Exercise 3

Fundamentals of Machine Learning 8 ECTS
Heidelberg University WiSe 20/21

Catherine Knobloch, Elias Olofsson, Julia Siegl

December 15, 2020

1 Computing $\hat{b}$ (4 points)

The optimal parameters $\hat{w}$ and $\hat{b}$ in Linear Discriminant Analysis are the ones that minimize the least squares error criterion

$$\hat{w}, \hat{b} = \text{argmin}_{w,b} \sum_{i=1}^{N} \left(w^T x_i + b - y_i\right)^2, \tag{1}$$

where $w, x \in \mathbb{R}^D$ are column vectors, and $b \in \mathbb{R}$ and $y \in \{-1, 1\}$ are scalars. In order to find these optimal parameters, we start by solving for the minimum of eq.(1) with respect to b as per

$$\frac{\partial}{\partial b} \sum_{i=1}^{N} \left(w^T x_i + b - y_i\right)^2 = 0. \tag{2}$$

Thus, carrying out the derivative for the left hand side, we have

$$\begin{aligned}
0 &= \frac{\partial}{\partial b} \sum_{i=1}^{N} \left(w^T x_i + b - y_i\right)^2 && (3)\\
&= \frac{\partial}{\partial b} \sum_{i=1}^{N} \left[(w^T x_i - y_i)^2 + 2(w^T x_i - y_i)b + b^2\right] && (4)\\
&= \sum_{i=1}^{N} \left[2(w^T x_i - y_i) + 2b\right] && (5)\\
&= 2Nb + 2\sum_{i=1}^{N}(w^T x_i - y_i) && (6)\\
\implies \quad \hat{b} &= -\frac{1}{N} \sum_{i=1}^{N}(w^T x_i - y_i). && (7)
\end{aligned}$$

We can furthermore continue rewriting this expression by noting that we can split up the sum for the two different classes as per

$$\begin{aligned}
\hat{b} &= -\frac{1}{N}\sum_{i=1}^{N}(w^T x_i - y_i) \qquad (8)\\
&= -\frac{1}{N}\left[\sum_{i:y_1=1}(w^T x_i - 1) + \sum_{i:y_1=-1}^{N}(w^T x_i + 1)\right] \qquad (9)\\
&= -\frac{1}{N}\left[w^T \sum_{i:y_1=1} x_i - N_1 + w^T \sum_{i:y_1=-1} x_i + N_{-1}\right] \qquad (10)\\
&= -\frac{1}{N}\left[N_1 w^T \mu_1 + N_{-1} w^T \mu_{-1}\right] \qquad (11)
\end{aligned}$$

where we used the two class means μ_1, μ_{-1} defined as

$$\begin{aligned}
\mu_1 &= \frac{1}{N_1}\sum_{i:y_i=1} x_i \qquad (12)\\
\mu_{-1} &= \frac{1}{N_{-1}}\sum_{i:y_i=-1} x_i. \qquad (13)
\end{aligned}$$

By assuming that the two classes are balanced, i.e. that

$$N_1 = N_{-1} = \frac{N}{2}, \tag{14}$$

where N_k denotes the number of instances in each class, we finally arrive at the expression

$$\hat{b} = -\frac{1}{2} w^T\left(\mu_1 + \mu_{-1}\right), \tag{15}$$

which is the optimal parameter $\hat{b}$ that minimizes the least squares error of LDA.

2 Intermediate equation (16 points)

To find the optimal parameter $\hat{w}$, we similarly try to find the minimum of the least squares error with respect to w, by solving

$$\frac{\partial}{\partial w}\sum_{i=1}^{N}\left(w^T x_i + \hat{b} - y_i\right)^2 = 0. \tag{16}$$

Thus, we start by substituting in eq.(15) into the left hand side of eq.(16) and carrying out the derivative using the chain rule as per

$$0 = \frac{\partial}{\partial w} \sum_{i=1}^{N} \left(w^T x_i - \frac{1}{2} w^T (\mu_1 + \mu_{-1}) - y_i \right)^2 \tag{17}$$

$$= 2 \sum_{i=1}^{N} \frac{\partial}{\partial w} \left[w^T x_i - \frac{1}{2} w^T (\mu_1 + \mu_{-1}) - y_i \right] \left(w^T x_i - \frac{1}{2} w^T (\mu_1 + \mu_{-1}) - y_i \right) \tag{18}$$

$$= 2 \sum_{i=1}^{N} \left(x_i - \frac{1}{2} (\mu_1 + \mu_{-1}) \right) \left(w^T x_i - \frac{1}{2} w^T (\mu_1 + \mu_{-1}) - y_i \right) \tag{19}$$

$$\begin{aligned} &= 2 \sum_{i=1}^{N} \left(x_i - \frac{1}{2} (\mu_1 + \mu_{-1}) \right) \left(w^T x_i - \frac{1}{2} w^T (\mu_1 + \mu_{-1}) \right) \\ &\quad - \quad 2 \sum_{i=1}^{N} x_i y_i \quad + \quad (\mu_1 + \mu_{-1}) \sum_{i=1}^{N} y_i. \end{aligned} \tag{20}$$

We can handle the two last terms in eq.(20) separately. For the first of the two, we have

$$2 \sum_{i=1}^{N} x_i y_i = 2 \sum_{i:y_i=1} x_i - 2 \sum_{i:y_i=-1} x_i = 2N_1 \mu_1 - 2N_{-1} \mu_{-1} = N(\mu_1 - \mu_{-1}), \tag{21}$$

and for the second term

$$(\mu_1 + \mu_{-1}) \sum_{i=1}^{N} y_i = (\mu_1 + \mu_{-1}) \left[\sum_{i:y_i=1} 1 + \sum_{i:y_i=1} (-1) \right] \tag{22}$$

$$= (\mu_1 + \mu_{-1}) [N_1 - N_{-1}] = 0, \tag{23}$$

where we used in the last expression that the two classes are balanced. Plugging the results from these two expressions back into eq.(20) and continuing by expanding the terms in the equation, we have

$$\begin{aligned} 0 = 2 \sum_{i=1}^{N} \Big[& x_i w^T x_i - \frac{1}{2} x_i w^T (\mu_1 + \mu_{-1}) - \frac{1}{2} (\mu_1 + \mu_{-1}) w^T x_i \\ & + \frac{1}{4} (\mu_1 + \mu_{-1}) w^T (\mu_1 + \mu_{-1}) \Big] - N(\mu_1 - \mu_{-1}) \end{aligned} \tag{24}$$

$$\begin{aligned} = \sum_{i=1}^{N} \Big[& x_i w^T x_i + x_i w^T x_i - x_i w^T (\mu_1 + \mu_{-1}) - (\mu_1 + \mu_{-1}) w^T x_i \\ & + \frac{1}{2} (\mu_1 + \mu_{-1}) w^T (\mu_1 + \mu_{-1}) \Big] - N(\mu_1 - \mu_{-1}) \end{aligned} \tag{25}$$

$$
\begin{aligned}
=&\sum_{i=1}^{N}\Bigg[x_i w^T(x_i-\mu_1)+x_i w^T(x_i-\mu_{-1})-(\mu_1+\mu_{-1})\,w^T x_i \\
&+\frac{1}{2}(\mu_1+\mu_{-1})\,w^T\,(\mu_1+\mu_{-1})\Bigg]-N(\mu_1-\mu_{-1})
\end{aligned}
\tag{26}
$$

$$
\begin{aligned}
=&\sum_{i=1}^{N}\Bigg[x_i w^T(x_i-\mu_1)+x_i w^T(x_i-\mu_{-1})-\mu_1 w^T(x_i-\mu_1+\mu_1) \\
&-\mu_{-1}w^T(x_i-\mu_{-1}+\mu_{-1})+\frac{1}{2}(\mu_1+\mu_{-1})\,w^T\,(\mu_1+\mu_{-1})\Bigg] \\
&-N(\mu_1-\mu_{-1})
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
=&\sum_{i=1}^{N}\Bigg[x_i w^T(x_i-\mu_1)+x_i w^T(x_i-\mu_{-1})-\mu_1 w^T(x_i-\mu_1)-\mu_{-1}w^T(x_i-\mu_{-1}) \\
&-\mu_1 w^T\mu_1-\mu_{-1}w^T\mu_{-1}+\frac{1}{2}(\mu_1+\mu_{-1})\,w^T\,(\mu_1+\mu_{-1})\Bigg]-N(\mu_1-\mu_{-1})
\end{aligned}
\tag{28}
$$

$$
\begin{aligned}
=&\sum_{i=1}^{N}\Bigg[(x_i-\mu_1)w^T(x_i-\mu_1)+(x_i-\mu_{-1})w^T(x_i-\mu_{-1})-\mu_1 w^T\mu_1 \\
&-\mu_{-1}w^T\mu_{-1}+\frac{1}{2}(\mu_1+\mu_{-1})\,w^T\,(\mu_1+\mu_{-1})\Bigg]-N(\mu_1-\mu_{-1})
\end{aligned}
\tag{29}
$$

$$
\begin{aligned}
=&\sum_{i:y_i=1}\big[(x_i-\mu_1)w^T(x_i-\mu_1)+(x_i-\mu_{-1})w^T(x_i-\mu_{-1})\big] \\
&+\sum_{i:y_i=-1}\big[(x_i-\mu_1)w^T(x_i-\mu_1)+(x_i-\mu_{-1})w^T(x_i-\mu_{-1})\big] \\
&+\sum_{i=1}^{N}\left[\frac{1}{2}(\mu_1+\mu_{-1})w^T(\mu_1+\mu_{-1})-\mu_1 w^T\mu_1-\mu_{-1}w^T\mu_{-1}\right]. \\
&-N(\mu_1-\mu_{-1})
\end{aligned}
\tag{30}
$$

Furthermore, by utilizing the relation given in the exercise sheet for the general vectors a, b and c

$$
a\cdot\left(b^T\cdot c\right)=\left(a\cdot c^T\right)\cdot b,
\tag{31}
$$

we can reshuffle eq.(30) into

$$\begin{aligned} 0 = &\left[\sum_{i:y_i=1} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{i:y_i=-1} (x_i - \mu_{-1})(x_i - \mu_{-1})^T\right] w \\ &+ \sum_{i:y_i=1} (x_i - \mu_{-1}) w^T (x_i - \mu_{-1}) + \sum_{i:y_i=-1} (x_i - \mu_1) w^T (x_i - \mu_1) \\ &+ \sum_{i=1}^{N} \left[\frac{1}{2}(\mu_1 + \mu_{-1}) w^T (\mu_1 + \mu_{-1}) - \mu_1 w^T \mu_1 - \mu_{-1} w^T \mu_{-1}\right] \\ &- N(\mu_1 - \mu_{-1}) \end{aligned} \tag{32}$$

$$\begin{aligned} = &\left[\sum_{i=1}^{N} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T\right] w \\ &+ \sum_{i:y_i=1} (x_i - \mu_{-1}) w^T (x_i - \mu_{-1}) + \sum_{i:y_i=-1} (x_i - \mu_1) w^T (x_i - \mu_1) \\ &+ N \left[\frac{1}{2}(\mu_1 + \mu_{-1}) w^T (\mu_1 + \mu_{-1}) - \mu_1 w^T \mu_1 - \mu_{-1} w^T \mu_{-1}\right] \\ &- N(\mu_1 - \mu_{-1}) \end{aligned} \tag{33}$$

$$\begin{aligned} = &N \cdot S_W w \\ &+ \sum_{i:y_i=1} (x_i - \mu_{-1}) w^T (x_i - \mu_{-1}) + \sum_{i:y_i=-1} (x_i - \mu_1) w^T (x_i - \mu_1) \\ &+ \frac{N}{2}(\mu_1 + \mu_{-1}) w^T (\mu_1 + \mu_{-1}) - N\left[\mu_1 w^T \mu_1 - \mu_{-1} w^T \mu_{-1}\right] \\ &- N(\mu_1 - \mu_{-1}) \end{aligned} \tag{34}$$

where S_W is the within-class covariance matrix defined as

$$S_W = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu_{y_i})(x_i - \mu_{y_i})^T, \tag{35}$$

where the notation μ_{y_i} means

$$\mu_{y_i} = \begin{cases} \mu_{-1} & \text{if} \quad y_i = -1 \\ \mu_1 & \text{if} \quad y_i = 1. \end{cases} \tag{36}$$

Taking each the two sums left in eq.(34) separately, we start by expanding the first term as per

$$\begin{aligned}
&\sum_{i:y_i=1} (x_i - \mu_{-1}) w^T (x_i - \mu_{-1}) = && (37)\\
&\quad = \sum_{i:y_i=1} \left[x_i w^T x_i - x_i w^T \mu_{-1} - \mu_{-1} w^T x_i + \mu_{-1} w^T \mu_{-1} \right] && (38)\\
&\quad = \sum_{i:y_i=1} x_i w^T x_i - \frac{N}{2} \mu_1 w^T \mu_{-1} - \frac{N}{2} \mu_{-1} w^T \mu_1 + \frac{N}{2} \mu_{-1} w^T \mu_{-1}, && (39)
\end{aligned}$$

where we used the fact that the set of instances are balanced. For the second sum in eq.(34) we similarly have

$$\begin{aligned}
&\sum_{i:y_i=-1} (x_i - \mu_1) w^T (x_i - \mu_1) = && (40)\\
&\quad = \sum_{i:y_i=-1} x_i w^T x_i - \frac{N}{2} \mu_{-1} w^T \mu_1 - \frac{N}{2} \mu_1 w^T \mu_{-1} + \frac{N}{2} \mu_1 w^T \mu_1, && (41)
\end{aligned}$$

and joining these two terms together, the expression simplifies to

$$\begin{aligned}
&\sum_{i:y_i=1} (x_i - \mu_{-1}) w^T (x_i - \mu_{-1}) + \sum_{i:y_i=-1} (x_i - \mu_1) w^T (x_i - \mu_1) =\\
&\quad = \sum_{i=1}^{N} x_i w^T x_i - N \left[\mu_{-1} w^T \mu_1 + \mu_1 w^T \mu_{-1} \right] + \frac{N}{2} \left[\mu_1 w^T \mu_1 + \mu_{-1} w^T \mu_{-1} \right]. && (42)
\end{aligned}$$

Furthermore, we can take the fourth term of eq.(34) in isolation and rewrite it as per

$$\begin{aligned}
&\frac{N}{2} (\mu_1 + \mu_{-1}) w^T (\mu_1 + \mu_{-1}) = && (43)\\
&\quad = \frac{N}{2} (\mu_1 - \mu_{-1} + 2\mu_{-1}) w^T (\mu_1 - \mu_{-1} + 2\mu_{-1}) && (44)\\
&\quad = \frac{N}{2} (\mu_1 - \mu_{-1}) w^T (\mu_1 - \mu_{-1}) + N \mu_{-1} w^T (\mu_1 - \mu_{-1})\\
&\quad\ \ + \ N (\mu_1 - \mu_{-1}) w^T \mu_{-1} + 2N \mu_{-1} w^T \mu_{-1} && (45)\\
&\quad = \frac{N}{2} (\mu_1 - \mu_{-1})(\mu_1 - \mu_{-1})^T w + N \mu_{-1} w^T \mu_1 - N \mu_{-1} w^T \mu_{-1}\\
&\quad\ \ + \ N \mu_1 w^T \mu_{-1} - N \mu_{-1} w^T \mu_{-1} + 2N \mu_{-1} w^T \mu_{-1} && (46)\\
&\quad = \frac{N}{2} S_B w + N \left[\mu_{-1} w^T \mu_1 + \mu_1 w^T \mu_{-1} \right], && (47)
\end{aligned}$$

where we again used the relation in eq.(31), and the definition of the between-class covariance matrix

$$S_B = (\mu_1 - \mu_{-1}) (\mu_1 - \mu_{-1})^T. \qquad (48)$$

Substituting the results of eq.(42) and eq.(47) back into eq.(34), we obtain

$$\begin{aligned}0 = {} & N \cdot S_W w \\ & + \sum_{i=1}^{N} x_i w^T x_i - N \left[\mu_{-1} w^T \mu_1 + \mu_1 w^T \mu_{-1}\right] + \frac{N}{2} \left[\mu_1 w^T \mu_1 + \mu_{-1} w^T \mu_{-1}\right] \\ & + \frac{N}{2} S_B w + N \left[\mu_{-1} w^T \mu_1 + \mu_1 w^T \mu_{-1}\right] - N \left[\mu_1 w^T \mu_1 - \mu_{-1} w^T \mu_{-1}\right] \\ & - N(\mu_1 - \mu_{-1})\end{aligned} \tag{49}$$

$$\begin{aligned}= {} & N \cdot S_W w \\ & + \sum_{i=1}^{N} x_i w^T x_i - \frac{N}{2} \left[\mu_1 w^T \mu_1 + \mu_{-1} w^T \mu_{-1}\right] \\ & + \frac{N}{2} S_B w \\ & - N(\mu_1 - \mu_{-1}).\end{aligned} \tag{50}$$

Then, we can take the second and third term in eq.(50) and rewrite them as

$$\sum_{i=1}^{N} x_i w^T x_i - \frac{N}{2} \left[\mu_1 w^T \mu_1 + \mu_{-1} w^T \mu_{-1}\right] = \tag{51}$$

$$\begin{aligned}= & \sum_{i:y_i=1} x_i w^T x_i - N_1 \mu_1 w^T \mu_1 - N_1 \mu_1 w^T \mu_1 + N_1 \mu_1 w^T \mu_1 \\ & + \sum_{i:y_i=-1} x_i w^T x_i - N_{-1} \mu_{-1} w^T \mu_{-1} - N_{-1} \mu_{-1} w^T \mu_{-1} + N_{-1} \mu_{-1} w^T \mu_{-1}\end{aligned} \tag{52}$$

$$\begin{aligned}= & \sum_{i:y_i=1} \left(x_i w^T x_i - x_i w^T \mu_1 - \mu_1 w^T x_i + \mu_1 w^T \mu_1\right) \\ & + \sum_{i:y_i=-1} \left(x_i w^T x_i - x_i w^T \mu_{-1} - \mu_{-1} w^T x_i + \mu_{-1} w^T \mu_{-1}\right)\end{aligned} \tag{53}$$

$$\begin{aligned}= & \sum_{i:y_i=1} \left(x_i w^T x_i - x_i w^T \mu_1 - \mu_1 w^T x_i + \mu_1 w^T \mu_1\right) \\ & + \sum_{i:y_i=-1} \left(x_i w^T x_i - x_i w^T \mu_{-1} - \mu_{-1} w^T x_i + \mu_{-1} w^T \mu_{-1}\right)\end{aligned} \tag{54}$$

$$\begin{aligned}= & \sum_{i:y_i=1} \left(x_i w^T (x_i - \mu_1) - \mu_1 w^T (x_i - \mu_1)\right) \\ & + \sum_{i:y_i=-1} \left(x_i w^T (x_i - \mu_{-1}) - \mu_{-1} w^T (x_i - \mu_{-1})\right)\end{aligned} \tag{55}$$

$$\tag{56}$$

$$= \sum_{i:y_i=1} (x_i - \mu_1) w^T (x_i - \mu_1) + \sum_{i:y_i=-1} (x_i - \mu_{-1}) w^T (x_i - \mu_{-1}) \tag{57}$$

$$= \sum_{i=1}^{N} (x_i - \mu_{y_i}) w^T (x_i - \mu_{y_i}) \tag{58}$$

$$= \sum_{i=1}^{N} (x_i - \mu_{y_i}) (x_i - \mu_{y_i})^T w \tag{59}$$

$$= N \cdot S_W w, \tag{60}$$

where we again used the relation in eq.(31), and the definition of the within-class covariance matrix from eq.(35). Thus, we can plug eq.(60) back into eq.(50) and write the full equation as

$$0 = N S_W w + N S_W w + \frac{N}{2} S_B w - N(\mu_1 - \mu_{-1}) \tag{61}$$

$$= 2N \left[\left(S_W + \frac{1}{4} S_B \right) w - \frac{\mu_1 - \mu_{-1}}{2} \right], \tag{62}$$

where we finally arrive at the intermediate equation

$$\left(S_W + \frac{1}{4} S_B \right) \hat{w} = \frac{\mu_1 - \mu_{-1}}{2}, \tag{63}$$

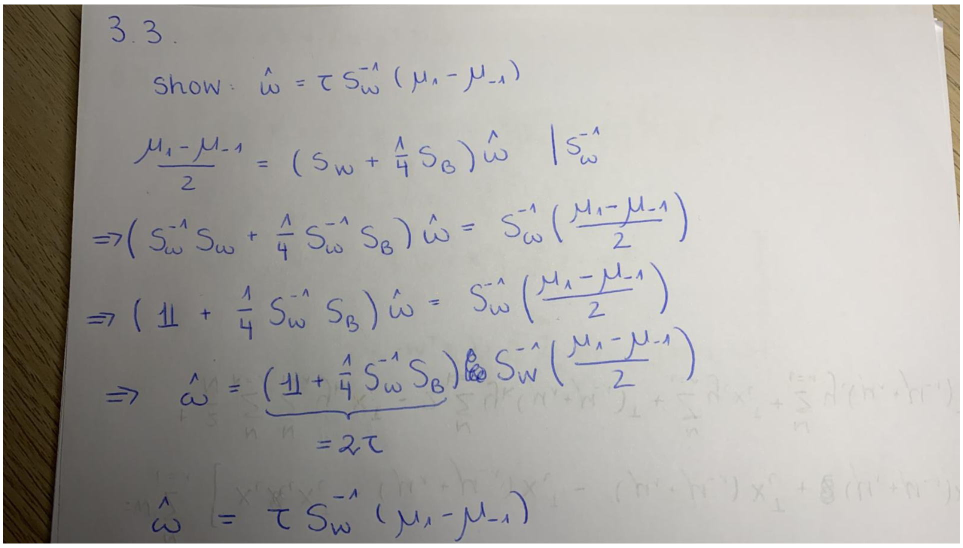
for the optimal parameter $\hat{w}$ for LDA. □

3 Final transformation

We did not really reach a conclusion for this part of the exercise, but here we present part of our effort to solve the problem in fig.(1). We believe that in order to perform the final transformation such that an explicit expression of $\hat{w}$ can be obtained, we need to show that

$$S_W^{-1} S_B \propto \mathbb{I}, \tag{64}$$

where $\mathbb{I}$ is the identity matrix. However, we are not sure on how to show this at this stage.



3.3.

Show: $\hat{\omega} = \tau S_\omega^{-1} (\mu_1 - \mu_{-1})$

$$\frac{\mu_1 - \mu_{-1}}{2} = (S_w + \frac{1}{4} S_B)\hat{\omega} \quad | S_\omega^{-1}$$

$$\Rightarrow (S_\omega^{-1} S_\omega + \frac{1}{4} S_\omega^{-1} S_B)\hat{\omega} = S_\omega^{-1}\left(\frac{\mu_1 - \mu_{-1}}{2}\right)$$

$$\Rightarrow (\mathbb{1} + \frac{1}{4} S_\omega^{-1} S_B)\hat{\omega} = S_\omega^{-1}\left(\frac{\mu_1 - \mu_{-1}}{2}\right)$$

$$\Rightarrow \hat{\omega} = \underbrace{(\mathbb{1} + \frac{1}{4} S_\omega^{-1} S_B)}_{=2\tau} S_W^{-1}\left(\frac{\mu_1 - \mu_{-1}}{2}\right)$$

$$\hat{\omega} = \tau S_W^{-1}(\mu_1 - \mu_{-1})$$

Figure 1