ML2 Home Assignment 3

id: 12179078

Volodymyr Medentsiy

Problem 1.

1. Given that $H(X) = \underset{p(x)}{E}(-\log p(x))$ and $H(Y|X) = \underset{p(x,y)}{E}(-\log p(y|x))$

show that $H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$.

1) $H(X,Y) = \underset{p(x,y)}{E}(-\log p(x,y)) = \underset{p(x,y)}{E}(-\log p(x)p(y|x)) = \underset{p(x,y)}{E}(-\log p(x)) +$

$+ \underset{p(x,y)}{E}(-\log p(y|x)) = \underset{p(x)}{E}(-\log p(x)) + \underset{p(x,y)}{E}(-\log p(y|x)) =$

$= H(X) + H(Y|X).$

2) $H(X,Y) = \underset{p(x,y)}{E}(-\log p(y)\cdot p(x|y)) = \underset{p(x,y)}{E}(-\log p(y)) + \underset{p(x,y)}{E}(-\log p(x|y)) =$

$= \underset{p(y)}{E}(-\log p(y)) + \underset{p(x,y)}{E}(-\log p(x|y)) = H(Y) + H(X|Y)$

2. Given that $I(X;Y|Z) = \underset{p(z)}{E}(KL(p(x,y|z)\,||\,p(x|z)\,p(y|z)))$

show that: $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(Y|Z) - H(Y|X,Z).$

1) $\underset{p(z)}{E} KL(p(x,y|z)\,||\,p(x|z)\,p(y|z)) =$

$= \underset{p(z)}{E}\ \underset{p(x,y|z)}{E} -\log \frac{p(x|z)p(y|z)}{p(x,y|z)}$ (=)

$\frac{p(x|z)p(y|z)}{p(x,y|z)} = \frac{p(x|z)\,p(y|z)\,p(z)}{p(x,y,z)} = \frac{p(x|z)\,p(y,z)}{p(x|y,z)\,p(y,z)} = \frac{p(x|z)}{p(x|y,z)}$

(=) $\underset{p(z)}{E}\ \underset{p(x,y|z)}{E} -\log \frac{p(x|z)}{p(x|y,z)} =$ [we consider X, Y, Z are continuous r.v.] =

$$= \int_{\mathbb{R}} p(z)\,dz \int_{\mathbb{R}^2} p(x,y|z)\left(-\log \frac{p(x|z)}{p(x|y,z)}\right)dx\,dy =$$

$$= \iiint_{\mathbb{R}^3} p(x,y,z)\left(-\log p(x|z)\right)dx\,dy\,dz - \iiint_{\mathbb{R}^3} p(x,y,z)\left(-\log p(x|y,z)\right)dx\,dy\,dz =$$

$$= \iint_{\mathbb{R}^2} p(x,z)\left(-\log p(x|z)\right)dx\,dz - \iiint_{\mathbb{R}^3} p(x,y,z)\left(-\log p(x|y,z)\right)dx\,dy\,dz =$$

$$= \underset{p(x,z)}{E}\left(-\log p(x|z)\right) - \underset{p(x,y,z)}{E}\left(-\log p(x|y,z)\right) = H(X|Z) - H(X|Y,Z).$$

(If X, Y, Z are discrete then integrals will transform to sum.)

2)

$$\underset{p(z)}{E}\, KL\left(p(x,y|z)\,\|\,p(x|z)\,p(y|z)\right) =$$

$$= \underset{p(z)}{E}\ \underset{p(x,y|z)}{E}\left(-\log \frac{p(x|z)\,p(y|z)}{p(x,y|z)}\right) \quad (=)$$

$$\frac{p(x|z)\,p(y|z)}{p(x,y|z)} = \frac{p(x|z)\,p(y|z)\,p(z)}{p(x,y,z)} = \frac{p(x,z)\,p(y|z)}{p(x,y,z)} = \frac{p(y|z)}{p(y|x,z)}$$

$$(=)\ \underset{p(z)}{E}\ \underset{p(x,y|z)}{E}\left(-\log \frac{p(y|z)}{p(y|x,z)}\right) = \iiint_{\mathbb{R}^3} p(x,y,z)\left(-\log p(y|z)\right)dx\,dy\,dz -$$

$$- \iiint_{\mathbb{R}^3} p(x,y,z)\left(-\log p(y|x,z)\right)dx\,dy\,dz = \iint_{\mathbb{R}^2} p(y,z)\left(-\log p(y|z)\right)dy\,dz -$$

$$- \iiint_{\mathbb{R}^3} p(x,y,z)\left(-\log p(y|x,z)\right)dx\,dy\,dz = \underset{p(y,z)}{E}\left(-\log p(y|z)\right) - \underset{p(x,y,z)}{E}\left(-\log p(y|x,z)\right) =$$

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

Problem 2

$$\text{Mult}(x|\pi) = \frac{M!}{\prod_{i=1}^{K} x_i!} \cdot \prod_{i=1}^{K} \pi_i^{x_i}, \quad \text{with} \quad \sum_{i=1}^{K} \pi_i = 1, \quad \sum_{i=1}^{K} x_i = M.$$

1. Show that it's a member of an exponential family:

$$\text{Mult}(x|\pi) = \frac{M!}{(M - \sum_{j=1}^{K-1} x_j)! \prod_{j=1}^{K-1} x_j!} \cdot \exp\Big(\sum_{j=1}^{K-1} x_j \log \pi_j + \Big(M - \sum_{j=1}^{K-1} x_j\Big) \cdot$$

$$\cdot \log\Big(1 - \sum_{j=1}^{K-1} \pi_j\Big)\Big) = \frac{M!}{(M - \sum_{j=1}^{K-1} x_j)! \prod_{j=1}^{K-1} x_j!} \exp\Big(\sum_{j=1}^{K-1} x_j \log \frac{\pi_j}{1 - \sum_{j=1}^{K-1} \pi_j} +$$

$$+ M \log\Big(1 - \sum_{j=1}^{K-1} \pi_j\Big)\Big) = \frac{M!}{(M - \sum_{j=1}^{K-1} x_j)! \prod_{j=1}^{K-1} x_j!} \Big(1 - \sum_{j=1}^{K-1} \pi_j\Big)^M \cdot$$

$$\cdot \exp\left(\sum_{j=1}^{K-1} x_j \log \frac{\pi_j}{1 - \sum_{j=1}^{K-1} \pi_j} \right)$$

So $$h(x) = \frac{M!}{(M - \sum_{j=1}^{K-1} x_j)! \prod_{j=1}^{K-1} x_j!}$$

$u(x) = (x_1 \dots x_{K-1})^T$ – sufficient statistics.

$\vec{\eta} \in \mathbb{R}^{K-1}$: $\eta_i = \log \frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j}$

$$g(\eta(\pi)) = \Big(1 - \sum_{j=1}^{K-1} \pi_j\Big)^M$$

Let us express π in terms of η:

$$\eta_i = \log\left(\frac{\pi_i}{1-\sum_{j=1}^{K-1}\pi_j}\right) \Rightarrow e^{\eta_i} = \frac{\pi_i}{1-\sum_{j=1}^{K-1}\pi_j} \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{K} e^{\eta_i} = \frac{1}{1-\sum_{j=1}^{K-1}\pi_j} \Rightarrow 1-\sum_{j=1}^{K-1}\pi_j = \frac{1}{e^{\eta_K}+\sum_{j=1}^{K-1}e^{\eta_j}} = \left[e^{\eta_K}=e^{\log 1}=0\right] =$$

$$= \frac{1}{1+\sum_{j=1}^{K-1}e^{\eta_j}}$$

So $\pi_i = \dfrac{e^{\eta_i}}{1+\sum_{j=1}^{K-1}e^{\eta_j}}$.

$$g(\eta) = \left(1-\sum_{j=1}^{K-1}\pi_j\right)^M = \left(1+\sum_{j=1}^{K-1}e^{\eta_j}\right)^{-M}$$

And $A(\eta) = -\log g(\eta) = M\log\left(1+\sum_{j=1}^{K-1}e^{\eta_j}\right)$

2. $E\,u(x) = -\nabla \log g(\eta) = \nabla A(\eta)$

$$E\,u(x)_i = \frac{\partial A(\eta)}{\partial \eta_i} = \frac{M\cdot e^{\eta_i}}{1+\sum_{j=1}^{K-1}e^{\eta_j}}$$

$$Cov\,u(x) = -\nabla^2 \log g(\eta) = \nabla^2 A(\eta)$$

$$Cov\,u(x)_{ij} = \frac{\partial^2 A(\eta)}{\partial\eta_i\partial\eta_j} = Me^{\eta_i}\frac{\partial}{\partial\eta_j}\left(1+\sum_{j=1}^{K-1}e^{\eta_j}\right)^{-1} = -Me^{\eta_i+\eta_j}\left(1+\sum_{j=1}^{K-1}e^{\eta_j}\right)^{-2}.$$

3. Construct the conjugate prior. According to 2.229 Bishop conjugate prior takes form:

$p(\eta(\pi)|\chi,\nu) = f(\chi,\nu)\, g(\eta(\pi))^{\nu} \exp(\nu\, \eta(\pi)^T \chi)$ with $f(\chi,\nu)$ - normalization const.

$$p(\,|\chi,\nu) \propto g(\eta(\pi))^{\nu} \exp(\nu\, \eta(\pi)^T \chi) =$$

$$= \left(1 - \sum_{j=1}^{K-1} \pi_j\right)^{M\nu} \exp\left(\nu \sum_{j=1}^{K-1} \log\left(\frac{\pi_j}{1-\sum_{j=1}^{K-1}\pi_j}\right) \cdot \chi_j\right) =$$

$$= \left(1 - \sum_{j=1}^{K-1} \pi_j\right)^{M\nu} \prod_{j=1}^{K-1} \left(\frac{\pi_j}{1-\sum_{j=1}^{K-1}\pi_j}\right)^{\chi_j \nu} =$$

$$= \left(1 - \sum_{j=1}^{K-1} \pi_j\right)^{\left(M - \sum_{j=1}^{K-1}\chi_j\right)\nu} \cdot \prod_{j=1}^{K-1} \pi_j^{\chi_j \nu} =$$

$$= \pi_K^{\left(M - \sum_{j=1}^{K-1}\chi_j\right)\nu} \cdot \prod_{j=1}^{K-1} \pi_j^{\chi_j \nu} \propto \mathrm{Dir}(\pi|\alpha) \text{ with } \alpha \in \mathbb{R}^K:$$

$$\alpha_i = \chi_i \nu + 1 \quad i = \overline{1, K-1}$$

$$\alpha_K = \left(M - \sum_{j=1}^{K-1} \chi_j\right)\nu$$

4. Prior - to - posterior update rule:

$$p(\pi|D,\chi,\nu) \propto p(D|\pi,\chi,\nu)\, p(\pi|\chi,\nu) \propto \prod_{j=1}^{K} \pi_j^{\alpha_j} \prod_{d=1}^{|D|} p(x^{(d)}|\eta,\chi,\nu) =$$

$$= \prod_{j=1}^{K} \pi_j^{\alpha_j} \prod_{d=1}^{|D|} \frac{M!}{\prod_{j=1}^{K} x_j^{(d)}} \cdot \prod_{j=1}^{K} \pi_j^{x_j^{(d)}} = \prod_{j=1}^{K} \pi_j^{\alpha_j} \prod_{j=1}^{K} \pi_j^{\sum_{d=1}^{|D|} x_j^{(d)}} \cdot \prod_{d=1}^{|D|} \frac{M!}{\prod_{j=1}^{K} x_j^{(d)}!} =$$

$$= \prod_{j=1}^{K} \pi_j^{\alpha_j + \sum_{d=1}^{|D|} x_j^{(d)}} \cdot \prod_{d=1}^{|D|} \frac{M!}{\prod_{j=1}^{K} x_j^{(d)}!} \Rightarrow p(\pi|D,\chi,\nu) = \mathrm{Dir}(\pi|\alpha_{\text{posterior}})$$

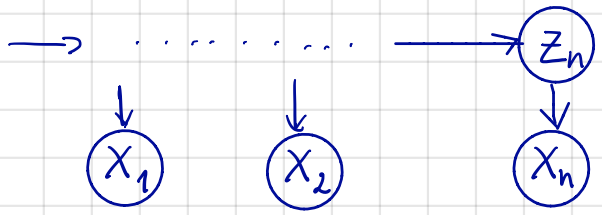
with update rule: $\alpha_i^{\text{posterior}} = \alpha_i + \sum_{d=1}^{|D|} x_i^{(d)}$.

Problem 4.

1. $p(x_1, \dots, x_{n-1} | x_n, z_n) = p(x_1, \dots, x_{n-1} | z_n)$

We need to show that $\{x_1, \dots, x_{n-1}\} \perp\!\!\!\perp x_n | z_n$

All paths from $\{x_1, \dots, x_n\}$ to x_n are blocked by z_n, because z_n is a head-to-tail node with respect to path from $\{x_1, \dots, x_{n-1}\}$ to x_n:



Thus $\{x_1, \dots x_{n-1}\}$ and x_n are d-separated by z_n which implies conditional independence.

2. $p(x_1, \dots, x_{n-1} | z_{n-1}, z_n) = p(x_1, \dots, x_{n-1} | z_{n-1})$

Let us show that $\{x_1, \dots, x_{n-1}\} \perp\!\!\!\perp z_n | z_{n-1}$

All paths from $\{x_1, \dots x_{n-1}\}$ to z_n are blocked by z_{n-1} because they either contain z_{n-1} as head-to-tail node (in case of $\{x_1, \dots, x_{n-2}\}$ and z_n) or as tail-to-tail node (x_{n-1} and z_n).

3. $p(x_{n+1}, \dots, x_N | z_n, z_{n+1}) = p(x_{n+1}, \dots, x_N | z_{n+1})$

$$p(x_{n+1}, \dots, x_N | z_n, z_{n+1}) = \frac{p(z_n, z_{n+1} | x_{n+1}, \dots, x_N) \cdot p(x_{n+1}, \dots, x_N)}{p(z_n, z_{n+1})}$$

$$= \frac{p(z_n | z_{n+1}, x_{n+1}, \dots x_N)\, p(z_{n+1} | x_{n+1}, \dots, x_N) \cdot p(x_{n+1}, \dots, x_N)}{p(z_n, z_{n+1})} = [\text{using 2.}]$$

$$= \frac{p(z_n | z_{n+1})}{p(z_n, z_{n+1})} \cdot p(z_{n+1} | x_{n+1}, \dots, x_N)\, p(x_{n+1}, \dots, x_N) =$$

$$= \frac{1}{p(z_{n+1})} \cdot p(z_{n+1}, x_{n+1}, \dots, x_N) = p(x_{n+1}, \dots, x_N | z_{n+1}).$$

4. $p(z_{N+1} | z_N, X) = p(z_{N+1} | z_N)$

$$p(z_{N+1} | z_N, X) = \frac{p(X, z_N | z_{N+1}) \, p(z_{N+1})}{p(z_n, X)} =$$

$$= \frac{p(z_{N+1}) \, p(z_N | X, z_{N+1}) \cdot p(X | z_{N+1})}{p(z_N, X)} =$$

$$= \frac{p(z_N) \cdot p(z_N | X) \cdot p(X | z_{N+1})}{p(z_N | X) \, p(X)} = \frac{p(X, z_{N+1})}{p(X)} = p(z_{N+1} | X)$$

PROBLEM 3

1. This is an ICA model because:
   - sources are independent of each other
   - sources are non-Gaussians.
   - each timepoint is an independent observation
   - observations are noisy linear combinations of the sources.

2. $$p(\{s_{1t}, s_{2t}, x_{1t}, x_{2t}, x_{3t}\}, t=\overline{1,T}) = \prod_{t=1}^{T} p(s_{1t} | \gamma_1) \cdot p(s_{2t} | \gamma_2) \cdot$$

$$\cdot p(x_{1t} | s_{1t}, s_{2t}, A_1, \delta_1) \cdot p(x_{2t} | s_{1t}, s_{2t}, A_2, \delta_2) \cdot$$

$$\cdot p(x_{3t} | s_{1t}, s_{2t}, A_3, \delta_3) = \prod_{t=1}^{T} St(0, \gamma_1) St(0, \gamma_2) \cdot N(A_{11} s_{1t} + A_{12} s_{2t}, \delta_1^2)$$

$$\cdot N(A_{21} s_{1t} + A_{22} s_{2t}, \delta_2^2) \cdot N(A_{31} s_{1t} + A_{32} s_{2t}, \delta_3^2).$$

3. The notion "explaining away" refers to the fact that in Directed graphical models observations of the child nodes do not block the path of the co-parent. Thus if we want path to be blocked we also

need to observe the parents of the child node.

Yes, this phenomenon is present in the model. For example if we want to construct Markov Blanket for s_{1t} we also need to include s_{2t}, parent of s_{1t} children (x_{1t}, x_{2t}, x_{3t}).

4. (a) False
(b) True
(c) False
(d) True
(e) False
(f) False
(g) False
(h) False

5. $MB(s_1) = \{x_1, x_2, x_3, s_2\}$

$MB(x_1) = \{s_1, s_2\}$.

6. $$p(\{x_{kt}\} | W, \{\nu_i\}) = \prod_{t=1}^{T} p_s(W x_t | \{\nu_i\}) \left| \det \frac{\partial s}{\partial x} \right| =$$

$$= \prod_{t=1}^{T} p_s(W x_t | \{\nu_i\}) |\det W| = \prod_{t=1}^{T} |\det W| \prod_{i=1}^{J} p_s(s_{it} | \nu_i) =$$

$$= \prod_{t=1}^{T} \prod_{i=1}^{J} |\det W| \, St(s_{it} | 0, \nu_i)$$

7. $$\log p(\{x_{kt}\} | W, \{\nu_i\}) = \log \prod_{t=1}^{T} \prod_{i=1}^{J} |\det W| \, St(s_{it} | 0, \nu_i) =$$

$$= \sum_{t=1}^{T} \sum_{i=1}^{J} \log St(s_{it} | 0, \nu_i) + T \log |\det(W)|$$

8. The goal of ICA is to recover source signals S given observations X and we know that $S = WX$. Thus we need to find W by the principle of maximum likelihood:

$$\max_W L(X, W)$$

We can solve it by applying gradient ascent algorithm, so by taking the gradient of the objective function $L(X, W)$ with respect to W and then updating the parameters in the direction of the gradient. When computing the gradient $\nabla_W L$ we substitute $\nabla_s \log p(s)$ by activation function $g p(s)$, so instead of choosing prior $p(s)$ we choose $\nabla_s \log p(s)$ which implicitly defines prior as well. This trick helps to easier optimization.

9. I expect overfit to happen when $K >> T$. The main reason is that we treat K as the number of features and T as a number of datapoints. If we consider ICA as a regression problem then it becomes obvious that model will probably overfit with $K >> T$.