8A

Machine Learning II

ID: 5684926 Tristan Scheidemann

8A-1

Auxiliary calculation

Introducing the following notation:

$$\mathrm{row_j}\left(\mathbf{A}_{-\mathrm{j}}\right) = (\mathbf{A}_{\mathrm{j}1} \quad \cdots \quad \mathbf{A}_{j,\mathrm{j}-1} \quad \mathbf{A}_{\mathrm{j,j}+1} \quad \cdots \quad \mathbf{A}_{j,\mathrm{n}}).$$

Above is the j-th row of A without the j-th element

$$\mathrm{col_j}\left(\mathbf{A}_{-j}\right) = (\mathbf{A}_{1\mathrm{j}} \quad \cdots \quad \mathbf{A}_{\mathrm{j}-1,\mathrm{j}} \quad \mathbf{A}_{\mathrm{j}+1,\mathrm{j}} \quad \cdots \quad \mathbf{A}_{\mathrm{n,j}})^{\mathrm{T}}$$

Removing everything which is not w_j from the matrices:

$$\begin{aligned}\mathbf{w}^T\mathbf{\Sigma}_0^{-1}\mathbf{w} \quad &\Rightarrow^{w_j} \; w_j \cdot \mathrm{row}_j(\mathbf{\Sigma}_0^{-1})\mathbf{w} + w_j \cdot \mathbf{w}_{-j}^T \cdot \mathrm{col_j}\left(\mathbf{\Sigma}_{0}{}_{-j}^{-1}\right) \\ &= \left[\mathrm{w_j w_j}\mathbf{\Sigma}_{jj}^{-1} + \mathrm{w_j row_j}\left(\mathbf{\Sigma}_{0}{}_{-\mathrm{j}}^{-1}\right)\mathbf{w}_{-\mathrm{j}}\right] + w_j \cdot \mathbf{w}_{-j}^T \cdot \mathrm{col_j}\left(\mathbf{\Sigma}_{0}{}_{-j}^{-1}\right)\end{aligned}$$

Next term:

$$\begin{aligned}\sum_{\mathrm{i}=1}^{\mathrm{N}} -2\mathrm{t_i}\mathbf{w}^T\mathbf{x}_\mathrm{i} \quad &\Rightarrow^{w_j} \sum_{\mathrm{i}=1}^{\mathrm{N}} -2t_i w_j x_{ij} \\ &= -2w_j \mathrm{col}_j(\mathbf{X})^T\mathbf{t}\end{aligned}$$

Next term:

$$\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathbf{w}^T\mathbf{x}_\mathrm{i})^2 \Rightarrow^{w_j} \sum_{\mathrm{i}=1}^{\mathrm{N}} \left[\sum_{\mathrm{k}=1}^{\mathrm{d}} [\mathrm{w_k} x_{ik} \cdot \mathrm{w_j} x_{ij}] + \sum_{\substack{\mathrm{k}=1 \\ \mathrm{k}\neq\mathrm{j}}}^{\mathrm{d}} [w_j x_{\mathrm{ij}} \cdot w_k x_{ik}]\right]$$

Splitting above:

$$\mathrm{w_j} \sum_{\mathrm{i}=1}^{\mathrm{N}} x_{ij} \sum_{\substack{\mathrm{k}=1 \\ \mathrm{k}\neq\mathrm{j}}}^{\mathrm{d}} [w_k x_{ik}] \quad = \mathrm{w_j} \cdot \mathbf{w}_{-\mathrm{j}}^T \mathbf{X}_{-j}^T \cdot \mathrm{col}_j(\mathbf{X})$$

$$
\begin{aligned}
\mathrm{w_j} \sum_{\mathrm{i=1}}^{\mathrm{N}} x_{ij} \sum_{\mathrm{k=1}}^{\mathrm{d}} \mathrm{w_k} x_{ik} \quad &= \mathrm{w_j} \cdot \mathbf{w}^T \mathbf{X}^T \cdot \mathrm{col}_j(\mathbf{X}) \\
&= \mathrm{w_j} \cdot \mathbf{w}_{-\mathrm{j}}^T \mathbf{X}_{-j}^T \cdot \mathrm{col}_j(\mathbf{X}) + w_j w_j \mathrm{col}_j(\mathbf{X})^T \mathrm{col}_j(\mathbf{X})
\end{aligned}
$$

Calculating $p\left(\mathrm{w_j} | \mathbf{w}_{-\mathrm{j}}, \mathbf{t}, \mathbf{X}, \tau, \Sigma_0\right)$ explicitly. Note: $\mathbf{X}, \Sigma_0$ are hyperparameters that we do not condition on in this model:

$$
\begin{aligned}
p\left(\mathrm{w_j} | \mathbf{w}_{-\mathrm{j}}, \mathbf{t}, \mathbf{X}, \tau, \mathbf{\Sigma}_0\right) \quad &\propto \mathrm{p}(\mathbf{w}, \mathbf{t}, \tau | \mathbf{X}, \mathbf{\Sigma}_0) \\
&= p(\mathbf{t} | \mathbf{X}, \tau, \mathbf{w})\ \mathrm{p}(\mathbf{w})\ \mathrm{p}(\tau | \mathbf{\Sigma}_0) \\
&= \prod_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{N}(t_i | \mathbf{w}^T \mathbf{x}_\mathrm{i}, \mathrm{r}^{-1}) \cdot \mathrm{N}(\mathbf{w} | \mathbf{0}, \mathbf{\Sigma}_0) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \tau^{\alpha-1} e^{-\beta\tau}.
\end{aligned}
$$

Taking the logarithm:

$$
\begin{aligned}
\log \mathrm{p}(\mathbf{w}, \mathbf{t}, \tau | \mathbf{X}, \mathbf{\Sigma}_0) \quad &= -\frac{N}{2}[\log 2\pi - \log \tau] - \frac{1}{2}\tau \sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathrm{t_i} - \mathbf{w}^T \mathbf{x}_\mathrm{i})^2 \\
&+ -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_0| - \frac{1}{2}\mathbf{w}^T \mathbf{\Sigma}_0^{-1} \mathbf{w} \\
&+ \alpha \log \beta - \log \Gamma(\alpha) + (\alpha - 1)\log \tau - \beta\tau.
\end{aligned}
$$

We now remove everything that does not depend on $\mathrm{w_j}$. Note: Because of the logarithm, proportionality to $\mathrm{p}(\mathbf{w}, \mathbf{t}, \tau | \mathbf{X}, \mathbf{\Sigma}_0)$ is now with respect to additive operations.

$$
\begin{aligned}
\log \mathrm{p}(\mathbf{w}, \mathbf{t}, \tau | \mathbf{X}, \mathbf{\Sigma}_0) \quad &\propto -\frac{1}{2}\tau \sum_{\mathrm{i=1}}^{\mathrm{N}} (\mathrm{t_i} - \mathbf{w}^T \mathbf{x}_\mathrm{i})^2 - \frac{1}{2}\ \mathbf{w}^T \mathbf{\Sigma}_0^{-1} \mathbf{w} \\
&\propto -\frac{1}{2}\tau \sum_{\mathrm{i=1}}^{\mathrm{N}} [-2t_i \mathbf{w}^T \mathbf{x}_\mathrm{i} + (\mathbf{w}^T \mathbf{x}_\mathrm{i})^2] - \frac{1}{2}\ \mathbf{w}^T \mathbf{\Sigma}_0^{-1} \mathbf{w} \\
&\propto -\frac{1}{2}\tau \left[\underbrace{-2w_j \mathrm{col}_j(\mathbf{X})^T \mathbf{t}}_{\sum_{\mathrm{i=1}}^{\mathrm{N}}[-2t_i \mathbf{w}^T \mathbf{x}_\mathrm{i}]} + \underbrace{2\mathrm{w_j} \cdot \mathbf{w}_{-\mathrm{j}}^T \mathbf{X}_{-j}^T \cdot \mathrm{col}_j(\mathbf{X}) + w_j w_j \mathrm{col}_j(\mathbf{X})^T \mathrm{col}_j(\mathbf{X})}_{\sum_{\mathrm{i=1}}^{\mathrm{N}}[(\mathbf{w}^T \mathbf{x}_\mathrm{i})^2]} \right] \\
&-\frac{1}{2} \underbrace{\left[\mathrm{w_j w_j} {\mathbf{\Sigma}_0}_{jj}^{-1} + \mathrm{w_j row_j}\left({\mathbf{\Sigma}_0}_{-\mathrm{j}}^{-1}\right)\mathbf{w}_{-\mathrm{j}}\right] + w_j \cdot \mathbf{w}_{-j}^T \cdot \mathrm{col_j}\left({\mathbf{\Sigma}_0}_{-j}^{-1}\right)}_{\mathbf{w}^T \mathbf{\Sigma}_0^{-1} \mathbf{w}}.
\end{aligned}
$$

Rearranging the terms and utilizing that $\mathrm{col_j}\left({\mathbf{\Sigma}_0}_{-j}^{-1}\right) = \mathrm{row_j}\left({\Sigma_0}_{-\mathrm{j}}^{-1}\right)$ because is symmetric the above becomes:

$$
-\frac{1}{2}\left[\mathrm{w_j^2}\left({\mathbf{\Sigma}_0}_{jj}^{-1} + \tau \mathrm{col}_j(\mathbf{X})^T \mathrm{col}_j(\mathbf{X})\right) - 2w_j\left(\tau\left(\mathrm{col}_j(\mathbf{X})^T \mathbf{t} - \mathbf{w}_{-\mathrm{j}}^T \mathbf{X}_{-j}^T \cdot \mathrm{col}_j(\mathbf{X})\right) - \mathrm{row_j}\left(\mathbf{\Sigma}_{\mathbf{0}_{-\mathbf{j}}}^{-\mathbf{1}}\right)\mathbf{w}_{-\mathrm{j}}\right)\right].
$$

We can now complete the square and get a normal distribution with parameters

$$\sigma^2 = \left(\boldsymbol{\Sigma_0}_{jj}^{-1} + \tau \operatorname{col}_j(\mathbf{X})^T \operatorname{col}_j(\mathbf{X})\right)^{-1}$$

$$\mu = \sigma^2 \cdot \left(\tau\left(\operatorname{col}_j(\mathbf{X})^T \mathbf{t} - \mathbf{w}_{-\mathrm{j}}^T \mathbf{X}_{-j}^T \cdot \operatorname{col}_j(\mathbf{X})\right) - \mathrm{row_j}\left(\boldsymbol{\Sigma_0}_{-\mathrm{j}}^{-1}\right)\mathbf{w}_{-\mathrm{j}}\right).$$