5 Classification

Helge Voss

Exercise 5.1: Derivation of a linear classifier using quadratic loss function and its equivalence to the Fisher discriminant

a)

$$y(\mathbf{x}) = w_0 + \sum_{k=1}^{D} x_k \cdot w_k = \mathbf{x'}^{\mathsf{T}}\mathbf{w'}\,.$$

b) $y_{signal} = 1$ and $y_{background} = -1$ as $y(\mathbf{x}) > 0$ for events on one side of the hyperplane given by $y(\mathbf{x}) = 0$ and $y(\mathbf{x}) < 0$ on the other side.

c)

$$\begin{aligned} L(\mathbf{w'}) &= (y - \mathbf{x'}^{\mathsf{T}}\mathbf{w'})^2\,; \\ E[L(\mathbf{w'})] &= N_s \int (p(\mathbf{x}|S)(y^S - \mathbf{x'}^{\mathsf{T}}\mathbf{w'})^2 dx + \\ & \quad N_B \int (p(\mathbf{x}|B)(y^B - \mathbf{x'}^{\mathsf{T}}\mathbf{w'})^2 dx\,. \end{aligned}$$

d) The estimator of the expectation value is given by the average over the sum of the events:

$$\begin{aligned} E[L(\mathbf{w'})] &= \frac{1}{N_S + N_B} \sum_{k}^{events} (y^{(k)} - \mathbf{x'}^{\mathbf{k}\mathsf{T}}\mathbf{w'})^2 \\ &= \frac{1}{N_S + N_B} \left(\sum_{k}^{N_S} (y^S - \mathbf{x'}^{\mathbf{k}\mathsf{T}}\mathbf{w'})^2 + \sum_{k}^{N_B} (y^B - \mathbf{x'}^{\mathbf{k}\mathsf{T}}\mathbf{w'})^2 \right)\,. \end{aligned}$$

One can nicely write the expectation value of the loss function in vector notation if one defines a "column Vector" $\mathbf{Y}$ which has in the first N_s lines a 1 and in the following N_B lines a (-1). Likewise a matrix $\mathbf{X}$, that has as line index, just like the vector $\mathbf{Y}$, the events, and as column index the coordinates:

$$\mathbf{X} = \begin{pmatrix} x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \cdots & x_D^{(1)} \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \cdots & x_D^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^{(N)} & x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{pmatrix} \quad ; \quad \mathbf{Y} = \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} . \tag{5.1}$$

Then, the (estimate) of the expectation of the loss function can be written as

$$E[L(\mathbf{w'})] = (\mathbf{Y} - \mathbf{w'}^\intercal \mathbf{X})(\mathbf{Y} - \mathbf{X}\mathbf{w'}) . \tag{5.2}$$

e)

$$\begin{aligned} \nabla_w E[L] &\stackrel{!}{=} \mathbf{0} , \\ \nabla_w \frac{1}{N}(Y - \mathbf{w}^\intercal \mathbf{X})^\intercal (Y - \mathbf{X}\mathbf{w}) &= \mathbf{0} , \\ \frac{1}{N}(\mathbf{X}\mathbf{Y} - \mathbf{X}^\intercal \mathbf{X}\mathbf{w}) &= \mathbf{0} , \\ \frac{1}{N}\mathbf{X}^\intercal \mathbf{X}\mathbf{w} &= \frac{1}{N}\mathbf{X}^\intercal \mathbf{Y} ., \end{aligned}$$

Note: From this equation you get immediately the weights or coefficents w of the linear classifier $\mathbf{w} = \frac{1}{N}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{Y}$. However the question here was to show that this is equivalent to the equations derived for the Fisher coefficients, and it would be difficult to infer something about the inverse matrix $(\mathbf{X}^\intercal\mathbf{X})^{-1}$. Therefore we continue to look individually at the two sides of the equation:

1) Rather than making use of extensive matrix calculations, which would probably allow the following derivation to be written much more concise, we will list here the full details by writing out the matrices.

$$\sum_k^{events} (\mathbf{w}\mathbf{x}^{(k)})\mathbf{x}^{(k)} =$$

$$\sum_k^{events} \begin{pmatrix} w_0 x_0^{(k)} x_0^{(k)} + w_1 x_1^{(k)} x_0^{(k)} + w_2 x_2^{(k)} x_0^{(k)} + \cdots + w_D x_D^{(k)} x_0^{(k)} \\ w_0 x_0^{(k)} x_1^{(k)} + w_1 x_1^{(k)} x_1^{(k)} + w_2 x_2^{(k)} x_1^{(k)} + \cdots + w_D x_D^{(k)} x_1^{(k)} \\ \vdots \\ w_0 x_0^{(k)} x_D^{(k)} + w_1 x_1^{(k)} x_D^{(k)} + w_2 x_2^{(k)} x_D^{(k)} + \cdots + w_D x_D^{(k)} x_D^{(k)} \end{pmatrix} =$$

$$\sum_k^{events} \begin{pmatrix} x_0^{(k)} x_0^{(k)} & x_1^{(k)} x_0^{(k)} & x_2^{(k)} x_0^{(k)} & \cdots & x_D^{(k)} x_0^{(k)} \\ x_0^{(k)} x_1^{(k)} & x_1^{(k)} x_1^{(k)} & x_2^{(k)} x_1^{(k)} & \cdots & x_D^{(k)} x_1^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^{(k)} x_D^{(k)} & x_1^{(k)} x_D^{(k)} & x_2^{(k)} x_D^{(k)} & \cdots & x_D^{(k)} x_D^{(k)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix} =$$

$$N_{events}\begin{pmatrix} \overline{x_0x_0} & \overline{x_1x_0} & \overline{x_2x_0} & \cdots & \overline{x_Dx_0} \\ \overline{x_0x_1} & \overline{x_1x_1} & \overline{x_2x_1} & \cdots & \overline{x_Dx_1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{x_0x_D} & \overline{x_1x_D} & \overline{x_2x_D} & \cdots & \overline{x_Dx_D} \end{pmatrix}\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix} = N_{events}(V+\boldsymbol{\mu}\boldsymbol{\mu}^{\intercal})\mathbf{w} \quad .$$

2) Assuming equal numbers of events $N_S = N_B = \frac{1}{2}N$, each element of $\mathbf{V}_{s+b}$, $(\mathbf{V}_{s+b})_{i,j}$ can be decomposed as follows:

$$\begin{aligned}
(\mathbf{V}_{s+b})_{i,j} &= \frac{1}{N}\sum_{k}^{N_S+N_B}(x_i^{(k)}-\mu_i)(x_j^{(k)}-\mu_j) \\
&= \frac{1}{N}\sum_{k}^{N_S+N_B}(x_i^{(k)}x_j^{(k)}-x_i^{(k)}\mu_j-x_j^{(k)}\mu_i+\mu_i\mu_j) \\
&= \frac{1}{N}\sum_{k}^{N_S+N_B}\Big(x_i^{(k)}x_j^{(k)}-x_i^{(k)}\frac{1}{2}(\mu_j^S+\mu_j^B)-x_j^{(k)}\frac{1}{2}(\mu_i^S+\mu_i^B) \\
&\qquad +\frac{1}{4}(\mu_i^S+\mu_i^B)(\mu_j^S+\mu_j^B)\Big) \\
&= \frac{1}{N}\sum_{k}^{N_S}\Big(x_i^{(k)}x_j^{(k)}-x_i^{(k)}\frac{1}{2}(\mu_j^S+\mu_j^B)-x_j^{(k)}\frac{1}{2}(\mu_i^S+\mu_i^B)\Big) \\
&\qquad +\frac{1}{N}\sum_{k}^{N_B}\Big(x_i^{(k)}x_j^{(k)}-x_i^{(k)}\frac{1}{2}(\mu_j^S+\mu_j^B)-x_j^{(k)}\frac{1}{2}(\mu_i^S+\mu_i^B)\Big) \\
&\qquad +\frac{1}{N}\sum_{k}^{N_S+N_B}\frac{1}{4}(\mu_i^S+\mu_i^B)(\mu_j^S+\mu_j^B) \\
&= \frac{1}{N}\sum_{k}^{N_S}\Big(x_i^{(k)}x_j^{(k)}-\frac{1}{2}x_i^{(k)}\mu_j^S-\frac{1}{2}x_j^{(k)}\mu_i^S\Big) \\
&\qquad +\frac{1}{N}\sum_{k}^{N_S}\Big(-\frac{1}{2}x_i^{(k)}\mu_j^B-\frac{1}{2}x_j^{(k)}\mu_i^B\Big) \\
&\qquad +\frac{1}{N}\sum_{k}^{N_B}\Big(x_i^{(k)}x_j^{(k)}-\frac{1}{2}x_i^{(k)}\mu_j^B-\frac{1}{2}x_j^{(k)}\mu_i^B\Big) \\
&\qquad +\frac{1}{N}\sum_{k}^{N_B}\Big(-\frac{1}{2}x_i^{(k)}\mu_j^S-\frac{1}{2}x_j^{(k)}\mu_i^S\Big) \\
&\qquad +\frac{1}{4}(\mu_i^S+\mu_i^B)(\mu_j^S+\mu_j^B)\,;
\end{aligned}$$

$$
\begin{aligned}
(\mathbf{V}_{s+b})_{i,j} &= \frac{1}{N}\sum_{k}^{N_S}\left(x_i^{(k)}x_j^{(k)} - x_i^{(k)}\mu_j^S - x_j^{(k)}\mu_i^S + \mu_i^S\mu_j^S + \frac{1}{2}x_i^{(k)}\mu_j^S + \frac{1}{2}x_j^{(k)}\mu_i^S - \mu_i^S\mu_j^S\right) \\
&\quad + \frac{1}{2}\left(-\frac{1}{2}\mu_i^S\mu_j^B - \frac{1}{2}\mu_j^S\mu_i^B\right) \\
&\quad + \frac{1}{N}\sum_{k}^{N_B}\left(x_i^{(k)}x_j^{(k)} - x_i^{(k)}\mu_j^B - x_j^{(k)}\mu_i^B + \mu_i^B\mu_j^B + \frac{1}{2}x_i^{(k)}\mu_j^B + \frac{1}{2}x_j^{(k)}\mu_i^B - \mu_i^B\mu_j^B\right. \\
&\quad + \frac{1}{2}\left(-\frac{1}{2}\mu_i^B\mu_j^S - \frac{1}{2}\mu_j^B\mu_i^S\right) \\
&\quad + \frac{1}{4}(\mu_i^S + \mu_i^B)(\mu_j^S + \mu_j^B) \\
&= \frac{1}{2}\left(V_{i,j}^S + \frac{1}{2}\mu_i^S\mu_j^S + \frac{1}{2}\mu_j^S\mu_i^S - \mu_i^S\mu_j^S\right) \\
&\quad + \frac{1}{4}\left(-\mu_i^S\mu_j^B - \mu_j^S\mu_i^B\right) \\
&\quad + \frac{1}{2}\left(V_{i,j}^B + \frac{1}{2}\mu_i^B\mu_j^B + \frac{1}{2}\mu_j^B\mu_i^B - \mu_i^B\mu_j^B\right) \\
&\quad + \frac{1}{4}\left(-\mu_i^B\mu_j^S - \mu_j^B\mu_i^S\right) \\
&\quad + \frac{1}{4}(\mu_i^S + \mu_i^B)(\mu_j^S + \mu_j^B) \\
&= \frac{1}{2}\left(V_{i,j}^S + V_{i,j}^B\right) \\
&\quad + \frac{1}{4}\left(\mu_i^S\mu_j^S + \mu_j^S\mu_i^S - 2\mu_i^S\mu_j^S - \mu_i^S\mu_j^B - \mu_j^S\mu_i^B\right. \\
&\quad \left. + \mu_i^B\mu_j^B + \mu_j^B\mu_i^B - 2\mu_i^B\mu_j^B - \mu_i^B\mu_j^S - \mu_j^B\mu_i^S + (\mu_i^S + \mu_i^B)(\mu_j^S + \mu_j^B)\right) \\
&= \frac{1}{2}\left(V_{i,j}^S + V_{i,j}^B\right)\frac{1}{4}\left(-2\mu_i^S\mu_j^B - \mu_j^S\mu_i^B + (\mu_i^S + \mu_i^B)(\mu_j^S + \mu_j^B)\right) \\
&= \frac{1}{2}\left(V_{i,j}^S + V_{i,j}^B\right) + \frac{1}{4}(\mu_i^S - \mu_i^B)(\mu_j^S - \mu_j^B)\,.
\end{aligned}
$$

3) For any vector $\boldsymbol{\mu}$ and a matrix $\boldsymbol{\mu}\boldsymbol{\mu}^\intercal$) we have:

$$
\begin{aligned}
(\boldsymbol{\mu}\boldsymbol{\mu}^\intercal)\mathbf{x}_i &= \sum_j(\boldsymbol{\mu}\boldsymbol{\mu}^\intercal)_{ij}x_j \\
&= \sum_j \mu_i\mu_j x_j \\
&= (\sum_j \mu_j x_j)\mu_i
\end{aligned}
$$

and then equally for $(\boldsymbol{\mu}_\mathbf{S} - \boldsymbol{\mu}_\mathbf{B})(\boldsymbol{\mu}_\mathbf{S} - \boldsymbol{\mu}_\mathbf{B})^\intercal$, by simply replacing $\boldsymbol{\mu} = \boldsymbol{\mu}_\mathbf{S} - \boldsymbol{\mu}_\mathbf{B}$
.

Now we can proceed with rewriting the two sides of the equation derived for: $\nabla_\mathbf{w}E \overset{!}{=} 0$ for $\mathbf{w}$:

$$
\frac{1}{N}\mathbf{X}^\intercal\mathbf{X}\mathbf{w} \quad =
$$

$$\frac{1}{N}\begin{pmatrix} x_0^{(1)} & x_0^{(2)} & x_0^{(3)} & \cdots & x_0^{(N)} \\ x_1^{(1)} & x_1^{(2)} & x_1^{(3)} & \cdots & x_1^{(N)} \\ x_2^{(1)} & x_2^{(2)} & x_2^{(3)} & \cdots & x_2^{(N)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_D^{(1)} & x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{pmatrix} \begin{pmatrix} x_0^{(1)} & x_1^{(1)} & x_2^{(1)} & \cdots & x_D^{(1)} \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \cdots & x_D^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^{(N)} & x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} =$$

$$\frac{1}{N}\begin{pmatrix} \sum_k^N x_0^{(k)} x_0^{(k)} & \sum_k^N x_0^{(k)} x_1^{(k)} & \sum_k^N x_0^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_0^{(k)} x_D^{(k)} \\ \sum_k^N x_1^{(k)} x_0^{(k)} & \sum_k^N x_1^{(k)} x_1^{(k)} & \sum_k^N x_1^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_1^{(k)} x_D^{(k)} \\ \sum_k^N x_2^{(k)} x_0^{(k)} & \sum_k^N x_2^{(k)} x_1^{(k)} & \sum_k^N x_2^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_2^{(k)} x_D^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_k^N x_D^{(k)} x_0^{(k)} & \sum_k^N x_D^{(k)} x_1^{(k)} & \sum_k^N x_D^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_D^{(k)} x_D^{(k)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} =$$

$$\frac{1}{N}\begin{pmatrix} \sum_k^N 1 & \sum_k^N x_1^{(k)} & \sum_k^N x_2^{(k)} & \cdots & \sum_k^N x_D^{(k)} \\ \sum_k^N x_1^{(k)} & \sum_k^N x_1^{(k)} x_1^{(k)} & \sum_k^N x_1^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_1^{(k)} x_D^{(k)} \\ \sum_k^N x_2^{(k)} & \sum_k^N x_2^{(k)} x_1^{(k)} & \sum_k^N x_2^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_2^{(k)} x_D^{(k)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_k^N x_D^{(k)} & \sum_k^N x_D^{(k)} x_1^{(k)} & \sum_k^N x_D^{(k)} x_2^{(k)} & \cdots & \sum_k^N x_D^{(k)} x_D^{(k)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} =$$

$$\frac{1}{N}\begin{pmatrix} N & N\mu_1 & N\mu_2 & \cdots & N\mu_D \\ N\mu_1 & N(V_{11}+\mu_1\mu_1) & N(V_{12}+\mu_1\mu_2) & \cdots & N(V_{1D}+\mu_1\mu_D) \\ N\mu_2 & N(V_{21}+\mu_2\mu_1) & N(V_{12}+\mu_2\mu_2) & \cdots & N(V_{2D}+\mu_2\mu_D) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ N\mu_D & N(V_{D1}+\mu_D\mu_1) & N(V_{D2}+\mu_D\mu_2) & \cdots & N(V_{DD}+\mu_D\mu_D) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} =$$

$$\begin{pmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_D \\ \mu_1 & (V_{11}+\mu_1\mu_1) & (V_{12}+\mu_1\mu_2) & \cdots & (V_{1D}+\mu_1\mu_D) \\ \mu_2 & (V_{21}+\mu_2\mu_1) & (V_{12}+\mu_2\mu_2) & \cdots & (V_{2D}+\mu_2\mu_D) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mu_D & (V_{D1}+\mu_D\mu_1) & (V_{D2}+\mu_D\mu_2) & \cdots & (V_{DD}+\mu_D\mu_D) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} =$$

$$\begin{pmatrix} w_0 + w_1\mu_1 + w_2\mu_2 + \cdots + w_D\mu_D \\ w_0\mu_1 + w_1(V_{11}+\mu_1\mu_1) + w_2(V_{12}+\mu_1\mu_2) + \cdots + w_D(V_{1D}+\mu_1\mu_D) \\ w_0\mu_2 + w_1(V_{21}+\mu_2\mu_1) + w_2(V_{12}+\mu_2\mu_2) + \cdots + w_D(V_{2D}+\mu_2\mu_D) \\ \vdots \\ w_0\mu_D + w_1(V_{D1}+\mu_D\mu_1) + w_2(V_{D2}+\mu_D\mu_2) + \cdots + w_D(V_{DD}+\mu_D\mu_D) \end{pmatrix}.$$

Let $\boldsymbol{\mu}^\mathsf{T} = (\mu_1, ..., \mu_D)$ and $\boldsymbol{\mu}'^\mathsf{T} = (1, \mu_1, ..., \mu_D)$, and this matrix can be rewritten such that we get for the right side of the equation:

$$\frac{1}{N}\mathbf{X}^\intercal\mathbf{X}\mathbf{w} \quad = \quad w_0\boldsymbol{\mu'} + \begin{pmatrix} \mu^\intercal \\ \mathbf{V}+\mu\mu^\intercal \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix}$$

The other side of the equation:

$$\frac{1}{N}\mathbf{X}^\intercal\mathbf{Y} \quad =$$

$$\frac{1}{N}\begin{pmatrix} x_0^{(1)} & x_0^{(2)} & x_0^{(3)} & \cdots & x_0^{(N)} \\ x_1^{(1)} & x_1^{(2)} & x_1^{(3)} & \cdots & x_1^{(N)} \\ x_2^{(1)} & x_2^{(2)} & x_2^{(3)} & \cdots & x_2^{(N)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_D^{(1)} & x_1^{(N)} & x_2^{(N)} & \cdots & x_D^{(N)} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} \quad = \quad \frac{N_s}{N}\begin{pmatrix} N_S \\ \boldsymbol{\mu}^{\mathbf{S}} \end{pmatrix} - \frac{N_B}{N}\begin{pmatrix} N_B \\ \boldsymbol{\mu}^{\mathbf{B}} \end{pmatrix},$$

where $\boldsymbol{\mu}^{\mathbf{S}\intercal} = (\mu_1^S, ..., \mu_D^S)$ and likewise for $\boldsymbol{\mu}^{\mathbf{B}}$.

Together we have:

$$w_0\boldsymbol{\mu'} + \begin{pmatrix} \mu^\intercal \\ \mathbf{V}+\mu\mu^\intercal \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} \quad = \quad \frac{N_s}{N}\begin{pmatrix} N_S \\ \boldsymbol{\mu}^{\mathbf{S}} \end{pmatrix} - \frac{N_B}{N}\begin{pmatrix} N_B \\ \boldsymbol{\mu}^{\mathbf{B}} \end{pmatrix}.$$

Use the very first row of this set of equations to “fix” w_0:

$$\begin{aligned} w_0 + \boldsymbol{\mu}^\intercal\mathbf{w} &= \frac{N_S^2 - N_B^2}{N} \\ w_0 &= \frac{N_S^2 - N_B^2}{N} - \boldsymbol{\mu}^\intercal\mathbf{w}\,. \end{aligned}$$

Plugging this into the “lower D” equations yields:

$$w_0\boldsymbol{\mu} + \begin{pmatrix} \mathbf{V}+\mu\mu^\intercal \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_D \end{pmatrix} \quad = \quad \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}},$$

$$(\frac{N_S^2 - N_B^2}{N} - \boldsymbol{\mu}^\intercal\mathbf{w})\boldsymbol{\mu} + (\mathbf{V}+\mu\mu^\intercal)\mathbf{w} \quad = \quad \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}}\,.$$

Note that $\boldsymbol{\mu} = \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} + \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}}$
and $(\mu\mu^{\intercal})\mathbf{w} = \boldsymbol{\mu}^{\intercal}\mathbf{w}\boldsymbol{\mu}$ which then gives:

$$
\begin{aligned}
\frac{N_S^2 - N_B^2}{N}\boldsymbol{\mu} + \mathbf{V}\mathbf{w} &= \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}} \\
&= \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}} - \frac{N_S^2 - N_B^2}{N}(\frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} + \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}})\,, \\
&= \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}} - \frac{(N_S - N_B)(N_S + N_B)}{N}(\frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} + \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}},\,, \\
&= \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}} - \frac{(N_S - N_B)N}{N}(\frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} + \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}},\,, \\
&= \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}} - (N_S - N_B)(\frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} + \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}})\,.
\end{aligned}
$$

Now, assuming that all background events are weighted by N_s/N_B, then:

$$
\begin{aligned}
\mathbf{V}\mathbf{w} &= \frac{N_s}{N}\boldsymbol{\mu}^{\mathbf{S}} - \frac{N_B}{N}\boldsymbol{\mu}^{\mathbf{B}}\,, \\
\frac{1}{2}(\mathbf{V}_S + \mathbf{V}_B) + \frac{1}{4}(\mu_S - \mu_B)(\mu_S - \mu_B)^{\intercal})\mathbf{w} &= \frac{N_s}{N}(\boldsymbol{\mu}^{\mathbf{S}} - \boldsymbol{\mu}^{\mathbf{B}})\,, \\
\frac{1}{2}(\mathbf{V}_S + \mathbf{V}_B)\mathbf{w} + \frac{1}{4}(\mu_S - \mu_B)(\mu_S - \mu_B)^{\intercal}\mathbf{w} &= \frac{N_s}{N}(\boldsymbol{\mu}^{\mathbf{S}} - \boldsymbol{\mu}^{\mathbf{B}})\,, \\
\frac{1}{2}(\mathbf{V}_S + \mathbf{V}_B)\mathbf{w} + const(|\mathbf{w}|)(\mu_S - \mu_B) &= \frac{N_s}{N}(\boldsymbol{\mu}^{\mathbf{S}} - \boldsymbol{\mu}^{\mathbf{B}})\,, \\
\frac{1}{2}(\mathbf{V}_S + \mathbf{V}_B)\mathbf{w} &= const'(|\mathbf{w}|)(\boldsymbol{\mu}^{\mathbf{S}} - \boldsymbol{\mu}^{\mathbf{B}})\,, \\
\frac{1}{2}\mathbf{W}\mathbf{w} &= const'(|\mathbf{w}|)(\boldsymbol{\mu}^{\mathbf{S}} - \boldsymbol{\mu}^{\mathbf{B}})\,, \\
\mathbf{w} &\propto \mathbf{W}^{-1}(\boldsymbol{\mu}^{\mathbf{S}} - \boldsymbol{\mu}^{\mathbf{B}})\,.
\end{aligned}
$$

Exercise 5.2: LDA (Linear Discriminant Ananlysis) and Gaussian probability densities

a)

$$
y(\mathbf{x}) = \frac{\exp(-(\mathbf{x} - \boldsymbol{\mu}_S)^{\intercal} V_S^{-1}(\mathbf{x} - \boldsymbol{\mu}_S))}{\exp(-(\mathbf{x} - \boldsymbol{\mu}_B)^{\intercal} V_B^{-1}(\mathbf{x} - \boldsymbol{\mu}_B))}\,. \tag{5.3}
$$

b)

$$
\begin{aligned}
y(\mathbf{x}) &= log(\frac{\exp(-(\mathbf{x}-\boldsymbol{\mu}_S)^{\intercal}V^{-1}(\mathbf{x}-\boldsymbol{\mu}_S))}{\exp(-(\mathbf{x}-\boldsymbol{\mu}_B)^{\intercal}V^{-1}(\mathbf{x}-\boldsymbol{\mu}_B))}) \\
&= -(\mathbf{x}-\boldsymbol{\mu}_S)^{\intercal}V^{-1}(\mathbf{x}-\boldsymbol{\mu}_S) - (-(\mathbf{x}-\boldsymbol{\mu}_S)^{\intercal}V^{-1}(\mathbf{x}-\boldsymbol{\mu}_S)) \\
&= -(\mathbf{x}-\boldsymbol{\mu}_S)^{\intercal}V^{-1}(\mathbf{x}-\boldsymbol{\mu}_S) + (\mathbf{x}-\boldsymbol{\mu}_S)^{\intercal}V^{-1}(\mathbf{x}-\boldsymbol{\mu}_S) \\
&= 2\mathbf{x}^{\intercal}V^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_B) - \boldsymbol{\mu}_S^{\intercal}V^{-1}\boldsymbol{\mu}_S + \boldsymbol{\mu}_B^{\intercal}V^{-1}\boldsymbol{\mu}_B \\
&= 2\mathbf{x}^{\intercal}V^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_B) - (\boldsymbol{\mu}_S-\boldsymbol{\mu}_B)^{\intercal}V^{-1}(\boldsymbol{\mu}_S+\boldsymbol{\mu}_B)\,, \\
&\quad \text{and as you do not care about the constant second term you may choose:} \\
y(\mathbf{x}) &= 2\mathbf{x}^{\intercal}V^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_B)
\end{aligned}
$$

c) The constant second term can be omitted, hence $y(\mathbf{x})$ may be reduced to:

$$y(\mathbf{x}) = 2\mathbf{x}^{\intercal}V^{-1}(\boldsymbol{\mu}_S-\boldsymbol{\mu}_B)\,, \tag{5.4}$$

and the Fisher Discriminant was given by

$$F_0 + \mathbf{x}^{\intercal}\mathbf{F} = \mathbf{x}^{\intercal}\mathbf{W}^{-1}\,(\boldsymbol{\mu}_{\mathbf{S}}-\boldsymbol{\mu}_{\mathbf{B}}) \tag{5.5}$$

with $\mathbf{W} = \mathbf{V}_S + \mathbf{V}_B$. Under the assumption that the covariances of signal and background distributions are the same, this reduces to $\mathbf{W} = 2\mathbf{V}$, which is exactly what we had derived in exercise 5.1 for the LDA.