Aleksei Maliutin
1224****
a*****n@student.uva.nl

Homework Assignment 2
Machine Learning 1, 18/19

2018-12-16

1 MAP solution with correlated responses

In practice assignment 2 you derived the maximum a posterior for linear regression assuming independent responses. In general, it is possible that a measurement device which is used to record data is influenced by the measurement process itself, ultimately leading to correlated measurements. We will therefore drop the assumption of independence and assume correlated measurements, quantified by a known, nonsingular covariance matrix $\mathbf{\Omega}$

a)

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

$$\begin{aligned} p(\mathcal{D}|\theta) &= \mathcal{N}(\mathbf{t}|\mathbf{\Phi}\mathbf{w}, \mathbf{\Omega}) \\ &= \frac{1}{2\pi^{N/2}\,|\Omega|^{1/2}} e^{-\frac{1}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^\top \mathbf{\Omega}^{-1}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})} \end{aligned}$$

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
ANSWER:
According to Spectral Theorem, any real symmetric matrix $\mathbf{\Omega}$ could be presented as $\mathbf{U}^\top \Lambda \mathbf{U}$ where $\mathbf{U}$ is a full rank orthogonal matrix containing of the eigenvectors of $\mathbf{\Omega}$ and $\mathbf{\Lambda}$ is a diagonal matrix containing $\mathbf{\Omega}$'s eigenvalues. Hence, we can make the following substitution $\mathbf{\Omega} = \mathbf{A}^\top \mathbf{D}\mathbf{A}$ what leads to:

$$\begin{aligned} p(\mathcal{D}|\theta) &= \frac{1}{2\pi^{N/2}\,|\Omega|^{1/2}} e^{-\frac{1}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^\top \mathbf{\Omega}^{-1}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})} \\ &= \frac{1}{2\pi^{N/2}\,|\mathbf{A}^\top\mathbf{D}\mathbf{A}|^{1/2}} e^{-\frac{1}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^\top (\mathbf{A}^\top\mathbf{D}\mathbf{A})^{-1}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})} \\ &= \frac{1}{2\pi^{N/2}\,|\mathbf{D}|^{1/2}} e^{-\frac{1}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})^\top (\mathbf{A}^{-1}\mathbf{D}^{-1}(\mathbf{A}^\top)^{-1})(\mathbf{t}-\mathbf{\Phi}\mathbf{w})} \\ &= \frac{1}{2\pi^{N/2}\,|\mathbf{D}|^{1/2}} e^{-\frac{1}{2}(\mathbf{A}(\mathbf{t}-\mathbf{\Phi}\mathbf{w}))^\top D^{-1}\mathbf{A}(\mathbf{t}-\mathbf{\Phi}\mathbf{w})} \\ &= \frac{1}{2\pi^{N/2}\,|\mathbf{D}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\tau}-\mathbf{\Psi}\mathbf{w}))^\top D^{-1}(\boldsymbol{\tau}-\mathbf{\Psi}\mathbf{w})} \end{aligned}$$

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

c)

Factorize the distribution, into a product of univariate Gaussians. Write ψ_i for the i-th row of the matrix $\mathbf{\Psi}$.
ANSWER:
Taking into account the fact that *The determinant of a diagonal matrix can be expressed as the product of its elements* as well as the diagonal nature of matrix $\mathbf{D}$

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= \frac{1}{2\pi^{N/2}\left|\mathbf{D}\right|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}))^{\mathsf{T}} D^{-1}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w})} \\
&= \frac{1}{2\pi^{N/2}\left|\mathbf{D}\right|^{1/2}} e^{-\frac{\boldsymbol{\beta}}{2}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}))^{\mathsf{T}}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w})} \\
&= \frac{1}{2\pi^{N/2}\left|\mathbf{D}\right|^{1/2}} e^{-\frac{\boldsymbol{\beta}}{2}\sum_{i=n}^{N}(\tau_n-\mathbf{w}^{\mathsf{T}}\psi_n)^2} \\
&= \frac{1}{2\pi^{N/2}\left|\mathbf{D}\right|^{1/2}} e^{-\frac{\boldsymbol{\beta}}{2}\sum_{i=n}^{N}(\tau_n-\mathbf{w}^{\mathsf{T}}\psi_n)^2} \\
&= \prod_{n=1}^{N} \frac{\beta_i^{1/2}}{2\pi^{1/2}} e^{-\frac{\beta_i}{2}(\tau_n-\mathbf{w}^{\mathsf{T}}\psi_n)^2} \\
&= \prod_{n=1}^{N} \mathcal{N}(\tau_n|\mathbf{w}^{\mathsf{T}}\psi_n, 1/\beta_n)
\end{aligned}
$$

d)

Write down the explicit form of the prior p(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(w).
ANSWER:

$$
\begin{aligned}
p(\mathbf{w}) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I}) \\
&= \frac{\alpha^{N/2}}{(2\pi)^{N/2}} e^{-\frac{\alpha}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}}
\end{aligned}
$$

Compute the logarithm of the prior ln p(w).

$$
\begin{aligned}
\ln p(\mathbf{w}) &= \ln \frac{\alpha^{N/2}}{(2\pi)^{N/2}} e^{-\frac{\alpha}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w}} \\
&= \underbrace{\frac{N}{2}\ln\frac{\alpha}{2\pi}}_{C} - \frac{\alpha}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} \\
&= -\frac{\alpha}{2}\mathbf{w}^{\mathsf{T}}\mathbf{w} + C
\end{aligned}
$$

e)

Write down an expression for the posterior $p(\mathbf{w}|\mathcal{D})$ over $\mathbf{w}$ by applying Bayes rule.
ANSWER:
Taking into account that data samples are not i.i.d., we should use integral for $p(\mathcal{D})$ rather than simple product of probabilities.

$$
\begin{aligned}
p(\mathbf{w}|\mathcal{D}) &= \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \\
&= \frac{\mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \boldsymbol{\Omega})\mathcal{N}(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I})}{\int d\mathbf{w}\mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \boldsymbol{\Omega})\mathcal{N}(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I})} \\
&= \frac{\mathcal{N}(\mathbf{t}|\boldsymbol{\Phi}\mathbf{w}, \boldsymbol{\Omega})\mathcal{N}(\mathbf{w}|\mathbf{0}, \frac{1}{\alpha}\mathbf{I})}{p(\mathbf{t}|\boldsymbol{\Phi}, \alpha, \boldsymbol{\Omega})}
\end{aligned}
$$

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.
ANSWER:

$$
\begin{aligned}
\ln p(\mathbf{w}|\mathcal{D}) &= \ln \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \\
&= \ln \frac{\frac{1}{2\pi^{N/2}|\mathbf{D}|^{1/2}} e^{-\frac{\boldsymbol{\beta}}{2}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}))^\mathsf{T}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w})} \cdot \frac{\alpha^{N/2}}{(2\pi)^{N/2}} e^{-\frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w}}}{p(\mathbf{t}|\boldsymbol{\Phi},\alpha,\boldsymbol{\Omega})} \\
&= -\frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} - \frac{\boldsymbol{\beta}}{2}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}))^\mathsf{T}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}) + C \\
&= -\frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} - \sum_{n=1}^{N} \frac{\beta_n}{2}(\tau_n - \mathbf{w}^\mathsf{T}\psi_n)^2 + C
\end{aligned}
$$

Q1: Which parts of the previous expression do not depend on $\mathbf{w}$?
A1: First of all, $p(\mathcal{D})$ does not depends on it, because it depends only on $\boldsymbol{\Phi}, \alpha, \boldsymbol{\Omega}$. Also, we can collect to constant C all multipliers in front exponents due to their independence regard $\mathbf{w}$.
Q2: Why is finding the MAP much simpler than finding the full posterior distribution?
A2: Because we don't need to calculate the hard integral :)))

g)

Solve for $\mathbf{w}_{MAP}$ by first taking the derivative of the log-posterior with respect to $\mathbf{w}$, then setting it to 0, and finally solving for $\mathbf{w}$.
ANSWER:

$$
\begin{aligned}
\ln p(\mathbf{w}|\mathcal{D}) &= -\frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} - \frac{\boldsymbol{\beta}}{2}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}))^\mathsf{T}(\boldsymbol{\tau}-\boldsymbol{\Psi}\mathbf{w}) + C \\
&= -\frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} - \frac{\boldsymbol{\beta}}{2}(\mathbf{w}^\mathsf{T}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\mathbf{w}) + \boldsymbol{\beta}\mathbf{w}^\mathsf{T}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\tau} + D
\end{aligned}
$$

Now calculate the derivative.

$$
\frac{\partial \ln p(\mathbf{w}|\mathcal{D})}{\partial \mathbf{w}} = -\alpha\mathbf{w}^\mathsf{T} - \boldsymbol{\beta}\mathbf{w}^\mathsf{T}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi} + \boldsymbol{\beta}\boldsymbol{\tau}^\mathsf{T}\boldsymbol{\Phi} = 0
$$

Transpose both sides:

$$
\begin{aligned}
\alpha\mathbf{w} + \boldsymbol{\beta}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\mathbf{w} &= \boldsymbol{\beta}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\tau} \\
\mathbf{w}_{\mathbf{MAP}} &= (\alpha\mathbf{I} + \boldsymbol{\beta}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi})^{-1}\boldsymbol{\beta}\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\tau}
\end{aligned}
$$

h)

Express the solution for $\mathbf{w}_{MAP}$ in terms of the original quantities $\mathbf{t}$ and $\boldsymbol{\Phi}$ to end up with the final solution stated above.
NOTE: $\boldsymbol{\Psi} = \mathbf{A}\boldsymbol{\Phi}$ and $\boldsymbol{\tau} = \mathbf{A}\boldsymbol{t}$

$$
\begin{aligned}
\mathbf{w}_{\mathbf{MAP}} &= (\alpha\mathbf{I} + \boldsymbol{\beta}(\mathbf{A}\boldsymbol{\Phi})^\mathsf{T}\mathbf{A}\boldsymbol{\Phi})^{-1}\boldsymbol{\beta}(\mathbf{A}\boldsymbol{\Phi})^\mathsf{T}\boldsymbol{\tau} \\
&= (\alpha\mathbf{I} + \boldsymbol{\beta}\boldsymbol{\Phi}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\boldsymbol{\Phi})^{-1}\boldsymbol{\beta}\boldsymbol{\Phi}^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\boldsymbol{t} \\
&= (\alpha\mathbf{I} + \boldsymbol{\Phi}^\mathsf{T}\underbrace{\mathbf{A}^\mathsf{T}\mathbf{D}^{-1}\mathbf{A}}_{\boldsymbol{\Omega}^{-1}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\mathsf{T}\underbrace{\mathbf{A}^\mathsf{T}\mathbf{D}^{-1}\mathbf{A}}_{\boldsymbol{\Omega}^{-1}}\boldsymbol{t} \\
&= (\alpha\mathbf{I} + \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Omega}^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\mathsf{T}\boldsymbol{\Omega}^{-1}\boldsymbol{t}
\end{aligned}
$$

what is equal the stated above $\mathbf{w}_{\mathbf{MAP}}$.

2 ML estimate of angle measurements

Find the maximum likelihood estimate of the angle θ, given that you have two independent noisy measurements, c and s, where s is a measure of the cosine, and s is a measure of the sine, of the angle θ. Assume each measurement has a known Gaussian standard deviation σ (same in both cases).
ANSWER:
Based on task we can make a conclusion, that $\mathbf{c}$ and $\mathbf{s}$ has Gaussian distribution with mean $\cos\theta$ and $\sin\theta$ respectively. Hence, we can use the following log-likelihood formula:

$$\ln p(\mathcal{D}|\theta) = -\frac{N}{2}\ln 2\pi\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(c_i - \cos\theta)^2 + (-\frac{N}{2}\ln 2\pi\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(s_i - \sin\theta)^2)$$

Hence, taking the derivative with respect to θ we are getting the next result:

$$\begin{aligned}
\frac{\partial \ln p(\mathcal{D}|\theta)}{\partial\theta} &= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(s_i - \sin\theta)\cos\theta - (c_i - \cos\theta)\sin\theta \\
&= \frac{1}{2\sigma^2}\sum_{n=1}^{N} s_i\cos\theta - \sin\theta\cos\theta - c_i\sin\theta + \cos\theta\sin\theta \\
&= \frac{1}{2\sigma^2}\sum_{n=1}^{N} s_i\cos\theta - c_i\sin\theta \\
&= 0
\end{aligned}$$

Hence,

$$\begin{aligned}
\sum_{n=1}^{N} s_i\cos\theta &= \sum_{n=1}^{N} c_i\sin\theta \\
\frac{\sin\theta}{\cos\theta} &= \frac{\sum_{n=1}^{N} s_i}{\sum_{n=1}^{N} c_i} \\
\theta_{ML} &= \arctan\frac{\sum_{n=1}^{N} s_i}{\sum_{n=1}^{N} c_i}
\end{aligned}$$