Gabriel Romon

Exercise 1

All computational details and further arguments are deferred to Page 5. Let $(z_1, x_1), \dots, (z_n, x_n)$ be an i.i.d sample from (z, x). The log-likelihood is given by $\displaystyle\sum_{m=1}^{M} \log \pi_m \sum_{i=1}^{n} \mathbb{1}_{z_i=m} + \sum_{m=1}^{M}\sum_{k=1}^{K} \log\theta_{mk} \sum_{i=1}^{n} \mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}$. Since $\sup_{x,y}(f(x)+g(y)) = \sup_x f(x) + \sup_y g(y)$, both summands in the previous equation can be maximized separately. Let $n_m = \sum_{i=1}^n \mathbb{1}_{z_i=m}$ and $n_{mk} = \sum_{i=1}^n \mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}$.

The first problem is $\max_\pi \sum_{m=1}^M n_m \log \pi_m$ s.t. $\sum_{m=1}^M \pi_m = 1$. It is amenable to Lagrange multipliers. Computations show that $\boxed{\pi_m = \frac{n_m}{n}}$ is optimal.
The second problem is $\max_{\theta_{mk}} \sum_{m=1}^M \sum_{k=1}^K n_{mk} \log\theta_{mk}$ s.t. $\forall m, \sum_{k=1}^K \theta_{mk} = 1$. This can also be solved using Lagrange multipliers. The optimum is given by $\boxed{\theta_{mk} = \frac{n_{mk}}{n_m}}$.

LDA

All computational details and further arguments are deferred to Page 5. Let $(y_1, x_1), \dots, (y_n, x_n)$ be an i.i.d sample from (y, x), $n_1 = |\{i, y_i = 1\}|$ and $n_0 = n - n_1$. The log-likelihood of the model is

$$\begin{aligned}\ell(\mu_0, \mu_1, \Sigma, \pi | y_i, x_i) = & -\frac{n}{2}\log((2\pi)^d) + n_1 \log\pi + (n-n_1)\log(1-\pi) \\ & -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\left(\sum_{\substack{i\\ y_i=1}} (x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) + \sum_{\substack{i\\ y_i=0}} (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\right)\end{aligned}$$

which may be maximized separately in π and (μ_0, μ_1, Σ). Computations show that the optimal parameters are

$$\begin{cases} \pi = \frac{n_1}{n} \qquad \mu_1 = \frac{1}{n_1}\sum_{i,y_i=1} x_i \qquad \mu_0 = \frac{1}{n_0}\sum_{i,y_i=0} x_i \\ \Sigma = \frac{1}{n}\left(\sum_{i,y_i=1}(x_i-\mu_1)(x_i-\mu_1)^T + \sum_{i,y_i=0}(x_i-\mu_0)(x_i-\mu_0)^T\right)\end{cases}$$

Using Bayes' theorem and a bit of algebra,

$$P(y=1|x) = \sigma\left(x^T \underbrace{\Sigma^{-1}(\mu_1-\mu_0)}_{\beta} + \underbrace{\frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_0-\mu_1) + \log\frac{\pi}{1-\pi}}_{\gamma}\right)$$

QDA

All computational details and further arguments are deferred to Page 7. Let $(y_1, x_1), \dots, (y_n, x_n)$ be an i.i.d sample from (y, x), $n_1 = |\{i, y_i = 1\}|$ and $n_0 = n - n_1$. The log-likelihood of the model is

$$\begin{aligned}\ell(\mu_0, \mu_1, \Sigma_0, \Sigma_1, \pi | y_i, x_i) = & -\frac{n}{2}\log((2\pi)^d) + n_1 \log\pi + (n-n_1)\log(1-\pi) \\ & -\frac{n_1}{2}\log|\Sigma_1| - \frac{n_0}{2}\log|\Sigma_0| - \frac{1}{2}\left(\sum_{\substack{i\\ y_i=1}} (x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1) + \sum_{\substack{i\\ y_i=0}} (x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)\right)\end{aligned}$$

Computations show that the optimal parameters are

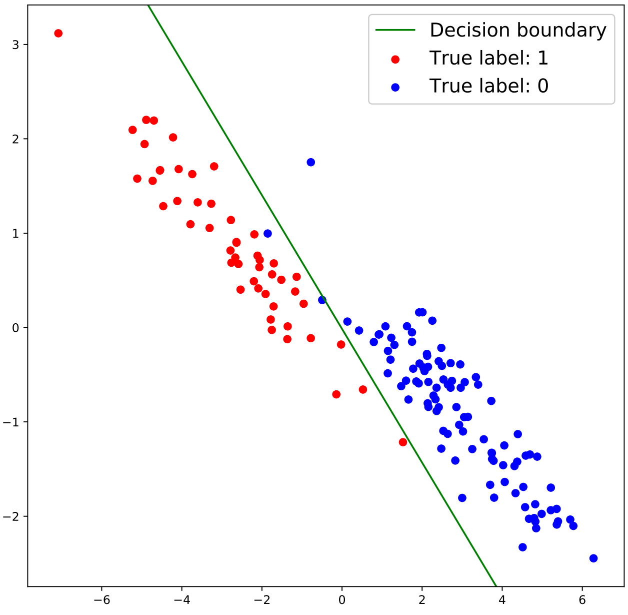
$$\begin{cases} \pi = \frac{n_1}{n} \qquad \mu_1 = \frac{1}{n_1}\sum_{i,y_i=1} x_i \qquad \mu_0 = \frac{1}{n_0}\sum_{i,y_i=0} x_i \\ \Sigma_1 = \frac{1}{n_1}\left(\sum_{i,y_i=1}(x_i-\mu_1)(x_i-\mu_1)^T\right) \qquad \Sigma_0 = \frac{1}{n_0}\left(\sum_{i,y_i=0}(x_i-\mu_0)(x_i-\mu_0)^T\right)\end{cases}$$

Using Bayes' theorem and a bit of algebra,

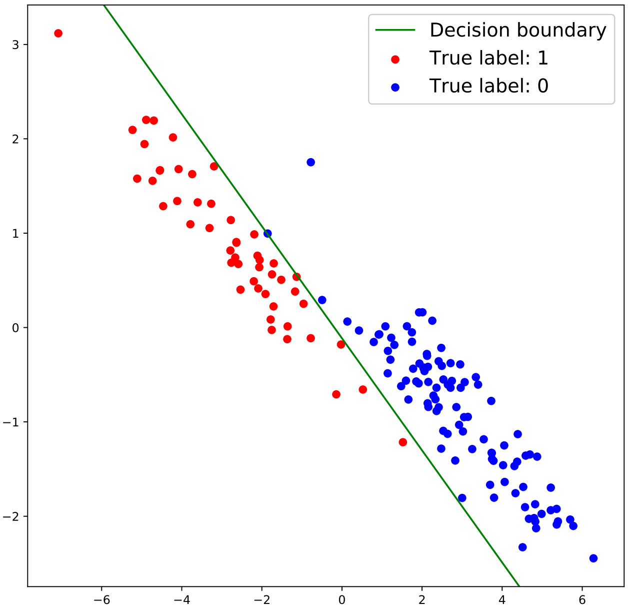
$$\begin{aligned}P(y=1|x) = \sigma\Bigg(&\frac{1}{2}x^T(\Sigma_0^{-1}-\Sigma_1^{-1})x + x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) \\ & + \log\frac{\pi}{1-\pi} + \frac{1}{2}\left[\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1 + \log|\Sigma_0| - \log|\Sigma_1|\right]\Bigg)\end{aligned}$$

Dataset A

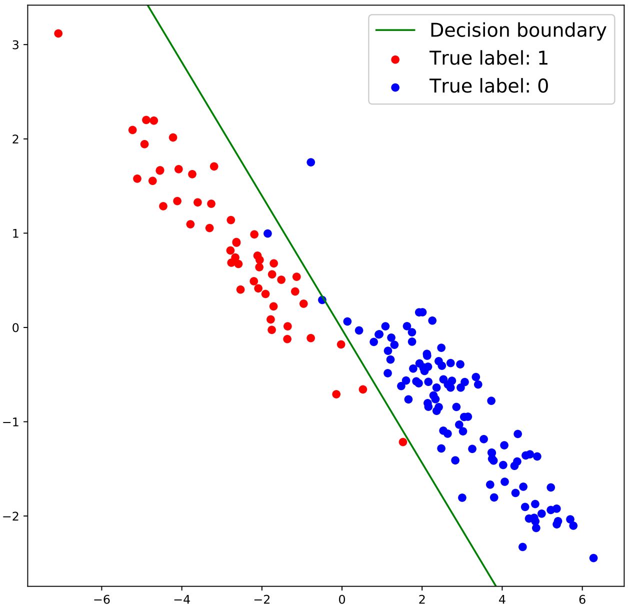
All the plots were done on the **training set**.



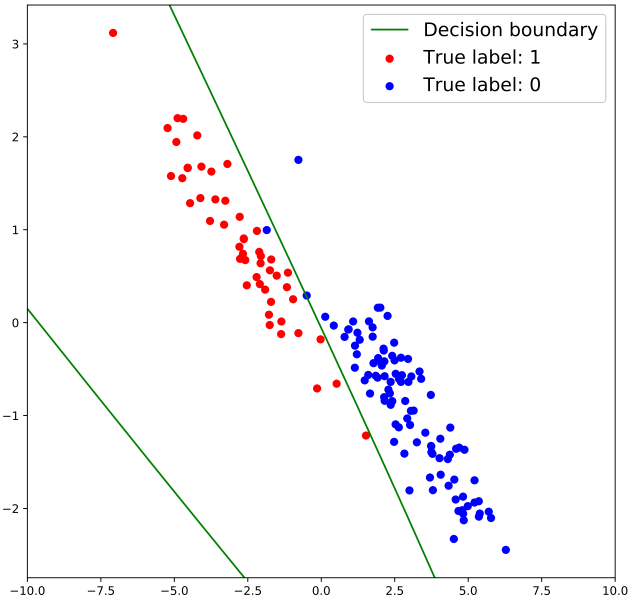
LDA



Logistic regression



Linear regression



QDA

	Train	Test
LDA	1.33	2.00
Logistic	0.00	3.27
Linear	1.33	2.07
QDA	0.67	2.00

Table 1: Missclassification rates (%) on the training and test sets for Dataset A

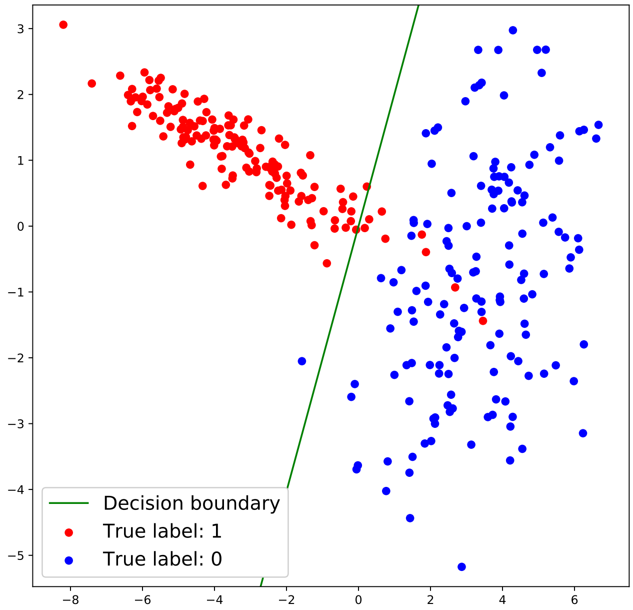
• The training data is made up of two parallel clusters with similar densities. All four methods yield satisfactory results. LDA, Linear regression and QDA exhibit similar performance on the training and test sets, while Logistic regression seems to overfit: it has perfect accuracy on the training set but performs worst on the test set.

• Because the training set is linearly separable, it can be shown that the log-loss can be made arbitrarily small by taking w with large norm. This is the reason for the numeric instability we have observed. Our stopping criterion for the IRLS algorithm is to stop as soon as the model classifies every point of the training set correctly.

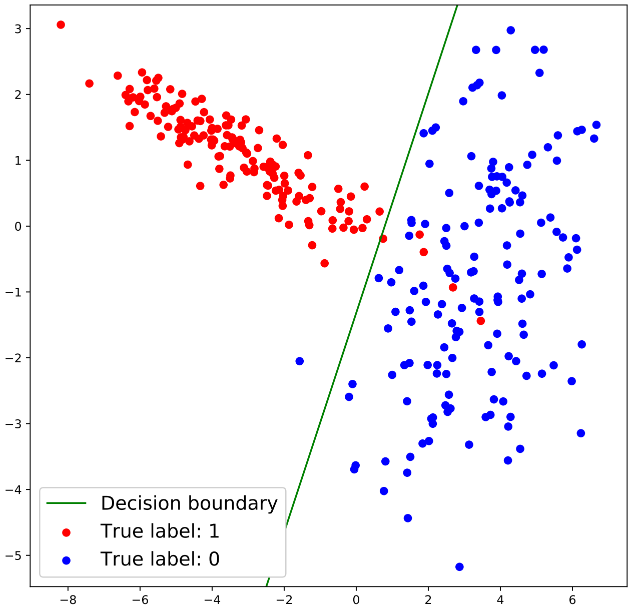
• In this linearly separable case, we have observed that the performance of the IRLS method depends greatly on the initialization.

Dataset B

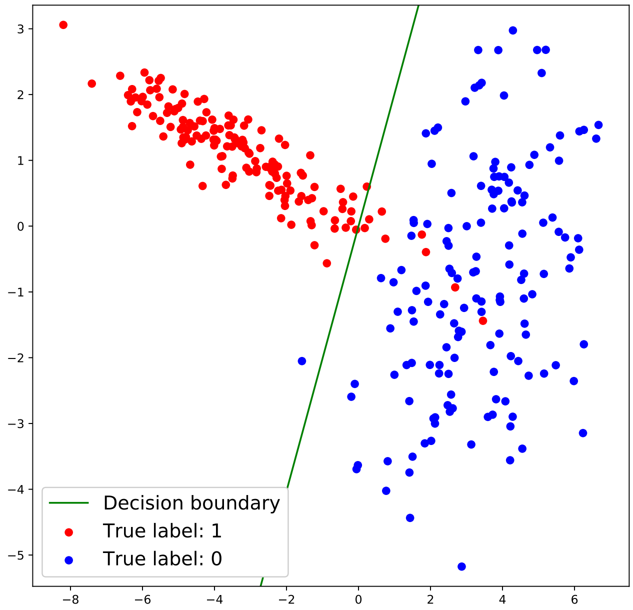
All the plots were done on the **training set**.



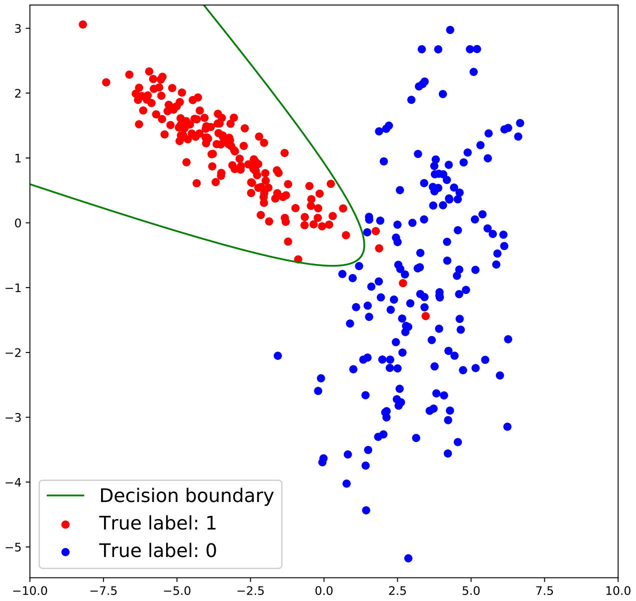
LDA



Logistic regression



Linear regression



QDA

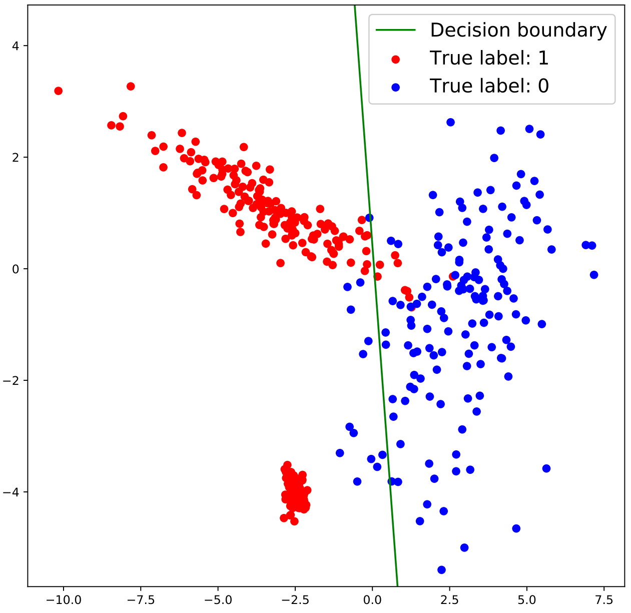
	Train	Test
LDA	3.00	4.15
Logistic	2.00	4.30
Linear	3.00	4.15
QDA	1.33	2.00

Table 2: Missclassification rates (%) on the training and test sets for Dataset B

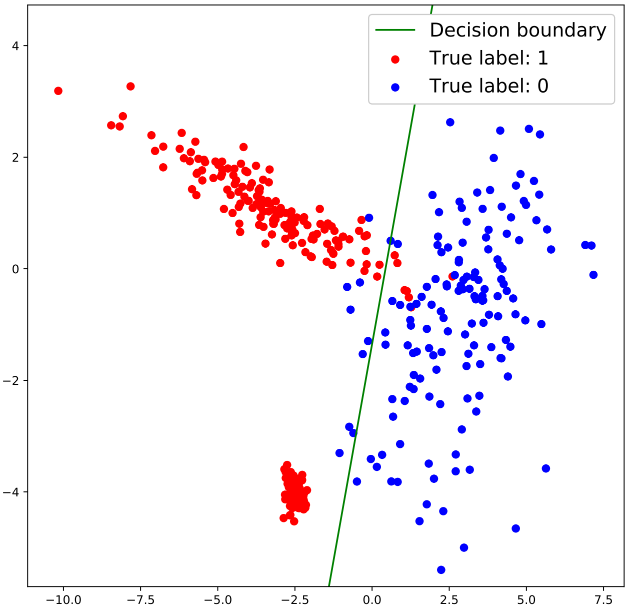
- The training data is made up of two orthogonal clusters with different densities. All four methods yield satisfactory results. LDA, Logistic and Linear regression exhibit similar performance on the training set, while QDA performs better on both sets.
- QDA takes into account the distinct cluster densities because it models the data with two different covariance matrices, hence the better accuracy.
- Besides, the curvature of the ellipsoid makes it easier for QDA to discriminate between the two classes, compared to the other three models with linear decision boundaries.

Dataset C

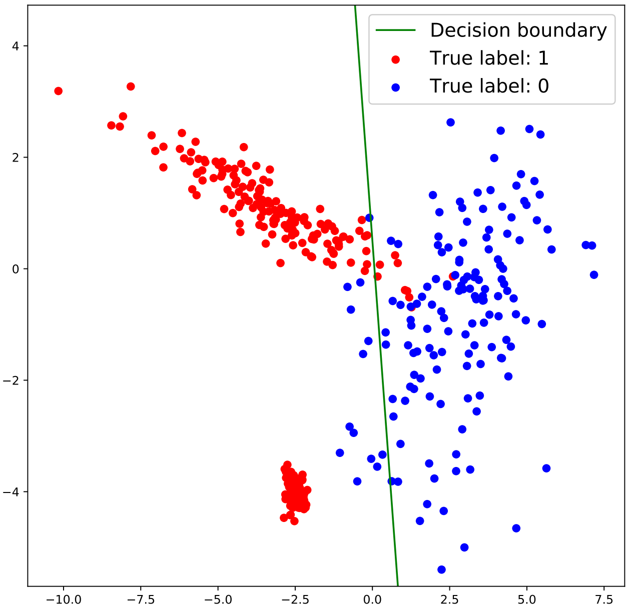
All the plots were done on the **training set**.



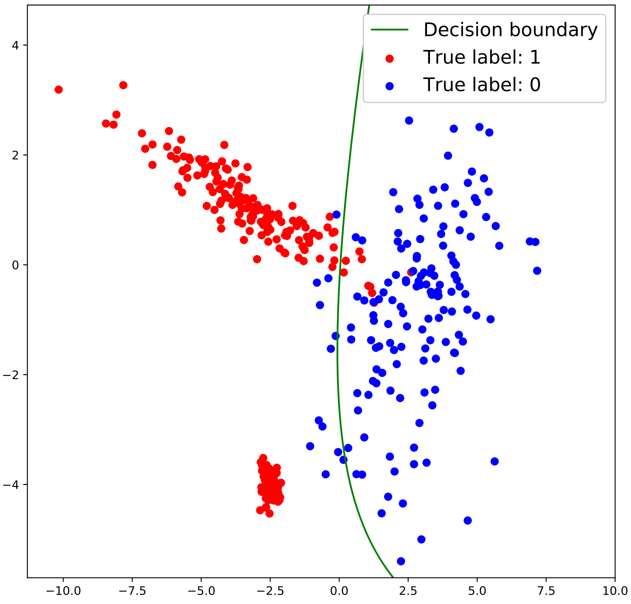
LDA



Logistic regression



Linear regression



QDA

	Train	Test
LDA	5.50	4.23
Logistic	4.00	2.27
Linear	5.50	4.23
QDA	5.25	3.83

Table 3: Missclassification rates (%) on the training and test sets for Dataset C

- A cluster of 1's has been added to the previous training set. The points with label 1 should therefore be considered as the result of a mixture of two Gaussians with different centers and covariances. LDA and QDA hypothesise that these points originate from exactly one Gaussian, hence their poor performance on this dataset.
- On the contrary, Logistic Regression performs better than the others. That is because it does not try to model the x's (it is a discriminative model, not a generative one).

Exercise 1

Here's a more detailed solution with all the arguments fleshed out.
For the conditional probability $P(x=k|z=m)$ to be well-defined, one must assume that $\forall m \in [\![1,M]\!], \pi_m > 0$. Let $(z_1,x_1),\dots,(z_n,x_n)$ be an i.i.d sample from (z,x). The likelihood of the model may be written as

$$\begin{aligned} L(\pi,\theta|z_i,x_i) &= \prod_{i=1}^n P(z=z_i \cap x=x_i) = \prod_{i=1}^n P(z=z_i)P(x=x_i|z=z_i) \\ &= \prod_{i=1}^n \prod_{m=1}^M \pi_m^{\mathbb{1}_{z_i=m}} \prod_{i=1}^n \prod_{m=1}^M \prod_{k=1}^K \theta_{mk}^{\mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}} \end{aligned}$$

Let $A = \{(m,k) \in [\![1,M]\!] \times [\![1,K]\!], \sum_{i=1}^n (\mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}) > 0\}$. If $(m,k) \notin A$, θ_{mk} does not appear in the likelihood, **so we may set $\boldsymbol{\theta_{mk}=0}$**. If $(m,k) \in A$ and $\theta_{mk}=0$, then the likelihood is 0. Thus we might restrict ourselves to $(m,k) \in A$ and $\theta_{mk} > 0$.

The log-likelihood is well-defined and given by $\sum\limits_{m=1}^M \log \pi_m \sum\limits_{i=1}^n \mathbb{1}_{z_i=m} + \sum\limits_{(m,k)\in A} \log \theta_{mk} \sum\limits_{i=1}^n \mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}$. Since the objective and the constraints are separable in π and θ, both summands in the previous equation can be maximized separately. Let $n_m = \sum_{i=1}^n \mathbb{1}_{z_i=m}$ and $n_{mk} = \sum_{i=1}^n \mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}$.

- The first problem is $\max_\pi \sum_{m=1}^M n_m \log \pi_m$ s.t. $\sum_{m=1}^M \pi_m = 1$. Since the objective is concave and the constraint is affine, KKT conditions yield solutions that are primal and dual optimal (see *KKT conditions for convex problems* p.244 in Boyd's book). Thus, it suffices to find a solution of $\begin{cases} \forall m, \pi_m \in (0,1] \\ \forall m, \frac{n_m}{\pi_m} + \lambda = 0 \\ \sum_{m=1}^M \pi_m = 1 \end{cases}$

If one of the n_m is 0, the system has no solutions. Otherwise, $n_m = -\lambda \pi_m$, hence $n = \sum_{m=1}^M n_m = -\lambda$, yielding $\boxed{\pi_m = \frac{n_m}{n}}$.

- The second problem is $\max\limits_{\substack{\theta_{mk} \\ (m,k)\in A}} \sum\limits_{(m,k)\in A} n_{mk} \log \theta_{mk}$ s.t. $\forall m, \sum\limits_{k=1}^K \theta_{mk} = 1$. Here too KKT conditions are sufficient. It suffices to find a solution of $\begin{cases} \forall (m,k) \in A, \theta_{mk} \in (0,1] \\ \forall (m,k) \in A, \frac{n_{mk}}{\theta_{mk}} + \lambda_m = 0 \\ \forall m, \sum_{k=1}^K \theta_{mk} = 1 \end{cases}$.

By definition of A, all the $(n_{mk})_{(m,k)\in A}$ are > 0. $n_{mk} = -\lambda_m \theta_{mk}$ hence $n_m = \sum_{k=1}^K n_{mk} = -\lambda_m$. Thus a solution is $\theta_{mk} = \frac{n_{mk}}{n_m}$ if $(m,k) \in A$ and 0 otherwise, which, by definition of A, rewrites more compactly as

$$\boxed{\forall (m,k) \in [\![1,M]\!] \times [\![1,K]\!], \theta_{mk} = \frac{n_{mk}}{n_m}}$$

LDA

- Let N be the counting measure with respect to $\{0,1\}$. By Bayes' theorem, the joint density of (y,x) with respect to $N \otimes \lambda_d$ is

$$f_{y,x}(i,a) = f_y(i) f_{x|y=i}(a) = \pi^i (1-\pi)^{1-i} \frac{1}{((2\pi)^d|\Sigma|)^{1/2}} \exp(-\frac{1}{2}(a-\mu_i)^T \Sigma^{-1}(a-\mu_i))$$

Let $(y_1,x_1),\dots,(y_n,x_n)$ be an i.i.d sample from (y,x). The likelihood of the model is

$$\begin{aligned} L(\mu_0,\mu_1,\Sigma,\pi|y_i,x_i) &= \frac{1}{((2\pi)^d|\Sigma|)^{n/2}} \prod_{\substack{i \\ y_i=1}}^n \pi \exp(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)) \prod_{\substack{i \\ y_i=0}}^n (1-\pi) \exp(-\frac{1}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)) \\ &= \frac{\pi^{n_1}(1-\pi)^{n-n_1}}{((2\pi)^d|\Sigma|)^{n/2}} \prod_{\substack{i \\ y_i=1}} \pi \exp(-\frac{1}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)) \prod_{\substack{i \\ y_i=0}} (1-\pi) \exp(-\frac{1}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0) \end{aligned}$$

- The log-likelihood of the model is

$$\begin{aligned}\ell(\mu_0, \mu_1, \Sigma, \pi | y_i, x_i) = & -\frac{n}{2}\log((2\pi)^d) + n_1 \log \pi + (n - n_1)\log(1-\pi) \\ & -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\left(\sum_{\substack{i \\ y_i=1}} (x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + \sum_{\substack{i \\ y_i=0}} (x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0)\right)\end{aligned}$$

which may be maximized separately in π and (μ_0, μ_1, Σ).

- $\pi \mapsto n_1 \log \pi + (n - n_1)\log(1-\pi)$ may be optimized simply by looking at its derivative: the maximum occurs at $\pi = \frac{n_1}{n} \in [0,1]$

- $(\mu_0, \mu_1, \Sigma) \mapsto n \log|\Sigma| + \left(\sum_{\substack{i \\ y_i=1}} (x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + \sum_{\substack{i \\ y_i=0}} (x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0)\right)$ is to be minimized.

For fixed Σ, the function $\mu \mapsto (x-\mu)^T \Sigma^{-1}(x-\mu)$ has Hessian $\Sigma^{-1} \in S_d^+(\mathbb{R})$ and is thus convex. The previous function is consequently convex in (μ_0, μ_1) as the separable sum of two convex functions. The gradient of $\mu \mapsto (x-\mu)^T\Sigma^{-1}(x-\mu)$ is $-2\Sigma^{-1}(x-\mu)$, hence equating the gradient with respect to μ_1 to 0 yields $\sum_{\substack{i \\ y_i=1}} (-2\Sigma^{-1}(x_i - \mu_1)) = 0$, that is $\Sigma^{-1} \sum_{\substack{i \\ y_i=1}} (x_i - \mu_1) = 0$ and since Σ is invertible, $\boxed{\mu_1 = \frac{1}{n_1} \sum_{\substack{i \\ y_i=1}} x_i}$ and similarly $\boxed{\mu_0 = \frac{1}{n_0} \sum_{\substack{i \\ y_i=0}} x_i}$.

For these values of μ_0 and μ_1, we're left with minimizing

$$\Sigma \mapsto n \log|\Sigma| + \left(\sum_{\substack{i \\ y_i=1}} (x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + \sum_{\substack{i \\ y_i=0}} (x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0)\right) \qquad (*)$$

By the change of variable $\Delta = \Sigma^{-1}$ this turns into

$$\Delta \mapsto -n \log|\Delta| + \left(\sum_{\substack{i \\ y_i=1}} (x_i - \mu_1)^T \Delta(x_i - \mu_1) + \sum_{\substack{i \\ y_i=0}} (x_i - \mu_0)^T \Delta(x_i - \mu_0)\right)$$

It is well-known that $\Delta \mapsto \log|\Delta|$ is concave with gradient Δ^{-1} and since

$$\begin{aligned}(x-\mu)^T\Delta(x-\mu) &= \operatorname{tr}((x-\mu)^T\Delta(x-\mu)) = \operatorname{tr}(\Delta(x-\mu)(x-\mu)^T)) \\ &= \langle \Delta, (x-\mu)(x-\mu)^T \rangle \quad \text{where } \langle \cdot, \cdot \rangle \text{ is Frobenius inner product,}\end{aligned}$$

the gradient of $\Delta \mapsto (x-\mu)^T\Delta(x-\mu)$ is $(x-\mu)(x-\mu)^T$. Consequently, a critical point for the previous function of Δ is such that $-n\Delta^{-1} + \left(\sum_{\substack{i \\ y_i=1}} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{\substack{i \\ y_i=0}} (x_i - \mu_0)(x_i - \mu_0)^T\right) = 0$, hence

$$\Delta^{-1} = \frac{1}{n}\left(\sum_{\substack{i \\ y_i=1}} (x_i - \mu_1)(x_i - \mu_1)^T + \sum_{\substack{i \\ y_i=0}} (x_i - \mu_0)(x_i - \mu_0)^T\right)$$

Finally, this value of Δ^{-1} minimizes $(*)$, so $\boxed{\text{it's the optimal } \Sigma}$.

- By Bayes' theorem,

$$\begin{aligned}P(y=1|x=a) &= \frac{f_{x|y=1}(a)f_y(1)}{f_X(a)}\\ &= \frac{\frac{\pi}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(a-\mu_1)^T\Sigma^{-1}(a-\mu_1)\right)}{\frac{\pi}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(a-\mu_1)^T\Sigma^{-1}(a-\mu_1)\right)+\frac{1-\pi}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(a-\mu_0)^T\Sigma^{-1}(a-\mu_0)\right)}\\ &= \sigma\left(a^T\Sigma^{-1}(\mu_1-\mu_0)+\frac{1}{2}(\mu_1+\mu_0)^T\Sigma^{-1}(\mu_0-\mu_1)+\log\frac{\pi}{1-\pi}\right)\end{aligned}$$

QDA

- The reasoning is very similar to that of LDA. Let $(y_1,x_1),\dots,(y_n,x_n)$ be an i.i.d sample from (y,x). The likelihood of the model is

$$L(\mu_0,\mu_1,\Sigma_0,\Sigma_1,\pi|y_i,x_i) = \frac{\pi^{n_1}(1-\pi)^{n-n_1}}{(2\pi)^{nd/2}|\Sigma_0|^{n_0/2}|\Sigma_1|^{n_1/2}}\prod_{\substack{i\\ y_i=1}}\pi\exp(-\frac{1}{2}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1))$$
$$\prod_{\substack{i\\ y_i=0}}(1-\pi)\exp(-\frac{1}{2}(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0))$$

- The log-likelihood consequently writes as

$$\begin{aligned}\ell(\mu_0,\mu_1,\Sigma_0,\Sigma_1,\pi|y_i,x_i) = -\frac{n}{2}\log((2\pi)^d) &+ n_1\log\pi+(n-n_1)\log(1-\pi)\\ &-\frac{n_1}{2}\log|\Sigma_1|-\frac{n_0}{2}\log|\Sigma_0|-\frac{1}{2}\left(\sum_{\substack{i\\ y_i=1}}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)+\sum_{\substack{i\\ y_i=0}}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\right)\end{aligned}$$

which may be maximized separately in π and $(\mu_0,\mu_1,\Sigma_0,\Sigma_1)$.

- Exactly the same steps as in LDA may be reproduced, yielding

$$\begin{cases}\pi=\frac{n_1}{n}\\ \mu_1=\frac{1}{n_1}\sum_{i,y_i=1}x_i\\ \mu_0=\frac{1}{n_0}\sum_{i,y_i=0}x_i\\ \Sigma_1=\frac{1}{n_1}\left(\sum_{i,y_i=1}(x_i-\mu_1)(x_i-\mu_1)^T\right)\\ \Sigma_0=\frac{1}{n_0}\left(\sum_{i,y_i=0}(x_i-\mu_0)(x_i-\mu_0)^T\right)\end{cases}$$