Julien Fer
10649441
julien.fer88@gmail.com

Homework Assignment 2
Machine Learning 1, 19/20

2019-09-25

1   MAP solution with correlated responses

1.1   Question 1

a) *Write down the likelihood $p(\mathcal{D} \mid \boldsymbol{\theta})$ in vector/matrix form, i.e. in terms of $\boldsymbol{t}, \boldsymbol{\Psi}, \boldsymbol{w}$ and $\boldsymbol{\Omega}$. Note that the distribution can not be factored into independent multiplicands in this basis.*

$$
\begin{aligned}
p(\mathbf{D} \mid \boldsymbol{\theta}) &= p(\mathbf{t} \mid \boldsymbol{\Psi}, \mathbf{w}, \boldsymbol{\Omega}) \\
&= \mathcal{N}(\mathbf{t} \mid \boldsymbol{\Psi}\mathbf{w}, \boldsymbol{\Omega}) \\
&= \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Omega})}} \exp\left(-\frac{1}{2}(\mathbf{t} - \boldsymbol{\Psi}\mathbf{w})^T \boldsymbol{\Omega}^{-1}(\mathbf{t} - \boldsymbol{\Psi}\mathbf{w})\right) \qquad (1)
\end{aligned}
$$

b) *Write the likelihood in terms of a Gaussian distribution with a diagonal covariance matrix by changing the basis of the space in which the targets are expressed. Specifically, express the covariance matrix in its eigenbasis, i.e. write it as $\boldsymbol{\Omega} = \boldsymbol{A}^T\boldsymbol{D}\boldsymbol{A}$ with $\boldsymbol{D} := diag(d_1, \ldots, d_N)$ being a diagonal matrix containing the eigenvalues of $\boldsymbol{\Omega}$ and $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$ being an $\underline{orthogonal}$ change of basis. This is possible in general since covariance matrices are symmetric.*

Let $\boldsymbol{\Omega} := \mathbf{A}^T\mathbf{D}\mathbf{A}$. From (1) we see that it is necessary to determine the determinant and the inverse of the 'new' defined matrix $\boldsymbol{\Omega}$:

$$
\begin{aligned}
\det(\boldsymbol{\Omega}) &= \det(\mathbf{A}^T\mathbf{D}\mathbf{A}) \\
&= \det(\mathbf{A}^{-1}\mathbf{D}\mathbf{A}) \quad \text{(as } \mathbf{A} \text{ is orthogonal)} \\
&= \det(\mathbf{A}^{-1})\det(\mathbf{D})\det(\mathbf{A}) = \det(\mathbf{A})^{-1}\det(\mathbf{D})\det(\mathbf{A}) \\
&= \det(\mathbf{D}) \qquad (2) \\
\boldsymbol{\Omega}^{-1} &= (\mathbf{A}^T\mathbf{D}\mathbf{A})^{-1} = (\mathbf{A}^{-1})(\mathbf{D}^{-1})(\mathbf{A}^T)^{-1} \\
&= \mathbf{A}^T\mathbf{D}^{-1}\mathbf{A} \quad \text{(as } \mathbf{A} \text{ is orthogonal)} \qquad (3)
\end{aligned}
$$

Substituting (2) and (3) in (1) gives the following likelihood:

$$
\begin{aligned}
p(\mathcal{D} \mid \boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{\Omega})}} \exp\left(-\frac{1}{2}\big[(\mathbf{t} - \boldsymbol{\Psi}\mathbf{w})^T \boldsymbol{\Omega}^{-1}(\mathbf{t} - \boldsymbol{\Psi}\mathbf{w})\right) \\
&= \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{D})}} \exp\left(-\frac{1}{2}(\mathbf{t} - \boldsymbol{\Psi}\mathbf{w})^T \mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}(\mathbf{t} - \boldsymbol{\Psi}\mathbf{w})\right) \\
&= \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{D})}} \exp\left(-\frac{1}{2}(\mathbf{t}^T\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}\mathbf{t} - \mathbf{t}^T\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}\boldsymbol{\Psi}\mathbf{w} - \mathbf{w}^T\boldsymbol{\Psi}^T\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}\mathbf{t} + \mathbf{w}^T\boldsymbol{\Psi}^T\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A}\boldsymbol{\Psi}\mathbf{w})\right) \\
&= \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{D})}} \exp\left(-\frac{1}{2}(\boldsymbol{\tau}^T\mathbf{D}^{-1}\boldsymbol{\tau} - \boldsymbol{\tau}^T\mathbf{D}^{-1}\boldsymbol{\Phi}\mathbf{w} - \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{D}^{-1}\boldsymbol{\tau} + \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{D}^{-1}\boldsymbol{\Phi}\mathbf{w})\right) \\
&= \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{D})}} \exp\left(-\frac{1}{2}(\boldsymbol{\tau}^T\mathbf{D}^{-1}\boldsymbol{\tau} - 2\boldsymbol{\tau}^T\mathbf{D}^{-1}\boldsymbol{\Phi}\mathbf{w} + \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{D}^{-1}\boldsymbol{\Phi}\mathbf{w})\right) \quad \text{(as } \mathbf{w}^T\boldsymbol{\Phi}^T\mathbf{D}^{-1}\boldsymbol{\tau} \text{ is scalar/symmetrical)} \\
&= \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{D})}} \exp\left(-\frac{1}{2}(\boldsymbol{\tau} - \boldsymbol{\Phi}\mathbf{w})^T \mathbf{D}^{-1}(\boldsymbol{\tau} - \boldsymbol{\Phi}\mathbf{w})\right) \\
&= \mathcal{N}(\boldsymbol{\tau} \mid \boldsymbol{\Phi}\mathbf{w}, \mathbf{D}) \qquad (4)
\end{aligned}
$$

with $\boldsymbol{\tau} := \mathbf{A}\mathbf{t}$ and $\boldsymbol{\Phi} := \mathbf{A}\boldsymbol{\Psi}$

c) *Factorize the distribution, into a product of univariate Gaussians. Write ϕ_i for the i-th row of the matrix Φ. Hint: The determinant of a diagonal matrix can be expressed as the product of its elements.*:

We can define the determinant of (2) as $\det(\mathbf{D}) = \prod_{i=1}^{N} d_{ii}$. Substiting this in (4) yields:

$$
\begin{aligned}
p(\mathcal{D} \mid \boldsymbol{\theta}) &= \frac{1}{\sqrt{(2\pi)^N \prod_{i=i}^{N} d_{ii}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{N}(\tau_i - \boldsymbol{\phi}_i \mathbf{w})^2 \frac{1}{d_{ii}}\right) \\
&= \prod_{i=i}^{N} \frac{1}{\sqrt{2\pi d_{ii}}} \exp\left(\frac{1}{2 d_{ii}}(\tau_i - \boldsymbol{\phi}_i \mathbf{w})^2\right) \\
&= \prod_{i=i}^{N} \mathcal{N}(\tau_i \mid \boldsymbol{\phi}_i \mathbf{w},\ d_{ii})
\end{aligned}
\tag{5}
$$

d) *Write down the explicit form of the prior $p(\boldsymbol{w})$, i.e. use the expression for a multivariate Gaussian distribution with the correct mean and covariance. Compute the logarithm of the prior $\ln p(\boldsymbol{w})$.*

$$
\begin{aligned}
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha \mathbf{I}) \\
&= \frac{1}{\sqrt{(2\pi)^M \det(\alpha \mathbf{I})}} \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{0})^T (\alpha \mathbf{I})^{-1}(\mathbf{w} - \mathbf{0})\right) \\
&= \frac{1}{\sqrt{(2\pi\alpha)^M}} \exp\left(-\frac{\alpha}{2}\mathbf{w}^T \mathbf{w}\right)
\end{aligned}
\tag{6}
$$

Subsequentely, we take the natural logarithm of (6) to obtain the logarithm of the prior:

$$
\begin{aligned}
\ln p(\mathbf{w}) &= -\frac{M}{2}\ln(2\pi\alpha) - \frac{\alpha}{2}\mathbf{w}^T \mathbf{w} \\
&= -\frac{\alpha}{2}\mathbf{w}^T \mathbf{w} + C
\end{aligned}
\tag{7}
$$

e) *Write down an expression for the posterior $p(\boldsymbol{w} \mid \boldsymbol{D})$ over $\boldsymbol{w}$ by applying Bayes rule. You do not need to write out the explicit form of the Gaussian distributions, instead use the form $\mathcal{N}(a \mid b, c^2)$ with appropriate means b and variances c^2. Show that the evidence will require an integral, which you do not need to solve analytically! However, you need to replace it with a probability distribution like $p(a \mid b, c^2)$ with thecorrect corresponding variables and conditioning variables. (Note that $p(a \mid b, c^2)$ is just an example, there might be more or less than 2 conditioning variables.)*

The Bayes rule is given by:

$$
\overbrace{p(\mathbf{w} \mid \mathcal{D})}^{posterior} = \frac{\overbrace{p(\mathcal{D} \mid \mathbf{w})}^{likelihood} \overbrace{p(\mathbf{w})}^{prior}}{\underbrace{p(\mathcal{D})}_{evidence}}
$$

From (5) we retrieve the *likelihood* and from (6) the *prior*. By applying the product rule with (5) and (6) and integrating over the entire domain of $\mathbf{w}$ we retrieve the *evidence*.

$$
\begin{aligned}
p(\mathbf{w} \mid \mathcal{D}) &= \frac{\prod_{i=i}^{N} \mathcal{N}(\tau_i \mid \boldsymbol{\phi}_i \mathbf{w},\ d_{ii})\, \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha \mathbf{I})}{\int \prod_{i=i}^{N} \mathcal{N}(\tau_i \mid \boldsymbol{\phi}_i \mathbf{w},\ d_{ii})\, \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha \mathbf{I}) \mathrm{d}\mathbf{w}} \\
&= \frac{\mathcal{N}(\boldsymbol{\tau} \mid \boldsymbol{\Phi}\mathbf{w}, \mathbf{D})\, \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha \mathbf{I})}{\int \mathcal{N}(\boldsymbol{\tau} \mid \boldsymbol{\Phi}\mathbf{w}, \mathbf{D})\, \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha \mathbf{I})} \\
&= \frac{\mathcal{N}(\boldsymbol{\tau} \mid \boldsymbol{\Phi}\mathbf{w}, \mathbf{D})\, \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \alpha \mathbf{I})}{p(\boldsymbol{\tau} \mid \boldsymbol{\Phi}, \mathbf{D}, \alpha)}
\end{aligned}
\tag{8}
$$

f) *Compute the log-posterior for both the matrix form of the likelihood as derived in 1.2 and the factorized component form of the likelihood as derived in 1.3. Collect all terms which are independent of $\boldsymbol{w}$ into a constant C. Which parts of the previous expression do not depend on $\boldsymbol{w}$? Why is finding the MAP much simpler than finding the full posterior distribution?*

If we apply Bayes rule as in exercise **e)** and take the logarithm one can find the log-posterior:

$$
\ln p(\mathbf{w} \mid \mathcal{D}) = \ln p(\mathcal{D} \mid \mathbf{w}) + \ln p(\mathbf{w}) - \ln p(\mathcal{D})
\tag{9}
$$