Machine Learning 2 - Homework 3

Sindy Löwe (11594969)

June 11, 2018

Collaborators: Pascal Esser, Gautier Dagan, Andrii Skliar, Linda Petrini

Problem 1

1.

$$\begin{aligned}
H(X,Y) &= \mathop{\mathbb{E}}_{p(x,y)}[-\log p(x,y)] \\
&= \iint -\log(p(x,y))p(x,y)dxdy \\
&= \iint -\log(p(x|y))p(x,y)dxdy + \iint -\log(p(y))p(x|y)p(y)dxdy \\
&= \iint -\log(p(x|y))p(x,y)dxdy + \int -\log(p(y))p(y)dy \\
&= \mathop{\mathbb{E}}_{p(x,y)}[-\log p(x|y)] + \mathop{\mathbb{E}}_{p(y)}[-\log p(y)] \\
&= H(X|Y) + H(Y)
\end{aligned}$$

Equivalently for $H(Y|X) + H(X)$.

2.

$$
\begin{aligned}
I(X,Y|Z) &= \iiint p(x,y|z)\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right)p(z)dxdydz \\
&= \iiint p(x,y,z)\log\left(p(x,y|z)\right)dxdydz \\
&\quad - \iiint p(x,y,z)\log\left(p(x|z)p(y|z)\right)dxdydz \\
&= \iiint p(x,y,z)\log\left(p(x|y,z)\right)dxdydz \\
&\quad + \iiint p(x,y,z)\log\left(p(y|z)\right)dxdydz \\
&\quad - \iiint p(x,y,z)\log\left(p(x|z)\right)dxdydz \\
&\quad - \iiint p(x,y,z)\log\left(p(y|z)\right)dxdydz \\
&= \iiint p(x,y,z)\log\left(p(x|y,z)\right)dxdydz \\
&\quad - \iiint p(x,y,z)\log\left(p(x|z)\right)dxdydz \\
&= \iint -\log(p(x|z))p(x,z)dxdz - \iiint -\log(p(x|y,z))p(x,y,z)dxdydz \\
&= \mathop{\mathbb{E}}_{p(x,z)}[-\log p(x|z)] - \mathop{\mathbb{E}}_{p(x,y,z)}[-\log p(x|y,z)] \\
&= H(X|Z) - H(X|Y,Z)
\end{aligned}
$$

Problem 2

1. In order to show that a distribution is an exponential family, we need to cast it into the form:

$$p(x|\eta) = h(x)\exp[\eta^T T(x) - A(\eta)]$$

Applying this to the multinomial distribution gives:

$$
\begin{aligned}
\mathrm{Mult}(\mathbf{x}|\pi) &= \frac{M!}{x_1!x_2!...x_K!}\pi_1^{x_1}\pi_2^{x_2}...\pi_K^{x_K} \\
&= \frac{M!}{x_1!x_2!...x_K!}\exp[\ln(\pi_1)x_1 + ... + \ln(\pi_K)x_K]
\end{aligned}
$$

$$
\begin{aligned}
h(x) &= \frac{M!}{x_1!x_2!...x_K!} \\
\eta &= \begin{pmatrix} \ln\pi_1 \\ ... \\ \ln\pi_K \end{pmatrix} \\
\pi &= \begin{pmatrix} \exp\eta_1 \\ ... \\ \exp\eta_K \end{pmatrix} \\
T(x) &= \begin{pmatrix} x_1 \\ ... \\ x_K \end{pmatrix} \\
A(\eta) &= 0
\end{aligned}
$$

Where $T(x)$ is the sufficient statistic and $A(\eta)$ is the log-partition function. However, this is not a minimal representation, since we can use the equations $\sum_{i=1}^{K} x_i = M$ and $\sum_{i=1}^{K}\pi_i = 1$ to represent the K-th parameter. We keep $h(x)$ and only consider the exponential term:

$$
\begin{aligned}
&\exp[\ln(\pi_1)x_1 + ... + \ln(\pi_K)x_K] \\
&= \exp[\sum_{i=1}^{K}\ln(\pi_i)x_i] \\
&= \exp[\sum_{i=1}^{K-1}\ln(\pi_i)x_i + \ln(1-\sum_{i=1}^{K-1}\pi_i))(M-\sum_{i=1}^{K-1}x_i)] \\
&= \exp[\sum_{i=1}^{K-1}x_i(\ln\pi_i - \ln(1-\sum_{i=1}^{K-1}\pi_i)) + M\ln(1-\sum_{i=1}^{K-1}\pi_i))]
\end{aligned}
$$

From this we get:

$$\eta = \begin{pmatrix} \ln \frac{\pi_1}{1-\sum_{i=1}^{K-1} \pi_i} \\ \dots \\ \ln \frac{\pi_{K-1}}{1-\sum_{i=1}^{K-1} \pi_i} \end{pmatrix}$$

$$\pi = \begin{pmatrix} \frac{\exp(\eta_1)}{1+\sum_{i=1}^{K-1} \exp(\eta_i)} \\ \dots \\ \frac{\exp(\eta_{K-1})}{1+\sum_{i=1}^{K-1} \exp(\eta_i)} \end{pmatrix}$$

$$T(x) = \begin{pmatrix} x_1 \\ \dots \\ x_{K-1} \end{pmatrix}$$

$$\begin{aligned} A(\eta) &= -M \ln(1 - \sum_{i=1}^{K-1} \pi_i)) \\ &= M \ln(1 + \sum_{i=1}^{K-1} \exp(\eta_i)) \end{aligned}$$

Therefore, the multinomial distribution is an exponential family and we have found its minimal representation.

2. In order to derive the mean and covariance from the log-partition function, we need to take its first and second order derivative.

$$\begin{aligned} A(\eta) &= M \ln(1 + \sum_{i=1}^{K-1} \exp(\eta_i)) \\ \frac{\partial A(\eta)}{\partial \eta_j} &= M \cdot \frac{\exp(\eta_j)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} = M \cdot \pi_j \end{aligned}$$

$$\begin{aligned} \text{cov} = \frac{\partial^2 A(\eta)}{\partial \eta_j \partial \eta_k} &= -M \cdot \frac{\exp(\eta_j)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \cdot \frac{\exp(\eta_k)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \\ &= -M \pi_j \pi_k \\ \text{var} = \frac{\partial^2 A(\eta)}{\partial^2 \eta_j} &= \dots \end{aligned}$$

3. The conjugate prior of an exponential family takes the form:

$$p(\eta|\tau, \nu) \propto \exp[\eta^T \tau - \nu A(\eta)]$$

Using the non-minimal representation of the multinomial distribution, we get:

$$
\begin{aligned}
&\propto \exp\left[\begin{pmatrix} \ln \pi_1 \\ ... \\ \ln \pi_K \end{pmatrix}^T \tau\right] \\
&= \exp\left[\sum_{i=1}^{K} \ln(\pi_i)\tau_i\right] \\
&= \sum_{i=1}^{K} \pi_i \exp[\tau_i] \\
&\propto \sum_{i=1}^{K} \pi_i^{\tau_i}
\end{aligned}
$$

When we set $\boldsymbol{\tau} = \boldsymbol{\alpha} - 1$, this is equivalent to the Dirichlet distribution (without its normalization constant).

4. The prior-to-posterior update rule for the hyperparameters is the same for all distributions that can be represented as an exponential family:

$$
\boldsymbol{\tau} \to \boldsymbol{\tau} + \sum_{n=1}^{N} \mathbf{x}_n
$$

Since our prior does not depend on ν, we do not need an update rule for this parameter (when using the non-minimal representation).

Problem 3

1. According to Bishop, models are known as ICA, when:

   - the observed variables are related linearly to the latent variables
   - the latent distribution is non-Gaussian
   - the distribution over the latent variables factorizes, i.e. $p(\mathbf{z}) = \prod_{j=1}^{M} p(z_j)$

   Here, the last point is equivalent to saying that the latent variables are independent from one another. All three points are fulfilled by the given setting. The observed variables x_{kt} depend linearly on the latent variables s_{it} given the equation. The latent distribution is a Student's T distribution and the sources are assumed to be generated independently. Therefore, the given setting describes an ICA model.

2.

$$
\begin{aligned}
&p(\{s_{1t}\},\{s_{2t}\},\{x_{1t}\},\{x_{2t}\},\{x_{3t}\}) \\
&= \prod_{t=1}^{T}\left(\prod_{j=1}^{2} p(\{s_{jt}\}|\nu_j)\cdot\prod_{j=1}^{3} p(\{x_{jt}\}|\{s_{1t}\},\{s_{2t}\},A_j,\sigma_j)\right) \\
&\text{with:}
\end{aligned}
$$

$$
\begin{aligned}
p(\{s_{jt}\}|\nu_j) &= \mathcal{T}_t(0,\nu_j) \\
p(\{x_{jt}\}|\{s_{1t}\},\{s_{2t}\},A_j,\sigma_j) &= \sum_{i=1}^{2} A_{ji}s_{it}+\epsilon_{jt} \\
&= \sum_{i=1}^{2} A_{ji}\mathcal{T}_t(0,\nu_i)+\mathcal{N}(0,\sigma_j^2) \\
&= \mathcal{N}\left(\sum_{i=1}^{2} A_{ji}\mathcal{T}_t(0,\nu_i),\sigma_j^2\right)
\end{aligned}
$$

3. “Explaining away” occurs in Bayesian networks, when the variables are connected such that they represent the collider case. Here, two variables are independent, but become dependent given a third variable. This phenomenon is also present in the ICA model. Here, the two sources s_1 and s_2 are independent. However, given one of the observed variables and one of the sources, we can directly infer the value of the second source.

4. (a) false
   (b) true
   (c) false
   (d) true
   (e) false
   (f) false
   (g) false
   (h) false

5.

$$
\begin{aligned}
MB^G(s_1) &= \{s_2,x_1,x_2,x_3\} \\
MB^G(x_1) &= \{s_1,s_2\}
\end{aligned}
$$

Since we did not define whether hyperparameters are counted as parents, we decided to exclude them from the Markov blanket.

6.

$$\begin{aligned}
p(\{x_{kt}\}|\mathbf{W},\{\nu_i\}) &= \prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} p(s_{it}) \\
&= \prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} p_i(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}) \\
&= \prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} \mathcal{T}(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}|0,\{\nu_i\})
\end{aligned}$$

7.

$$\log p(\{x_{kt}\}|\mathbf{W},\{\nu_i\}) = \sum_{t=1}^{T} \log(|\det \mathbf{W}|) + \sum_{i=1}^{K_s} \log(\mathcal{T}(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}|0,\{\nu_i\}))$$

8. For the "stochastic gradient ascent" algorithm in ICA, one has to perform the following steps:

Algorithm 1 Stochastic gradient ascent for ICA

Initialize learning rate α
Randomly initialize $\mathbf{W}$

while $||\mathbf{W}^{(\tau-1)} - \mathbf{W}^{(\tau)}|| > \varepsilon$ (i.e. until convergence) **do**
  for each data point $\mathbf{x}(t)$ **do**
    Put $\mathbf{x}$ through a linear mapping:
      $\mathbf{c}(t) = \mathbf{W} \cdot \mathbf{x}(t)$
    Put $\mathbf{c}$ through a nonlinear mapping (popular choice for ϕ is $-\tanh$):
      $z_k(t) = \phi_k(c_k(t))$
    Put $\mathbf{x}$ back through $\mathbf{W}$:
      $\tilde{\mathbf{x}}(t) = \mathbf{W}^T \cdot \mathbf{c}(t)$
    Adjust the weights in accordance with:
      $\mathbf{W}^{(\tau+1)} = \mathbf{W}^{(\tau)} + \alpha \left[\mathbf{W} + \mathbf{z}(t) \cdot \tilde{\mathbf{x}}(t)^T\right]$
After convergence, we can use $\mathbf{W}$ to retrieve the separated signals:
  $\mathbf{s}(t) = \mathbf{W} \cdot \mathbf{x}(t)$

9. For $K \gg T$ the model will overfit, since the number of parameters is much larger than the number of data points.

Problem 4

1. We can write $p(\mathbf{x}_1, ..., \mathbf{x}_{n-1}|\mathbf{x}_n, \mathbf{z}_n) = p(\mathbf{x}_1, ..., \mathbf{x}_{n-1}|\mathbf{z}_n)$, if $\mathbf{x}_1, ..., \mathbf{x}_{n-1}$ are independent of $\mathbf{x}_n$ given $\mathbf{z}_n$ (i.e. $\mathbf{x}_1, ..., \mathbf{x}_{n-1} \perp\!\!\!\perp \mathbf{x}_n|\mathbf{z}_n$).

   This is given, if $\mathbf{x}_1, ..., \mathbf{x}_{n-1}$ are d-separated from $\mathbf{x}_n$ by $\mathbf{z}_n$. We can see that this is indeed the case, as $\mathbf{z}_n$ is an intermediate node between $\mathbf{x}_1, ..., \mathbf{x}_{n-1}$ and $\mathbf{x}_n$, a non-collider and given.

2. We can write $p(\mathbf{x}_1, ..., \mathbf{x}_{n-1}|\mathbf{z}_{n-1}, \mathbf{z}_n) = p(\mathbf{x}_1, ..., \mathbf{x}_{n-1}|\mathbf{z}_{n-1})$, if $\mathbf{x}_1, ..., \mathbf{x}_{n-1}$ are independent of $\mathbf{z}_n$ given $\mathbf{z}_{n-1}$ (i.e. $\mathbf{x}_1, ..., \mathbf{x}_{n-1} \perp\!\!\!\perp \mathbf{z}_n|\mathbf{z}_{n-1}$).

   This is given, if $\mathbf{x}_1, ..., \mathbf{x}_{n-1}$ are d-separated from $\mathbf{z}_n$ by $\mathbf{z}_{n-1}$. We can see that this is indeed the case, as $\mathbf{z}_{n-1}$ is an intermediate node between $\mathbf{x}_1, ..., \mathbf{x}_{n-1}$ and $\mathbf{z}_n$, a non-collider and given.