Machine Learning 2 - Homework 3

Pascal M. Esser

April 22, 2018

Collaborators: Sindy Löewe, Linda Petrini, Alex Geenen, Andrew Sklyar, Gabriele Ce, Davide Belli

1

1.1

$$
\begin{aligned}
H(x,y) &= \mathbb{E}_{p(x,y)}[-\log p(x,y)] && (1)\\
&= \iint -\log[p(x,y)]p(x,y)dxdy && (2)\\
&= \iint (-\log[p(x|y)]) - \log[p(y)]p(x,y)dxdy && (3)\\
&= \iint (-\log[p(x|y)]) - \log[p(y)])p(x|y)p(y)dxdy && (4)\\
&= \iint -\log[p(x|y)])]p(x|y)p(y)dxdy - \iint -\log[p(y)]p(x|y)p(y)dxdy && (5)\\
&= \iint -\log[p(x|y)])]p(x|y)p(y)dxdy - \int -\log[p(y)]p(y)dy && (6)\\
&= H(x|y) - H(y) && (7)\\
&= H(y|x) - H(x) && (8)
\end{aligned}
$$

1.2

$$
\begin{aligned}
I(x,y|z) &= \iiint p(x,y|z)\ln\frac{p(x,y|z)}{p(x|z)p(y|z)}p(z)dxdydz && (9)\\
&= \iiint p(x|y,z)\ln(p(x|y,z))p(y|z)p(z)dxdydz && (10)\\
&\quad - \iiint p(x|y,z)p(y|z)\ln(p(x|z))p(z)dxdydz && (11)\\
&= \iiint p(x,y,z)\ln(p(x|y,z))dxdydz - \iiint p(x,y,z)\ln(p(x|z))dxdydz && (12)\\
&= -H(x|y,z) - \iint \ln(p(x|z)) \int p(x,y,z)dydxdz && (13)\\
&= -H(x|y,z) - \iint \ln(p(x|z))p(x,z)dxdz && (14)\\
&= -H(x|y,z) - H(x|z) && (15)\\
&= H(x|z) - H(x|y,z) && (16)\\
&= H(y|z) - H(y|x,z) && (17)\\
& && (18)
\end{aligned}
$$

2

2.1

$$\begin{aligned}
Mult(x|\pi) =& \frac{M!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \pi_i^{x_i} && (19)\\
=& \frac{M!}{\prod_{i=1}^{k} x_i!} exp\left[\ln\left(\prod_{i=1}^{k} \pi_i^{x_i}\right)\right] && (20)\\
=& \frac{M!}{\prod_{i=1}^{k} x_i!} exp\left[\sum x_i \ln \pi_i\right] && (21)\\
& \text{not minimal. Therefore:} && (22)\\
=& \frac{M!}{\prod_{i=1}^{k} x_i!} exp\left[\sum^{k-1} x_i \ln \pi_i + \left(1 - \sum^{k-1} x_i\right) \ln \left(1 - \sum^{k-1} x_i\right)\right] && (23)\\
=& \frac{M!}{\prod_{i=1}^{k} x_i!} exp\left[\sum^{k-1} x_i \ln \frac{\pi_i}{\pi_k} + M \ln \left(1 - \sum^{k-1} \pi_i\right)\right] && (24)\\
& && (25)
\end{aligned}$$

This gives us:

$$\begin{aligned}
h(x) =& \frac{M!}{\prod_{i=1}^{k} x_i!} && (26)\\
M(x) =& (x_1, ...x_{k-1})^\top && (27)\\
\eta(\pi) =& \left(\ln \frac{\pi_1}{\left(1+\sum^{k-1} exp[\eta_i]\right)}, ... \ln \frac{\pi_{k-1}}{\left(1+\sum^{k-1} exp[\eta_i]\right)}\right)^\top && (28)\\
\eta_i =& \ln(\pi_i) - \ln\left(1 + \sum^{k-1} exp[\eta_i]\right) && (29)\\
A(\eta) =& M \ln\left(1 + \sum^{k-1} exp[\eta_i]\right) = M \ln\left(1 + \sum^{k-1} exp[\eta_i]\right) && (30)\\
g(\eta) =& \left(1 + \sum^{k-1} exp[\eta_i]\right)^M && (31)
\end{aligned}$$

2.2

we will use $A(\eta) = -M \ln(1 + \sum^{k-1} exp[\eta_i])$ to calculate the mean:

$$\begin{aligned}
\frac{\partial A}{\partial \eta_i} =& \frac{exp(\eta_i)}{1 + \sum^{k-1} exp[\eta_i]} M && (32)\\
=& \frac{\pi_i/\pi_k}{1 + \sum^{k-1} \pi_i/\pi_k} M && (33)\\
=& \frac{\pi_j}{\pi_k + \sum \pi_j} M && (34)\\
=& \pi_i M && (35)
\end{aligned}$$

$$\mathbb{E}[x_k] = [M - \sum^{k-1} x_i] \tag{36}$$
$$= M - \sum^{k-1} \mathbb{E}[x_i] \tag{37}$$
$$= M - M \sum \pi_i \tag{38}$$
$$= M\pi_k \tag{39}$$

similar we can calculate the covariance:

$$cov = \frac{\partial^2 A}{\partial \eta_i^2 \partial \eta_j^2} \tag{40}$$
$$= -M \frac{\exp \eta_j \left(1 + \sum^{k-1} exp[\eta_i]\right) - exp[\eta_j] exp[\eta_j]}{\left(1 + \sum^{k-1} exp[\eta_i]\right)^2} \tag{41}$$
$$- M \frac{\exp \eta_j \left(1 + \sum^{k-1} exp[\eta_i] - exp[\eta_j]\right)}{\left(1 + \sum^{k-1} exp[\eta_i]\right)^2} \tag{42}$$
$$= -M \frac{\frac{pi_i}{\pi_k} \frac{1}{\pi_k} - \frac{pi_i^2}{\pi_k^2}}{\frac{1}{\pi_k}} \tag{43}$$
$$= -M\pi_i \pi_j \tag{44}$$

2.3

$$p(\eta|x, v) \propto (\pi_k^M)^v \exp\left[v \sum^{k-1} \ln \frac{\pi_i}{\pi_k} x_i\right] \tag{45}$$
$$= (\pi_k^M)^v \exp\left[v \sum^{k-1} \ln[\pi_i] x_i\right] \exp\left[-v \ln[\pi_k] \sum^{k-1} x_i\right] \tag{46}$$
$$= (\pi_k^M)^v \prod^{k-1} \pi_i^{v x_i} \pi_K^{v\left(1 - \sum_j^{k-1} x_j\right)} \tag{47}$$
$$= \pi_K^{v\left(M - \sum_j^{k-1} x_j\right)} \prod^{k-1} \pi_i^{v x_i} \tag{48}$$
$$\tag{49}$$

This means that the conjugate prior belongs to the family of Dirichlet distributions with parameters:

$$a_j - 1 = v x_j \Rightarrow a_j = 1 + v x_j \text{ iff } j < K \tag{50}$$
$$a_K - 1 = v\left(M - \sum_j^{k-1} x_j\right) \Rightarrow a_K = 1 + v\left(M - \sum_j^{k-1} x_j\right) \text{ iff } j = K \tag{51}$$

2.4

Using the results from the section above and donate (j) as the value of the j-th observation:

$$x_i \to x_i + \sum_{j}^{n} x_i^{(j)} \tag{52}$$

$$v \to v + n \tag{53}$$

3

3.1

Explain why this is an ICA model. In ICA we attempt to decompose a multivariate signal into independent non-Gaussian signals, where the observed signal is a mixture of some unknown sources signals. For the model to be an ICA model two assumptions must hold:

1. "The source signals are independent of each other." This is given in the exercise description.
2. "The values in each source signal have non-Gaussian distributions." This holds as s_{it} is distributed as a zero mean Student's T distribution and only the noise random variable is drawn from a zero mean normal (Gaussian) distribution.

As both assumptions hold we can conclude that the model is indeed an ICA model.

3.2

Write a general (Bayesian network) expression for the joint probability distribution $p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}), t = 1..T$ Factorize the distribution into smaller conditional and marginal distributions as much as possible. Use explicit (conditional) distributions such as Normal and Student's T distributions instead of a generic form p as much as possible.

$$
\begin{aligned}
p(\{s_{1t}\}, \{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}) &= \prod_t^T \prod_i^2 p(\{s_{it}\}|v_i) \prod_i^3 p(x_{it}|\{s_{1t}\}, \{s_{2t}\}, A_i, \sigma_i) && (54) \\
&= \prod_t^T \prod_i^2 \mathcal{T}(s_{it}|0, v_i) \prod_j^3 \sum_i^{K_s} A_{ki}\mathcal{T}(s_{it}|0, v_{ij}) + \mathcal{N}(0, \sigma_{ij}^2) && (55) \\
&= \prod_t^T \prod_i^2 \mathcal{T}(s_{it}|0, v_i) \prod_j^3 \mathcal{N}\left(\sum_i^{K_s} A_{ki}\mathcal{T}(s_{it}|0, v_{ij}), \sigma_{ij}^2\right) && (56)
\end{aligned}
$$

3.3

Explain what the term "explaining away" means and indicate if this explaining away phenomenon is present in the ICA model under discussion.

We have a case of a collider, $A \to B \leftarrow C$, and B, which we observe, could be caused by A or B. Furthermore we we observe a value for B. Now if we see that B is true this automatically reduces our probability that C is true and is causing B.

In the given example we could look at $s_{1t} \to x_{2t} \leftarrow s_{2t}$. Ee could explain away the effect of for example s_{1t} if we see that x_{2t} can be explained by s_{2t}.

3.4

Since samples across time t are independent, we will ignore the index t in the following two questions (you may imagine t = 1). For all of the (conditional) independence expressions below, state if they are true or (typically) false:

1. false
2. true
3. false

4. true
5. false
6. false
7. false
8. false

3.5

What is the Markov blanket of s_1? What is the Markov blanket of x_1?

As we did not discuss MB throughly during the lectures and general descriptions only refer to nodes without specifying if parameters are included we will not include them in the following.

$$MB^G(s_1) = \{\{s_{2t}\}, \{x_{1t}\}, \{x_{2t}\}, \{x_{3t}\}\} \tag{57}$$
$$MB^G(x_1) = \{\{s_{1t}\}, \{s_{2t}\}\} \tag{58}$$

3.6

Write an explicit expression in terms of W and the sources' student's T distributions $T(s_i|0, v_i)$ of the probability: $p(\{x_{kt}\}|W, \{v_i\})$ $t = 1..T, k = 1..K$

$$\prod_t^T p_X(x) = \prod^T p_S(s(x))|detJac(s \to x)| \tag{59}$$
$$\text{with } Jac(s \to x) = Jac\left(\sum^K W_{ik}x_{kt}\right) p(\{x_{xt}\}|\boldsymbol{W}, \{v_i\}) = \frac{\partial \sum^K W_{ik}x_{kt}}{\partial(x_1, ...x_k)} = \boldsymbol{W} \tag{60}$$
$$= \prod_t^T \prod_i^K p(s_{it})det(\boldsymbol{W})| \tag{61}$$
$$= \prod_t^T \prod_i^K \mathcal{T}(0, v_i)|det(\boldsymbol{W})| \tag{62}$$
$$= \prod_t^T |det(\boldsymbol{W})| \prod_i^K \mathcal{T}(0, v_i) \tag{63}$$

3.7

Write down the log-likelihood of the complete deterministic ICA model above.

$$\log p(\{x_{kt}\}|W, \{v_i\}) = \sum_t^T \ln|det(\boldsymbol{W})| + \sum_i^{K_s} \ln \mathcal{T}(0, v_i) \tag{64}$$

3.8

Explain in detail the "stochastic gradient ascent" optimization algorithm to maximize the log- likelihood of the previous question. Note: you do not have to derive or provide the expression of the gradient; instead you can provide a general description of the algorithm.

In line with the question above our goal is to update $\boldsymbol{W}$ by iterative stochastically improving the value of an objective function until convergence by going over all data points.

The steps of the algorithm are described in more detail in the following.

Algorithm 1 stochastic gradient

initialize learning rate η
initialize $\boldsymbol{W}^{(0)}$

while $||\boldsymbol{W}^{(\tau-1)} - \boldsymbol{W}^{(\tau)}|| > \varepsilon$ // until convergence **do**
  for every data point $\boldsymbol{x}$ **do**
    put $\boldsymbol{x}$ through linear mapping: $\boldsymbol{a} = \boldsymbol{W}\boldsymbol{x}$
    put $\boldsymbol{a}$ through a non linear mapping: $z_{-i} = \phi_i(a_i)$
    put $\boldsymbol{a}$ back through $\boldsymbol{W}$: $\boldsymbol{x}' = \boldsymbol{W}^\top \boldsymbol{a}$
    adjust the weights: $\Delta \boldsymbol{W} \propto \boldsymbol{W}^{-1} + \boldsymbol{z}\boldsymbol{x}'^\top$

3.9

In which limit do you expect overfitting: K» T or T» K? Explain your answer.

$K >> T$ because in this case we have an under-constraint problem where we have to many features we want to estimate and not enough data points to estimate them.

4

4.1

From the given conditioning on z_n we see that every path connecting $x_1, ..., x_{n-1}$ to x_n is blocked by z_n which is not an end node or a non-collider which satisfies $z_n \in \{z_n\}$. $x_1, ..., x_{n-1}$ are d-separation from x_n given z_n. Therefore $\{x_1, ..., x_{n-1}\} \perp^d \{x_n\} | z_n \Rightarrow \{x_1, ..., x_{n-1}\} \perp\!\!\!\perp_p \{x_n\} | z_n$ and it follows directly: $p(x_1, ..., x_{n-1} | x_n, z_n) = p(x_1, ..., x_{n-1} | z_n)$

4.2

From the given conditioning on z_{n-1} we see that every path connecting $x_1, ..., x_{n-1}$ to z_n is blocked by z_{n-1} which is not an end node or a non-collider which satisfies $z_{n-1} \in \{z_{n-1}\}$. $x_1, ..., x_{n-1}$ are d-separation from z_n given z_{n-1}. Therefore $\{x_1, ..., x_{n-1}\} \perp^d \{n-1\} | z_n \Rightarrow \{x_1, ..., x_{n-1}\} \perp\!\!\!\perp_p \{n-1\} | z_n$ and it follows directly: $p(x_1, ..., x_{n-1} | z_{n-1}, z_n) = p(x_1, ..., x_{n-1} | z_{n-1})$

4.3

Write out Bayesian Network as the full joint probability distribution:

$$p(x_1, ..., x_N, z_1, ..., z_N) = p(z_1)p(x_1|z_1) \prod_{i=2}^{N} p(z_i|z_{i-1})p(x_i|z_i) \tag{65}$$

This gives:

$$\begin{aligned}
p(x_{n+1}, ..., x_N | z_n, z_{n+1}) =& \frac{p(z_n, z_{n+1} | x_{n+1}, ..., x_N) p(x_{n+1}, ..., x_N)}{p(z_n, z_{n+1})} && (66)\\
=& \frac{p(z_n, z_{n+1}) p(z_{n+1} | x_{n+1}, ..., x_N) p(x_{n+1}, ..., x_N)}{p(z_n | z_{n+1}) p(z_{n+1})} && (67)\\
=& \frac{p(z_{n+1} | x_{n+1}, ..., x_N) p(x_{n+1}, ..., x_N)}{p(z_{n+1})} && (68)\\
& \text{with Bayes rule} && (69)\\
=& p(x_{n+1}, ..., x_N | z_{n+1}) && (70)
\end{aligned}$$

4.4

Since z_{N+1} is not a node in the given graph lets assume its a new node.

$$p(x_1, ..., x_N, z_1, ..., z_N, z_{N+1}) = p(z_1)p(x_1|z_1) \prod_{i=2}^{N} p(z_i|z_{i-1})p(x_i|z_i)p(z_{N+1}|z_N) \tag{71}$$

This gives:

$$\begin{aligned}
p(z_{N+1} | z_n, z_{n+1}) =& \frac{p(z_{N+1}, X | z_N) p(z_{N+1})}{p(z_N, X)} && (72)\\
=& \frac{p(X | z_N, z_{N+1}) p(z_N | z_{N+1}) p(z_{N+1})}{p(X, z_N)} && (73)\\
=& \frac{p(X | z_N) p(z_N | z_{N+1}) p(z_{N+1})}{p(X, z_N) p(z_N)} && (74)\\
& \text{with Bayes rule} && (75)\\
=& p(z_{N+1} | z_N) && (76)
\end{aligned}$$