4A

Machine Learning II

ID: 5684926 Tristan Scheidemann

4A-1.

Auxiliary calculation

Given a scalar field $f(\boldsymbol{\mu}): \mathbb{R}^D \to \mathbb{R}$ we define

$$\frac{\partial}{\partial \boldsymbol{\mu}} = (\partial \mu_1, \dots, \partial \mu_D).$$

First, we transform via logarithm:

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \Sigma) = -\frac{D\ N}{2} \log 2\pi - \frac{N}{2} \log|\Sigma| - \frac{1}{2} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}).$$

Which leads to

$$\frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu}, \Sigma) \propto \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}).$$

For a given $\mathbf{x}_i \in \mathbb{R}^D$ and single component $\mu_j \in \mathbb{R}$ we get:

$$\begin{aligned} \frac{\partial}{\partial \mu_j} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \quad &= -2\Sigma_{jj}^{-1}\left(x_j - \mu_j\right) - \sum_{\substack{i=1 \\ i \neq j}}^{d} \Sigma_{ji}^{-1}(x_i - \mu_i) - \sum_{\substack{i=1 \\ i \neq j}}^{d} -\Sigma_{ij}^{-1}(x_i - \mu_i) \\ &= -\sum_{i=1}^{d} \Sigma_{ji}^{-1}(x_i - \mu_i) - \sum_{i=1}^{d} \Sigma_{ij}^{-1}(x_i - \mu_i) \\ &= \left\langle -\mathrm{row}_j \Sigma^{-1} \middle| (\mathbf{x}_i - \boldsymbol{\mu}) \right\rangle \left\langle -\mathrm{col_j} \Sigma^{-1} \middle| (\mathbf{x}_i - \boldsymbol{\mu}) \right\rangle, \end{aligned}$$

where $\langle \cdot \mid \cdot \rangle$ denotes the usual dot product in $\mathbb{R}^d$.

Generally, this leads to

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) &= \sum_{i=1}^{N} -\Sigma^{-1}(\mathbf{x}_i - \boldsymbol{\mu}) - {\Sigma^{-1}}^T (\mathbf{x}_i - \boldsymbol{\mu}) \\
&= \sum_{i=1}^{N} [-2\Sigma^{-1}\mathbf{x}_i + 2\Sigma^{-1}\boldsymbol{\mu}].
\end{aligned}
$$

where we utilized that ${\Sigma^{-1}}^T = \Sigma^{-1}$ for a symmetric matrix.

Solving:

$$
\begin{aligned}
& \sum_{i=1}^{N} [-2\Sigma^{-1}\mathbf{x}_i + 2\Sigma^{-1}\boldsymbol{\mu}] &= 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} -2\Sigma^{-1}\mathbf{x}_i &= \sum_{i=1}^{N} -2\Sigma^{-1}\boldsymbol{\mu} \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \Sigma^{-1}\mathbf{x}_i &= \sum_{i=1}^{N} \Sigma^{-1}\boldsymbol{\mu} \\
\Leftrightarrow \quad & \Sigma \sum_{i=1}^{N} \Sigma^{-1}\mathbf{x}_i &= \Sigma \sum_{i=1}^{N} \Sigma^{-1}\boldsymbol{\mu} \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \mathbf{x}_i &= \sum_{i=1}^{N} \boldsymbol{\mu} \\
\Leftrightarrow \quad & \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i &= \boldsymbol{\mu}.
\end{aligned}
$$

Note: Because Σ^{-1} is positive definite it has always an inverse Σ.

4A-2.

Using Bayes' theorem and the given definitions for the probabilities:

$$
\begin{aligned}
p(\mathbf{z}_k = 1|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathbf{z}_k = 1)p(\mathbf{z}_k = 1)}{p(\mathbf{x})} \\
&= \frac{N(\mathbf{x}|\mu_k, \Sigma_k)\, \pi_k}{\sum_{i=1}^{K} N(\mathbf{x}|\mu_i, \Sigma_i)\, \pi_i}.
\end{aligned}
$$

We can interpret $p(\mathbf{z}_k = 1|\mathbf{x})$ as the relative share of a slice $p(\mathbf{x}|\mathbf{z}_k = 1)p(\mathbf{z}_k = 1)$ of the pie $p(\mathbf{x})$. This becomes more apparent by rewriting the above into joint probabilities:

$$
\frac{N(\mathbf{x}|\mu_k, \Sigma_k)\, \pi_k}{\sum_{i=1}^{K} N(\mathbf{x}|\mu_i, \Sigma_i)\, \pi_i} = \frac{p(\mathbf{x}, \mathbf{z}_k = 1)}{\sum_{i=1}^{K} p(\mathbf{x}, \mathbf{z}_i = 1)}.
$$

4A-3.1.

The log likelihood is:

$$\begin{aligned}\log p(\mathbf{X}|\pi, \mathrm{M}, \Sigma) &= \log\left[\prod_{i=1}^{N} p(\mathbf{x}_i|\pi)\right] \\ &= \sum_{i=1}^{N} \log \sum_{j=1}^{K} \pi_j\, N\big(\mathbf{x}_i|\boldsymbol{\mu}_j, \Sigma_j\big).\end{aligned}$$

Implicit relation for maximum likelihood:

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \log p(\mathbf{X}|\pi, \mathrm{M}, \Sigma) = 0$$

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\mu}_k} \log p(\mathbf{X}|\pi, \mathrm{M}, \Sigma) && = 0 \\
\Leftrightarrow \quad & \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{i=1}^{N} \log \sum_{j=1}^{K} \pi_j \; N\left(\mathbf{x}_i | \boldsymbol{\mu}_j, \Sigma_j\right) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \frac{\pi_k \; (2\pi)^{-\frac{D}{2}} \, |\Sigma_\mathrm{k}|^{-\frac{D}{2}} \; \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_\mathrm{j}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right] e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_\mathrm{j}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)}}{\sum_{j=1}^{K} \pi_j \; (2\pi)^{-\frac{D}{2}} \left|\Sigma_\mathrm{j}\right|^{-\frac{D}{2}} e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^T \Sigma_\mathrm{j}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j)}} && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \left(\frac{\pi_k \; N(\mathbf{x}_n | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \; N\left(\mathbf{x}_n | \boldsymbol{\mu}_j, \Sigma_j\right)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_\mathrm{j}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right] \right) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \left(p(\mathbf{z}_k = 1 | \mathbf{x}_i) \cdot -\frac{1}{2} \underbrace{\frac{\partial}{\partial \boldsymbol{\mu}_k} \left[(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_\mathrm{j}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right]}_{see\ Exercise\ 4A-1\ for\ single\ \mathbf{x}_i} \right) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \left(p(\mathbf{z}_k = 1 | \mathbf{x}_i) \cdot -\frac{1}{2} \left[-2\Sigma^{-1} \mathbf{x}_i + 2\Sigma^{-1} \boldsymbol{\mu}_k \right] \right) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) \left[\Sigma^{-1} \mathbf{x}_i - \Sigma^{-1} \boldsymbol{\mu}_k \right] && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} \left(p(\mathbf{z}_k = 1 | \mathbf{x}_i) [\Sigma^{-1} \mathbf{x}_i] - p(\mathbf{z}_k = 1 | \mathbf{x}_i) [\Sigma^{-1} \boldsymbol{\mu}_k] \right) && = 0 \\
\Leftrightarrow \quad & \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) [\Sigma^{-1} \mathbf{x}_i] && = \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) [\Sigma^{-1} \boldsymbol{\mu}_k] \\
\Leftrightarrow \quad & \Sigma \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) [\Sigma^{-1} \mathbf{x}_i] && = \Sigma \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) [\Sigma^{-1} \boldsymbol{\mu}_k] \\
\Leftrightarrow \quad & \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) \mathbf{x}_i && = \sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) \boldsymbol{\mu}_k \\
\Leftrightarrow \quad & \frac{\sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_i)} && = \boldsymbol{\mu}_k \\
\Leftrightarrow \quad & \frac{1}{N_k} \sum_{i=1}^{N} \gamma(\mathbf{z}_{ik}) \mathbf{x}_i && = \boldsymbol{\mu}_k.
\end{aligned}
$$

4A-3.2.

We define

$$
\ln p(\mathbf{X}|\pi, \mathrm{M}, \Sigma) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) = L(\pi_1, \dots, \pi_K, \lambda).
$$

We calculate

$$\left(\frac{\partial}{\partial \pi_1}, \dots, \frac{\partial}{\partial \pi_K}, \frac{\partial}{\partial \lambda}\right) L(\pi_1, \dots, \pi_K, \lambda) = \mathbf{0}.$$

First, we derive with respect to π_i:

$$\begin{aligned}\left(\frac{\partial}{\partial \pi_i}\right) L(\pi_1, \dots, \pi_K, \lambda) &= \left(\frac{\partial}{\partial \pi_i}\right)\left[\sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(\mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k) + \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] \\ &= \sum_{n=1}^{N} \frac{N(\mathbf{x}_n, \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k)} + \lambda.\end{aligned}$$

Setting above to zero yields:

$$\begin{aligned} & \sum_{n=1}^{N} \frac{N(\mathbf{x}_n, \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k)} + \lambda &&= 0 \\ \Leftrightarrow\quad & \sum_{n=1}^{N} \frac{N(\mathbf{x}_n, \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k)} &&= -\lambda \\ \Leftrightarrow\quad & \pi_i \sum_{n=1}^{N} \frac{N(\mathbf{x}_n, \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k)} &&= -\lambda \pi_i \\ \Leftrightarrow\quad & -\frac{1}{\lambda} \sum_{n=1}^{N} \frac{\pi_i N(\mathbf{x}_n, \boldsymbol{\mu}_i, \Sigma_i)}{\sum_{k=1}^{K} \pi_k N(\mathbf{x}_n, \boldsymbol{\mu}_k, \Sigma_k)} &&= \pi_i \\ \Leftrightarrow\quad & -\frac{1}{\lambda} \sum_{n=1}^{N} \frac{p(\mathbf{x}_n | \mathbf{z}_i = 1)\, p(\mathbf{z}_i = 1)}{p(\mathbf{x}_n)} &&= \pi_i \\ \Leftrightarrow\quad & \underbrace{-\frac{1}{\lambda} \sum_{n=1}^{N} p(\mathbf{z}_i = 1 | \mathbf{x}_n)}_{(*)} &&= \pi_i. \end{aligned}$$

Equivalently:

$$\left(\frac{\partial}{\partial \lambda}\right) L(\pi_1, \dots, \pi_K, \lambda) = \sum_{k=1}^{K} \pi_k - 1\ .$$

Substituting $\pi_i = (*)$**:**

$$
\begin{aligned}
& \sum_{k=1}^{K} \pi_k - 1 && = 0 \\
\Leftrightarrow \quad & \sum_{k=1}^{K} -\frac{1}{\lambda} \sum_{n=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_n) - 1 && = 0 \\
\Leftrightarrow \quad & -\sum_{k=1}^{K} \sum_{n=1}^{N} p(\mathbf{z}_k = 1 | \mathbf{x}_n) && = \lambda. \\
\Leftrightarrow \quad & -\underbrace{\sum_{n=1}^{N} \underbrace{\sum_{k=1}^{K} p(\mathbf{z}_k = 1 | \mathbf{x}_n)}_{=\ 1}}_{\mathit{Fubini/Tonelli}} && = \lambda \\
\Leftrightarrow \quad & -N && = \lambda.
\end{aligned}
$$

Substituting $\lambda = -N$ back into $\left(\frac{\partial}{\partial \pi_i}\right) L(\pi_1, \dots, \pi_K, \lambda)$:

$$
\begin{aligned}
-\frac{1}{\lambda} \sum_{n=1}^{N} p(\mathbf{z}_i = 1 | \mathbf{x}_n) \quad & = \frac{1}{N} \sum_{n=1}^{N} p(\mathbf{z}_i = 1 | \mathbf{x}_n) \\
& = \frac{1}{N} \sum_{n=1}^{N} \gamma(\mathbf{z}_{ni}) \\
& = \frac{N_i}{N} \\
& = \pi_i.
\end{aligned}
$$