Graphical models homework 1

Léo Alberge, Romain Petit

January 9, 2019

Exercise 1 : learning in discrete graphical models

$$l(\boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \sum_{m=1}^{M} x_i^k z_i^m log(\theta_{m,k}) + \sum_{m=1}^{M} z_i^m log(\pi_m) \right] \quad \boxed{\hat{\pi}_m = \frac{N_m}{n} = \frac{1}{n} \sum_{i=1}^{n} z_i^m \quad \hat{\theta}_{m,k} = \frac{N_{m,k}}{N_m} = \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\sum_{i=1}^{n} z_i^m}}$$

Exercise 2.1.(a) : LDA formulas

$$l(\theta, \mu_0, \mu_1, \Sigma) = -\frac{nd}{2} log(2\pi) + N_y \, log(\theta) + (n - N_y) \, log(1-\theta) - \frac{n}{2} log(|\Sigma|) - \frac{n}{2} \mathrm{Tr} \left(\Sigma^{-1} \left(\frac{n - N_y}{n} \, \tilde{\Sigma}_0(\mu_0) + \frac{N_y}{n} \, \tilde{\Sigma}_1(\mu_1) \right) \right)$$

with $N_y = \sum_{i=1}^{n} y_i$ and $\tilde{\Sigma}_0(\mu_0) = \frac{1}{n - N_y} \sum_{\substack{i=1 \\ y_i=0}}^{n} (x_i - \mu_0)(x_i - \mu_0)^T$ and $\tilde{\Sigma}_1(\mu_1) = \frac{1}{N_y} \sum_{\substack{i=1 \\ y_i=1}}^{n} (x_i - \mu_1)(x_i - \mu_1)^T$

$$\boxed{\hat{\theta} = \frac{N_y}{n} \quad \hat{\mu}_0 = \frac{1}{n - N_y} \sum_{\substack{i=1 \\ y_i=0}}^{n} x_i \quad \hat{\mu}_1 = \frac{1}{N_y} \sum_{\substack{i=1 \\ y_i=1}}^{n} x_i \quad \hat{\Sigma} = \frac{n - N_y}{n} \tilde{\Sigma}_0(\hat{\mu}_0) + \frac{N_y}{n} \tilde{\Sigma}_1(\hat{\mu}_1)}$$

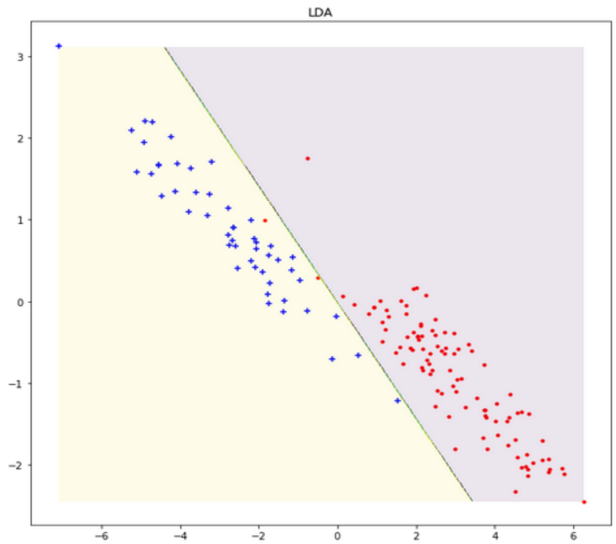
Exercise 2.5.(a) : QDA formulas

$$\begin{aligned} l(\theta, \mu_0, \mu_1, \Sigma) = -\frac{nd}{2} log(2\pi) + N_y \, log(\theta) + (n - N_y) \, log(1 - \theta) - \frac{n}{2} \left(\frac{n - N_y}{n} log(|\Sigma_0|) + \frac{N_y}{n} log(|\Sigma_1|) \right) \\ - \frac{n}{2} \mathrm{Tr} \left(\frac{n - N_y}{n} \Sigma_0^{-1} \tilde{\Sigma}_0(\mu_0) + \frac{N_y}{n} \Sigma_1^{-1} \tilde{\Sigma}_1(\mu_1) \right) \end{aligned}$$

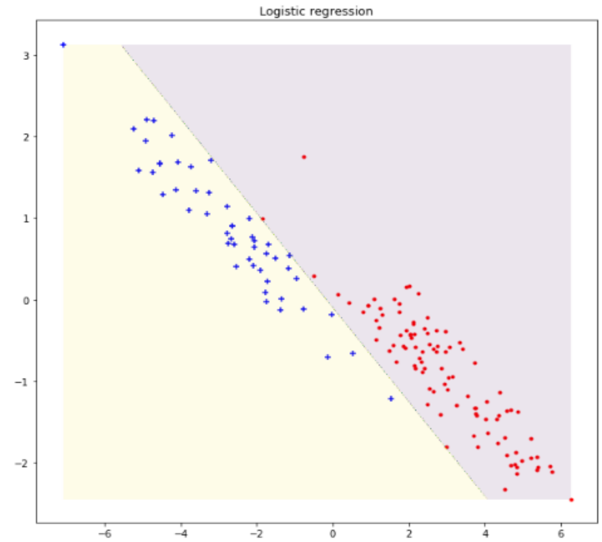
with $N_y = \sum_{i=1}^{n} y_i$ and $\tilde{\Sigma}_0(\mu_0) = \frac{1}{n - N_y} \sum_{\substack{i=1 \\ y_i=0}}^{n} (x_i - \mu_0)(x_i - \mu_0)^T$ and $\tilde{\Sigma}_1(\mu_1) = \frac{1}{N_y} \sum_{\substack{i=1 \\ y_i=1}}^{n} (x_i - \mu_1)(x_i - \mu_1)^T$

$$\boxed{\hat{\theta} = \frac{N_y}{n} \quad \hat{\mu}_0 = \frac{1}{n - N_y} \sum_{\substack{i=1 \\ y_i=0}}^{n} x_i \quad \hat{\mu}_1 = \frac{1}{N_y} \sum_{\substack{i=1 \\ y_i=1}}^{n} x_i \quad \hat{\Sigma}_0 = \tilde{\Sigma}_0(\hat{\mu}_0) \quad \hat{\Sigma}_1 = \tilde{\Sigma}_1(\hat{\mu}_1)}$$

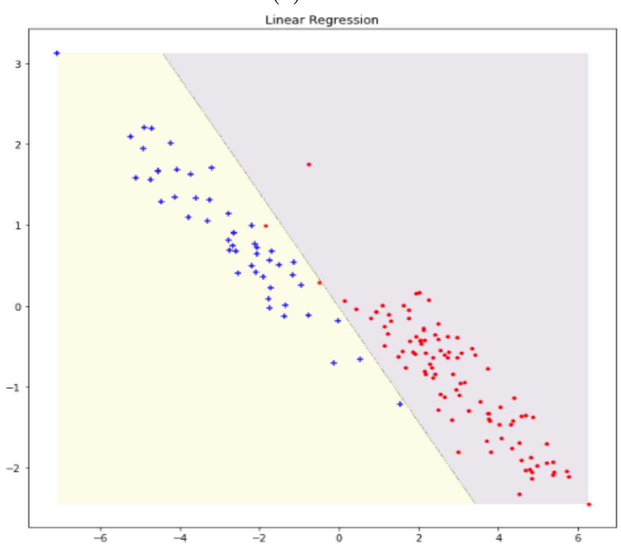
Dataset A



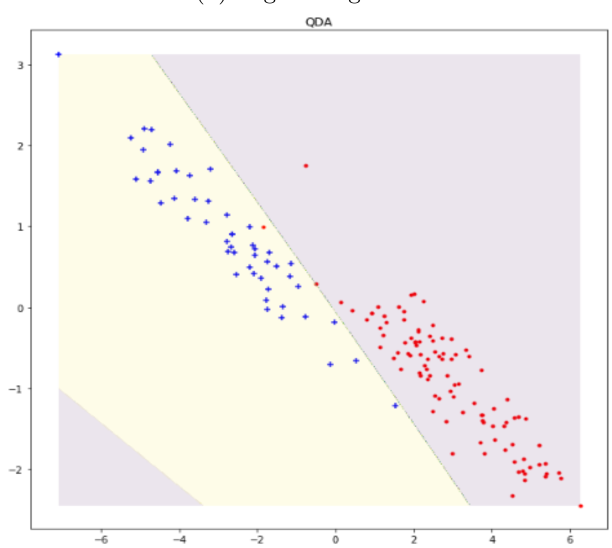
(a) LDA



(b) Logistic regression



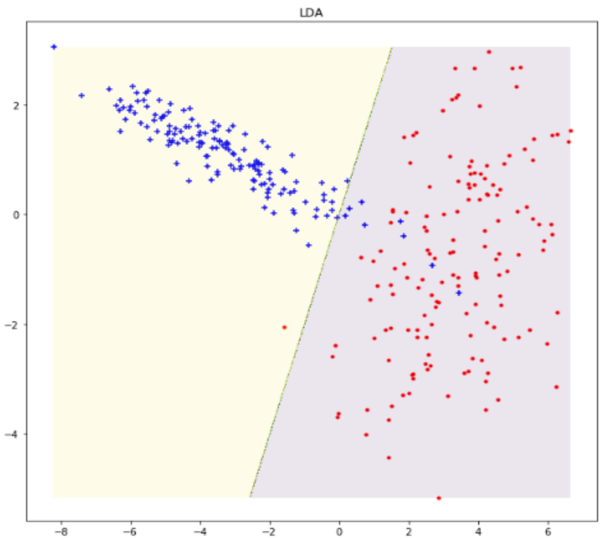
(c) Linear regression



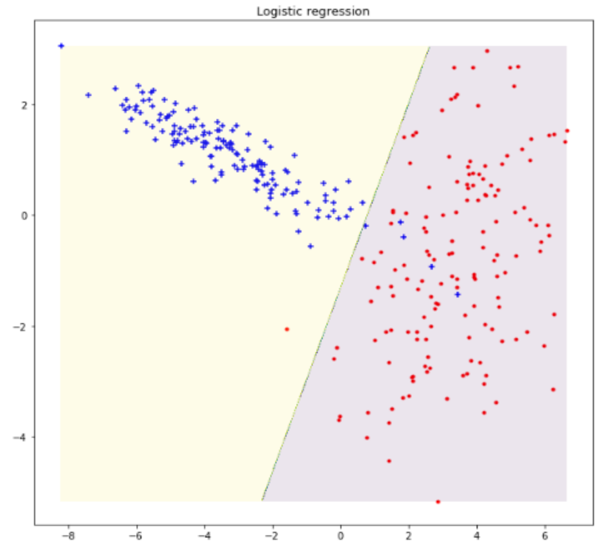
(d) QDA

Table of errors			Comments
Model	Training error	Test error	QDA and LDA models seem to perform well because the a priori on the data seem to be "valid" i.e. the data are normally distributed with the same covariance matrix. We can notice some overfitting for the logistic regression.
Linear reg.	1.3%	2.1%	
Log reg.	0%	3.4%	
LDA	1.3%	2%	
QDA	0.7%	2%	

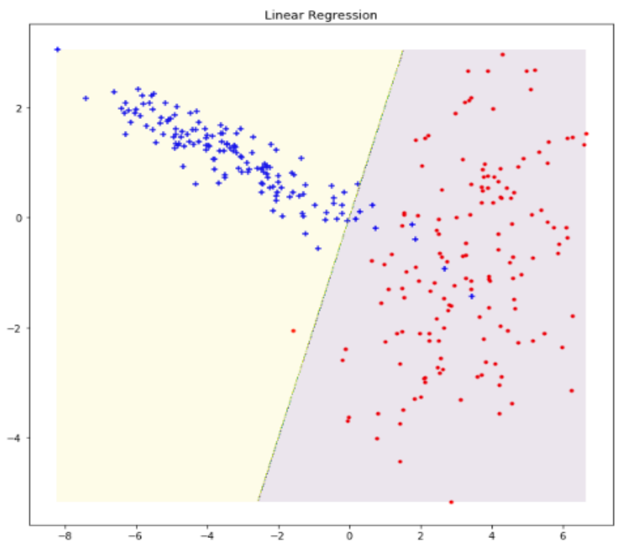
Dataset B



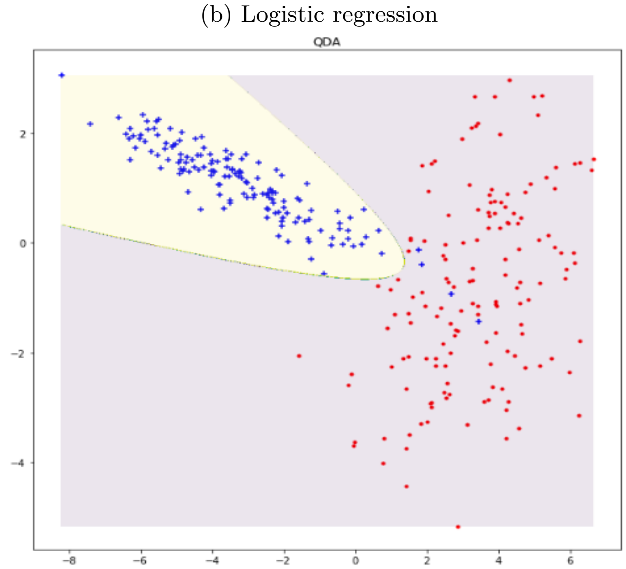
(a) LDA



(b) Logistic regression



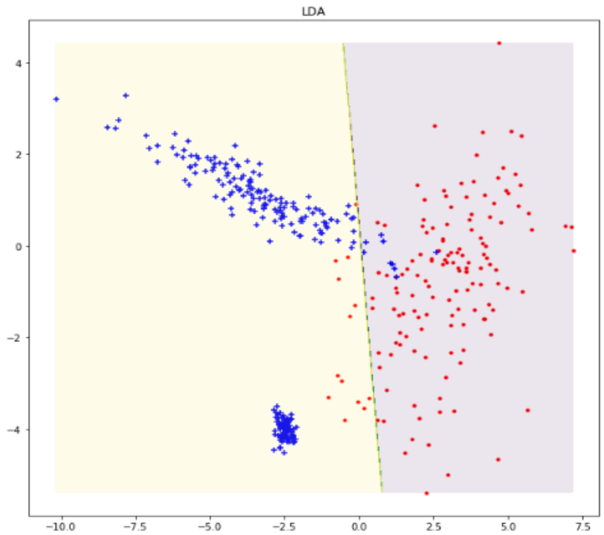
(c) Linear regression



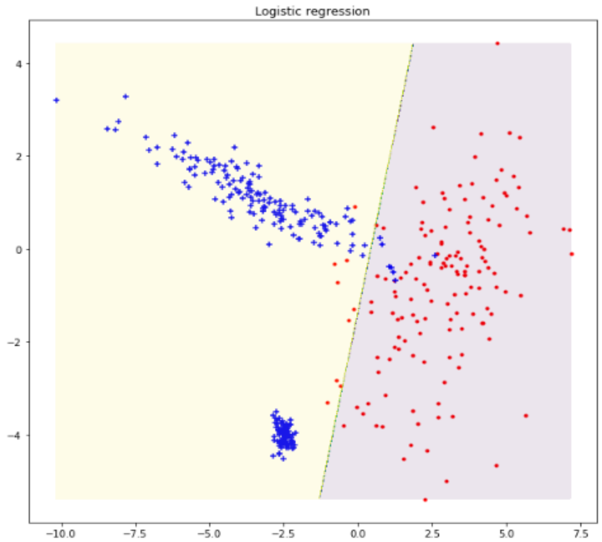
(d) QDA

Table of errors			Comments
Model	Training error	Test error	Here even if each class seems to be normally distributed, covariance matrices are no longer similar. It is therefore not surprising that QDA outperforms LDA.
Linear reg.	3%	4.2%	
Log reg.	2%	4.3%	
LDA	3%	4.2%	
QDA	1.3%	2%	

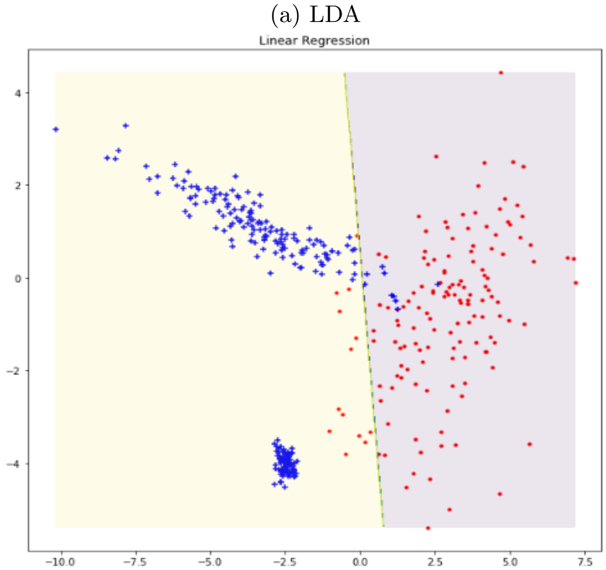
Dataset C



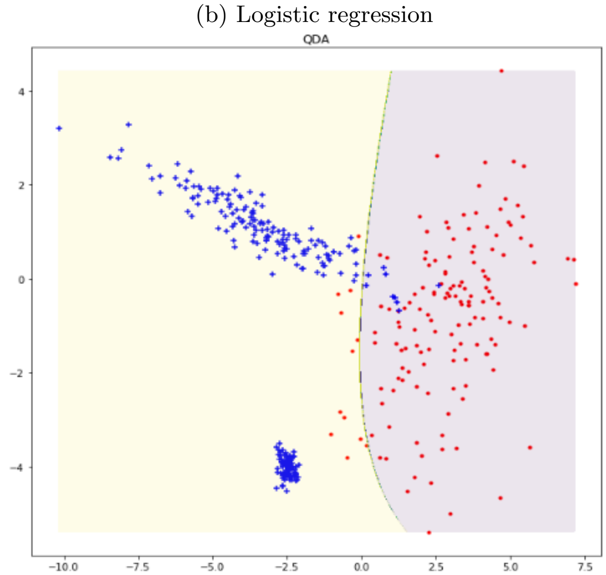
(a) LDA



(b) Logistic regression



(c) Linear regression



(d) QDA

Table of errors		
Model	Training error	Test error
Linear reg.	5.5%	4.2%
Log reg.	4%	2.3%
LDA	5.5%	4.2%
QDA	5.3%	3.8%

Comments

Here, the gaussian assumption is no longer valid. Since we do not make any gaussian assumption in the logistic regression model, it is not surprising that it outperforms the other models. On this dataset only, the test error is smaller than the training error for all models. This is surprising, and could be caused by the data generation process. It should be noted that on all datasets, LDA and linear regression yield the exact same results. This suggests a strong link between the two methods in the binary classification setting.

Detailed proofs

Exercise 1 : learning in discrete graphical models

Let us denote $\mathbf{x}_i \in \{0,1\}^K$ and $\mathbf{z}_i \in \{0,1\}^M$ the vectors representing realizations of the discrete random variables X and Z. The log-likelihood of the model writes :

$$l(\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{i=1}^{n} log(p_{\boldsymbol{\theta}}(\mathbf{x}_i|\mathbf{z}_i)) + log(p_{\boldsymbol{\pi}}(\mathbf{z_i})) = \sum_{i=1}^{n}\left[\sum_{k=1}^{K}\sum_{m=1}^{M} log(\theta_{m,k}^{x_i^k z_i^m}) + \sum_{m=1}^{M} log(\pi_m^{z_i^m})\right]$$
$$l(\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[\sum_{k=1}^{K}\sum_{m=1}^{M} x_i^k z_i^m log(\theta_{m,k}) + \sum_{m=1}^{M} z_i^m log(\pi_m)\right]$$

The maximum likelihood problem is therefore :

$$\begin{aligned} \min_{\boldsymbol{\theta},\boldsymbol{\pi}\ \geq 0} \quad & -l(\boldsymbol{\pi},\boldsymbol{\theta}) \\ \text{s.t.} \quad & \mathbf{1}^T\boldsymbol{\pi} = 1 \text{ and } \mathbf{1}^T\boldsymbol{\theta_m} = 1 \text{ for all } m \end{aligned} \tag{1}$$

If for a certain m, $z_i^m = 0$ for all i, then at the optimum we necessarily have $\pi_m = 0$. The maximum likelihood problem is therefore equivalent to solving problem (1) after setting π_m to 0. We can therefore restrict ourselves to the case where for all m, there is at least one i such that $z_i^m > 0$. The same reasoning can be applied to restrict ourselves to the case where for all k, there is at least one i such that $x_i^k > 0$.

In this case, $-l$ is equal to $+\infty$ on the boundary of $\{\boldsymbol{\pi},\boldsymbol{\theta} \geq 0,\ \mathbf{1}^T\boldsymbol{\pi} = 1 \text{ and } \mathbf{1}^T\boldsymbol{\theta_m} = 1 \text{ for all } m\}$. We therefore have a convex minimization problem with the objective differentiable on the domain, and Slater's condition is satisfied. A feasible point $(\boldsymbol{\pi},\boldsymbol{\theta})$ is therefore optimal if and only if there exist $\boldsymbol{\nu}$ such that $(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\nu})$ satisfies Karush Kuhn Tucker conditions (we consider the positivity constraints implicit).

$$\mathcal{L}(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\nu}) = -l(\boldsymbol{\pi},\boldsymbol{\theta})\ + \nu_0\ (\sum_{m=1}^{M}\pi_m - 1) + \sum_{m=1}^{M}\nu_m\ (\sum_{k=1}^{K}\theta_{m,k} - 1)$$
$$\frac{\partial\mathcal{L}}{\partial\pi_m}(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\nu}) = -\sum_{i=1}^{n}\frac{z_i^m}{\pi_m} + \nu_0 \qquad \frac{\partial\mathcal{L}}{\partial\theta_{m,k}}(\boldsymbol{\pi},\boldsymbol{\theta},\boldsymbol{\nu}) = -\sum_{i=1}^{n}\frac{x_i^k z_i^m}{\theta_{m,k}} + \nu_m$$

The first order condition therefore writes $\hat{\nu}_0\ \hat{\pi}_m = \sum_{i=1}^{n} z_i^m \stackrel{def}{=} N_m$ and $\hat{\theta}_{m,k}\ \hat{\nu}_m = \sum_{i=1}^{n} x_i^k z_i^m \stackrel{def}{=} N_{m,k}$.

This gives $\hat{\nu}_0 = n$, $\hat{\nu}_m = N_m$, $\hat{\pi}_m = \frac{N_m}{n}$, and $\hat{\theta}_{m,k} = \frac{N_{m,k}}{N_m}$. Since $(\boldsymbol{\pi},\boldsymbol{\theta})$ is feasible for the primal, we have that the maximum likelihood estimator is unique and given by :

$$\boxed{\hat{\pi}_m = \frac{N_m}{n} = \frac{1}{n}\sum_{i=1}^{n} z_i^m \qquad \hat{\theta}_{m,k} = \frac{N_{m,k}}{N_m} = \frac{\sum_{i=1}^{n} x_i^k z_i^m}{\sum_{i=1}^{n} z_i^m}}$$

Exercise 2.1.(a) : Generative model (LDA)

⚠ In this section, we replaced the π parameter of the Bernoulli law by θ to avoid confusion with the constant π appearing in likelihood computations.

We have that : $p(y) = \theta^y(1-\theta)^{1-y}$
$$p(x|y) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} exp(-\frac{1}{2}((1-y)(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + y(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)))$$

The log-likelihood therefore writes :

$$l(\theta,\mu_0,\mu_1,\Sigma) = -\frac{nd}{2}log(2\pi)+N_y\,log(\theta)+(n-N_y)\,log(1-\theta)-\frac{n}{2}log(|\Sigma|)-\frac{n}{2}\text{Tr}\left(\Sigma^{-1}\left(\frac{n-N_y}{n}\;\tilde{\Sigma}_0(\mu_0)+\frac{N_y}{n}\;\tilde{\Sigma}_1(\mu_1)\right)\right)$$

$$\text{with } N_y = \sum_{i=1}^{n} y_i \text{ and } \tilde{\Sigma}_0(\mu_0) = \frac{1}{n-N_y}\sum_{\substack{i=1\\y_i=0}}^{n}(x_i-\mu_0)(x_i-\mu_0)^T \text{ and } \tilde{\Sigma}_1(\mu_1) = \frac{1}{N_y}\sum_{\substack{i=1\\y_i=1}}^{n}(x_i-\mu_1)(x_i-\mu_1)^T$$

If for all i, $y_i = 0$, then for all (μ_0,μ_1,Σ) the function $-l(\bullet,\mu_0,\mu_1,\Sigma)$ is increasing and its minimum is reached at $\hat{\theta}=0$. Similarly, if for all i, $y_i=1$, $-l(\bullet,\mu_0,\mu_1,\Sigma)$ is decreasing and its minimum is reached at $\hat{\theta}=1$. In these cases, we are therefore left with a maximum likelihood problem for the classical multivariate gaussian model. We will therefore restrict ourselves to the case where there exsits i such that $y_i \neq 0$.

In this case, $\lim\limits_{\theta\to 0^+} -l(\theta,\mu_0,\mu_1,\Sigma) = \lim\limits_{\theta\to 1^-} -l(\theta,\mu_0,,\mu_1,\Sigma) = +\infty$. The domain of our objective is therefore $]0,1[\;\times\mathbb{R}^d\times\mathbb{R}^d\times S_n^{++}(\mathbb{R})$.

$-l(\bullet,\mu_0,\mu_1,\Sigma)$ is a differentiable convex function with a single stationary point $\theta=\frac{N_y}{n}$. Its minimum is therefore reached only at $\hat{\theta}=\frac{N_y}{n}$, which does not depend on (μ_0,μ_1,Σ).

$-l(\theta,\bullet,\mu_1,\Sigma)$ is a differentiable convex function whose gradient is $\sum\limits_{\substack{i=1\\y_i=0}}^{n}\Sigma^{-1}(\mu_0-x_i)$. It therefore has a single stationary point, and therefore reaches its minimum at $\hat{\mu}_0=\frac{1}{n-N_y}\sum\limits_{\substack{i=1\\y_i=0}}^{n}x_i$, which does not depend on (θ,μ_1,Σ). Similarly, $\hat{\mu}_1=\frac{1}{N_y}\sum\limits_{\substack{i=1\\y_i=1}}^{n}x_i$.

Our maximum likelihood problem is therefore equivalent to minimizing $-l(\hat{\theta},\hat{\mu}_0,\hat{\mu}_1,\bullet)$ over $S_n^{++}(\mathbb{R})$, which is equivalent to minimizing $-log(det(\Lambda))+Tr(\Lambda\tilde{\Sigma})$ over $S_n^{++}(\mathbb{R})$ and then let $\hat{\Sigma}=\hat{\Lambda}^{-1}$ (with $\tilde{\Sigma}=\frac{n-N_y}{n}\tilde{\Sigma}_0(\hat{\mu}_0)+\frac{N_y}{n}\tilde{\Sigma}_1(\hat{\mu}_1)$). $A\mapsto log(det(A))$ is concave and $A\mapsto Tr(A\tilde{\Sigma})$ is linear. We therefore have a convex optimization problem, with the gradient of the objective equal to $-\Lambda^{-1}+\tilde{\Sigma}$. If $\tilde{\Sigma}\in S_n^{++}(\mathbb{R})$, then the objective have a unique stationary point which is also the optimum, and $\hat{\Sigma}=\tilde{\Sigma}$. Otherwise, we can show that the objective is unbounded below, by taking Λ diagonal with its coefficients equal to 1 except a $\lambda>0$ at the index for which we have a zero eigen value for $\tilde{\Sigma}$. We therefore have $Tr(\Lambda\tilde{\Sigma})=Tr(\tilde{\Sigma})$ and $log(det(\Lambda))=log(\lambda)$. By letting λ go to plus infinity, we therefore have that the objective is unbounded below.

We finally get (in the case where $\tilde{\Sigma}\in S_n^{++}(\mathbb{R})$, and $0<N_y<n$) :

$$\boxed{\hat{\theta}=\frac{N_y}{n}\qquad \hat{\mu}_0=\frac{1}{n-N_y}\sum_{\substack{i=1\\y_i=0}}^{n}x_i\qquad \hat{\mu}_1=\frac{1}{N_y}\sum_{\substack{i=1\\y_i=1}}^{n}x_i\qquad \hat{\Sigma}=\frac{n-N_y}{n}\tilde{\Sigma}_0(\hat{\mu}_0)+\frac{N_y}{n}\tilde{\Sigma}_1(\hat{\mu}_1)}$$

In the case $N_y=0$, the value of μ_1 does not have an impact on the likelihood, and $\hat{\Sigma}=\tilde{\Sigma}_0(\hat{\mu}_0)$, with $\tilde{\Sigma}_0$ and $\hat{\mu}_0$ given by the formulas above. For $N_y=n$, we also recover the same kind of estimators, which can be seen as a limit of the ones found in the general case.

Exercise 2.5.(a) : QDA model

⚠ In this section, we replaced the π parameter of the Bernoulli law by θ to avoid confusion with the constant π appearing in likelihood computations.

We have that : $p(y) = \theta^y(1-\theta)^{1-y} = \frac{1}{2}$
$p(x|y) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_0|^{1-y}|\Sigma_1|^y}} exp(-\frac{1}{2}((1-y)(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + y(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)))$

The log-likelihood therefore writes :

$$l(\theta,\mu_0,\mu_1,\Sigma) = -\frac{nd}{2}log(2\pi) + N_y\ log(\theta) + (n-N_y)\ log(1-\theta) - \frac{n}{2}\left(\frac{n-N_y}{n}log(|\Sigma_0|) + \frac{N_y}{n}log(|\Sigma_1|)\right)$$
$$-\frac{n}{2}\mathrm{Tr}\left(\frac{n-N_y}{n}\Sigma_0^{-1}\tilde{\Sigma}_0(\mu_0) + \frac{N_y}{n}\Sigma_1^{-1}\tilde{\Sigma}_1(\mu_1)\right)$$

with $N_y = \sum_{i=1}^n y_i$ and $\tilde{\Sigma}_0(\mu_0) = \frac{1}{n-N_y}\sum_{\substack{i=1\\y_i=0}}^n (x_i-\mu_0)(x_i-\mu_0)^T$ and $\tilde{\Sigma}_1(\mu_1) = \frac{1}{N_y}\sum_{\substack{i=1\\y_i=1}}^n (x_i-\mu_1)(x_i-\mu_1)^T$

By the same reasoning than in the previous section, we have values of θ, μ_0 and μ_1 at the optimal identical to those for the LDA model, and are therefore left with the problem of minimizing

$$f(\Sigma_0,\Sigma_1) = \frac{n}{2}\left(\frac{n-N_y}{n}log(|\Sigma_0|) + \frac{N_y}{n}log(|\Sigma_1|)\right)\frac{n}{2}\mathrm{Tr}\left(\frac{n-N_y}{n}\Sigma_0^{-1}\tilde{\Sigma}_0(\mu_0) + \frac{N_y}{n}\Sigma_1^{-1}\tilde{\Sigma}_1(\mu_1)\right)$$

over $(S_n^{++}(\mathbb{R}))^2$. f is actually the sum of two independant functions of Σ_0 and Σ_1. The same reasoning than in the case of LDA can therefore be applied to find $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$.

$$\boxed{\hat{\theta} = \frac{N_y}{n} \quad \hat{\mu}_0 = \frac{1}{n-N_y}\sum_{\substack{i=1\\y_i=0}}^n x_i \quad \hat{\mu}_1 = \frac{1}{N_y}\sum_{\substack{i=1\\y_i=1}}^n x_i \quad \hat{\Sigma}_0 = \tilde{\Sigma}_0(\hat{\mu}_0) \quad \hat{\Sigma}_1 = \tilde{\Sigma}_1(\hat{\mu}_1)}$$

Decision boundaries

In this section, we detail the classification boundary's equations for the different models. In our case, we only have two classes so the decision boundary is given by the equation : $p(y=1|x) = p(y=0|x) = \frac{1}{2}$

LDA

For the LDA model, $p(y|x) \propto \theta^y(1-\theta)^{(1-y)}\ exp(-\frac{1}{2}((1-y)(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + y(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)))$
Therefore, the decision boundary is the line :

$$\boxed{x^T\Sigma^{-1}(\mu_0-\mu_1) = log(\frac{\theta}{1-\theta}) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1)}$$

We also have : $p(y=1|x) = \frac{\theta exp((\Sigma^{-1}\mu_1)^Tx - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1)}{\theta(exp(\Sigma^{-1}\mu_1)^Tx - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1) + (1-\theta)exp((\Sigma^{-1}\mu_0)^Tx - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0)}$.

Which shows that :

$$\boxed{p(y=1|x) \;=\; \sigma((\mu_0-\mu_1)^T\Sigma^{-1}x + log(\frac{1-\theta}{\theta}) - \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1))}$$

with σ the sigmoid function. We therefore have the same kind of expression than in the case of logistic regression.

Linear regression

For linear regression the decision boundary is the line :

$$\boxed{w^Tx + b = \frac{1}{2}}$$

Logistic regression

For logistic regression $p(y=1|x) = \frac{e^{w^Tx+b}}{1+e^{w^Tx+b}}$. The decision boundary is therefore the line :

$$\boxed{w^Tx + b = 0}$$

QDA

For the QDA model, we have that :

$p(y|x) \propto \theta^y(1-\theta)^{(1-y)}\frac{1}{\sqrt{|\Sigma_0|^y|\Sigma_1|^{1-y}}}exp(-\frac{1}{2}((1-y)(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + y(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)))$

The decision boundary is therefore quadratic :

$$\boxed{\frac{1}{2}x^T(\Sigma_0^{-1}+\Sigma_1^{-1})x + x^T(\Sigma_0^{-1}\mu_0 - \Sigma_1^{-1}\mu_1) = log(\frac{\theta\sqrt{|\Sigma_0|}}{(1-\theta)\sqrt{|\Sigma_1|}}) + \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1)}$$