)p(z)}{p(y,z)}\right) \\ &= -\sum_{y\in Y, z\in Z} p(y,z)\log\left(\frac{p(y)p(z)}{p(y)p(z)}\right) \\ &= -\sum_{y\in Y, z\in Z} p(y,z)\log 1 \\ &= 0 \end{aligned}$$

Hence, $MI(Y,Z) \geq 0$ with equality if and only if Y and Z are conditionally independent

(e) Left hand side,

$$\begin{aligned} \log L(G|D) &= \sum_{j=1}^{m} \log p(\mathbf{x}_j|G) \\ &= m\log(\prod_{i=1}^{n} \hat{P}(x_i|x_{\pi_i})) \\ &= m\sum_{i=1}^{n} \log \hat{P}(x_i|x_{\pi_i})) \\ &= m\sum_{i=1}^{n} \sum_{x_i\in X_i, x_{\pi_i}\in X_{\pi_i}} \hat{P}(x_i, x_{\pi_i}) \log \frac{\hat{P}(x_i, x_{\pi_i})}{\hat{P}(x_{\pi_i})} \end{aligned}$$

Right hand side,

$$\begin{aligned} m\sum_{i=1}^{n} MI_{\hat{P}}(X_i, X_{\pi_i}) - m\sum_{i=1}^{n} H_{\hat{P}}(X_i) &= m\sum_{i=1}^{n} H_{\hat{P}}(X_i) + m\sum_{i=1}^{n} H_{\hat{P}}(X_{\pi_i}) \\ &\quad - m\sum_{i=1}^{n} H_{\hat{P}}(X_i, X_{\pi_i}) - m\sum_{i=1}^{n} H_{\hat{P}}(X_i) \\ &= m\sum_{i=1}^{n} H_{\hat{P}}(X_{\pi_i}) - m\sum_{i=1}^{n} H_{\hat{P}}(X_i, X_{\pi_i}) \\ &= -m\sum_{i=1}^{n} \sum_{x_{\pi_i}\in X_{\pi_i}} \hat{P}(x_{\pi_i}) \log \hat{P}(x_{\pi_i}) \\ &\quad + m\sum_{i=1}^{n} \sum_{x_i\in X_i, x_{\pi_i}\in X_{\pi_i}} \hat{P}(x_i, x_{\pi_i}) \log \hat{P}(x_i, x_{\pi_i}) \\ &= -m\sum_{i=1}^{n} \sum_{x_i\in X_i, x_{\pi_i}\in X_{\pi_i}} \hat{P}(x_i, x_{\pi_i}) \log \hat{P}(x_{\pi_i}) \\ &\quad + m\sum_{i=1}^{n} \sum_{x_i\in X_i, x_{\pi_i}\in X_{\pi_i}} \hat{P}(x_i, x_{\pi_i}) \log \hat{P}(x_i, x_{\pi_i}) \\ &= m\sum_{i=1}^{n} \sum_{x_i\in X_i, x_{\pi_i}\in X_{\pi_i}} \hat{P}(x_i, x_{\pi_i}) \log \frac{\hat{P}(x_i, x_{\pi_i})}{\hat{P}(x_{\pi_i})} \end{aligned}$$

Hence,

$$\log L(G|D) = m\sum_{i=1}^{n} MI_{\hat{P}}(X_i, X_{\pi_i}) - m\sum_{i=1}^{n} H_{\hat{P}}(X_i)$$

(f) G2 is identical to G1 except for the fact that it has one extra arc. This means one node of G2 (say, x) has one extra parent (say, $x_{\pi'}$) than G1.

Maximum likelihood score for G1 and G2 are:

$$\log L(G1|D) = m\sum_{i=1}^{n} MI_{\hat{P}}(X_i, X_{\pi_i}) - m\sum_{i=1}^{n} H_{\hat{P}}(X_i)$$

$$\log L(G2|D) = m\sum_{i=1}^{n} MI_{\hat{P}}(X_i, X_{\pi_i}) + MI_{\hat{P}}(x, x_{\pi'}) - m\sum_{i=1}^{n} H_{\hat{P}}(X_i)$$

$$\log L(G2|D) - \log L(G1|D) = MI_{\hat{P}}(x, x_{\pi'})$$

$\log L(G2|D) - \log L(G1|D) \geq 0$ (since, mutual information cannot be negative)

$\therefore$ G1 has a lower score than G2. [Proved]

2. Sigmoid Bayes nets

Let, H = layer of hidden nodes

V = layer of visible nodes

$$p(\mathbf{V}) = \prod_{i=1}^{m}\sum_{H} p(H)p(V_i|H)$$

$$\begin{aligned}\mathcal{L} = \log p(\mathbf{V}) &= \sum_{i=1}^{m}\log\sum_{H} p(H)p(V_i|H)\\ &= \sum_{i=1}^{m}\log(\sum_{H} p(H)\sigma(\sum_{j\in H} w_{ij}H_j))\\ &= \sum_{i=1}^{m}\sum_{H}\log p(H) + \sum_{i=1}^{m}\sum_{H}\log\sigma(\sum_{j\in H} w_{ij}H_j))\end{aligned}$$

$$\frac{\partial\mathcal{L}}{\partial w_{ik}} = \frac{\sigma(\sum_{j\in H} w_{ij}H_j)(1-\sigma(\sum_{j\in H} w_{ij}H_j))H_k}{\sum_H \sigma(\sum_{j\in H} w_{ij}H_j)}$$

Gradient based learning rule:

$$w_{ik}^{(t+1)} = w_{ik}^{(t)} + \eta\frac{\partial\mathcal{L}}{\partial w_{ik}}$$

where, η = learning rate

3. Markov Random Fields

(a) Let $X_{i,j}$ be a node in a 2D spin glass model.

8-Neighborhood of the node consists of the nodes :

$X_{i-1,j-1}, X_{i,j-1}, X_{i+1,j-1}, X_{i-1,j}, X_{i+1,j}, X_{i+1,j-1}, X_{i+1,j}, X_{i+1,j+1}$

Node $X_{i,j}$ has 4 maximal cliques each of size 4.

Potential corresponding to each of the cliques are:

$\psi(X_{i-1,j-1}, X_{i,j-1}, X_{i-1,j}, X_{i,j})$
$\psi(X_{i,j-1}, X_{i+1,j-1}, X_{i,j}, X_{i+1,j})$
$\psi(X_{i-1,j}, X_{i,j}, X_{i-1,j+1}, X_{i,j+1})$
$\psi(X_{i,j}, X_{i+1,j}, X_{i,j+1}, X_{i+1,j+1})$

$$
\begin{aligned}
p(X_{i-1,j-1}, X_{i,j-1}, X_{i+1,j-1}, X_{i-1,j}, X_{i,j},& \\
X_{i+1,j}, X_{i+1,j-1}, X_{i+1,j}, X_{i+1,j+1}) &\propto \psi(X_{i-1,j-1}, X_{i,j-1}, X_{i-1,j}, X_{i,j}) \\
&\quad \psi(X_{i,j-1}, X_{i+1,j-1}, X_{i,j}, X_{i+1,j}) \\
&\quad \psi(X_{i-1,j}, X_{i,j}, X_{i-1,j+1}, X_{i,j+1}) \\
&\quad \psi(X_{i,j}, X_{i+1,j}, X_{i,j+1}, X_{i+1,j+1}) \\
&\propto e^{E(X_{i-1,j-1},X_{i,j-1},X_{i-1,j},X_{i,j})+E(X_{i,j-1},X_{i+1,j-1},X_{i,j},X_{i+1,j})} \\
&\quad +E(X_{i-1,j},X_{i,j},X_{i-1,j+1},X_{i,j+1}) \\
&\quad +E(X_{i,j},X_{i+1,j},X_{i,j+1},X_{i+1,j+1}) \\
&= \frac{1}{Z} e^{E(X_{i-1,j-1},X_{i,j-1},X_{i-1,j},X_{i,j})+E(X_{i,j-1},X_{i+1,j-1},X_{i,j},X_{i+1,j})} \\
&\quad +E(X_{i-1,j},X_{i,j},X_{i-1,j+1},X_{i,j+1}) \\
&\quad +E(X_{i,j},X_{i+1,j},X_{i,j+1},X_{i+1,j+1})
\end{aligned}
$$

where, Z = partition function

E = energy function

$$E(\mathbf{x}) = \sum_{i,j} \beta_{i,j} x_i x_j + \sum_i \alpha_i x_i$$

(b) **Connecting a pixel to 8 neighborhood instead of 4 neighborhood :**

Advantages :-

1) It performs better with respect to noise

2) It gives better information about the image

As we are considering 8 neighborhood, images tend to be smoother than 4 neighborhood. Any effect of noise also cancels out.

Disadvantages :-

1) Higher complexity of algorithm as compared to 4-neighborhood

2) Computationally more expensive due to more computation per node

Natural scenes tend to be more regular (eg. stretch of blue sky or green grass) than other images. So, even if we consider the 4 neighborhood, the model captures the scene well. So, we don't gain much at the cost of the extra computation.

(c) **Gibbs Sampling Algorithm for 2D Ising Model:-**

- Initialization : Initialize each node of first column of the lattice to evidence.
  $x_{1,1} = x_{2,1} = ... = x_{n,1} =$ evidence
- For τ = 1, . . . , T :
  For every non-evidence node $x_{i,j}$:
  – Sample $x_{i,j}^{(\tau+1)} \sim p(x_{i,j}|x_{N'_{i,j}}^{(\tau+1)}, x_{N''_{i,j}}^{(\tau)})$

where, $x_{N'_{i,j}}$ = neighborhood of node $x_{i,j}$ that have been sampled at step $(\tau + 1)$
$x_{N''_{i,j}}$ = neighborhood of node $x_{i,j}$ that have been sampled at step τ but not yet at $(\tau + 1)$

Note : For Markov Random field, each node is conditionally independent of all other nodes given its Markov Blanket (parent, children, spouse). So, neighborhood of each node consists of only its Markov Blanket.

4. EM Algorithms

(a) Each document consists of N words, which have been drawn i.i.d.

$$p(\mathbf{w}; \pi_k, \mu_k) = \prod_{i=1}^{N} p(w^{(i)}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k p(w^{(i)}|\mu_k)$$

$$\mathcal{L} = \log p(\mathbf{w}; \pi_k, \mu_k) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} \pi_k \prod_{j=1}^{M} (\mu_k(j))^{w^{(i)}(j)}$$

We define indicator variable $\delta_{d,k}$ as below :-

$$\delta_{d,k} = \begin{cases} 1, & \text{if document } d \text{ belongs to topic } k \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{N} \log \sum_{k=1}^{K} \delta_{d,k} \pi_k \prod_{j=1}^{M} (\mu_k(j))^{w^{(i)}(j)} \\ &= \sum_{i=1}^{N} \log(\pi_k \prod_{j=1}^{M} (\mu_k(j))^{w^{(i)}(j)})) \\ &= \sum_{i=1}^{N} \log \pi_k + \sum_{i=1}^{N} \sum_{j=1}^{M} w^{(i)}(j) \log \mu_k(j) \end{aligned}$$

Subjected to the constraints :-

$$\sum_{k=1}^{K} \pi_k = 1$$

$$\sum_{j=1}^{M} \mu_k(j) = 1 \text{(for each topic k)}$$

Using Lagrange multipliers constraints were incorporated in the optimization equation.

$$\mathcal{L} = \sum_{i=1}^{N} \log \pi_k + \sum_{i=1}^{N}\sum_{j=1}^{M} w^{(i)}(j) \log \mu_k(j) + \lambda_\pi(1 - \sum_{k=1}^{K} \pi_k) + \sum_{k=1}^{K} \lambda_{\mu_k}(1 - \sum_{j=1}^{M} \mu_k(j))$$

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \frac{N}{\pi_k} - \lambda_\pi = 0 \implies \pi_k = \frac{N}{\lambda_\pi}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_\pi} = 1 - \sum_{k=1}^{K} \pi_k = 0$$

$$\implies 1 - \sum_{k=1}^{K} \frac{N}{\lambda_\pi} = 0$$

$$\implies 1 - \frac{N}{\lambda_\pi} = 0$$

$$\implies \lambda_\pi = KN$$

$$\therefore \boxed{\pi_k = \frac{1}{K}}$$

(where, K is total number of topics)

$$\frac{\partial \mathcal{L}}{\partial \mu_k(j)} = \frac{count_{w(j)}}{\mu_k(j)} - \lambda_{\mu_k} = 0 \implies \mu_k(j) = \frac{count_{w(j)}}{\lambda_{\mu_k}}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda_{\mu_k}} = 1 - \sum_{j=1}^{M} \mu_k(j) = 0$$

$$\implies 1 - \sum_{j=1}^{M} \frac{count_{w(j)}}{\lambda_{\mu_k}} = 0$$

$$\implies 1 - \frac{N}{\lambda_{\mu_k}} = 0$$

$$\implies \lambda_{\mu_k} = N$$

$$\therefore \boxed{\mu_k(j) = \frac{count_{w(j)}}{N}}$$

(where, $count_{w(j)}$is the number of times j^{th} word of vocabulary appears in a document)

(b) In this case one document may cover multiple topics and topics (t_k) are latent variable.

Expectation step -

$$p(t_k|w) = \frac{p(t_k)p(w|t_k)}{p(w)}$$

$$= \frac{\pi_k \prod_{i=1}^{M} (\mu_k(i))^{w(i)}}{\sum_{k=1}^{K} \pi_k p(w|\mu_k)}$$

Maximization step -

We define indicator variable $\delta_{w^{(i)},t_k}$ as below :-

$$\delta_{w^{(i)},t_k} = \begin{cases} 1, & \text{if } i^{th} \text{ word } (w^{(i)}) \text{ belongs to topic k } (t_k) \\ 0, & \text{otherwise} \end{cases}$$

$$p(\mathbf{w};\pi_k,\mu_k) = \prod_{i=1}^{N}\sum_{k=1}^{K}\pi_k p(w^{(i)}|\mu_k)$$

$$\mathcal{L} = \log p(\mathbf{w};\pi_k,\mu_k) = \sum_{i=1}^{N}\log\sum_{k=1}^{K}\delta_{w^{(i)},t_k}\pi_k\prod_{j=1}^{M}(\mu_k(j))^{w^{(i)}(j)}$$

$$= \sum_{i=1}^{N}\log\pi_k + \sum_{i=1}^{N}\sum_{j=1}^{M}w^{(i)}(j)\log\mu_k(j)$$

Subjected to the constraints :-

$$\sum_{k=1}^{K}\pi_k = 1$$

$$\sum_{j=1}^{M}\mu_k(j) = 1\text{(for each topic k)}$$

Using Lagrange multipliers constraints were incorporated in the optimization equation.

$$\mathcal{L} = \sum_{i=1}^{N}\log\pi_k + \sum_{i=1}^{N}\sum_{j=1}^{M}w^{(i)}(j)\log\mu_k(j) + \lambda_\pi(1-\sum_{k=1}^{K}\pi_k) + \sum_{k=1}^{K}\lambda_{\mu_k}(1-\sum_{j=1}^{M}\mu_k(j))$$

$$\frac{\partial\mathcal{L}}{\partial\pi_k} = \frac{count_{w_k}}{\pi_k} - \lambda_\pi = 0 \implies \pi_k = \frac{count_{w_k}}{\lambda_\pi}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda_\pi} = 1 - \sum_{k=1}^{K}\pi_k = 0$$

$$\implies 1 - \sum_{k=1}^{K}\frac{count_{w_k}}{\lambda_\pi} = 0$$

$$\implies 1 - \frac{N}{\lambda_\pi} = 0$$

$$\implies \lambda_\pi = N$$

$$\therefore \boxed{\pi_k = \frac{count_{w_k}}{N}}$$

(where, $count_{w_k}$is the number of words belonging to topic k in a given document)

$$\frac{\partial\mathcal{L}}{\partial\mu_k(j)} = \frac{count_{w(j)}}{\mu_k(j)} - \lambda_{\mu_k} = 0 \implies \mu_k(j) = \frac{count_{w(j)}}{\lambda_{\mu_k}}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda_{\mu_k}} = 1 - \sum_{j=1}^{M}\mu_k(j) = 0$$

$$\implies 1 - \sum_{j=1}^{M}\frac{count_{w(j)}}{\lambda_{\mu_k}} = 0$$

$$\implies 1 - \frac{N}{\lambda_{\mu_k}} = 0$$

$$\implies \lambda_{\mu_k} = N$$

$$\therefore \boxed{\mu_k(j) = \frac{count_{w(j)}}{N}}$$

(where, $count_{w(j)}$is the number of times j^{th} word of vocabulary appears in a document)

(c) Here probability of i^{th} word (w^i) is conditioned on the topic (T_k) as well as the previous work in the

document ($w^i, w^2, ..., w^{(i-1)}$). Assuming one document may cover multiple topics, we have a hidden markov model where words are the observed variables and topics are the hidden variables.
Hence, we have K x K topic to topic state transition probabilities and K x M emission probabilities.
Where, K = no. of topics and M = size of vocabulary

Bias variance Trade off : Since we are no more assuming words are drawn i.i.d from a topic, bias of our model has decreased, but at the cost of increasing the variance.