HMWK1

Bouabid - Bounmy
Master M2 MVA 2018/2019
Probabilistic Graphical Models

October 23, 2018

MLE of π

Let $(z_i)_{i\in[\![1,n]\!]}$ iid, the log-likelighood is given by : $\ell(\pi) = \sum_{m=1}^{M} n_m \log \pi_m\,,\ \ n_m = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=m\}}$

$-\ell$ being convex and as $\exists \pi \in [0,1]^M / \pi^T 1_M = 1$, by Slater's constraints qualification we have strong duality and we can address its dual problem given by : $\max_{\lambda} \min_{\pi} \mathcal{L}(\lambda, \pi)\,,$ where $\mathcal{L}(\lambda,\pi) = -\ell(\pi) + \lambda(\pi^T 1_M - 1)$

$\mathcal{L}$ being convex w.r.t to π we can minimize it through its gradient : $\forall m,\ \dfrac{\partial \mathcal{L}}{\partial \pi_m} = 0 \Rightarrow -\dfrac{n_m}{\pi_m} + \lambda = 0 \Rightarrow \pi_m = \dfrac{n_m}{\lambda}$

Plus, $\pi^T 1_M = 1 \Rightarrow \sum_{m=1}^{M} \frac{n_m}{\lambda} = 1 \Rightarrow \lambda = \sum_{m=1}^{M} n_m = n$, hence : $\forall m,\ \hat{\pi}_m = \dfrac{n_m}{n}$

MLE of Θ

Let $(x_i)_{i\in[\![1,n]\!]}$ and $(z_i)_{i\in[\![1,n]\!]}$ iid and $\Theta = [\theta_{mk}] \in [0,1]^{M\times K}$. Conditional probability allow us to write the log-likelihood as : $\ell(\Theta,\pi) = \sum_{m=1}^{M} n_m \log \pi_m + \sum_{k=1}^{K}\sum_{m=1}^{M} n_{mk} \log \theta_{mk}\,,\ \ n_m = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=m\}}\,,\ \ n_{mk} = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=m, x_i=k\}}$

Samely, $\mathcal{L}(\lambda,\Theta,\pi) = -\ell(\Theta,\pi) + \begin{pmatrix}\pi^T\Theta 1_K - 1 & \pi^T 1_M - 1\end{pmatrix}\lambda\,,\ \ \lambda \in \mathbb{R}_+^2$

Derivating w.r.t to π we obtain the same estimtor as previously.

For Θ, the derivation goes : $\forall m,k,\ \dfrac{\partial \mathcal{L}}{\partial \theta_{mk}} = 0 \Rightarrow -\dfrac{n_{mk}}{\theta_{mk}} + \lambda_1 \pi_m = 0 \Rightarrow \theta_{mk} = \dfrac{n_{mk}}{\lambda_1 \pi_m}$

Once again, the constraints gives us $\lambda_1 = n$, hence : $\forall m,k,\ \hat{\theta}_{mk} = \dfrac{n_{mk}}{n\hat{\pi}_m} = \dfrac{n_{mk}}{n_m}$

LDA formulas

$$Y \sim \mathcal{B}(\pi),\ \ X|\{Y=i\} \sim \mathcal{N}(\mu_i, \Sigma).$$

$$\hat{\pi} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

$$\forall j \in \{0,1\},\ \ \hat{\mu}_j = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}} x_i}{\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i=0\}}(x_i-\hat{\mu}_0)(x_i-\hat{\mu}_0)^T + \mathbb{1}_{\{y_i=1\}}(x_i-\hat{\mu}_1)(x_i-\hat{\mu}_1)^T$$

$$p(y=1|x) = \frac{1}{2} \Leftrightarrow \left(\Sigma^{-1}(\mu_1-\mu_0)\right)^T x + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 = \log\left(\frac{\pi}{1-\pi}\right)$$

QDA formulas

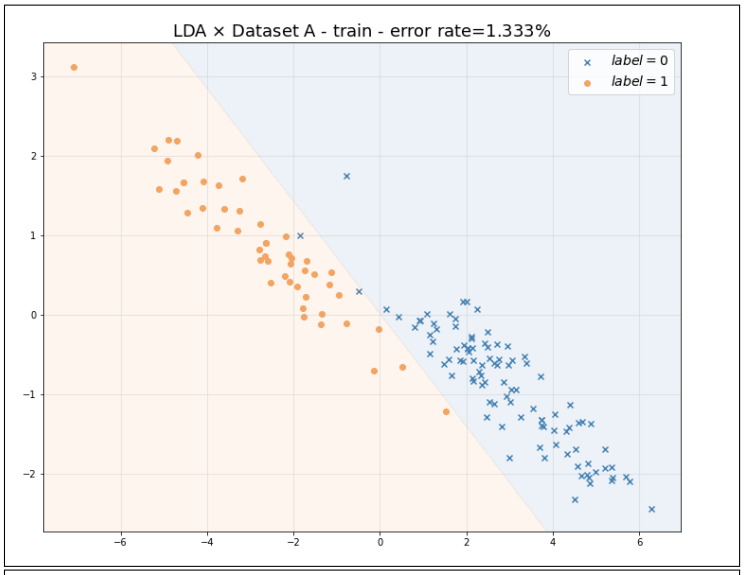
$$Y \sim \mathcal{B}(\pi),\ \ X \mid \{Y=i\} \sim \mathcal{N}(\mu_i, \Sigma_i)$$

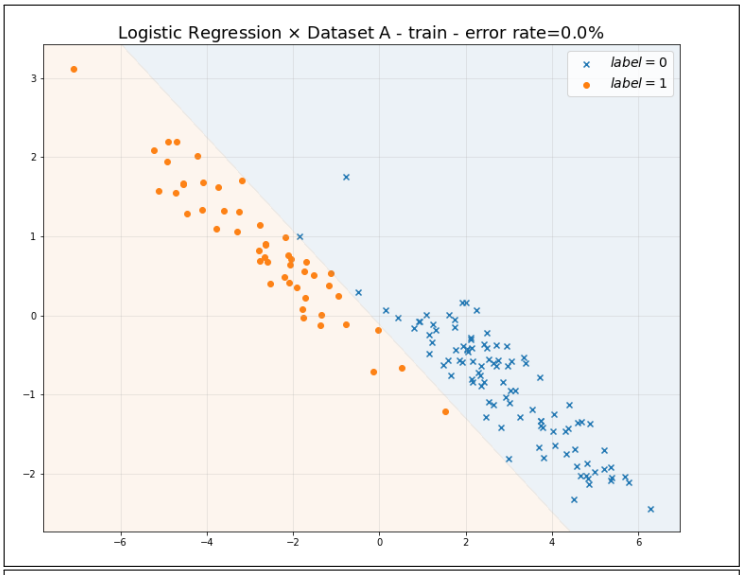
$$\hat{\pi} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

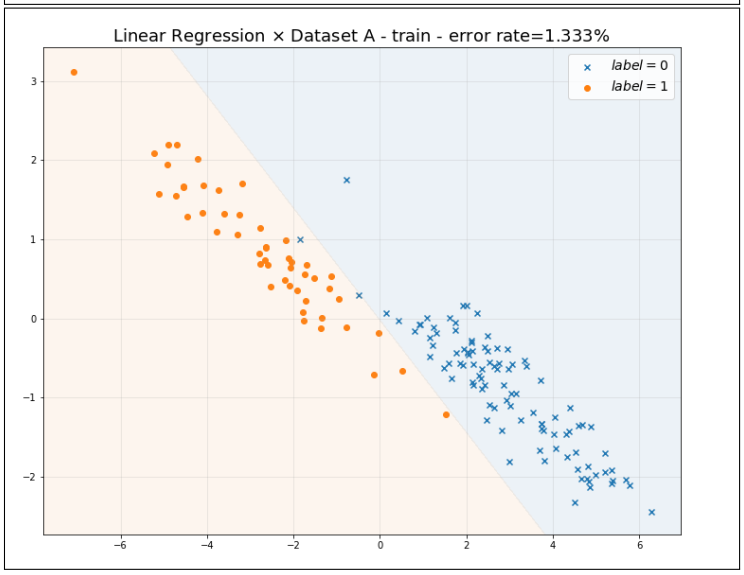
$$\forall j \in \{0,1\},\ \ \hat{\mu}_j = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}} x_i}{\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}}$$

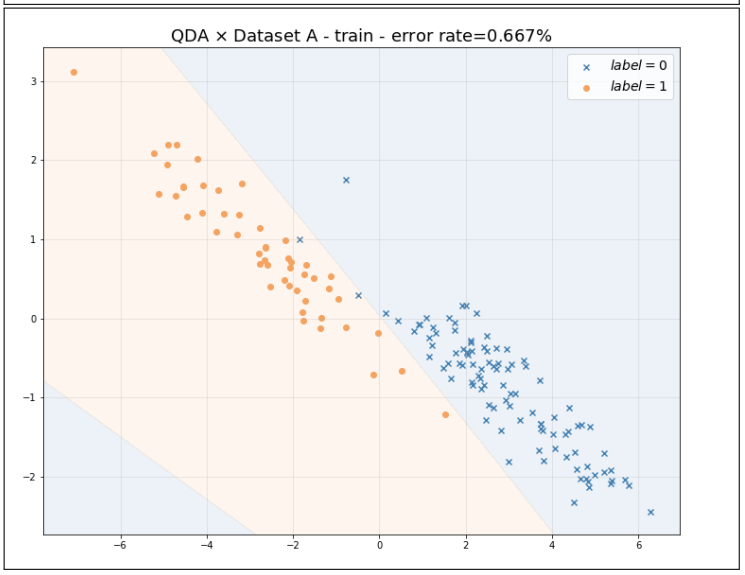
$$\forall j \in \{0,1\},\ \ \hat{\Sigma}_j = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}(x_i-\hat{\mu}_j)(x_i-\hat{\mu}_j)^T}{\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}}$$

$$p(y=1|x) = \frac{1}{2} \Leftrightarrow \frac{1}{2}\log\left(\frac{\det\Sigma_1^{-1}}{\det\Sigma_0^{-1}}\right) + \frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right] = \log\left(\frac{\pi}{1-\pi}\right)$$









Missclassification on Dataset A (%)

	test	train
LDA	2.000	1.333
Linear Regression	2.067	1.333
Logistic Regression	3.467	0.000
QDA	2.000	0.667

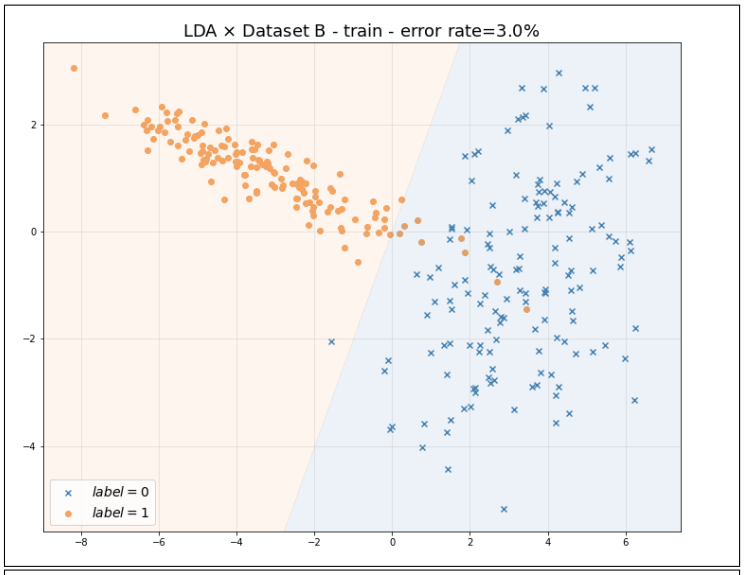
Comments:

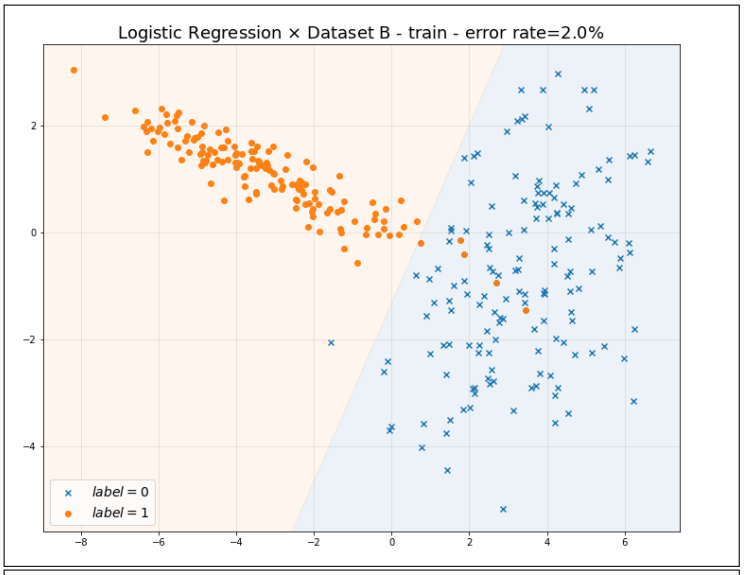
Globally, data is linearly separable and all classifier provide satisfying results. Logistic regression even returns a perfect score on the training set.

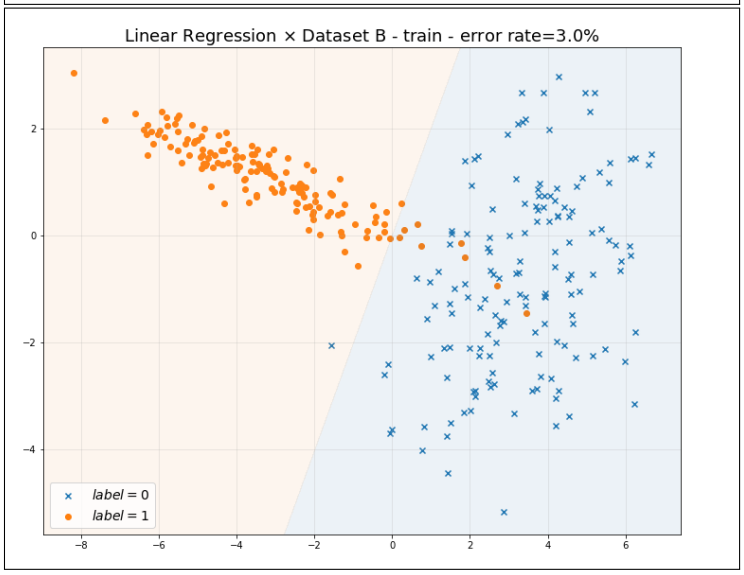
The best performer here is QDA, but LDA yields similar performances as the "scattering" of the two classes are similar (supporting the assumption of a unique covariance matrix).

Missclassification tends to be larger on the testing set than on the training set which is an expected behavior as our classifiers learn from the train data and should hence be better at predicting from this dataset. The testing set being 10 times as large as the training set emphasises this trend by testing the model robustness.

For instance, Logistic Regression yields very dissimilar results, pointing out a lack of robustness.







Missclassification on Dataset B (%)

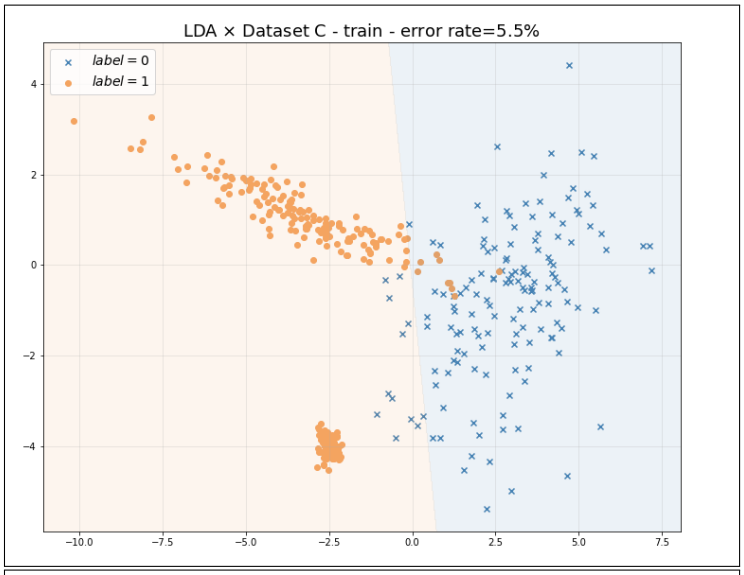
	test	train
LDA	4.15	3.000
Linear Regression	4.15	3.000
Logistic Regression	4.30	2.000
QDA	2.00	1.333

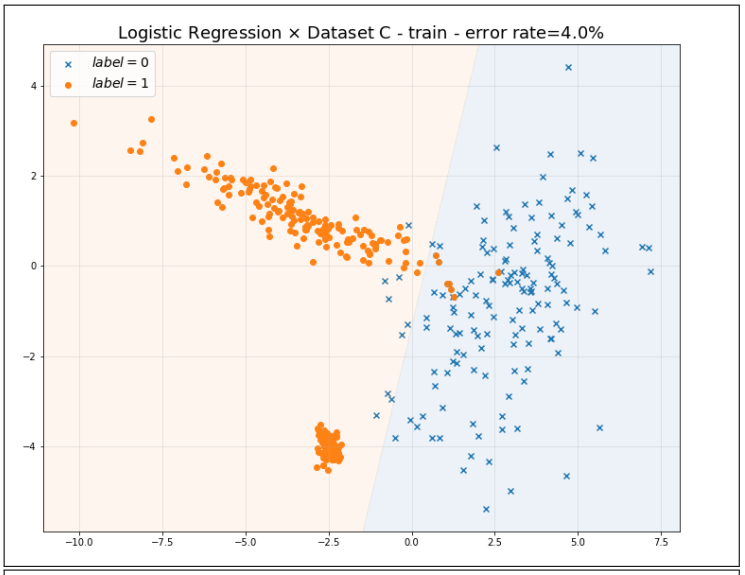
Comments:

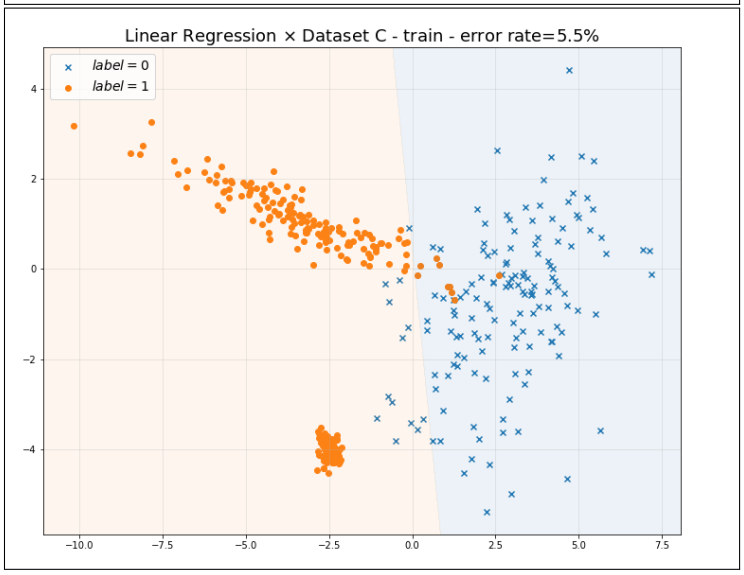
In this dataset, there is an overlapping area between the two classes making it harder to provide a separation and even harder with a linear model. Moreover, classes "scattering" being different, we also expect QDA to provide better performances which is the case.

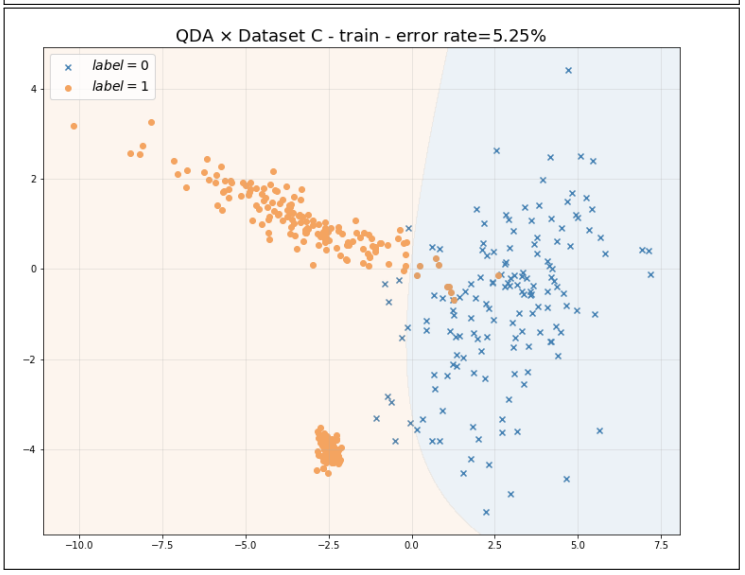
As classes distribution is homogeneous though, linear model still yield satisfying results.

Once again, missclassification is larger on the testing set than on the training. We note that the testing set is more than 6 times as large as the training set.









Missclassification on Dataset C (%)

	test	train
LDA	4.233	5.50
Linear Regression	4.233	5.50
Logistic Regression	2.267	4.00
QDA	3.833	5.25

Comments:

This dataset presents a dense and separate cluster of 1. It's interesting to see how linear regression tries to maximize the distance between the decision boundary and this cluster even if it means missclassifying other samples. Same goes for LDA and QDA as this dense cluster tends to shift the mean estimator for class 1. Logistic Regression for one seems to give it less importance in its weighting process.

Oddly enough, classification on testing set works better than on training set. The dense cluster being always completely well classified, the more samples we provide for classification, the more are gonna be in this cluster and the more their correct classification is going to pull the missclassification score downwards.

Learning in discrete graphic models

MLE of π

Let $(z_i)_{i\in[\![1,n]\!]}$ iid, the log-likelighood is given by :

$$\ell(\pi) = \sum_{m=1}^{M} n_m \log \pi_m \,, \;\; n_m = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=m\}}$$

$-\ell$ being convex and as $\exists \pi \in [0,1]^M / \pi^T 1_M = 1$, by Slater's constraints qualification we have strong duality and we can address its dual problem given by :

$$\max_{\lambda} \min_{\pi} \mathcal{L}(\lambda, \pi) \,, \;\; \text{where } \mathcal{L}(\lambda, \pi) = -\ell(\pi) + \lambda(\pi^T 1_M - 1)$$

$\mathcal{L}$ being convex w.r.t to π we can minimize it through its gradient :

$$\forall m, \; \frac{\partial \mathcal{L}}{\partial \pi_m} = 0 \Rightarrow -\frac{n_m}{\pi_m} + \lambda = 0 \Rightarrow \pi_m = \frac{n_m}{\lambda}$$

Plus, $\pi^T 1_M = 1 \Rightarrow \sum_{m=1}^{M} \frac{n_m}{\lambda} = 1 \Rightarrow \lambda = \sum_{m=1}^{M} n_m = n$, hence :

$$\forall m, \; \hat{\pi}_m = \frac{n_m}{n}$$

MLE of Θ

Let $(x_i)_{i\in[\![1,n]\!]}$ and $(z_i)_{i\in[\![1,n]\!]}$ iid and $\Theta = [\theta_{mk}] \in [0,1]^{M\times K}$.

Conditional probability allow us to write the log-likelihood as :

$$\begin{aligned}
\ell(\Theta, \pi) &= \sum_{i=1}^{n} \log \left(p_\Theta(x_i|y_i) p_\pi(y_i) \right) && \text{(Conditional probabilty)} \\
&= \sum_{i=1}^{n} \sum_{m=1}^{M} \log \pi_m \mathbb{1}_{\{y_i=m\}} + \sum_{i=1}^{n} \sum_{m=1}^{M} \sum_{k=1}^{K} \log \theta_{mk} \mathbb{1}_{\{x_i=k, y_I=m\}} \\
&= \sum_{m=1}^{M} n_m \log \pi_m + \sum_{k=1}^{K} \sum_{m=1}^{M} n_{mk} \log \theta_{mk}
\end{aligned}$$

where,

$$n_m = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=m\}} \,, \;\; n_{mk} = \sum_{i=1}^{n} \mathbb{1}_{\{z_i=m, x_i=k\}}$$

As log is concave, and $\forall m, k \;\; n_m \geq 0$ and $n_{mk} \geq 0$, $-\ell$ is convex.

Also, we can trivially find π_0 and Θ_0 satisfying the constraints given by : $\begin{cases} \pi^T \Theta 1_K = 1 \\ \pi^T 1_M = 1 \end{cases}$

By Slaters's constraints qualification, we hence have strong duality and can address its dual problem stated by :

$$\max_{\lambda \in \mathbb{R}_+^2} \min_{\Theta, \pi} \mathcal{L}(\lambda, \Theta, \pi)$$

where $\mathcal{L}(\lambda, \Theta, \pi) = -\ell(\Theta, \pi) + \begin{pmatrix} \pi^T \Theta 1_K - 1 & \pi^T 1_M - 1 \end{pmatrix} \lambda$

$\mathcal{L}$ being convex w.r.t to π and Θ we can minimize it through its gradient :

Derivating w.r.t to π, we obtain the same estimtor as previously : $\forall m, \; \hat{\pi}_m = \frac{n_m}{n}$

For Θ, the derivation goes :

$$\forall m,k,\ \frac{\partial \ell}{\partial \theta_{mk}} = \frac{n_{mk}}{\theta_{mk}}$$

And, $\pi^T \Theta 1_K = \mathrm{Tr}\left(\pi^T \Theta 1_K\right) = \mathrm{Tr}\left(\Theta 1_K \pi^T\right) =< \Theta, \pi 1_K^T > \Rightarrow \nabla_\Theta \left(\pi^T \Theta 1_K\right) = \pi 1_K^T$

$$\begin{aligned}\forall m,k,\ \frac{\partial \mathcal{L}}{\partial \theta_{mk}} = 0 \Rightarrow -\frac{n_{mk}}{\theta_{mk}} + \lambda_1 \pi_k = 0 \\ \Rightarrow \theta_{mk} = \frac{n_{mk}}{\lambda_1 \pi_m}\end{aligned}$$

Once again, the constraints gives us :
$\lambda_1 = n$, hence : $\forall m,k,\ \hat{\theta}_{mk} = \frac{n_{mk}}{n\hat{\pi}_m} = \frac{n_{mk}}{n_m}$

Linear classification

MLE for LDA

Hypothesis:

$$Y \sim \mathcal{B}(\pi),\ \ \forall j \in \{0,1\}\ X|\{Y=j\} \sim \mathcal{N}(\mu_j, \Sigma)$$

MLE of π : We computed in the previous part the MLE of a Multinomial law with parameter $\pi \in [0,1]^M$, $M \in \mathbb{N}^*$. A Bernoulli law is nothing more than a bidimensional Multinomial law, hence :
Let $(y_i)_{i\in[\![1,n]\!]}$ n observations,

$$\boxed{\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{y_i=1\}} = \frac{1}{n}\sum_{i=1}^{n} y_i}$$

MLE of μ_j , Σ :
Let $((x_i.y_i))_{i\in[\![1,n]\!]}$ a set of n iid observations
Then, if we note $\theta = (\mu_0, \mu_1, \Sigma)$

$$\begin{aligned}\ell(\theta) = \log p_\theta(x) &= \sum_{i=1}^{n} \log p_\theta(x_i) \\ &= \sum_{i=1}^{n} \log p_\theta(x_i|y_i) + \log p_\theta(y_i) \\ &= \sum_{i=1}^{n} y_i \left[\log \pi - \frac{1}{2}\left(d\log 2\pi + \log(\det \Sigma) + (x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)\right)\right] \\ &+ (1-y_i)\left[\log(1-\pi) - \frac{1}{2}\left(d\log 2\pi + \log(\det \Sigma) + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\right)\right]\end{aligned}$$

We remind that the MLE of the multivariate Gaussian model is given by :

$$\ell_{\mathcal{N}}(\mu,\Sigma) = \sum_{i=1}^{n} \underbrace{-\frac{1}{2}\left(d\log 2\pi + \log(\det\Sigma) + (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right)}_{\ell_{\mathcal{N}}^{(i)}(\mu,\Sigma)}$$

and that

$$\nabla_\mu \ell_{\mathcal{N}}^{(i)} = \Sigma^{-1}(x_i-\mu) \qquad \nabla_{\Sigma^{-1}}\ell_{\mathcal{N}}^{(i)} = \Sigma + (x_i-\mu)(x_i-\mu)^T$$

Let $j \in \{0,1\}$, ℓ being concave and differentiable w.r.t to μ_j we can maximize it by maximizing its gradient.

$$\begin{aligned}
\nabla_{\mu_j}\ell(\theta) = 0 &\Rightarrow \sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}\nabla_{\mu_j}\ell_{\mathcal{N}_j}^{(i)} = 0 \\
&\Rightarrow \sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}\Sigma^{-1}(x_i-\mu_j) = 0 \\
&\Rightarrow \Sigma^{-1}\left(\sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}(x_i-\mu_j)\right) = 0 \\
&\Rightarrow \sum_{i=1}^{n} \mathbb{1}_{\{y_i=j\}}(x_i-\mu_j) = 0 \qquad (\Sigma^{-1} \text{ injective})
\end{aligned}$$

Thus,

$$\boxed{\hat{\mu}_j = \frac{\sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}x_i}{\sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}}}$$

Samely,

$$\begin{aligned}
\nabla_{\Sigma^{-1}}\ell(\theta) = 0 &\Rightarrow \sum_{i=1}^{n} y_i\nabla_{\Sigma^{-1}}\ell_{\mathcal{N}_1}^{(i)} + (1-y_i)\nabla_{\Sigma^{-1}}\ell_{\mathcal{N}_0}^{(i)} = 0 \\
&\Rightarrow n\Sigma + \sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T + (1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T = 0
\end{aligned}$$

Thus,

$$\boxed{\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{y_i=0\}}(x_i-\hat{\mu}_0)(x_i-\hat{\mu}_0)^T + \mathbb{1}_{\{y_i=1\}}(x_i-\hat{\mu}_1)(x_i-\hat{\mu}_1)^T}$$

Decision boundary

$$
\begin{aligned}
p(y=1|x) =& \frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)+p(x|y=1)p(y=1)} \qquad (Bayes)\\
&= \frac{1}{1+\frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}\\
&= \frac{1}{1+\frac{1-\pi}{\pi}\exp\left(\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}\\
&= \frac{1}{1+\frac{1-\pi}{\pi}e^{-(w^Tx+b)}}\\
&= \sigma\left(\log\left(\frac{1-\pi}{\pi}\right)-(w^Tx+b)\right)
\end{aligned}
$$

with $w=\Sigma^{-1}(\mu_1-\mu_0)$ and $b=\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1$

The boundary decision is thus an affine boundary and we notice here that for $\pi=\frac{1}{2}$, it matches the boundary decision of a logistic regression.

Logitic Regression regularisation

By adding a ridge penalty in the logisitc regression, we get :

$$\mathcal{L}(\beta,w)=\ell(w)+\frac{\beta}{2}\|w\|^2$$

$$\nabla\mathcal{L}(\beta,w)=\nabla\ell(w)+\beta w$$

$$H\mathcal{L}(\beta,w)=H\ell(w)+\beta I$$

MLE for QDA

Let $\theta=(\mu_0.\mu_1,\Sigma_0,\Sigma_1)$, keeping the notation introduced for LDA's MLE computation, the log-likelihood is given by :

$$
\begin{aligned}
\ell(\theta) =& \sum_{i=1}^{n} y_i\left[\log\pi-\frac{1}{2}\left(d\log 2\pi+\log(\det\Sigma_1)+(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)\right)\right]\\
&+(1-y_i)\left[\log(1-\pi)-\frac{1}{2}\left(d\log 2\pi+\log(\det\Sigma_0)+(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)\right)\right]
\end{aligned}
$$

We can see here that this doesn't change anything for the maximisation w.r.t to μ_j and the MLE would be the same.
Regarding the covariance matrix, maximization goes :
$\forall j\in\{0,1\}$,

$$
\begin{aligned}
\forall j\in\{0,1\}\ \nabla_{\Sigma_j^{-1}}\ell(\theta)=0 &\Rightarrow \sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}\nabla_{\Sigma_j^{-1}}\ell_{\mathcal{N}_j}^{(i)}=0\\
&\Rightarrow \sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}\Sigma_j+\sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}(x_i-\mu_j)(x_i-\mu_j)^T=0
\end{aligned}
$$

Thus,

$$\boxed{\hat{\Sigma}_j=\frac{\sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}(x_i-\hat{\mu}_j)(x_i-\hat{\mu}_j)^T}{\sum_{i=1}^{n}\mathbb{1}_{\{y_i=j\}}}}$$

And any value in the decision boundary satisfies :

$$
\begin{aligned}
p(x|y=1) = \frac{1}{2} = p(x|y=0) &\Leftrightarrow \pi\mathcal{N}(x|\mu_1, \Sigma_1) = (1-\pi)\mathcal{N}(x|\mu_0, \Sigma_0) \\
&\Leftrightarrow \log\left(\frac{\pi}{1-\pi}\right) = \log\mathcal{N}(x|\mu_0, \Sigma_0) - \log\mathcal{N}(x|\mu_1, \Sigma_1) \\
&\Leftrightarrow \frac{1}{2}\log\left(\frac{\det\Sigma_1^{-1}}{\det\Sigma_0^{-1}}\right) + \frac{1}{2}\left[(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right] = \log\left(\frac{\pi}{1-\pi}\right)
\end{aligned}
$$

Once developped, such relationship boils down to a conic equation that can be plotted.