Machine Learning 2 — Homework 3

Maurice Frank
11650656
maurice.frank@posteo.de

September 23, 2019

Problem 1.

1.

$$
\begin{aligned}
H(X,Y) &= \mathbb{E}_{p(x,y)}[-\log p(x,y)] \\
&= \iint_{X,Y} p(x,y) - \log p(x,y) dxdy \\
&= -\iint_{X,Y} p(x,y) \cdot \log p(x)p(y|x) dxdy \\
&= -\int_X p(x) \cdot \log p(x) \int_Y p(y|x) dydx - \iint_{X,Y} p(x,y) \log p(y|x) dxdy \\
&= -\int_X p(x) \cdot \log p(x) dx - \iint_{X,Y} p(x,y) \log p(y|x) dxdy \\
&= \mathbb{E}_{p(x)}[-\log p(x)] + \mathbb{E}_{p(x,y)}[-\log p(y|x)] \\
&= H(X) - H(Y|X) \\
H(X,Y) &= \mathbb{E}_{p(x,y)}[-\log p(x,y)] \\
&= \iint_{X,Y} p(x,y) - \log p(x,y) dxdy \\
&= -\iint_{X,Y} p(x,y) \cdot \log p(y)p(x|y) dxdy \\
&= -\int_Y p(y) \cdot \log p(y) \int_X p(x|y) dxdy - \iint_{X,Y} p(x,y) \log p(x|y) dxdy \\
&= -\int_Y p(x) \cdot \log p(y) dy - \iint_{X,Y} p(x,y) \log p(y|x) dxdy \\
&= \mathbb{E}_{p(y)}[-\log p(y)] + \mathbb{E}_{p(x,y)}[-\log p(y|x)] \\
&= H(Y) - H(X|Y)
\end{aligned}
$$

2.

$$
\begin{aligned}
I(X;Y|Z) &= \mathbb{E}_{p(z)}[\mathcal{KL}(p(x,y|z)||p(x|z)p(y|z))] \\
&= \int_Z p(z) \iint_{X,Y} p(x,y|z) \log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right) dxdydz \\
&= \int_Z p(z) \iint_{X,Y} p(x,y|z) \log\left(\frac{p(y|z)p(x|y,z)}{p(x|z)p(y|z)}\right) dxdydz \\
&= \iiint_{X,Y,Z} p(x,y,z) \log p(x|y,z) dxdydz \\
&\quad - \iiint_{X,Y,Z} p(x,y,z) \log p(x|z) dxdydz \\
&= -\mathbb{E}_{p(x,y,z)}[-\log p(x|y,z)] + \mathbb{E}_{p(x,z)}[-\log p(x|z)] \\
&= -H(X|Y,Z) + H(X|Z) \\
I(X;Y|Z) &= \mathbb{E}_{p(z)}[\mathcal{KL}(p(x,y|z)||p(x|z)p(y|z))] \\
&= \iiint_{X,Y,Z} p(x,y,z) \log\left(\frac{p(x|z)p(y|x,z)}{p(x|z)p(y|z)}\right) dxdydz \\
&= \iiint_{X,Y,Z} p(x,y,z) \log p(y|x,z) dxdydz \\
&\quad - \iint_{Y,Z} p(y,z) \log p(y|z) dydz \\
&= -\mathbb{E}_{p(x,y,z)}[-\log p(y|x,z)] + \mathbb{E}_{p(y,z)}[-\log p(y|z)] \\
&= -H(Y|X,Z) + H(Y|Z)
\end{aligned}
$$

Problem 2.

We have:

$$
\mathrm{Mult}(\boldsymbol{x}|\boldsymbol{\pi}) = \frac{M!}{x_1!x_2!\cdots x_K!} \cdot \pi_1^{x_1}\pi_2^{x_2}\cdots\pi_K^{x_K}
$$

$$
\sum_{i=1}^{K} x_i = M \quad \wedge \quad \sum_{i=1}^{K} \pi_i = 1
$$

1.

For minimal amount of parameters we have: $x_K = M - \sum_{i=1}^{K-1} x_i$ and $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$

$$
\begin{aligned}
\mathrm{Mult}(\boldsymbol{x}|\boldsymbol{\pi}) &= \frac{M!}{x_1! \cdots x_K!} \cdot \exp\log[\prod_{i=1}^{K} \pi_i^{x_K}] \\
&= \frac{M!}{x_1! \cdots x_K!} \cdot \exp[\sum_{i=1}^{K} x_i \log \pi_i] \\
&= \frac{M!}{x_1! \cdots x_K!} \cdot \exp[\sum_{i=1}^{K-1} x_i \log \pi_i + (M - \sum_{i=1}^{K-1} x_i) \cdot \log(1 - \sum_{i=1}^{K-1} \pi_i)] \\
&= \frac{M!}{x_1! \cdots x_K!} \cdot \exp[\sum_{i=1}^{K-1} x_i \log \frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j} + M \cdot \log(1 - \sum_{i=1}^{K-1} \pi_i)] \\
&= h(\boldsymbol{x}) \cdot \exp[\boldsymbol{\eta}^T \cdot T(\boldsymbol{x}) - A(\boldsymbol{\eta})] \\
&\implies \\
h(x) &= \frac{M!}{x_1! \cdots x_K!} \\
\mathbf{T}(\boldsymbol{x}) &= \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \\
\eta &= \begin{bmatrix} \log \frac{\pi_1}{1 - \sum_{j=1}^{K-1} \pi_j} \\ \vdots \\ \log \frac{\pi_{K-1}}{1 - \sum_{j=1}^{K-1} \pi_j} \end{bmatrix} \\
\pi_i &= \frac{e^{\eta_i}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}} \\
A(\boldsymbol{\eta}) &= -M \cdot \log(1 - \sum_{i=1}^{K-1} \pi_i) \\
&= -M \cdot \log(1 - \sum_{i=1}^{K-1} \frac{e^{\eta_i}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}}) \\
&= -M \cdot \log(\frac{1}{1 + \sum_{j=1}^{K-1} e^{\eta_j}}) \\
&= M \cdot \log(1 + \sum_{j=1}^{K-1} e^{\eta_j})
\end{aligned}
$$

2.

$$
\begin{aligned}
\mathbb{E}[x_i] &= \frac{\partial A(\boldsymbol{\eta})}{\partial \eta_i} \\
&= M \frac{\partial}{\partial \eta_i} \log\left(1 + \sum_{j=1}^{K-1} e^{\eta_j}\right) \\
&= M \frac{e^{\eta_i}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}} \\
&= M\pi_i \\
\mathrm{cov}(x_i, x_j) &= \frac{\partial^2 A(\boldsymbol{\eta})}{\partial \eta_i \partial \eta_j} \\
&= -M \frac{e^{\eta_i + \eta_j}}{(1 + \sum_{j=1}^{K-1} e^{\eta_j})^2} \\
&= -M \frac{e^{\eta_i}}{(1 + \sum_{j=1}^{K-1} e^{\eta_j})^2} \frac{e^{\eta_j}}{(1 + \sum_{j=1}^{K-1} e^{\eta_j})^2} \\
&= -M\pi_i\pi_j
\end{aligned}
$$

3.

The canonical conjugate prior of the exponential family is:

$$
p(\eta|\chi, v) \propto \exp\left[v \cdot \chi^T \cdot \eta - v \cdot A(\eta)\right]
$$

thus:

$$
\begin{aligned}
p(\eta|\chi, v) &\propto \exp\left[v \cdot \sum_{i=0}^{K-1} \chi_i \cdot \log \frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j} + v \cdot M \cdot \log\left(1 - \sum_{i=1}^{K-1} \pi_i\right)\right] \\
&= \prod_{i=0}^{K-1} \exp\left[v \cdot \chi_i \cdot \log \frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j}\right] \cdot \exp\left[v \cdot M \cdot \log\left(1 - \sum_{j=1}^{K-1} \pi_j\right)\right] \\
&= \prod_{i=0}^{K-1} \left(\frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j}\right)^{v \cdot \chi_i} \cdot \left(1 - \sum_{j=1}^{K-1} \pi_j\right)^{v \cdot M} \\
&= \prod_{i=0}^{K-1} {\pi_i}^{v \cdot \chi_i} \cdot \left(1 - \sum_{j=1}^{K-1} \pi_j\right)^{v \cdot M - v \cdot \sum_{j=0}^{K-1} \chi_j} \\
&= \prod_{i=0}^{K-1} {\pi_i}^{v \cdot \chi_i} \cdot {\pi_K}^{v \cdot M - v \cdot \sum_{j=0}^{K-1} \chi_j} \\
&\propto \mathrm{Dir}(\{\pi_1, \ldots, \pi_K\}, \{v \cdot \chi_1 + 1, \ldots v \cdot \chi_{K-1} + 1, v \cdot (M - \sum_{j=0}^{K-1} \chi_j) + 1\}) \\
&= \mathrm{Dir}(\{\pi_1, \ldots, \pi_K\}, \{\alpha_1, \ldots, \alpha_K\})
\end{aligned}
$$

The conjugate prior is a Dirichlet distribution.

4.

$$
\begin{aligned}
p(\boldsymbol{\pi}|\boldsymbol{x}, \boldsymbol{\chi}, v) &= p(\boldsymbol{x}|\boldsymbol{\pi}) \cdot p(\boldsymbol{\pi}|\boldsymbol{\chi}, v) \\
&= \left(\frac{M!}{x_1! \cdots x_K!}\right)^n \cdot \prod_{j=0}^{n} \exp\left[\sum_{i=1}^{K} x_i^j \log \pi_i\right] \cdot \prod_{i=0}^{K} \pi_i^{\alpha_i} \\
&= \left(\frac{M!}{x_1! \cdots x_K!}\right)^n \cdot \exp\left[\sum_{j=0}^{n} \sum_{i=1}^{K} x_i^j \log \pi_i\right] \cdot \exp\left[\sum_{i=0}^{K} \alpha_i \log \pi_i\right] \\
&= \left(\frac{M!}{x_1! \cdots x_K!}\right)^n \cdot \exp\left[\sum_{i=1}^{K} (\alpha_i + \sum_{j=0}^{n} x_i^j) \log \pi_i\right] \\
&= \left(\frac{M!}{x_1! \cdots x_K!}\right)^n \cdot \prod_{i=1}^{K} \pi_i^{(\alpha_i + \sum_{j=0}^{n} x_i^j)}
\end{aligned}
$$

With that we see that the update after n datapoints ist:

$$
\alpha_i^j \leftarrow \alpha_i + \sum_{j=0}^{n} x_i^j
$$

Problem 3.

We have independent sources $\{s_{it} = (s_{i1}, \ldots, s_{iT})\}$ with measurements $x_{kt} = \sum_{i=1}^{K_s} A_{ki} s_{it} + \epsilon_{kt}$, noise ϵ_{kt} $\mathcal{N}(0, {}^2_k)$.

1.

This model is an ICA model as it fullfills the characteristics of such. Our measurement is a noisy linear mixture of signals. The original signals are independent and not Gaussian distributed. The signals are independent with respect to time.

2.

$$
\begin{aligned}
& p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}) \\
& = \prod_{t=1}^{T} p(s_{1t}|v_1) p(s_{2t}|v_2) p(x_{1t}|v_1, v_2, A_1, \sigma_1) p(x_{2t}|v_1, v_2, A_2, \sigma_2) p(x_{3t}|v_1, v_2, A_3, \sigma_3) \\
& = \prod_{t=1}^{T} \mathcal{T}(s_{1t}|0, v_1) \mathcal{T}(s_{2t}|0, v_2) \\
& \quad \cdot \left(\sum_{i=1}^{2} A_{1i} \mathcal{T}(s_{it}|0, v_i) + \mathcal{N}(\epsilon_{1t}|0, \sigma_1^2) \right) \\
& \quad \cdot \left(\sum_{i=1}^{2} A_{2i} \mathcal{T}(s_{it}|0, v_i) + \mathcal{N}(\epsilon_{2t}|0, \sigma_2^2) \right) \\
& \quad \cdot \left(\sum_{i=1}^{2} A_{3i} \mathcal{T}(s_{it}|0, v_i) + \mathcal{N}(\epsilon_{3t}|0, \sigma_3^2) \right)
\end{aligned}
$$

3.

Explaining away is a phenomena seen in a BN when we have a variable dependent on two (or more) causes. If we than have information about one of the sources and also about the derivative variable we gain information about the other source variable, changing its distribution. In our case we might have have two audio sources and its mixture, if now know one of the inputs and the mixed signal we can retrieve/explain the other audio input. As such this is present in the given ICA model.

4.

(a) False

(b) True

(c) False

(d) True

(e) False

(f) False

(g) False

(h) False

5.

Markov blankets for

s_1 gives $\{x_1, x_2, x_3, s_2\}$

x_1 gives $\{s_1, s_2\}$

6.

$$\begin{aligned}
p(\{x_{kt}\}|W, \{v_i\}) &= \prod_{i=1}^{T} p(\boldsymbol{W}\boldsymbol{x}_t|\{v_i\})|\det Jac(s \to x)| \\
&= \prod_{i=1}^{T} \left[\prod_{j=1}^{I} \mathcal{T}(s_{jt}|0, \{v_i\}) \right] |\det Jac(s \to x)| \\
&= \prod_{i=1}^{T} \left[\prod_{j=1}^{I} \mathcal{T}(s_{jt}|0, \{v_i\}) \right] |\det \boldsymbol{W}|
\end{aligned}$$

7.

$$\begin{aligned}
\log p(\{x_{kt}\}|W, \{v_i\}) &= \log \prod_{i=1}^{T} \left[\prod_{j=1}^{I} \mathcal{T}(s_{jt}|0, \{v_i\}) \right] |\det \boldsymbol{W}| \\
&= \sum_{i=1}^{T} \left[\sum_{j=1}^{I} \log \mathcal{T}(s_{jt}|0, \{v_i\}) \right] |\det \boldsymbol{W}| \\
&= T \cdot |\det \boldsymbol{W}| + \sum_{i=1}^{T} \sum_{j=1}^{I} \log \mathcal{T}(s_{jt}|0, \{v_i\})
\end{aligned}$$

8.

SGD maximizes the log-likelihood. In contrast to full batch gradient descent we update the weights (here the de-mixing matrix) only with one datapoint in each iteration. As a first step in ICA the data should centered and applied whitening. The de-mixing matrix is initalized to a non-singular random matrix. Then we iterate until convergence (no significant change of the de-mixing matrix anymore). In each iteration we update with the gradient of the weight scaled by a learning rate η. Additionally we use use a non-linear activation function for the weight update. The activation function corresponds to a prior distribution over the sources. Typical choice for this problem would be $\tanh(\cdot)$ as an activation function. After convergence the estimated de-mixing matrix will approximately give the original sources if multiplied to the mixed signals.

9.

Overfitting is expected at the limit $K >> T$. With more sources K we have a more complicated model for less data T. The over-paramatrization will lead to overfitting.

Problem 4.

1.

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

$\{x_1, \dots, x_{n-1}\}$ is d-seperated from $\{x_n\}$ by $\{z_n\}$. Every path to x_n goes by $x_{n-1} \to z_n \to x_n$. This is non-collider and reaches x_n thus as as a single blocking trace it d-seperated the two sets.

2.

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

$\{x_1, \dots, x_{n-1}\}$ is d-seperated from $\{z_n\}$ by $\{z_{n-1}\}$. Every path to z_n goes by $z_{n-2} \to z_{n-1} \to z_n$ or $x_{n-1} \leftarrow z_{n-1} \to z_n$. Both are non-colliders and reach x_n thus as as z_{n-1} is the single blocking node it d-seperated the two sets.

3.

From the factorization properties of this Markov chain we know that $z_n \perp\!\!\!\perp x_{n+1}, \dots, x_N | z_{n+1}$. Using this we can write:

$$
\begin{aligned}
p(x_{n+1},\ldots,x_N|z_n,z_{n+1}) &= \frac{p(z_n,z_{n+1}|x_{n+1},\ldots,x_N)\cdot p(x_{n+1},\ldots,x_N)}{p(z_n,z_{n+1})} \\
&= \frac{p(z_n|z_{n+1},x_{n+1},\ldots,x_N)\cdot p(z_{n+1}|x_{n+1},\ldots,x_N)\cdot p(x_{n+1},\ldots,x_N)}{p(z_n|z_{n+1})p(z_{n+1})} \\
&= \frac{p(z_n|z_{n+1})\cdot p(z_{n+1}|x_{n+1},\ldots,x_N)\cdot p(x_{n+1},\ldots,x_N)}{p(z_n|z_{n+1})p(z_{n+1})} \\
&= \frac{p(z_{n+1}|x_{n+1},\ldots,x_N)\cdot p(x_{n+1},\ldots,x_N)}{p(z_{n+1})} \\
&= p(x_{n+1},\ldots,x_N|z_{n+1})
\end{aligned}
$$

4.

z_{N+1} does not exist in Figure 1, thus we assume an extension $z_N \to z_{N+1}$. Again from the factorization properties we know that $\boldsymbol{X} \perp\!\!\!\perp z_{N+1}|z_N$:

$$
\begin{aligned}
\mathbf{X} &= \{x_1,\ldots,x_N\} \\
p(z_{N+1}|z_N,\boldsymbol{X}) &= \frac{p(z_N,\boldsymbol{X}|z_{N+1})\cdot p(z_{N+1})}{p(z_N,\boldsymbol{X})} \\
&= \frac{p(\boldsymbol{X}|z_N,z_{N+1})p(z_N|z_{N+1})\cdot p(z_{N+1})}{p(z_N)p(\boldsymbol{X}|z_N)} \\
&= \frac{p(\boldsymbol{X}|z_N)p(z_N|z_{N+1})\cdot p(z_{N+1})}{p(z_N)p(\boldsymbol{X}|z_N)} \\
&= \frac{p(z_N|z_{N+1})\cdot p(z_{N+1})}{p(z_N)} \\
&= p(z_{N+1}|z_N)
\end{aligned}
$$