Problem Sheet I

3.1 LDA Derivation from the Least Squares Error

We are looking for the global minimum of

$$\Delta : \mathbb{R}^{d+1} \to \mathbb{R} \quad (\mathbf{m}, b) \mapsto \sum_{i=1}^{N} (\mathbf{w}^{\mathbf{T}} \mathbf{x_i} + b - y_i)^2 = \sum_{i=1}^{N} (\mathbf{x_i^T} \mathbf{w} + b - y_i)^2 \tag{1}$$

First, we take a closer look at the summands. Let $i \in \{1, ..., N\}$.
The function def. by $f(x) := x^2$ is in $C^\infty(\mathbb{R})$ with derivative $f'(x) = 2x$. For the function

$$g_i : \mathbb{R}^{d+1} \to \mathbb{R} \quad (\mathbf{m}, b) \mapsto \mathbf{w}^{\mathbf{T}} \mathbf{x_i} + b - y_i \tag{2}$$

holds for $k \in 1, ..., d$, $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$:

$$\partial_{w_k} g_i(\mathbf{w}, b) = \partial_{w_k} \left(\sum_{j=1}^{d} x_{ij} w_j + b - y_i \right) = \sum_{j=1}^{d} x_{ij} \delta_{jk} = x_{ik} \tag{3}$$

$$\partial_b g_i(\mathbf{w}, b) = 1 \tag{4}$$

The partial derivatives are continuous, thus $g_i\ inC^1(\mathbb{R}^{+\not\Vdash}$ As a composition/sum of C^1 functions, Δ is a C^1 function as well and

$$\begin{aligned} D\Delta(\mathbf{w}, b) = D \left(\sum_{i=1}^{N} f \circ g_i \right) (\mathbf{w}, b) &= \sum_{i=1}^{N} Df(g_i(\mathbf{w}, b)) \cdot Dg_i(\mathbf{w}, b) \\ & \sum_{i=1}^{N} 2 g_i(\mathbf{w}, b) \cdot (\nabla_{\mathbf{w}} g_i(\mathbf{w}, b)^T, \partial_b g_i(\mathbf{w}, b)) \\ & \sum_{i=1}^{N} 2(\mathbf{x_i^T} \mathbf{w} + b - y_i)(\mathbf{x_i^T}, 1) \end{aligned}$$

$$\Rightarrow \nabla_{(\mathbf{w}, b)} = 2 \sum_{i=1}^{N} (\mathbf{x_i^T} \mathbf{w} + b - y_i) \begin{pmatrix} \mathbf{x_i^T} \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \sum_{i=1}^{N} (\mathbf{x_i^T} \mathbf{w} + b - y_i) \mathbf{x_i^T} \\ 2 \sum_{i=1}^{N} (\mathbf{x_i^T} \mathbf{w} + b - y_i) \end{pmatrix}$$

Because $\Delta \in C^1(\mathbb{R}^{d+1})$ and global maxima in an open set are local maxima, it holds for the argmax $(\hat{\mathbf{w}}, \hat{b})$:

$$\nabla_{(\mathbf{w}, b)} \Delta(\hat{\mathbf{w}}, \hat{\mathbf{b}}) = 0$$

This implies

$$
\begin{aligned}
\partial_b \Delta(\hat{\mathbf{w}}, \hat{\mathbf{b}}) = 0 &\Rightarrow 0 = \sum_{i=1}^{N} (\mathbf{x_i^T} \hat{\mathbf{w}} + \hat{b} - y_i) \\
&\Rightarrow 0 = N\hat{b} + \sum_{i=1}^{N} (\mathbf{x_i^T} \hat{\mathbf{w}} - y_i) \\
&\Rightarrow \hat{b} = \frac{1}{N} \sum_{i=1}^{N} (-\mathbf{x_i^T} \hat{\mathbf{w}} + y_i) = \frac{-1}{N} \sum_{i=1}^{N} \mathbf{x_i^T} \hat{\mathbf{w}} + \sum_{i:y_i=1} 1 - \sum_{i:y_i=-1} 1 \overset{\text{balanced}}{=} -\frac{1}{N} \sum_{i=1}^{N} \mathbf{x_i^T} \hat{\mathbf{w}}
\end{aligned}
$$

Furthermore $\Delta(\hat{\mathbf{w}}, \hat{\mathbf{b}}) = 0$ implies

$$
0 = \sum_{i=1}^{N} (\mathbf{x_i^T} \hat{\mathbf{w}} + \hat{b} - y_i) \mathbf{x_i}
$$

We insert our result for $\hat{b}$ into this equation:

$$
\begin{aligned}
0 &= \sum_{i=1}^{N} \left[\mathbf{x_i^T} \hat{\mathbf{w}} - \frac{1}{N} \sum_{j=1}^{N} \mathbf{x_j^T} \hat{\mathbf{w}} - y_i \right] \mathbf{x_i} \\
\Rightarrow \underbrace{\frac{1}{N} \sum_{i=1}^{N} y_i \mathbf{x_i}}_{\text{a)}} &= \underbrace{-\frac{1}{N} \sum_{i=1}^{N} \frac{1}{N} \sum_{j=1}^{N} (\mathbf{x_j^T} \hat{\mathbf{w}}) \mathbf{x_i}}_{\text{b)}} + \underbrace{\frac{1}{N} \sum_{i=1}^{N} (\mathbf{x_i^T} \hat{\mathbf{w}}) \mathbf{x_i}}_{\text{c)}}
\end{aligned}
$$

We will separately discuss the three terms a), b) and c):

a)

$$
\begin{aligned}
\frac{1}{N} \sum_{i=1}^{N} y_i \mathbf{x_i} &= \frac{1}{N} \sum_{i:y_i=1} \mathbf{x_i} - \frac{1}{N} \sum_{i:y_i=-1} \mathbf{x_i} \\
&= \frac{1}{2} \left(\frac{1}{N/2} \sum_{i:y_i=1} \mathbf{x_i} - \frac{1}{N/2} \sum_{i:y_i=-1} \mathbf{x_i} \right) \\
&\overset{\text{balanced}}{=} \frac{1}{2} \left(\frac{1}{N_1} \sum_{i:y_i=1} \mathbf{x_i} - \frac{1}{N_2} \sum_{i:y_i=-1} \mathbf{x_i} \right) \\
&= (\mu_{\mathbf{1}} - \mu_{-\mathbf{1}})/2
\end{aligned}
$$

b)

$$-\frac{1}{N}\sum_{i=1}\frac{1}{N}\sum_{j=1}^{N}(\mathbf{x_j^T}\hat{\mathbf{w}})\mathbf{x_i} = \left[-\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}\right]\left[\left(\frac{1}{N}\sum_{j=1}^{N}\mathbf{x_j^T}\right)\hat{\mathbf{w}}\right]$$

$$= -\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}\right)\left(\frac{1}{N}\sum_{j=1}^{N}\mathbf{x_j^T}\right)\right]\hat{\mathbf{w}}$$

$$= -\left(\left[\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x_i}y_i\right)+\left(\frac{2}{N}\sum_{i:y_i=1}\mathbf{x_i}y_i\right)\right]\left[\left(\frac{1}{N}\sum_{j=1}^{N}\mathbf{x_j^T}y_j\right)+\left(\frac{2}{N}\sum_{j:y_j=1}\mathbf{x_j^T}y_j\right)\right]\right)$$

$$= -(\frac{1}{2}(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})+\mu_{\mathbf{-1}})(\frac{1}{2}(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})^T+\mu_{\mathbf{-1}}{}^T)\hat{\mathbf{w}}$$

$$= -\left[\frac{1}{4}(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})^T+(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})\mu_{\mathbf{-1}}{}^T\right]\hat{\mathbf{w}}$$

$$= -\left[\frac{S_B}{4}+(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})\mu_{\mathbf{-1}}{}^T\right]\hat{\mathbf{w}}$$

c)

$$\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x_i^T}\hat{\mathbf{w}})\mathbf{x_i} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x_i}\mathbf{x_i^T})\hat{\mathbf{w}}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x_i}-\mu_{\mathbf{y_i}}+\mu_{\mathbf{y_i}})(\mathbf{x_i}-\mu_{\mathbf{y_i}}+\mu_{\mathbf{y_i}})^T\hat{\mathbf{w}}$$

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

$$= \left[S_W+\frac{1}{N/2}\sum_{i=1}^{N}\mathbf{x_i}\mu_{\mathbf{y_i}}^{\mathbf{T}}-\frac{2}{N}\sum_{i=1}^{N}\mu_{\mathbf{y_i}}\mu_{\mathbf{y_i}}^{\mathbf{T}}+\frac{1}{N}\sum_{i=1}^{N}\mu_{\mathbf{y_i}}\mu_{\mathbf{y_i}}^{\mathbf{T}}\right]\hat{\mathbf{w}}$$

$$= \left[S_W+\underbrace{\frac{1}{N/2}\sum_{i:y_i=1}\mathbf{x_i}\mu_{\mathbf{y_i}}^{\mathbf{T}}}_{=\mu_{\mathbf{1}}\mu_{\mathbf{1}}^{\mathbf{T}}}+\underbrace{\frac{1}{N/2}\sum_{i:y_i=-1}\mathbf{x_i}\mu_{\mathbf{y_{-1}}}^{\mathbf{T}}}_{=\mu_{-1}\mu_{\mathbf{-1}}^{\mathbf{T}}}-\mu_{\mathbf{1}}\mu_{\mathbf{1}}^{\mathbf{T}}-\mu_{\mathbf{-1}}\mu_{\mathbf{-1}}^{\mathbf{T}}+\frac{1}{2}\mu_{\mathbf{1}}\mu_{\mathbf{1}}^{\mathbf{T}}+\frac{1}{2}\mu_{\mathbf{-1}}\mu_{\mathbf{-1}}^{\mathbf{T}}\right]\hat{\mathbf{w}}$$

$$= \left[S_W+\frac{1}{2}(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})^T+(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})\mu_{\mathbf{1}}{}^T\right]\hat{\mathbf{w}}$$

$$= \left[S_W+\frac{S_B}{2}+(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})\mu_{\mathbf{1}}{}^T\right]\hat{\mathbf{w}}$$

Now we insert these results into the equation from last page.

$$(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})/2 = \left[-\frac{S_B}{4}-(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})\mu_{\mathbf{1}}{}^T+S_W+\frac{S_B}{2}+(\mu_{\mathbf{1}}-\mu_{\mathbf{-1}})\mu_{\mathbf{1}}{}^T\right]\hat{\mathbf{w}} = \left[S_W+\frac{S_B}{4}\right]\hat{\mathbf{w}}$$

This is equivalent to

$$S_W \hat{\mathbf{w}} = \frac{\boldsymbol{\mu_1} - \boldsymbol{\mu_{-1}}}{2} + \frac{S_B}{4}\hat{\mathbf{w}} \tag{5}$$

Because $\mathbb{R}^d$ is a finite dimensional vector space, we can choose $v_2, ..., v_d \in \mathbb{R}^d$ such that $\{(\mu_1 - \mu_{-1}), v_2, ..., v_d\}$ is an orthonormal basis of $\mathbb{R}^d$. Thus, we can write: $\hat{\mathbf{w}} = \lambda_1(\mu_1 - \mu_{-1}) + \sum_{i=2}^d \lambda_i v_i$ for $\lambda_1, ..., \lambda_d \in \mathbb{R}$. This way we can show:

$$\begin{aligned}
\frac{S_B}{4}\hat{\mathbf{w}} &= \frac{1}{4}(\mu_1 - \mu_{-1})(\mu_1 - \mu_{-1})^T \left(\lambda_1(\mu_1 - \mu_{-1}) + \sum_{i=2}^d \lambda_i v_i \right) \\
&= \frac{1}{4}\lambda_1(\mu_1 - \mu_{-1})(\mu_1 - \mu_{-1})^T(\mu_1 - \mu_{-1}) \\
&= \frac{1}{4}\lambda_1(\mu_1 - \mu_{-1})||\mu_1 - \mu_{-1}||^2
\end{aligned}$$

The second equality holds because the scalar product of $\mu_1 - \mu_{-1}$ and v_i vanishes for all $i \in \{2, ..., d\}$ (ONB). Thus, we obtain with the equality from above and $\tau := \frac{1}{2} + \frac{1}{4}\lambda_1||\mu_1 - \mu_{-1}||^2$:

$$\exists \tau \in \mathbb{R} : S_W \hat{\mathbf{w}} = \tau(\mu_1 - \mu_{-1})$$

Under the assumption that S_W is invertible (which is true if (x_i) are not located on a common $(d-1)$-dimensional hyperplane) we get:

$$\exists \tau \in \mathbb{R} : \hat{\mathbf{w}} = \tau S_W^{-1}(\mu_1 - \mu_{-1})$$