Philipp Lintl
12152498
philipp.lintl@student.uva.nl

Homework Assignment 3

Machine Learning 2, 19/20

2019-09-23

I worked on it alone, Group E.

1 Problem 1

1.1

$$\begin{aligned}
\text{(Entropy definition)} \qquad H(X,Y) &= \mathbb{E}_{p(x,y)}[-\log p(x,y)]\\
\text{(Expectation rule)} \qquad &= \iint -\log(p(x,y))p(x,y)dxdy\\
(p(x,y) = p(x|y)p(y)) \qquad &= \iint -\log(p(x|y)p(y))p(x,y)dxdy\\
&= \iint -(\log p(x|y) + \log p(y))p(x,y)dxdy\\
&= \iint -\log(p(x|y))p(x,y)dxdy + \iint -\log(p(y))p(x,y)dxdy\\
(\int p(x,y)dx = p(y)) \qquad &= \iint -\log(p(x|y))p(x,y)dxdy + \int -\log(p(y))p(y)dy\\
&= \mathbb{E}_{p(x,y)}[-\log p(x|y)] + \mathbb{E}_{p(y)}[-\log p(y)]\\
&= H(X|Y) + H(Y)
\end{aligned}$$

Analogously, using $p(x,y) = p(y|x)p(x)$ in row three, the other equality can be shown: $H(X,Y) = H(Y|X) + H(X)$

1.2

$$\begin{aligned}
I(X,Y|Z) &= \mathbb{E}_{p(z)}[\mathcal{KL}(p(x,y|z)\|p(x|z)p(y|z))]\\
&= \mathbb{E}_{p(z)}\left[\iint p(x,y|z)\cdot\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right)dxdy\right]\\
&= \iiint p(x,y|z)\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right)p(z)dxdydz\\
(p(x,y|z)p(z) = p(x,y,z)) \qquad &= \iiint p(x,y,z)\log\left(\frac{p(x,y|z)}{p(x|z)p(y|z)}\right)dxdydz\\
(p(x,y|z) = p(x|y,z)p(y|z)) \qquad &= \iiint p(x,y,z)\log(p(x,y|z))dxdydz - \iiint p(x,y,z)\log(p(x|z)p(y|z))dxdydz\\
&= \iiint p(x,y,z)\log(p(x|y,z))dxdyd + \iiint p(x,y,z)\log(p(y|z))dxdydz\\
&- \iiint p(x,y,z)\log p(x|z)dxdydz - \iiint p(x,y,z)\log p(y|z)dxdydz\\
\text{(integrate out y)} \qquad &= \iint -\log(p(x|z))p(x,z)dxdz - \iiint -\log(p(x|y,z))p(x,y,z)dxdydz\\
&= \mathbb{E}_{p(x,z)}[-\log p(x|z)] - \mathbb{E}_{p(x,y,z)}[-\log p(x|y,z)]\\
&= H(X|Z) - H(X|Y,Z)
\end{aligned}$$

2 Problem 2

2.1

1. Similar to the last assignment, a distribution is part of the exponential family when brought to this form

$$p(x|\eta) = h(x)\exp\left(\eta^\top T(x) - A(\eta)\right)$$

Therefore, we take the given distribution function and apply the common exp(log()) trick:

$$\begin{aligned}
Mult(\boldsymbol{x}|\pi) &= \frac{M!}{x_1! \cdot \ldots \cdot x_K!} \pi_1^{x_1} \cdot \ldots \cdot \pi_K^{x_k} \\
&= \frac{M!}{x_1! \cdot \ldots \cdot x_K!} \exp\left(x_1 \ln(\pi_1) + \ldots + x_K \ln(\pi_K)\right) \\
\Rightarrow h(x) &= \frac{M!}{x_1! \cdot \ldots \cdot x_K!} \\
\eta &= \begin{bmatrix} \ln \pi_1 \\ \vdots \\ \ln \pi_K \end{bmatrix} \\
\Rightarrow \pi = \exp(\eta) &= \begin{bmatrix} \exp \eta_1 \\ \vdots \\ \exp \eta_K \end{bmatrix} \\
T(x) &= \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \\
A(\eta) &= 0
\end{aligned}$$

This is not expressed in the minimal number of parameters, as the given conditions are not utilized yet. With $\sum_{i=1}^{K} x_i = M$ and $\sum_{i=1}^{K} \pi_i = 1$:

$$\begin{aligned}
& h(x) \cdot \exp\left(x_1 \ln(\pi_1) + \ldots + x_K \ln(\pi_K)\right) \\
& = h(x) \cdot \exp\left[\sum_{i=1}^{K} x_i \ln(\pi_i)\right] \\
& = h(x) \cdot \exp\left[\sum_{i=1}^{K-1} x_i \ln(\pi_i) + x_K \ln(\pi_K)\right] \\
& = h(x) \cdot \exp\left[\sum_{i=1}^{K-1} x_i \ln(\pi_i) + \left(M - \sum_{i=1}^{K-1} x_i\right) \ln\left(1 - \sum_{i=1}^{K-1} \pi_i\right)\right] \\
& = h(x) \cdot \exp\left[\sum_{i=1}^{K-1} x_i \left(\ln(\pi_i) - \ln(1 - \sum_{i=1}^{K-1} \pi_i)\right) + M \ln\left(1 - \sum_{i=1}^{K-1} \pi_i\right)\right] \\
\Rightarrow h(x) &= \frac{M!}{x_1! \cdot \ldots \cdot x_K!} \\
\eta &= \begin{bmatrix} \ln \frac{\pi_1}{1 - \sum_{i=1}^{K-1} \pi_i} \\ \vdots \\ \ln \frac{\pi_{K-1}}{1 - \sum_{i=1}^{K-1} \pi_i} \end{bmatrix} \\
\Rightarrow \pi = \exp(\eta) &= \begin{bmatrix} \frac{\exp(\eta_1)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \\ \vdots \\ \frac{\eta_{K-1}}{1 + \sum_{i=1}^{K-1} \eta_i} \end{bmatrix} \\
T(x) &= \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
A(\eta) &= -M \ln \left(1 - \sum_{i=1}^{K-1} \pi_i \right) \\
&= -M \cdot \log \left(1 - \sum_{i=1}^{K-1} \frac{e^{\eta_i}}{1 + \sum_{j=1}^{K-1} e^{\eta_j}} \right) \\
&= -M \cdot \log \left(\frac{1}{1 + \sum_{j=1}^{K-1} e^{\eta_j}} \right) \\
&= M \ln \left(1 + \sum_{j=1}^{K-1} \exp(\eta_i) \right)
\end{aligned}$$

Thus, the Multinomial distribution is exponential family with the sufficienct statistic T(x) and the log partition function $A(\eta)$.

2. One ends up with the mean and covariance from the previously derived log-partition function by taking the first and second derivative wrt η:

$$\begin{aligned}
\mathbb{E}(x_j) &= \frac{\partial A(\eta)}{\partial \eta_j} = M \cdot \frac{\exp(\eta_j)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} = M \cdot \pi_j \\
\text{(quotient rule)} \qquad \mathrm{Cov}(x_j, x_k) &= \frac{\partial^2 A(\eta)}{\partial \eta_j \partial \eta_k} = -M \cdot \frac{\exp(\eta_j)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \cdot \frac{\exp(\eta_k)}{1 + \sum_{i=1}^{K-1} \exp(\eta_i)} \\
&= -M \pi_j \pi_k
\end{aligned}$$

Where the first term if the quotient rule disappears, as the numerator is wrt η_j, but we take the derivative for η_k.

3. Once again, taking advantage of the properties of exponential families (Bishop 2.229 with alternative form) shows that a conjugate prior of an exponential family follows the form:

$$p(\eta | \mathcal{X}, \nu) \propto \exp \left[\nu \mathcal{X}^\top \eta - \nu A(\eta) \right]$$

Considering the exponential family representation in minimal representation to this formulation, we end up with:

$$\begin{aligned}
p(\pi | \chi, \nu) &\propto \exp \left[\sum_{i=1}^{K-1} \chi_i \nu \cdot \log \frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j} + \nu \cdot M \cdot \log \left(1 - \sum_{i=1}^{K-1} \pi_i \right) \right] \\
&= \prod_{i=1}^{K-1} \exp \left[\nu \cdot \chi_i \cdot \log \frac{\pi_i}{1 - \sum_{j=1}^{K} \pi_j} \right] \cdot \exp \left[\nu \cdot M \cdot \log \left(1 - \sum_{j=1}^{K-1} \pi_j \right) \right] \\
&= \prod_{i=1}^{K-1} \left(\frac{\pi_i}{1 - \sum_{j=1}^{K-1} \pi_j} \right)^{\nu \cdot X_i} \cdot \left(1 - \sum_{j=1}^{K-1} \pi_j \right)^{M\nu} \\
&= \prod_{i=1}^{K-1} \left(\frac{\pi_i}{\pi_K} \right)^{\chi_i \nu} (\pi_K)^{M\nu} \\
&= \prod_{i=1}^{K-1} \pi_i^{\nu \cdot \chi_i} \cdot \pi_K^{\nu \cdot M - \nu \cdot \Sigma_{j=1}^{K} - 1 \chi_i} \\
&\propto \mathrm{Dir}(\boldsymbol{\pi}, \boldsymbol{\tau}) \\
&\Rightarrow \mathrm{Dir} \left(\{\pi_i\}_1^K, \{ \{\tau_j\}_1^{K-1}, \tau_K \} \} \right)
\end{aligned}$$

$$\text{with} \qquad \tau_j = \nu \cdot \chi_j + 1$$
$$\tau_K = \nu \cdot \left(M - \sum_{j=1}^{K-1} \chi_j \right) + 1$$

Considering, we are disregarding the normalizing constant this is a Dirichlet Distribution. The conjugate prior to the Multinomial distribution is the Dirichlet Distribution: $\Rightarrow \sim Dir(\boldsymbol{\pi}, \boldsymbol{\tau})$

4. For n iid. multinomial observations x we get the prior to posterior update rule by firstly building the posterior with the building blocks derived before:

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

Therefore, the update rule after n datapoints arises as:

$$\tau_i^{(n)} = \tau_i + \sum_{j=1}^{n} x_i^{(j)}$$

3 Problem 3

3.1

1. To be an ICA model, these assumptions must be satisfied (Bishop section 12.4.1):

   - sources/latent distribution are independent (distribution over the latent variables can be factorized) and non Gaussian
   - Time delay is not part of the model
   - The latent variables are linearly related to the observed variables

   All of these points are fulfilled, as the sources are Students'T distributed (thus non Gaussian), the sources are assumed to be generated independently and the observed variables x_{kt} depend linearly on the sources s_{it}.

2. In terms of the general bayesian network epression, the plate notation yields the following joint distribution of latent and observed variables:

$$p\left(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}\right) = \prod_{t=1}^{T} \left(\prod_{j=1}^{2} p\left(\{s_{jt}\} | \nu_j\right) \cdot \prod_{j=1}^{3} p\left(\{x_{jt}\} | \{s_{1t}\}, \{s_{2t}\}, A_j, \sigma_j\right) \right)$$

   The individual distributions were given as:

(sources are each students T) $$p\left(\{s_{jt}\}\,|\nu_j\right) = \mathcal{T}\left(0, \nu_j\right)$$

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

3. ”Explaining away” is related to Bayesian networks and appears when variables are connected such that they represent the collider case, e.g. a variable is dependent on two or more variables. The two otherwise independent variables are dependent given the third variable. Given this, if we now observe information about one of the independent and the third variable, information is gained with respect to the other independent variable.

   This phenomenon appears in the ICA model, too, as the two sources s1 and s2 are independent. Given one of the observed and one of the source variables, the value of the second second source variable can be inferred.

4.
   - False
   - True
   - False
   - True
   - False
   - False
   - False
   - False

5. Given a node X, the Markov blanket is comprised of variables that are parents, children or parents of its children to X. So, for $s_1 = \{s_2, x_1, x_2, x_3\}$ and for $x_1 = \{s_1, s_2\}$. In this scenario, hyperparameters are not considered as variables and thus excluded.

6.

$$\begin{aligned} p\left(\{x_{kt}\}\,|\mathbf{W}, \{\nu_i\}\right) &= \prod_{t=1}^{T} p_S\left(\boldsymbol{W}\boldsymbol{x}_t |\, \{\nu_i\}\right) \left|\det \frac{\partial \boldsymbol{S}}{\partial \boldsymbol{X}}\right| \\ &= \prod_{t=1}^{T} p_S\left(\boldsymbol{W}\boldsymbol{x}_t |\, \{\nu_i\}\right) |\det(\boldsymbol{W})| \\ &= \prod_{i=1}^{T} p_S\left(s_t |\, \{\nu_i\}\right) |\det(\boldsymbol{W})| \\ &= \prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} p\left(s_{it}\right) \\ &= \prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} p_i\left(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}\right) \\ &= \prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} \mathcal{T}\left(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}|0, \{\nu_i\}\right) \end{aligned}$$

7.

$$\begin{aligned} \log p\left(\{x_{kt}\}\,|W, \{\nu_i\}\right) &= \log\left[\prod_{t=1}^{T} |\det \mathbf{W}| \cdot \prod_{i=1}^{K_s} \mathcal{T}\left(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}|0, \{\nu_i\}\right)\right] \\ &= T \log|\det(\boldsymbol{W})| + \sum_{i=1}^{K_s} \log \mathcal{T}\left(\sum_{k=1}^{K_x} \mathbf{W}_{ik}\mathbf{x}_{kt}|0, \{\nu_i\}\right) \end{aligned}$$

8. Generally speaking, the Stochastic Gradient Ascent algorithm maximizes the log-likelihood. It differs from an ordinary gradient ascent algorithm, in that the weight update is conducted after each datapoint and not after iterating through the entire dataset. The parameters in this scenario stem from the mixture matrix, which is randomly initialized. Until convergence is reached, each weight update is applied with respect to the weight gradient and a specified learning rate. A special feature of the SGA for the ICA model is the deployment of a non linear activation function ϕ, which is often chosen to be $tanh()$. After a convergence criterion with regards to the change within the mixture matrix is fulfilled, the sources can be retrieved by multiplying the converged mixture matrix to the signals.

9. I expect overfitting to appear, when $K >> T$, as in general more parameters/degrees of freedom allow to approximate the datapoints almost perfectly. A much smaller number of datapoints than parameters especially allows the almost perfect 'learning' of the few datapoints, which indeed is considered overfitting.

4 Problem 4

1. To proove $p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{z}_n) = p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}|\boldsymbol{z}_n)$, we have to show that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1} \perp\!\!\!\perp \boldsymbol{x}_n|\boldsymbol{z}_n$. This conditional independence is given, if $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}$ are d-separated by $\boldsymbol{z}_n$ from $\boldsymbol{x}_n$. To prove this in this scenario, we show that all paths from $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}$ to $\boldsymbol{x}_n$ pass by $\boldsymbol{z}_n$. Now, this intermediate node blocks the two sets, as it is not an end node, it is a non-collider and given.

2. Analogously, $p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}|\boldsymbol{x}_n, \boldsymbol{z}_n) = p(\boldsymbol{x}_1, \dots, \boldsymbol{z}_{n-1}|\boldsymbol{z}_n)$ holds, if $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}$ are independent of $\boldsymbol{z}_n$ given $\boldsymbol{z}_{n-1}$. This in fact is the case, as $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}$ are d-separated from $\boldsymbol{z}_n$ by $\boldsymbol{z}_{n-1}$. As in 1), $\boldsymbol{z}_{n-1}$ is passed by for all paths from $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n-1}$ to $\boldsymbol{z}_n$ (intermediate node), is a non-collider and already given.

3.

(Bayes Theorem)

$$\begin{aligned}
p(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N|\boldsymbol{z}_n, \boldsymbol{z}_{n+1}) &= \frac{p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1}|\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N) \cdot p(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N)}{p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1})} \\
&= \frac{p(\boldsymbol{z}_n, \boldsymbol{z}_{n+1}|\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N) \cdot p(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N)}{p(\boldsymbol{z}_n|\boldsymbol{z}_{n+1})p(\boldsymbol{z}_{n+1})} \\
&= \frac{p\left(\boldsymbol{z}_n|\boldsymbol{z}_{n+1}\right) \cdot p\left(\boldsymbol{z}_{n+1}|\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N\right) \cdot p\left(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N\right)}{p\left(\boldsymbol{z}_n|\boldsymbol{z}_{n+1}\right) p\left(\boldsymbol{z}_{n+1}\right)} \\
&= \frac{p\left(\boldsymbol{z}_{n+1}|\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N\right) \cdot p\left(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N\right)}{p\left(\boldsymbol{z}_{n+1}\right)} \\
&= p\left(\boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N|\boldsymbol{z}_{n+1}\right)
\end{aligned}$$

Here, we drew on the factorization property (Bishop 8.5), such that $\boldsymbol{z}_n \perp\!\!\!\perp \boldsymbol{x}_{n+1}, \dots, \boldsymbol{x}_N|\boldsymbol{z}_{n+1}$

4. As $\boldsymbol{z}_{N+1}$ is not present in the graph, an extension for this node is assumed. As before, the factorization property yields the independence relation: $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{z}_{N+1}|\boldsymbol{z}_N$, which is drawn on in line two of:

$$\begin{aligned}
p\left(\boldsymbol{z}_{N+1}|\boldsymbol{z}_N, \boldsymbol{X}\right) &= \frac{p\left(\boldsymbol{z}_{N+1}\right) p\left(\boldsymbol{z}_N, \boldsymbol{X}|\boldsymbol{z}_{N+1}\right)}{p\left(\boldsymbol{z}_N, \boldsymbol{X}\right)} \\
&= \frac{p\left(\boldsymbol{z}_{N+1}\right) p\left(\boldsymbol{z}_N|\boldsymbol{X}\right) p\left(\boldsymbol{X}|\boldsymbol{z}_{N+1}\right)}{p\left(\boldsymbol{z}_N|\boldsymbol{X}\right) p(\boldsymbol{X})} \\
&= \frac{p\left(\boldsymbol{z}_{N+1}\right) p\left(\boldsymbol{X}|\boldsymbol{z}_{N+1}\right)}{p(\boldsymbol{X})} \\
&= p\left(\boldsymbol{z}_{N+1}|\boldsymbol{X}\right)
\end{aligned}$$