Probabilistic Graphical Models
Homework 1

Antonin Berthon

October 2018

1 Exercice 1: learning in discrete graphical model

Let z and x be two discrete random variables such that $p(z=m)=\pi_m$ and $p(x=k|z=m)=\theta_{k,m}$ with $k \in [\![1,K]\!]$, $m \in [\![1,M]\!]$. The maximization of the log-likelihood (see Appendix) gives :

$$\forall m \in [\![1,M]\!], \boxed{\hat{\pi}_m = \frac{n(z)_m}{n}}$$

$$\forall m \in [\![1,M]\!], k \in [\![1,K]\!], \boxed{\hat{\theta}_{m,k} = \frac{n(z,x)_{m,k}}{n(z)_m}}$$

where $n(z)_m = \sum\limits_{i=1}^{n} \mathbb{1}_{(z^i=m)}$ and $n(z,x)_{m,k} = \sum\limits_{i=1}^{n} \mathbb{1}_{(z^i=m)}\mathbb{1}_{(x^i=k)}$.

2 Exercice 2.1(a): LDA

We consider the following model:

$$y \sim \mathcal{B}(\pi) \text{ and } x|y=j \sim \mathcal{N}(\mu_j, \Sigma)$$

The maximization of the log likelihood yields :

$$\boxed{\hat{\pi} = \frac{\sum\limits_{i=1}^{n} \mathbb{1}_{(y_i=1)}}{n}}, \quad \boxed{\hat{\mu_j} = \frac{\sum\limits_{i=1}^{n} \mathbb{1}_{(y_i=j)} x_i}{\sum\limits_{i=1}^{n} \mathbb{1}_{(y_i=j)}}} \text{ for } j=0,1, \quad \boxed{\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0,1} \mathbb{1}_{(y_i=j)} (x_i - \hat{\mu_j})^T (x_i - \hat{\mu_j})}$$

Also, $p(y=1|x)$ is analogous to a logistic regression : $p(y=1|x) = \sigma(b^T x + a)$ with:

$$\boxed{a = \log \frac{\pi}{1-\pi} - \frac{1}{2}\Big(\mu_1^T \Sigma \mu_1 - \mu_0^T \Sigma \mu_0\Big)}, \quad \boxed{b = {\Sigma^{-1}}^T (\mu_0 - \mu_1)}$$

3 Exercice 2.5(a): QDA

We now consider different covariance matrix for each class:

$$y \sim \mathcal{B}(\pi) \text{ and } x|y=j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

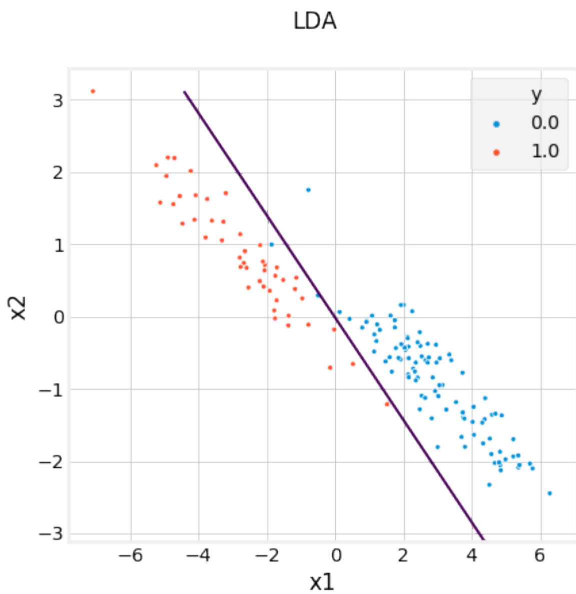
The MLE estimators are now:

$$\boxed{\hat{\pi} = \frac{\sum\limits_{i=1}^{n} \mathbb{1}_{(y_i=1)}}{n}}, \quad \boxed{\hat{\mu_j} = \frac{\sum\limits_{i=1}^{n} \mathbb{1}_{(y_i=j)} x_i}{\sum\limits_{i=1}^{n} \mathbb{1}_{y_i=j}}} \text{ for } j=0,1, \quad \boxed{\hat{\Sigma_j} = \frac{\sum\limits_{i|y_i=j} (x_i - \hat{\mu_j})^T (x_i - \hat{\mu_j})}{\sum\limits_{i=1}^{n} \mathbb{1}_{y_i=j}}} \text{ for } j=0,1$$

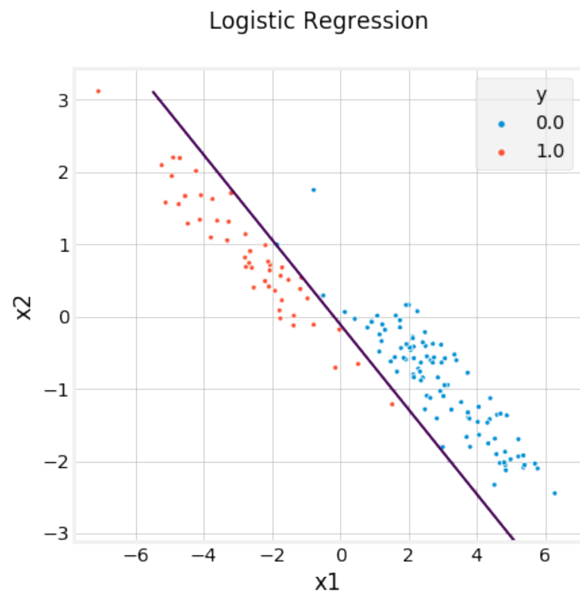
Also we have $p(y=1|x) = \sigma(-\frac{1}{2}x^T M x + b^T x + a)$ with

$$\boxed{M = \Sigma_1^{-1} - \Sigma_0^{-1}}, \quad \boxed{b^T = \mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1}}, \quad \boxed{a = \log \frac{\pi \sqrt{\det \Sigma_0}}{(1-\pi)\sqrt{\det \Sigma_1}} - \frac{1}{2}\Big(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0\Big)}$$

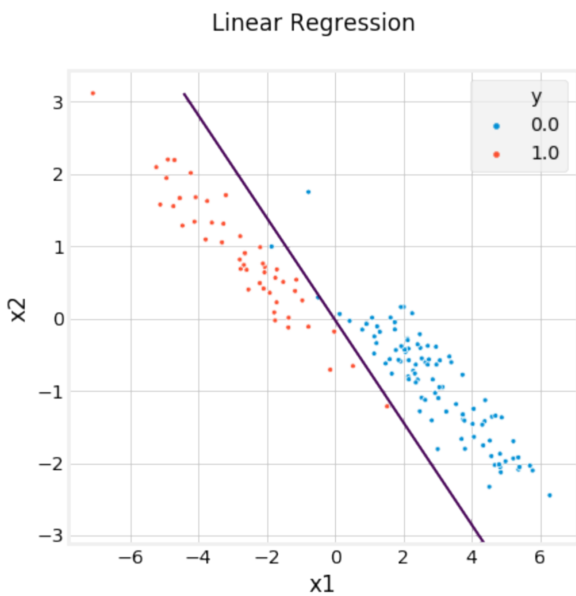
4 Dataset A



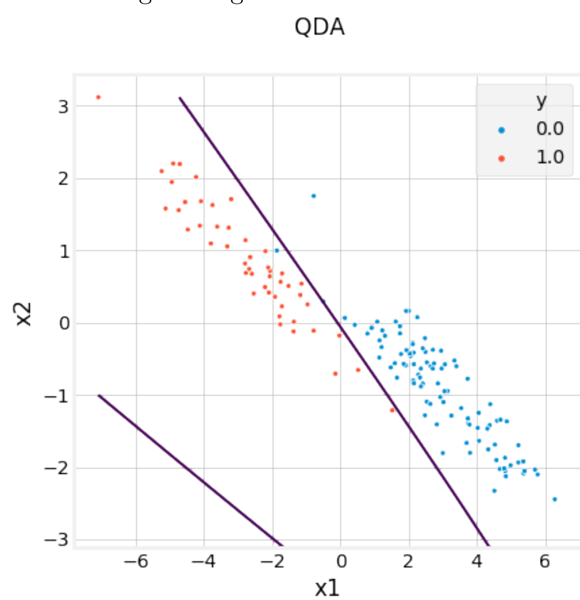
LDA for dataset A



Logistic regression for dataset A



Linear regression for dataset A



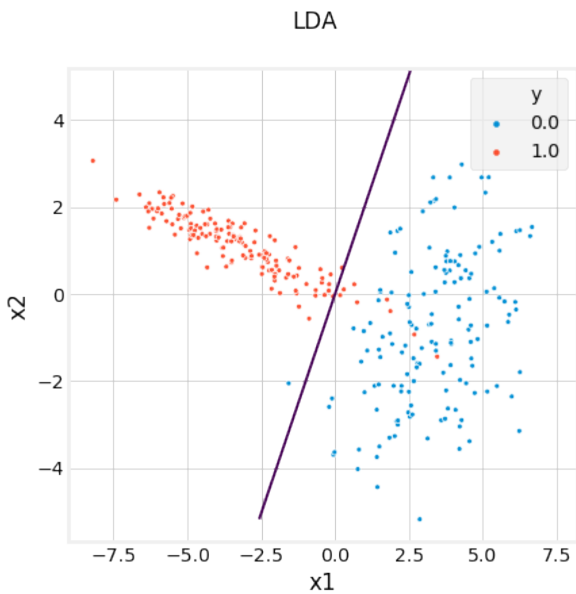
QDA for dataset A

Classification error (in %)		
Model	Train	Test
LDA	1.33	2.00
Logistic reg.[a]	0.00	3.53
Linear reg.	1.33	2.07
QDA	0.67	2.00

[a]The logistic regression is compute using the IRLS algorithm with an error term $= 10^{-3}$.

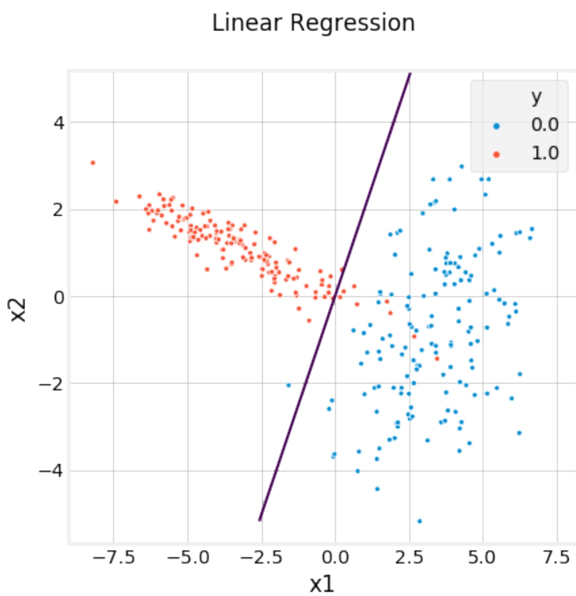
- All four methods perform better on the training data than on the test data.
- The logistic regression seems to overfit the data (very low training error but relatively high test error). To avoid this we can add a regularization term.
- The QDA seems to overfit the data as well since it assumes a more complex distribution compared to the actual one.
- The LDA and linear methods have very similar performances.

5 Dataset B

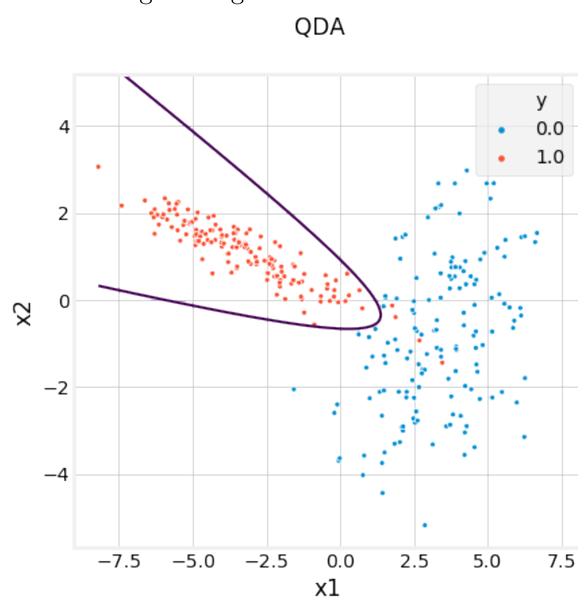


LDA for dataset B

Logistic regression for dataset B



Linear regression for dataset B

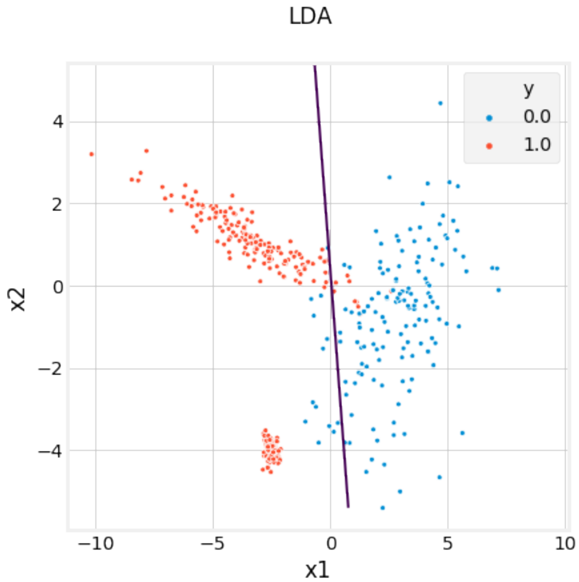


QDA for dataset B

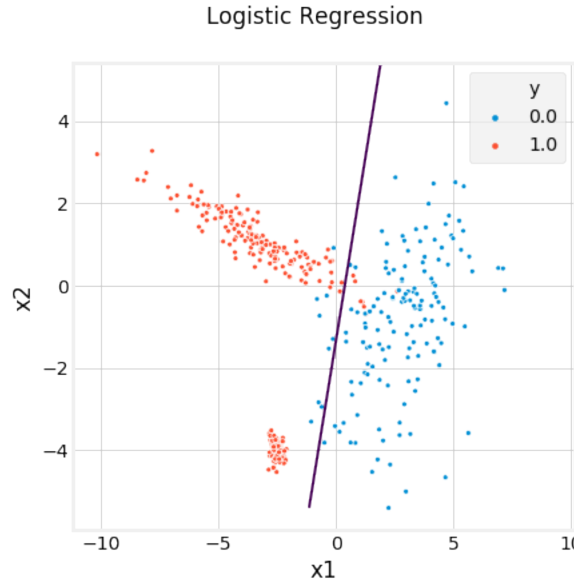
Classification error (in %)		
Model	Train	Test
LDA	3.00	4.15
Logistic reg.	2.00	4.25
Linear reg.	3.00	4.15
QDA	1.33	2.00

- Again, all four methods perform better on the training data than on the test data.
- The QDA performs better than the LDA since the hypothesis of equal variance do not hold here.
- Generative modelling with QDA outperforms the logistic regression in this context where the assumptions it makes hold and there is relatively few data to learn from.
- The LDA and Linear Regression perform similarly.

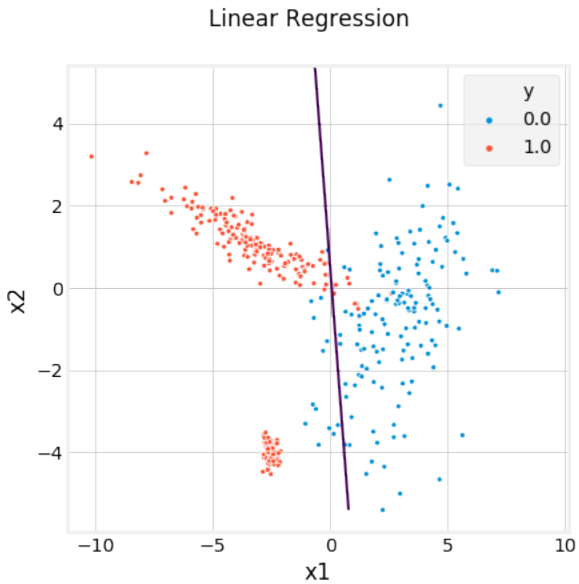
6 Dataset C



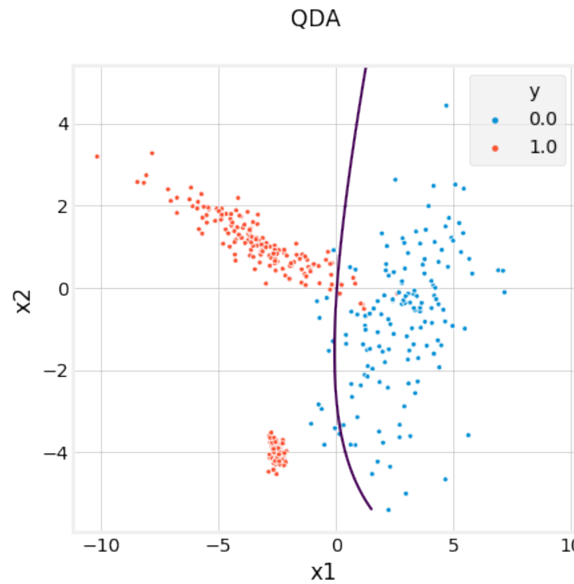
LDA for dataset C



Logistic regression for dataset C



Linear regression for dataset C



QDA for dataset C

Classification error (in %)		
Model	Train	Test
LDA	5.50	4.23
Logistic reg.	4.00	2.30
Linear reg.	5.50	4.23
QDA	5.25	3.83

- Here, all four methods performs better on the test data than on the training data. While this is surprising, it is probably due to a different proportion of data points found in the small bottom cluster that are easier to classify (1/4 of data points in training data vs. 1/3 in test data).
- As for the other data sets, the LDA and Linear method have identical performances which shows the similarity in their underlying assumptions.
- The Logistic Regression performs better than the other methods since it is the only method that do not rely on a Gaussian assumption, which do not hold here.

7 Appendix

7.1 Exercise 1: Discrete model

Let z and x be two discrete random variables such that $p(z=m)=\pi_m$ and $p(x=k|z=m)=\theta_{k,m}$ with $k \in [\![1,K]\!]$, $m \in [\![1,M]\!]$. We also introduce the variables $(Z_m)_m$ and $(X_k)_k$ such that $Z_m = \mathbb{1}_{z=m}$ and $X_k = \mathbb{1}_{x=k}$

Given $\big((x_1,z_1),...,(x_n,z_n)\big)$ n i.i.d observations, the likelihood of such observations is:

$$p(\pi,\theta) = \prod_{i=1}^{n} \mathbb{P}_\pi(z=z_i)\mathbb{P}_\theta(x=x_i|z=z_i) = \prod_{i=1}^{n}\prod_{m=1}^{M} \Big(\pi^{Z^i_m} \prod_{k=1}^{K} \theta_{m,k}^{X^i_k Z^i_m}\Big)$$

We compute the log-likelihood:

$$\begin{aligned}
\log p(\pi,\theta) &= \sum_{i=1}^{n} \Big(\sum_{m=1}^{M} Z^i_m \log \pi_m + \sum_{m=1}^{M}\sum_{k=1}^{K} X^i_k Z^i_m \log \theta_{m,k}\Big) \\
&= \sum_{m=1}^{M} n(z)_m \log \pi_m + \sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k} \log \theta_{m,k}
\end{aligned}$$

where $n(z)_m = \sum\limits_{i=1}^{n} \mathbb{1}_{(z^i=m)}$ and $n(z,x)_{m,k} = \sum\limits_{i=1}^{n} \mathbb{1}_{(z^i=m)}\mathbb{1}_{(x^i=k)}$.
Since $\log p$ is strictly concave w.r to π_m and $\theta_{m,k}$ for all $k \in [\![1,K]\!]$, $m \in [\![1,M]\!]$, it has a unique global maximum on the set $\mathcal{D} = \Big\{\pi,\theta | \sum\limits_{i=m}^{M} \pi_m = 1, \sum\limits_{i=m}^{M}\sum\limits_{i=1}^{K} \theta_{m,k} = 1\Big\}$
The two terms above depend only of the $(\pi_m)_m$ and the $(\theta_{m,k})_{m,k}$ respectively, so we optimize then seperatly:

$$\begin{aligned}
\sum_{m=1}^{M} n(z)_m \log \pi_m &= \sum_{m=1}^{M} n(z)_m \log\big(\frac{\pi_m}{n(z)_m} n \frac{n(z)_m}{n}\big) \\
&= \sum_{m=1}^{M} n(z)_m \log\big(\frac{\pi_m}{n(z)_m} n\big) + C \\
&\leq \sum_{m=1}^{M} n(z)_m \big(\frac{\pi_m}{n(z)_m} n - 1\big) + C
\end{aligned}$$

with $C = \sum_{m=1}^{M} n(z)_m \log \frac{n(z)_m}{n}$ a constant.
Also,

$$\sum_{m=1}^{M} n(z)_m \big(\frac{\pi_m}{n(z)_m} n - 1\big) = n \underbrace{\Big(\sum_{m=1}^{M} \pi_m\Big)}_{=1} - \underbrace{\sum_{m=1}^{M} n(z)_m}_{=n} = 0$$

Therefore $\sum_{m=1}^{M} n(z)_m \log \pi_m \leq C$ with equality if $\forall m \in [\![1,M]\!], \pi_m = \hat{\pi_m} = \frac{n(z)_m}{n}$
Since $\sum_{m=1}^{M} n(z)_m \log \pi_m$ is strictly concave in π_m, $\hat{\pi} = (\hat{\pi_m})_m$ is its unique global optimum.
Similarly,

$$\begin{aligned}
\sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k} \log \theta_{m,k} &= \sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k} \log\big(\frac{\theta_{m,k}}{n(z,x)_{m,k}} n(z)_m\big) + C' \\
&\leq \sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k} \big(\frac{\theta_{m,k}}{n(z,x)_{m,k}} n(z)_m - 1\big) + C'
\end{aligned}$$

with $C' = \sum\limits_{m=1}^{M}\sum\limits_{k=1}^{K} n(z,x)_{m,k}\log\left(\frac{n(z,x)_{m,k}}{n(z)_m}\right)$
Since

$$\sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k} = n$$

and

$$\sum_{m=1}^{M}\sum_{k=1}^{K}\theta_{m,k}n(z)_m = \sum_{m=1}^{M} n(z)_m \underbrace{\sum_{k=1}^{K}\theta_{m,k}}_{=1} = \sum_{m=1}^{M} n(z)_m = n$$

We have $\sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k}\left(\frac{\theta_{m,k}}{n(z,x)_{m,k}}n(z)_m - 1\right) = 0$
Therefore,

$$\sum_{m=1}^{M}\sum_{k=1}^{K} n(z,x)_{m,k}\log\theta_{m,k} \leq C'$$

with equality if $\forall k \in [\![1,K]\!], \forall m \in [\![1,M]\!], \theta_{m,k} = \frac{n(x,z)_{m,k}}{n(z)_m}$, which is the optimal point because of the strict concavity of the objective function.
In conclusion, the MLE estimators for this model are:

$$\forall k \in [\![1,K]\!], \forall m \in [\![1,M]\!], \ \boxed{\widehat{\pi_m} = \frac{n(z)_m}{n}}, \ \boxed{\widehat{\theta_{m,k}} = \frac{n(x,z)_{m,k}}{n(z)_m}}$$

NB: Here we make the assumption that all classes (m,k) are found at least once in the data. If it is not the case for a given pair (m_0,k_0), we simply consider the set $[\![1,M]\!] \times [\![1,K]\!] \backslash \{(m_0,k_0)\}$

7.2 Exercise 2.1: LDA model

We consider the following model:

$$y \sim \mathcal{B}(\pi) \text{ and } x|y=j \sim \mathcal{N}(\mu_j, \Sigma)$$

Given $\big((x_1,y_1),...,(x_n,y_n)\big)$ n i.i.d observations, the log-likelihood of such observations is:

$$\log p(\pi,\mu_0,\mu_1,\Sigma) = \sum_{i=1}^{n}\Bigg[\overbrace{y_i\log(\pi) + (1-y_i)\log(1-\pi)}^{A} \\ \underbrace{-\log(\sqrt{(2\pi)^d|\Sigma|}) \underbrace{-\frac{1}{2}\sum_{j=0,1}(x_i-\mu_j)^T\Sigma^{-1}(x_i-\mu_j)\mathbb{1}_{(y_i=j)}}_{B}}_{C}\Bigg]$$

In order to maximize with respect to π, μ and Σ, we will consider respectively A, B and C.

- $\sum_{i=1}^{n} y_i\log(\pi) + (1-y_i)\log(1-\pi)$ is strictly concave in π because of the strict concavity of the log function. Therefore, the optimization problem

$$\begin{array}{ll} \max\limits_{\pi} & \log p(\pi,\mu,\Sigma) \\ \text{s.t.} & \pi \in]0,1[\end{array}$$

can be solved by setting the gradient of the objective function to 0:

$$\frac{\partial}{\partial\pi}\log p(\pi,\mu,\Sigma) = \frac{n(y)}{\pi} + \frac{n-n(y)}{1-\pi} = 0 \Longleftrightarrow \boxed{\pi = \hat{\pi} = \frac{n(y)}{n}}$$

with $n(y)$ the number of y^i equal to 1.

- The optimization process in order to find μ_j is the same for $j = 0$ or $j = 1$. Let's consider $j = 0$:

$$\sum_{i=1}^{n} -\frac{1}{2}(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0)\mathbb{1}_{(y_i=0)} = -\frac{1}{2}\sum_{i=1}^{n} Tr\big(\Sigma^{-1}(x_i - \mu_0)(x_i - \mu_0)^T\big)\mathbb{1}_{(y_i=0)}$$

  This function is differentiable and concave in μ_0 as a quadratic function with a negative leading coefficient.

  Therefore, the problem

$$\begin{aligned} \max_{\mu_0} \quad & \log p(\pi, \mu_0, \mu_1, \Sigma) \\ s.t \quad & \mu_0 \in [0, 1] \end{aligned}$$

  admits a unique optimal solution $\hat{\mu_0}$ such that $\frac{\partial}{\partial \mu_0} \log p(\pi, \hat{\mu_0}, \mu_1, \Sigma) = 0$

$$\Longrightarrow \sum_{i=1}^{n} \Sigma^{-1}(x_i - \hat{\mu_0})\mathbb{1}_{(y_i=0)} = 0 \Longleftrightarrow \boxed{\hat{\mu_0} = \frac{1}{n(y)_0} \sum_{\substack{i=1 \\ y_i=0}}^{n} x_i}$$

  Respectively for μ_1:

$$\boxed{\hat{\mu_1} = \frac{1}{n(y)_1} \sum_{\substack{i=1 \\ y_i=1}}^{n} x_i}$$

- Since the MLE estimators $\hat{\pi}$, $\hat{\mu_0}$ and $\hat{\mu_1}$ that we found so far do not depend on Σ, the MLE estimator for Σ is found by maximizing $\log p(\hat{\pi}, \hat{\mu_0}, \hat{\mu_1}, \cdot)$, which is equivalent to the following problem:

$$\begin{aligned} \max_{\Sigma} \quad & -\frac{n}{2}\log\det\Sigma - \frac{1}{2}Tr(\Sigma^{-1}\hat{M}) \\ \text{s.t.} \quad & \Sigma \in \mathcal{S}^d_{++} \end{aligned}$$

  where $\hat{M} = \sum_{i=1}^{n} \sum_{j=0,1} \mathbb{1}_{(y_i=j)}(x_i - \hat{\mu_j})(x_i - \hat{\mu_j})^T$

  The objective function of this optimization problem is not concave because $A \mapsto -\log\det A$ is convex. However, we introduce $\Lambda = \Sigma^{-1}$. The problem is now equivalent to

$$\begin{aligned} \max_{\Lambda} \quad & \frac{n}{2}\log\det\Lambda - \frac{1}{2}Tr(\Lambda\hat{M}) \\ \text{s.t.} \quad & \Lambda \in \mathcal{S}^d_{++} \end{aligned}$$

  where the objective function is concave as the sum of a concave function and a linear function, and the feasible set is convex.

  Therefore, it admits an optimal point in $\hat{\Lambda}$ such that, with $f : A \mapsto \frac{n}{2}\log\det A - \frac{1}{2}Tr(A\hat{M})$,

$$\frac{\partial}{\partial\Lambda} f(\hat{\Lambda}) = \frac{n}{2}\hat{\Lambda}^{-1} - \frac{1}{2}\hat{M} = 0 \Longleftrightarrow \hat{\Lambda}^{-1} = \frac{1}{n}\hat{M}$$

  Therefore,

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

  NB: here we assume that $\hat{M}$ is invertible.

Link to logistic regression

Here we detail how, in such a model, $p(y=1|x)$ compares to a logistic regression. We express $p(y|x)$ by keeping only the terms that depend on y:

$$\begin{aligned}
p(y|x) &= p(y)p(x,y) \\
&\propto \pi^y(1-\pi)^{1-y}\exp\big(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)y - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)(1-y)\big) \\
&\propto \pi^y(1-\pi)^{1-y}\exp\big(-\frac{1}{2}y(2(\mu_0-\mu_1)^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)\big) \\
&\propto \exp(ya + yb^Tx)
\end{aligned}$$

with

$$\boxed{a = \log\frac{\pi}{1-\pi} - \frac{1}{2}(\mu_1\Sigma^{-1}\mu_1 - \mu_0\Sigma^{-1}\mu_0)}\text{, and }\boxed{b = -(\Sigma^{-1})^T(\mu_0-\mu_1)}$$

Finally, we obtain $p(y=1|x)$ by normalizing (thus cancelling all the terms of the product that are constant in y):

$$\boxed{p(y=1|x) = \frac{\exp(a+b^Tx)}{1+\exp(a+b^Tx)} = \sigma(a+b^Tx)}$$

7.3 Exercise 2.5: QDA model

We consider different covariance matrix for each class:

$$y \sim \mathcal{B}(\pi) \text{ and } x|y=j \sim \mathcal{N}(\mu_j, \Sigma_j)$$

Given $\big((x_1,y_1),...,(x_n,y_n)\big)$ n i.i.d observations, the log-likelihood of such observations is:

$$\begin{aligned}
\log p(\pi,\mu_0,\mu_1,\Sigma) = \sum_{i=1}^{n}\Bigg[& y_i\log(\pi) + (1-y_i)\log(1-\pi) \\
& - \mathbb{1}_{(y_i=1)}\Big(\log(\sqrt{(2\pi)^d|\Sigma_1|}) + \frac{1}{2}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1)\Big) \\
& - \mathbb{1}_{(y_i=0)}\Big(\log(\sqrt{(2\pi)^d|\Sigma_0|}) + \frac{1}{2}(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)\Big)\Bigg]
\end{aligned}$$

- The optimization with respect to π is the same as in the Linear Discriminant Analysis, *i.e*:

$$\boxed{\hat{\pi} = \frac{n(y)}{n}}$$

- The same goes for each μ_j :

$$\text{For } j=0,1, \quad \boxed{\hat{\mu_j} = \frac{1}{n(y)_j}\sum_{\substack{i=1\\ y_i=j}}^{n} x_i}$$

- Again, the MLE estimators $\hat{\pi}$, $\hat{\mu_0}$ and $\hat{\mu_1}$ that we found so far do not depend of Σ_0 and Σ_1, we can compute $\hat{\Sigma}_0$ (resp. $\hat{\Sigma}_1$) by maximizing $-\log p(\hat{\pi},\hat{\mu_0},\hat{\mu_1},\cdot,\Sigma_1)$ (resp. $-\log p(\hat{\pi},\hat{\mu_0},\hat{\mu_1},\Sigma_0,\cdot)$), under the condition that both $\hat{\Sigma}$ do not depend of each other. Because of the symmetry of the terms depending on the Σ_j in the log-likelihood function, the optimization process in order to find each $\hat{\Sigma_j}$ is the same for both $j=0$ and $j=1$. For $j=0$, we have the following optimization problem:

$$\begin{aligned}
\max_{\Sigma_0} \quad & -\frac{n}{2}\log\det\Sigma_0 - \frac{1}{2}Tr(\Sigma_0^{-1}\hat{M}_0) \\
\text{s.t.} \quad & \Sigma_0 \in \mathcal{S}_{++}^d
\end{aligned}$$

where $\hat{M_0} = \sum\limits_{\substack{i=1 \\ y_i=0}}^{n} (x_i - \hat{\mu_0})(x_i - \hat{\mu_0})^T$.

Similarly to the LDA, we introduce $\Lambda_0 = \Sigma_0^{-1}$ so that the objective function $f_0 : A \mapsto \frac{n(y)_0}{2} \log\det A - \frac{1}{2} Tr(A\hat{M_0})$ is differentiable and convex. Since the feasible set $\mathcal{S}_{++}^d$ is also convex, we compute $\hat{\Lambda}$ the global maximum with:

$$\frac{\partial}{\partial \Lambda_0} f_0(\hat{\Lambda_0}) = \frac{n(y)_0}{2} \hat{\Lambda_0}^{-1} - \frac{1}{2}\hat{M_0} = 0 \Longleftrightarrow \hat{\Lambda_0}^{-1} = \frac{1}{n(y)_0}\hat{M_0}$$

Hence,

$$\boxed{\hat{\Sigma_0} = \frac{1}{n(y)_0} \sum_{\substack{i=1 \\ y_i=0}}^{n} (x_i - \hat{\mu_0})^T (x_i - \hat{\mu_0})} \text{ and similarly } \boxed{\hat{\Sigma_1} = \frac{1}{n(y)_1} \sum_{\substack{i=1 \\ y_i=1}}^{n} (x_i - \hat{\mu_1})^T (x_i - \hat{\mu_1})}$$

Then, we have:

$$\begin{aligned}
p(y|x) &\propto \pi^y (1-\pi)^{1-y_i} \exp\big(-\frac{y}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1) - \frac{1-y}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\big) \\
&\qquad \cdot \exp(-\frac{y}{2}\log\det\Sigma_1 - \frac{1-y}{2}\log\det\Sigma_0) \\
&\propto \pi^y (1-\pi)^{1-y_i} \exp\big(-\frac{1}{2} y x^T(\Sigma_1^{-1} - \Sigma_0^{-1})x + y(\mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1})x - \frac{1}{2}y(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0)\big) \\
&\qquad \cdot \exp(-\frac{y}{2}\log\frac{\det\Sigma_1}{\det\Sigma_0}) \\
&\propto \exp(-\frac{1}{2}yx^T Mx + yb^T x + ya)
\end{aligned}$$

where

$$\boxed{M = \Sigma_1^{-1} - \Sigma_0^{-1}}, \quad \boxed{b^T = \mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1}}, \quad \boxed{a = \log\frac{\pi\sqrt{\det\Sigma_0}}{(1-\pi)\sqrt{\det\Sigma_1}} - \frac{1}{2}\Big(\mu_1^T\Sigma_1^{-1}\mu_1 - \mu_0^T\Sigma_0^{-1}\mu_0\Big)}$$

Finally,

$$\boxed{p(y=1|x) = \frac{\exp(-\frac{1}{2}x^T Mx + b^T x + a)}{1 + \exp(-\frac{1}{2}x^T Mx + b^T x + a)} = \sigma(-\frac{1}{2}x^T Mx + b^T x + a)}$$