University of
Stavanger

Stavanger, December 17, 2024

Solutions to theoretical exercise 2

ELE520 Machine learning

Problem 1

a) Assume gaussian densityu functions and use the ML-method to estimate the necessary parameters. The number of elements N_1 and N_2 in class ω_1 and class ω_2 are $N_1 = N_2 = 7$ so that the total number is $n = N_1 + N_2 = 14$.

$$P(\omega_1) = \frac{4}{8} = \frac{1}{2}$$

and

$$P(\omega_2) = \frac{4}{8} = \frac{1}{2}$$

. Furthermore we find the density function parameters:

$$\begin{aligned} \hat{\boldsymbol{\mu}}_1 &= \frac{1}{N_1}\sum_{k=1}^{N_1}\mathbf{x}_k \\ &= \frac{1}{4}\left\{\begin{pmatrix}2\\6\end{pmatrix} + \begin{pmatrix}3\\4\end{pmatrix} + \begin{pmatrix}3\\8\end{pmatrix} + \begin{pmatrix}4\\6\end{pmatrix}\right\} \\ &= \begin{pmatrix}3\\6\end{pmatrix} \end{aligned} \tag{1}$$

and

$$\begin{aligned} \hat{\boldsymbol{\Sigma}}_1 &= \frac{1}{N_1}\sum_{k=1}^{N_1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_1)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_1)^t \\ &= \frac{1}{4}\left\{\begin{pmatrix}-1\\0\end{pmatrix}\begin{pmatrix}-1 & 0\end{pmatrix} + \begin{pmatrix}0\\-2\end{pmatrix}\begin{pmatrix}0 & -2\end{pmatrix} + \begin{pmatrix}0\\2\end{pmatrix}\begin{pmatrix}0 & 2\end{pmatrix} + \begin{pmatrix}1\\0\end{pmatrix}\begin{pmatrix}1 & 0\end{pmatrix}\right\} \\ &= \frac{1}{4}\left\{\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix} + \begin{pmatrix}0 & 0\\0 & 4\end{pmatrix} + \begin{pmatrix}0 & 0\\0 & 4\end{pmatrix} + \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}\right\} \\ &= \begin{pmatrix}0.5 & 0\\0 & 2\end{pmatrix} \end{aligned} \tag{2}$$

for class ω_1. Correspondingly we get

$$\begin{aligned}\hat{\boldsymbol{\mu}}_2 &= \frac{1}{N_2}\sum_{k=1}^{N_2}\mathbf{x}_k \\ &= \begin{pmatrix} 3 \\ -2 \end{pmatrix}\end{aligned} \tag{3}$$

and

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_2 &= \frac{1}{N_2}\sum_{k=1}^{N_2}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_2)(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_2)^t \\ &= \begin{pmatrix} 2.045 & 0.3 \\ 0.3 & 2 \end{pmatrix}\end{aligned} \tag{4}$$

b) The contour lines of the two class specific probability density functions are shown in figure 1. Compared to the corresponding figure from the proposed solution to exercise 2 we see that the estimates are unreliable due to few samples so that:

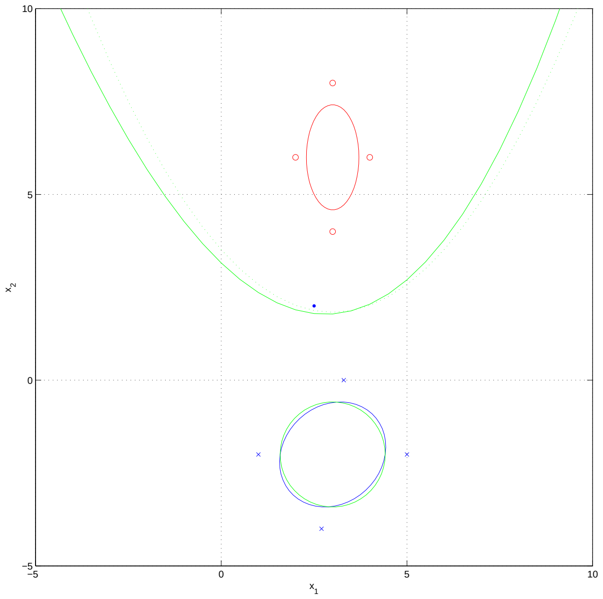

Figure 1: Desisjonsgrenser og konturlinjer for klassebetingede tetthetsfunksjoner.

- The eigenvectors are no longer parallell to the axes, and the orientation of the density functions will thus get different.

- The relationship between the eigenvalues for each of the two classes has also been changed so that the contour line for class two is no longer circular.

Due to this difference between estimated and true parameters, the decision boundary will also be different. Due to rotation of principal axes we get nonlinear terms in the discriminant functions as will be seen.

c) Bayes decision rule can be formulated as

$$\text{Decide} \begin{cases} \omega_1, & \text{if } P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}) \\ \omega_2, & \text{otherwise} \end{cases} \tag{5}$$

The decision border can be found by solving the equation $P(\omega_1|\boldsymbol{x}) = P(\omega_2|\boldsymbol{x})$. We will rather use the discriminant functions for unequal covariance matrices and solve the equation $g_1(\boldsymbol{x}) = g_2(\boldsymbol{x})$ where $g_1(\boldsymbol{x})$ is the same as the one we found in the previous exercise that you can compare.

$$\begin{aligned} g_i(\boldsymbol{x}) &= -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \\ &= \boldsymbol{x}^t\boldsymbol{\Theta}_i\boldsymbol{x} + \boldsymbol{\theta}_i^t\boldsymbol{x} + \theta_{i0} \\ &\text{where} \\ \boldsymbol{\Theta}_i &= -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1} \\ \boldsymbol{\theta}_i &= \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \\ \theta_{i0} &= -\frac{1}{2}\boldsymbol{\mu}_i^t\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \end{aligned} \tag{6}$$

Compute

$$\begin{aligned} \boldsymbol{\Theta}_1 &= -\frac{1}{2}\boldsymbol{\Sigma}_1^{-1} \\ &= -\frac{1}{2}\begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix} \\ &= \begin{pmatrix} -1 & 0 \\ 0 & -0.25 \end{pmatrix}, \end{aligned} \tag{7}$$

$$\begin{aligned} \boldsymbol{\theta}_1 &= \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}\begin{pmatrix} 3 \\ 6 \end{pmatrix} \\ &= \begin{pmatrix} 6 \\ 3 \end{pmatrix} \end{aligned} \tag{8}$$

og

$$\begin{aligned}\theta_{10} &= -\frac{1}{2}\boldsymbol{\mu}_1^t\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_1) \\ &= -\frac{1}{2}\begin{pmatrix}3 & 6\end{pmatrix}\begin{pmatrix}2 & 0\\ 0 & 0.5\end{pmatrix}\begin{pmatrix}3\\ 6\end{pmatrix} - 0 - \ln 1/2 \\ &= -18.693\end{aligned} \tag{9}$$

which gives

$$\begin{aligned}g_1(\boldsymbol{x}) &= \boldsymbol{x}^t\boldsymbol{\Theta}_1\boldsymbol{x} + \boldsymbol{\theta}_1^t\boldsymbol{x} + \theta_{10} \\ &= \begin{pmatrix}x_1 & x_2\end{pmatrix}\begin{pmatrix}-1 & 0\\ 0 & -0.25\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} + \begin{pmatrix}6 & 3\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} - 18.693 \\ &= -x_1^2 - 0.25x_2^2 + 6x_1 + 3x_2 - 18.69\end{aligned} \tag{10}$$

for class ω_1.

Compute

$$\begin{aligned}\boldsymbol{\Theta}_2 &= -\frac{1}{2}\boldsymbol{\Sigma}_2^{-1} \\ &= -\frac{1}{2}\begin{pmatrix}0.5 & -0.075\\ -0.075 & 0.51125\end{pmatrix} \\ &= \begin{pmatrix}-0.25 & 0.0375\\ 0.0375 & -0.255625\end{pmatrix},\end{aligned} \tag{11}$$

$$\begin{aligned}\boldsymbol{\theta}_2 &= \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 \\ &= \begin{pmatrix}0.5 & -0.075\\ -0.075 & 0.51125\end{pmatrix}\begin{pmatrix}3\\ -2\end{pmatrix} \\ &= \begin{pmatrix}1.65\\ -1.2475\end{pmatrix}\end{aligned} \tag{12}$$

and

$$\begin{aligned}\theta_{20} &= -\frac{1}{2}\boldsymbol{\mu}_2^t\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2 - \frac{1}{2}\ln|\boldsymbol{\Sigma}_2| + \ln P(\omega_2) \\ &= -\frac{1}{2}\begin{pmatrix}3 & -2\end{pmatrix}\begin{pmatrix}-0.25 & 0.0375\\ 0.0375 & -0.255625\end{pmatrix}\begin{pmatrix}3\\ -2\end{pmatrix} - 1/2\ln 4 - \ln 1/2 \\ &= -5.1088\end{aligned} \tag{13}$$

which gives

$$\begin{aligned} g_2(\boldsymbol{x}) &= \boldsymbol{x}^t \boldsymbol{\Theta}_2 \boldsymbol{x} + \boldsymbol{\theta}_2^t \boldsymbol{x} + \theta_{20} \\ &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} -0.25 & 0.0375 \\ 0.0375 & -0.255625 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} 1.65 & -1.2475 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 5.1088 \\ &= -0.25x_1^2 - 0.2556x_2^2 + 0.075x_1x_2 + 1.65x_1 - 1.248x_2 - 5.109 \end{aligned} \tag{14}$$

for class ω_2.

We then find the decision border by solving

$$\begin{aligned} g_1(\boldsymbol{x}) &= g_2(\boldsymbol{x}) \\ &\Downarrow \\ g_1(\boldsymbol{x}) - g_2(\boldsymbol{x}) &= 0 \\ -x_1^2 - 0.25x_2^2 + 6x_1 + 3x_2 - 18.69 & \\ -(-0.25x_1^2 - 0.2556x_2^2 + 0.075x_1x_2 + 1.65x_1 - 1.248x_2 - 5.109) &= 0 \\ -x_1^2 - 0.25x_2^2 + 6x_1 + 3x_2 - 18.69 & \\ +0.25x_1^2 + 0.2556x_2^2 - 0.075x_1x_2 - 1.65x_1 + 1.248x_2 + 5.109 &= 0 \\ -0.75x_1^2 + 0.0056x_2^2 + 4.35x_1 + 4.248x_2 - 13.58 - 0.075x_1x_2 &= 0 \end{aligned} \tag{15}$$

This wil ne a complicated expression to solve, but the main point is that compared to the solutiopn from the previous exercise we get cross-terms, ax_1x_2.

$$x_2 = -377.6 + 6.67x_1 + 0.26 \times 10^{-7} \pm \sqrt{0.21 \times 10^{21} - 0.83 \times 10^{19} x_1 + 0.25 \times 10^{18} x_1^2)} \tag{16}$$

The decision border is shown in solid line in figure 1. Compared to the true decision border shown in dotted line we can see the differences, mainly that the decision border is not symmetric.

d) The best way to improve the estimates will be by increasing the amount of data.

Problem 2

The clasification is done by computing $g_i(\boldsymbol{x}), i = 1, 2$ using the discriminant functions that we found in the previous problem and choose class corresponding to the largest value, where $\boldsymbol{x} = (2.5\ 2.0)^t$.

a) For the ML classifier we compute

$$\begin{aligned} g_1(\boldsymbol{x}) &= -x_1^2 - 0.25x_2^2 + 6x_1 + 3x_2 - 18.69 \\ &= -2.5^2 - 0.25 \times 2.0^2 + 6 \times 2.5 \times 2.0 - 18.69 \\ &= -4.94 \end{aligned} \tag{1}$$

and similarly we compute

$$g_2(\boldsymbol{x}) \;=\; -5.69 \tag{2}$$

As $g_1(\boldsymbol{x}) > g_2(\boldsymbol{x})$ we classify $\boldsymbol{x}$ as belonging to ω_1.

b) For the Parzen-classifier with $h_1 = 0.5$ which is computed with $h_N = h_1/\sqrt{N_1} = 0.5/\sqrt{4} = 0.25$: It might be useful to show that the estimate $p_n(\boldsymbol{x}|\omega_1)$ correspond to summing gaussian densities with covariance matrices $\mathbf{\Sigma} = h_N^2\mathbf{I}$ and mean vectors in the training samples, $\boldsymbol{x}_i$, when the window function is chosen to be a normalised gaussian.

$$\phi(\mathbf{u}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{I}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\mathbf{u})^t\mathbf{I}^{-1}(\mathbf{u})}. \tag{3}$$

We then have

$$\begin{aligned}
g_1(\boldsymbol{x}) &= P(\omega_1)p_n(\boldsymbol{x}|\omega_1) \\
&= P(\omega_1)\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{1}{V_N}\phi(\mathbf{u}) \\
&= P(\omega_1)\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{1}{h_N^d}\frac{1}{(2\pi)^{\frac{d}{2}}|\mathbf{I}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h_N})^t\mathbf{I}^{-1}(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h_N})} \\
&= P(\omega_1)\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{1}{(2\pi)^{\frac{d}{2}}|h_N^2\mathbf{I}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_i)^t(h_N^2\mathbf{I})^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)} \\
&= P(\omega_1)\frac{1}{N_1}\sum_{i=1}^{N_1}\frac{1}{(2\pi h_N^2)^{\frac{d}{2}}} \cdot e^{-\frac{1}{2h_N^2}(\boldsymbol{x}-\boldsymbol{x}_i)^t(\boldsymbol{x}-\boldsymbol{x}_i)} \\
&= \frac{1}{2}\frac{1}{4}(8.86\times10^{-57} + 4.36\times10^{-15} + 2.89\times10^{-126} + 9.98\times10^{-64}) \\
&= 1.46\times10^{-16}
\end{aligned} \tag{4}$$

and correspondingly

$$\begin{aligned}
g_2(\boldsymbol{x}) &= P(\omega_2)p_n(\boldsymbol{x}|\omega_2) \\
&= \frac{1}{2}\frac{1}{4}(9.98\times10^{-64} + 1.55\times10^{-125} + 1.93\times10^{-16} + 1.26\times10^{-77}) \\
&= 2.41\times10^{-17}
\end{aligned} \tag{5}$$

As $g_1(\boldsymbol{x}) > g_2(\boldsymbol{x})$, $\boldsymbol{x}$ is classified to ω_1.

c) For the Parzen-classifier with $h_1 = 5$ we do similar computations (with $h_N = h_1/\sqrt{N_1} = 5/\sqrt{4} = 2.5$)

$$\begin{aligned}
g_1(\boldsymbol{x}) &= P(\omega_1)p_n(\boldsymbol{x}|\omega_1) \\
&= \frac{1}{2}\frac{1}{4}(0.0069 + 0.0181 + 0.0014 + 0.0059) \\
&= 0.0040
\end{aligned} \tag{6}$$

and correspondingly

$$\begin{aligned} g_2(\boldsymbol{x}) &= P(\omega_2)p_n(\boldsymbol{x}|\omega_2) \\ &= 0.0037 \end{aligned} \tag{7}$$

As $g_1(\boldsymbol{x}) > g_2(\boldsymbol{x})$, $\boldsymbol{x}$ is classified to ω_1.

The classification has not changed. The question survived the revision from a former version were another data set was used when the wider window functions made the contribution from the nearest cluster of data (three samples in former version) for class ω_1 bigger than the closer cluster og data with only one sample from class ω_2 . With smaller window width, the cluster of three samples were too far away to give a signioficant contribution, thus giving classification to class ω_2.

d) For the k_N-nearest-neighbor classifier with $k_N = 1$ the set of distances to the feature vectors are computed to be

$$\begin{aligned} R_1 &= \left\{ \left\| \begin{pmatrix} 2 \\ 6 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\|, \left\| \begin{pmatrix} 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\|, \left\| \begin{pmatrix} 3 \\ 8 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\|, \left\| \begin{pmatrix} 4 \\ 6 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\| \right\} \\ &= \{4.03, 2.06, 6.02, 4.27\} \end{aligned} \tag{8}$$

and

$$\begin{aligned} R_2 &= \left\{ \left\| \begin{pmatrix} 1 \\ -2 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\|, \left\| \begin{pmatrix} 2.7 \\ -4 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\|, \left\| \begin{pmatrix} 3.3 \\ 0 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\|, \left\| \begin{pmatrix} 5 \\ -2 \end{pmatrix} - \begin{pmatrix} 2.5 \\ 2 \end{pmatrix} \right\| \right\} \\ &= \{4.27, 6.00, 2.15, 4.71\} \end{aligned} \tag{9}$$

For ω_1 the $k_N = 1$ nearest sample is chosen, giving $r = 2.06$ and furthermore $V_N = \pi r^2 = 13.25$

$$\begin{aligned} g_1(\boldsymbol{x}) &= P(\omega_1)p_n(\boldsymbol{x}|\omega_1) \\ &= P(\omega_1)\frac{k_N/N_1}{V_N} \\ &= 0.5\frac{1/4}{13.35} \\ &= 0.0094 \end{aligned} \tag{10}$$

and correspondingly for ω_2 we choose the $k_N = 1$ nearest sample, giving $r = 2.15$ and furthermore $V_N = \pi r^2 = 14.58$

$$\begin{aligned} g_2(\boldsymbol{x}) &= P(\omega_2)p_n(\boldsymbol{x}|\omega_2) \\ &= 0.0086. \end{aligned} \tag{11}$$

As $g_1(\boldsymbol{x}) > g_2(\boldsymbol{x})$, $\boldsymbol{x}$ is classified to class ω_1.

e) For the k_N-nearest neighbor-classifier with $k_N = 3$ we use the same distances R_1 and R_2. For ω_1 we then choose the $k_N = 3$ nearest sample giving $r = 4.27$ and furthermore $V_N = \pi r^2 = 57.33$

$$\begin{aligned} g_1(\boldsymbol{x}) &= P(\omega_1)p_n(\boldsymbol{x}|\omega_1) \\ &= 0.0065 \end{aligned} \tag{12}$$

and similarly for ω_2 we choose the $k_N = 3$ nearest sample, which gives $r = 4.71$ and furthermore $V_N = \pi r^2 = 69.90$

$$\begin{aligned} g_2(\boldsymbol{x}) &= P(\omega_2)p_n(\boldsymbol{x}|\omega_2) \\ &= 0.0053 \end{aligned} \tag{13}$$

As $g_1(\boldsymbol{x}) > g_2(\boldsymbol{x})$, $\boldsymbol{x}$ is classified to class ω_1.

Problem 3

The d-dimensional multivariate normal distribution is given as:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{l}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \cdot e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{1}$$

We select $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ independent samples from $p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The likelihood is then

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{l}{2}}|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \cdot e^{-\frac{1}{2}\sum_{k=1}^{N}(\boldsymbol{x}_k-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k-\boldsymbol{\mu})} \tag{2}$$

The log-likelihood-function to $\boldsymbol{\mu}$ og $\boldsymbol{\Sigma}$ is

$$\begin{aligned} l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -\frac{n}{2}\ln(2\pi)^l - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{k=1}^{N}(\boldsymbol{x}_k - \boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_k - \boldsymbol{\mu}) \\ &= -\frac{n}{2}\ln(2\pi)^l - \frac{n}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}[\sum_{k=1}^{N}\boldsymbol{x}_k^t\boldsymbol{\Sigma}^{-1}\boldsymbol{x}_k - 2\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{N}\boldsymbol{x}_k + n\boldsymbol{\mu}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]. \end{aligned} \tag{3}$$

We find $\hat{\boldsymbol{\mu}}$ by differentiating and setting to zero according to

$$\frac{\partial l(\boldsymbol{\mu}, \boldsymbol{\Sigma})}{\partial \boldsymbol{\mu}} = -\frac{1}{2}[-2\boldsymbol{\Sigma}^{-1}\sum_{k=1}^{N}\boldsymbol{x}_k + n2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}] = \boldsymbol{0}, \tag{4}$$

and find

$$\hat{\boldsymbol{\Sigma}}^{-1}\sum_{k=1}^{N}\boldsymbol{x}_k = N\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}. \tag{5}$$

This gives the maximum-likelihood-solution

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{k=1}^{N} \boldsymbol{x}_k, \tag{6}$$

as we expected.