EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

Statistical Machine Learning

Exercise sheet 5

Exercise 5.1 (Leave-one-out cross-validation for *linear smoothers*) In this exercise we consider *linear smoothers*, i.e., learning scheme producing decision functions $\widehat{f}$ for which the fitted values $\widehat{y}_i := \widehat{f}(\boldsymbol{x}_i)$ on the training set satisfy $\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}$, where $\mathbf{S}$ is an $n \times n$ matrix whose values only depend on the inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and $\widehat{\boldsymbol{y}} = (y_i)_{i=1\ldots n}$.

We consider the leave-one-out CV error

$$\mathrm{CV}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\right\}^2,$$

where $\widehat{f}^{-i}$ denote the model fitted to the original training sample with the ith observation $(y_i, \boldsymbol{x}_i)$ removed.

The goal of this exercise is to derive a fast way of computing the leave-one-out (or n-fold) cross-validation (CV) error for *linear smoothers* which produce leave-one-out decision functions with a particular form (given by Equation (1) below).

(a) Show that linear regression is a linear smoother in the sense that the obtained prediction function $\widehat{f}$ satisfies the property above. In particular specify $\mathbf{S}$.

Solution: We have seen in the course that the identity holds for $\mathbf{S} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ the so-called hat matrix.

(b) Assume that the leave-ith-out fit at $\boldsymbol{x}_i$ is given by

$$\widehat{f}^{-i}(\boldsymbol{x}_i) = \sum_{j\neq i}\frac{\mathbf{S}_{ij}}{1-\mathbf{S}_{ii}}y_j. \tag{1}$$

With this regularity assumption, show that

$$y_i - \widehat{f}^{-i}(\boldsymbol{x}_i) = \frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1-\mathbf{S}_{ii}}. \tag{2}$$

Solution: Equation (2) holds because

$$\begin{aligned} y_i - \widehat{f}^{-i}(\boldsymbol{x}_i) &= y_i - \sum_{j\neq i}\frac{\mathbf{S}_{ij}}{1-\mathbf{S}_{ii}}y_j \\ &= \frac{1}{1-\mathbf{S}_{ii}}\left\{y_i(1-\mathbf{S}_{ii}) - \sum_{j\neq i}\mathbf{S}_{ij}y_j\right\} \\ &= \frac{1}{1-\mathbf{S}_{ii}}\left\{y_i - \sum_{j=1}^{n}\mathbf{S}_{ij}y_j\right\} \\ &= \frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1-\mathbf{S}_{ii}}. \end{aligned}$$

(c) Explain why (2) may be used to compute the CV error more efficiently.

Solution: We have that

$$\mathrm{CV}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\right\}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1-\mathbf{S}_{ii}}\right\}^2,$$

which allows fast computation given $\widehat{f}$ and the diagonal elements of $\mathbf{S}$, removing the need to calculate each $\widehat{f}^{-i}(\boldsymbol{x}_i)$ separately.

(d) Our goal in the rest of this exercise is to identify some conditions that imply that $\widehat{f}^{-i}$ is of the form (1). We consider the squared loss $\ell(a,y) = (a-y)^2$ and we focus on the decision function minimizing the empirical risk in a hypothesis class S, that is

$$\widehat{f} = \arg\min_{f\in S} \tfrac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{x}_i) - y_i)^2,$$

assuming that the latter is unique. Assume that $\widehat{f}^{-i}$ has been computed and that we define a new dataset $\tilde{D}_n = \{(\boldsymbol{x}_j, \tilde{y}_j)\}_{j=1\ldots n}$ with $\tilde{y}_j = y_j$ for all $j \neq i$ and $\tilde{y}_i = \widehat{f}^{-i}(\boldsymbol{x}_i)$. Show that the minimizer of the empirical risk on this new dataset is $\widehat{f}^{-i}$.

Solution: Note that

$$\sum_{j=1}^{n}(f(\boldsymbol{x}_j) - \tilde{y}_j)^2 = \sum_{j\neq i}(f(\boldsymbol{x}_j) - y_j)^2 + (f(\boldsymbol{x}_i) - \tilde{y}_i)^2$$

Given that the first terms is minimized over S at $f = \widehat{f}^{-i}$ by definition and that the second term is equal to 0 at $f = \widehat{f}^{-i}$ by construction given that $\tilde{y}_i = \widehat{f}^{-i}(\boldsymbol{x}_i)$, we necessarily have that

$$\widehat{f}^{-i} = \arg\min_{f\in S} \tfrac{1}{n}\sum_{j=1}^{n}(f(\boldsymbol{x}_j) - \tilde{y}_j)^2.$$

(e) Given that the linear regression estimator is a linear smoother, there is a matrix $\mathbf{S}$ such that $\widehat{\boldsymbol{y}} = \mathbf{S}\boldsymbol{y}$. Use the previous question to show that $(\mathbf{S}\tilde{\boldsymbol{y}})_i = \widehat{f}^{-i}(\boldsymbol{x}_i)$ and use the form of $\tilde{\boldsymbol{y}}$ to prove that $\widehat{f}^{-i}$ takes the form of (1).

Solution: Given that

$$\tilde{\boldsymbol{y}} = \boldsymbol{y} - \left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\right\}\boldsymbol{e}_i,$$

we have

$$\left(\mathbf{S}\left[\boldsymbol{y} - \left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\right\}\boldsymbol{e}_i\right]\right)_i = \widehat{f}^{-i}(\boldsymbol{x}_i),$$

where

$$\begin{aligned}\left(\mathbf{S}\left[\boldsymbol{y} - \left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\boldsymbol{e}_i\right\}\right]\right)_i &= \sum_{j=1}^{n}\mathbf{S}_{ij}y_j - \left\{y_i - \widehat{f}^{-i}(\boldsymbol{x}_i)\right\}\mathbf{S}_{ii} \\ &= \mathbf{S}_{ii}\widehat{f}^{-i}(\boldsymbol{x}_i) + \sum_{j\neq i}\mathbf{S}_{ij}y_j,\end{aligned}$$

so that $\widehat{f}^{-i}(\boldsymbol{x}_i) = \mathbf{S}_{ii}\widehat{f}^{-i}(\boldsymbol{x}_i) + \sum_{j\neq i}\mathbf{S}_{ij}y_j$ and the result is obtained by isolating $\widehat{f}^{-i}(\boldsymbol{x}_i)$ on the LHS.

(f) Deduce from the previous questions the form of the LOO CV error for linear regression.

Solution: We have

$$\begin{aligned}
\mathrm{CV}(\widehat{f}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \widehat{f}(\mathbf{x}_i)}{1-\mathbf{S}_{ii}}\right\}^2, \\
&= \frac{1}{n}(\boldsymbol{y}-\mathbf{S}\boldsymbol{y})^\top \mathrm{diag}\left[(1-\mathbf{S}_{ii})^{-2}\right](\boldsymbol{y}-\mathbf{S}\boldsymbol{y}) \\
&= \frac{1}{n}\boldsymbol{y}^\top(\mathbf{I}-\mathbf{S})^\top \mathrm{diag}\left[(1-\mathbf{S}_{ii})^{-2}\right](\mathbf{I}-\mathbf{S})\boldsymbol{y}
\end{aligned}$$

where $\mathbf{S} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ and $\mathrm{diag}\left[(1-\mathbf{S}_{ii})^{-2}\right]$ denotes the $n\times n$ diagonal matrix with the iith entry given by $(1-\mathbf{S}_{ii})^{-2}$.

(g) Can a similar approach be used to obtain an expression of the LOO CV error for ridge regression?

Solution: No, because the form of the risk for ridge regression is different than the one in (d).

(h) Show that all local averaging methods are linear smoothers.

Solution: Define $\boldsymbol{S}_{ij} := \omega_j(x_i)$, then $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$, and this satisfies the definition of linear smoothers. Therefore, all local averaging methods are linear smoothers.

(i) Show that (1) holds for the Nadaraya-Watson estimator, and deduce the LOO CV error for it.

Solution: We have,

$$\begin{aligned}
\widehat{f}^{-i}(\boldsymbol{x}_i) &= \sum_{j\neq i}\omega_j^{-i}(\boldsymbol{x}_i)y_j \\
&= \sum_{j\neq i}\tilde{s}^{-i}(\boldsymbol{x}_i,\boldsymbol{x}_j)\boldsymbol{y}_j \\
&= \sum_{j\neq i}\frac{s(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{k\neq i}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}\boldsymbol{y}_j.
\end{aligned}$$

Now, we just have to show that $\dfrac{\tilde{s}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{1-\tilde{s}(\boldsymbol{x}_i,\boldsymbol{x}_i)} = \dfrac{s(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{k\neq i}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}$. We have,

$$\begin{aligned}
\frac{\tilde{s}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{1-\tilde{s}(\boldsymbol{x}_i,\boldsymbol{x}_i)} &= \frac{s(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{k=1}^{n}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}\left[1-\frac{s(\boldsymbol{x}_i,\boldsymbol{x}_i)}{\sum_{k=1}^{n}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}\right]^{-1} \\
&= \frac{s(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{k=1}^{n}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}\left[\frac{\sum_{k=1}^{n}s(\boldsymbol{x}_i,\boldsymbol{x}_k)-s(\boldsymbol{x}_i,\boldsymbol{x}_i)}{\sum_{k=1}^{n}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}\right]^{-1} \\
&= \frac{s(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{k\neq i}s(\boldsymbol{x}_i,\boldsymbol{x}_k)}.
\end{aligned}$$

Therefore, (1) holds for the Nadaraya-Watson estimator. Similarly, we also have

$$\mathrm{CV}(\widehat{f}^{\mathrm{NW}}) = \frac{1}{n}\boldsymbol{y}^\top(I-\boldsymbol{\Omega})^\top\mathrm{diag}(I-\boldsymbol{\Omega})(I-\boldsymbol{\Omega})\boldsymbol{y}$$

where $\boldsymbol{\Omega}_{ij} = \omega_i(\boldsymbol{x}_j) = \tilde{s}(\boldsymbol{x}_i,\boldsymbol{x}_j)$.

(j) Does (1) hold for histogram estimators? For the k nearest-neighbors?

Solution: Yes, for histogram estimators because the similarity measure

$$s(x,y) = \sum_{k=1}^{K} \mathbf{1}_{\{x \in A_k\}} \mathbf{1}_{\{y \in A_k\}}$$

is exclusively a function of x and y because $\{A_k\}$ are fixed. Thus, the similarity measure does not depend on the dataset which is why the reasoning of the previous subquestions applies. But **not** for k-nearest neighbours, where

$$s(x,y) = \mathbf{1}_{\{x \in V_k(y)\}}$$

which means x has to be among of the k inputs x_j which are closest to y, implying that it depends on the data set and therefore 1 does not hold.

Exercise 5.2 (Fisher Discriminant) Logistic regression was introduced in class as an optimization problem which is obtained by applying the maximum likelihood principle to a model of $p(y = 1|x)$ in which the log-odd ratio is an affine function of the input feature vector. This type of model is often called *conditional model* or *discriminative* model because it only models the conditional distribution of y given x and not the marginal distribution of x. By contrast, we consider here what is called a *generative* model, a model in which both a model of $p(y)$ and $p(x|y)$ are estimated and from which $p(y|x)$ can be deduced (and also $p(x)$ of course). The particular models that we will consider are due to Fisher and are called *linear discriminant analysis* (LDA) and *quadratic discriminant analysis* (QDA). We will focus on the binary classification setting, although the method generalizes immediately to the multiclass classification setting.

(a) We first consider the QDA model. Given the class variable $y \in \{0, 1\}$, the data are assumed to be Gaussian with different means and different covariance matrices for the two different classes but with the same covariance matrix.

$$y \sim \text{Bernoulli}(\pi), \quad x|\{y = k\} \sim \text{Normal}(\mu_k, \Sigma_k),$$

with $x, \mu_k \in \mathbb{R}^p$ and $\Sigma_k \in \mathbb{R}^{p \times p}$. Derive the form of the maximum likelihood estimators for the parameters in this model, i.e. for $\pi, \mu_1, \mu_0, \Sigma_1$ and Σ_0.

Solution: Of course, one can reason through conditional distributions and use the well-known expressions for the MLE of a Gaussian distribution in $\mathbb{R}^n$ to solve the problem in a jiffy. But we shall take this opportunity to work out the solution by ourselves and in full. Note that **this is extra material** and you are only expected to remember the MLE of the Gaussian distribution for the purpose of the exams.

We begin by writing the likelihood functions as follows:

$$p(\{(\mathbf{x}_j, y_j)\}_{j=1}^n | \pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) = \prod_{j=1}^{n} \left[\pi \mathcal{N}(\mathbf{x}_j, \mu_1, \Sigma_1)\right]^{y_j} \left[(1-\pi)\mathcal{N}(\mathbf{x}_j, \mu_0, \Sigma_2)\right]^{1-y_j}$$

Now, omitting all the irrelevant constant terms the log-likelihood function is given by:

$$
\begin{aligned}
\ell(\{(\mathbf{x}_j, y_j)\}_{j=1}^n | \pi, \mu_0, \mu_1, \Sigma_0, \Sigma_1) = & \left[\sum_{j=1}^n y_j \log \pi + (1-y_j)\log(1-\pi)\right] \\
& - \tfrac{1}{2}\left[\sum_{j=1}^n y_j \log\det\Sigma_1 + (1-y_j)\log\det\Sigma_0\right] \\
& - \tfrac{1}{2}\sum_{j=1}^n y_j(\mathbf{x}_j-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}_j-\mu_1) + (1-y_j)(\mathbf{x}_j-\mu_0)^\top\Sigma_0^{-1}(\mathbf{x}_j-\mu_0)
\end{aligned}
$$

Let $p = \sum_{j=1}^n y_j$ and $q = n - \sum_{j=1}^n y_j$. The first term is maximum when π is given by

$$\widehat{\pi} = p/n = 1 - q/n$$

Differentiating with respect to μ_1 and μ_0 gives

$$
\begin{aligned}
0 &= \sum_{j=1}^n y_j\Sigma_1^{-1}(\mathbf{x}_j-\mu_1) = p\Sigma_1^{-1}\left[\tfrac{1}{p}\sum_{j=1}^n y_j\mathbf{x}_j - \mu_1\right] \\
0 &= \sum_{j=1}^n (1-y_j)\Sigma_0^{-1}(\mathbf{x}_j-\mu_0) = q\Sigma_0^{-1}\left[\tfrac{1}{q}\sum_{j=1}^n (1-y_j)\mathbf{x}_j - \mu_0\right].
\end{aligned}
$$

It follows that $\widehat{\mu}_1 = \frac{1}{p}\sum_{j=1}^n y_j\mathbf{x}_j$ and $\widehat{\mu}_0 = \frac{1}{q}\sum_{j=1}^n (1-y_j)\mathbf{x}_j$.
Moreover, $\Lambda_0 = \Sigma_0$ and $\Lambda_1 = \Sigma_1$. Notice that for

$$
\begin{aligned}
P &= \tfrac{1}{2}\sum_{j=1}^n y_j(\mathbf{x}_j-\mu_1)(\mathbf{x}_j-\mu_1)^\top \text{ and} \\
Q &= \tfrac{1}{2}\sum_{j=1}^n (1-y_j)(\mathbf{x}_j-\mu_0)(\mathbf{x}_j-\mu_0)^\top
\end{aligned}
$$

we can write the previous expression simply as:

$$= \tfrac{p}{2}\log\det\Lambda_1 + \tfrac{q}{2}\log\det\Lambda_0 - \operatorname{tr}(P\Lambda_1) - \operatorname{tr}(Q\Lambda_0)$$

Differentiating with respect to μ_1, μ_0, Λ_1 and Λ_0 gives (see section 0.1 for details),

$$
\begin{aligned}
-\tfrac{p}{2}\Lambda_1^{-1} + P = 0 \\
-\tfrac{q}{2}\Lambda_0^{-1} + Q = 0.
\end{aligned}
$$

Solving for Σ_0 and Σ_1 gives,

$$
\begin{aligned}
\widehat{\Sigma}_1 &= \tfrac{1}{p}\sum_{j=1}^n y_j(\mathbf{x}_j-\widehat{\mu}_1)(\mathbf{x}_j-\widehat{\mu}_1)^\top \\
\widehat{\Sigma}_0 &= \tfrac{1}{q}\sum_{j=1}^n (1-y_j)(\mathbf{x}_j-\widehat{\mu}_0)(\mathbf{x}_j-\widehat{\mu}_0)^\top
\end{aligned}
$$

0.1 Differentiation of the log-likelihood.

To differentiate $g(A) = \mathrm{tr}(B^\top A)$, notice that

$$g(A+H) - g(A) = \mathrm{tr}(B^\top H) = \langle B, H\rangle_F$$

Therefore, $\nabla_A g(A) = B$. And to differentiate the function $f(A) = \log\det A$, notice that using the Laplace expansion of $\det A$ and the chain rule we can derive

$$\begin{aligned}\tfrac{\partial}{\partial a_{ij}}\left[\det A\right] &= \tfrac{\partial}{\partial a_{ij}}\left[\sum_{k=1}^{n}(-1)^{i+j}a_{ij}M_{ij}\right] = (-1)^{i+j}M_{ij}\\ \tfrac{\partial}{\partial a_{ij}}\left[\log\det A\right] &= \tfrac{1}{\det A}(-1)^{i+j}M_{ij} = (A^{-1})_{ij}\end{aligned}$$

where M_{ij} denotes the ij-minor of A, that is, the determinant of the submatrix of A formed by removing the ith row and the jth column. Using these partial derivatives, we can write the gradient in the matrix formalism as follows:

$$\nabla_A f = \left[\tfrac{\partial f}{\partial a_{ij}}\right]_{i,j=1}^{n} = A^{-1}.$$

Alternatively, using the total derivative we can write

$$\begin{aligned}f(A+H) - f(A) &= \sum_{i,j=1}^{n}\tfrac{\partial}{\partial a_{ij}}\left[\log\det A\right]h_{ij} + o(\|H\|_F)\\ &= \langle A^{-1}, H\rangle_F + o(\|H\|_F)\\ &= \mathrm{tr}\big[(A^{-1})^\top H\big] + o(\|H\|_F)\\ &= \mathrm{tr}\big[(\nabla_A f)^\top H\big] + o(\|H\|_F)\end{aligned}$$

since $\sum_{i,j=1}^{n} A_{ij}B_{ij} = \mathrm{tr}(A^\top B)$. Either way, it follows that $\nabla_A f(A) = A^{-1}$.

(b) Give an expression of the conditional distribution $p(y=1|x)$ as a function of $\pi, \mu_1, \mu_2, \Sigma_1$ and Σ_2.

Solution:

$$\mathbb{P}(Y=1 \mid X=\mathbf{x}) = \left(1+\tfrac{f_{X|Y}(\mathbf{x}|Y=0)\mathbb{P}(Y=0)}{f_{X|Y}(\mathbf{x}|Y=1)\mathbb{P}(Y=1)}\right)^{-1} = \left(1+\tfrac{1-\pi}{\pi}\sqrt{\tfrac{|\Sigma_1|}{|\Sigma_0|}\tfrac{\exp\left((\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1)\right)}{\exp\left((\mathbf{x}-\mu_0)^\top\Sigma_0^{-1}(\mathbf{x}-\mu_0)\right)}}\right)^{-1}$$

(c) What is the equation of the classification boundary, i.e., of the set of points for which $p(y=1|x) = 0.5$?

Solution: The conic with equation

$$(\mathbf{x}-\mu_1)^\top\Sigma_1^{-1}(\mathbf{x}-\mu_1) - (\mathbf{x}-\mu_0)^\top\Sigma_0^{-1}(\mathbf{x}-\mu_0) = 2\log\tfrac{\pi}{1-\pi} + \log\tfrac{|\Sigma_0|}{|\Sigma_1|}.$$

(d) LDA model. Given the class variable $y \in \{0,1\}$, the data is now assumed to be Gaussian with different means for different classes but with the same covariance matrix.

$$y \sim \text{Bernoulli}(\pi), \quad x|\{y=i\} \sim \text{Normal}(\mu_k, \Sigma)$$

What is the maximum likelihood estimator for Σ now?

Solution: The solution is a little tricky. If one works out the pdf of $\mathbf{x}$ and then tries applying MLE, things do not work out. So instead, we shall work with the joint pdf of $\mathbf{x}$ and y. We write the likelihood as:

$$p(\{(\mathbf{x}_j, y_j)\}_{j=1}^n | \pi, \mu_0, \mu_1, \Sigma) = \prod_{j=1}^n \left[\pi \mathcal{N}(\mathbf{x}_j, \mu_1, \Sigma)\right]^{y_j} \left[(1-\pi)\mathcal{N}(\mathbf{x}_j, \mu_0, \Sigma)\right]^{1-y_j}$$

And therein lies the trick. Now, for Σ the relevant terms in the log-likelihood $\ell(\{(\mathbf{x}_j, y_j)\}_{j=1}^n | \pi, \mu_0, \mu_1, \Sigma)$ are:

$$\begin{aligned}
&= \left[\sum_{j=1}^n y_j \log \pi + (1-y_j)\log(1-\pi)\right] \\
&- \tfrac{n}{2} \log \det \Sigma \\
&- \tfrac{1}{2} \sum_{j=1}^n y_j (\mathbf{x}_j - \mu_1)^\top \Sigma^{-1} (\mathbf{x}_j - \mu_1) + (1-y_j)(\mathbf{x}_j - \mu_0)^\top \Sigma^{-1} (\mathbf{x}_j - \mu_0)
\end{aligned}$$

The terms μ_0, μ_1 and π can be dealt with in the usual way. So let $\Lambda = \Sigma^{-1}$. Maximizing with respect to Σ is equivalent to maximizing with respect to Λ. We can write the last two terms of the above expression as

$$= \tfrac{n}{2} \log \det \Lambda - \text{tr}(M\Lambda)$$

for

$$M = \tfrac{1}{2} \sum_{j=1}^n y_j (\mathbf{x}_j - \mu_1)(\mathbf{x}_j - \mu_1)^\top + (1-y_j)(\mathbf{x}_j - \mu_0)(\mathbf{x}_j - \mu_0)^\top$$

Differentiating with respect to Λ gives:

$$-\tfrac{n}{2}\Lambda^{-1} + M = 0$$

Solving for Σ gives:

$$\Sigma = \tfrac{1}{n} \sum_{j=1}^n y_j (\mathbf{x}_j - \mu_1)^\top (\mathbf{x}_j - \mu_1) + (1-y_j)(\mathbf{x}_j - \mu_0)^\top (\mathbf{x}_j - \mu_0)$$

And thus, $\widehat{\Sigma} = (1-\widehat{\pi})\widehat{\Sigma}_0 + \widehat{\pi}\widehat{\Sigma}_1$.

(e) What is the equation of the classification boundary, i.e., of the set of points for which $p(y=1|x) = 0.5$? Compare the obtained predictor with the form of the logistic regression predictor.

Solution: From (b), we have

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = \mathbf{x}) &= \left(1 + \tfrac{1-\pi}{\pi}\sqrt{\tfrac{\exp\left((\mathbf{x}-\mu_1)^\top \Sigma^{-1}(\mathbf{x}-\mu_1)\right)}{\exp\left((\mathbf{x}-\mu_0)^\top \Sigma^{-1}(\mathbf{x}-\mu_0)\right)}}\right)^{-1} \\
&= \left(1 + \exp\left((\mu_0 - \mu_1)^\top \Sigma^{-1}\mathbf{x} + b\right)\right)^{-1} \\
&= \sigma(w^\top \mathbf{x} + b)
\end{aligned}$$

where $w = \Sigma^{-1}(\mu_0 - \mu_1)$ and $b = \log \frac{1-\pi}{\pi} + \frac{1}{2}\mu_1^\top \Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^\top \Sigma^{-1}\mu_0$. Now, $\sigma(w^\top \mathbf{x} + b) = 1/2$, implies that $w^\top x + b = 0$. Thus the classification boundary is given by the hyperplane of equation

$$(\mu_0 - \mu_1)^\top \Sigma^{-1}\mathbf{x} + b = 0$$

Notice, by the way, that Fisher's linear discriminant has the same logistic function form as in linear regression.