PGM - HW1 - Mehdi Boubnan & Amine Sadeq

Exercise 1: Learning in discrete graphical models

One hot encoding : we put $z_i^j = 1$ if $z_i = j$ and $x_i^p = 1$ if $x_i = p$

$$L(\pi,\theta) = L(\pi,\theta) = \prod_{i=1}^{n} p(x_i,z_i) = \prod_{i=1}^{n} p(x_i|z_i)p(z_i) + \prod_{i=1}^{n}(\prod_{j=1}^{M} p(z_i=j)^{z_i^j} \prod_{j=1}^{M}\prod_{p=1}^{k} p(x_i=p|z_i=j)^{x_i^p z_i^j}) = \prod_{i=1}^{n}(\prod_{j=1}^{M} \pi_j^{z_i^j} \prod_{j=1}^{M}\prod_{p=1}^{k} \theta_{p,j}^{x_i^p z_i^j})$$

$$\mathcal{L}(\pi,\theta) = \sum_{i=1}^{n}(\sum_{j=1}^{M} z_i^j log(\pi_j) + \sum_{j=1}^{M}\sum_{p=1}^{k} z_i^j x_i^p log(\theta_{p,j})) = \sum_{j=1}^{M}(\sum_{i=1}^{n} z_i^j) log(\pi_j) + \sum_{j=1}^{M}\sum_{p=1}^{k}(\sum_{i=1}^{n} z_i^j x_i^p) log(\theta_{p,j})$$

we put $n_j = \sum_{i=1}^{n} z_i^j$ the number of $z_i = j$ and $n_{p,j} = \sum_{i=1}^{n} z_i^j x_i^p$ the number of couples (x_i, z_i) such that $x_i = p$ and $z_i = j$

$$\mathcal{L}(\pi,\theta) = \sum_{j=1}^{M} n_j log(\pi_j) + \sum_{j=1}^{M}\sum_{p=1}^{k} n_{p,j} log(\theta_{p,j})$$

We need to maximize this log-likelihood with respect to the constraints $\sum_{j=1}^{M} \pi_j = 1$ (1) and $\sum_{j=1}^{M}\sum_{p=1}^{k} \theta_{p,j} = 1$ (2). Since the log-likelihood is concave and the constraints are affine, we'll compute the lagrangians to find the maximum for our log-likelihood.

$$Lag(\pi,\lambda) = -\sum_{j=1}^{M} n_j log(\pi_j) + \lambda((\sum_{j=1}^{M} \pi_j) - 1)$$
$$\nabla_{\pi_j} Lag = -\frac{n_j}{\pi_j} + \lambda = 0 \iff \hat{\pi}_j = \frac{n_j}{\lambda}$$
$$(1) \implies \sum_{j=1}^{M} \frac{n_j}{\lambda} = 1 \implies \lambda = n \implies \hat{\pi}_j = \frac{n_j}{n}$$

$$Lag(\theta,\lambda) = -\sum_{j=1}^{M}\sum_{p=1}^{k} n_{p,j} log(\theta_{p,j}) + \lambda((\sum_{j=1}^{M}\sum_{p=1}^{k} \theta_{p,j}) - 1)$$
$$\nabla_{\theta_{p,j}} Lag = -\frac{n_{p,j}}{\theta_{p,j}} + \lambda = 0 \iff \hat{\theta}_{p,j} = \frac{n_{p,j}}{\lambda}$$
$$(2) \implies \sum_{j=1}^{M}\sum_{p=1}^{k} \frac{n_{p,j}}{\lambda} = 1 \implies \lambda = n \implies \hat{\theta}_{p,j} = .\frac{n_{p,j}}{n}$$

Exercise 2.1(a): LDA Formulas

$$\nabla_{\pi}\mathcal{L} = -\frac{n}{1-\pi} + \sum_{i=1}^{n} y_i \frac{1-\pi}{\pi}\left(\frac{1}{1-\pi} + \frac{\pi}{(1-\pi)^2}\right) \qquad \hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\nabla_{\mu_0}\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n}(1-y_i)(2\Sigma^{-1}\mu_0 - 2\Sigma^{-1}x_i) \qquad \hat{\mu_0} = \frac{\sum_{i=1}^{n}(1-y_i)x_i}{\sum_{i=1}^{n}(1-y_i)}$$

$$\nabla_{\mu_1}\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n}(1-y_i)(2\Sigma^{-1}\mu_1 - 2\Sigma^{-1}x_i) \qquad \hat{\mu_1} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i}$$

$$\nabla_{\Sigma^{-1}}\mathcal{L} = \sum_{i=1}^{n} \frac{\Sigma}{2} - \frac{y_i}{2}(x_i-\mu_1)(x_i-\mu_1)^T - \frac{1-y_i}{2}(x_i-\mu_0)(x_i-\mu_0)^T \qquad \hat{\Sigma} = \frac{\sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T + (1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T}{n}$$

Exercise 2.5(a): QDA Formulas

$$\nabla_{\pi}\mathcal{L} = -\frac{n}{1-\pi} + \sum_{i=1}^{n} y_i \frac{1-\pi}{\pi}\left(\frac{1}{1-\pi} + \frac{\pi}{(1-\pi)^2}\right) \qquad \hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}$$

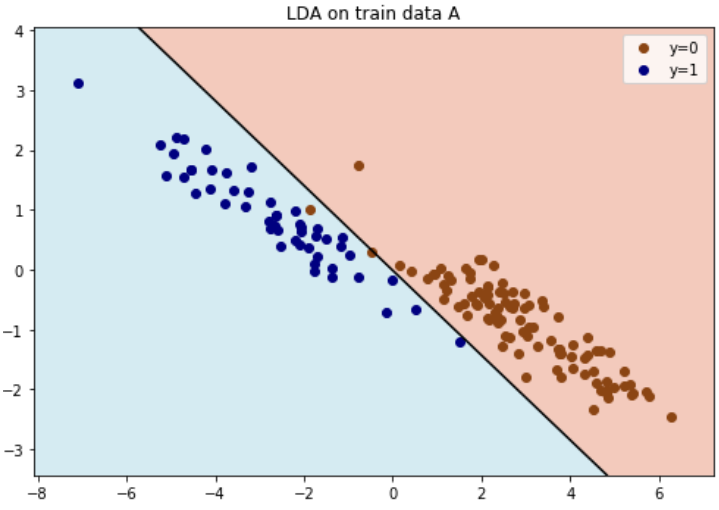
$$\nabla_{\mu_0}\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n}(1-y_i)(2\Sigma^{-1}\mu_0 - 2\Sigma^{-1}x_i) \qquad \hat{\mu_0} = \frac{\sum_{i=1}^{n}(1-y_i)x_i}{\sum_{i=1}^{n}(1-y_i)}$$

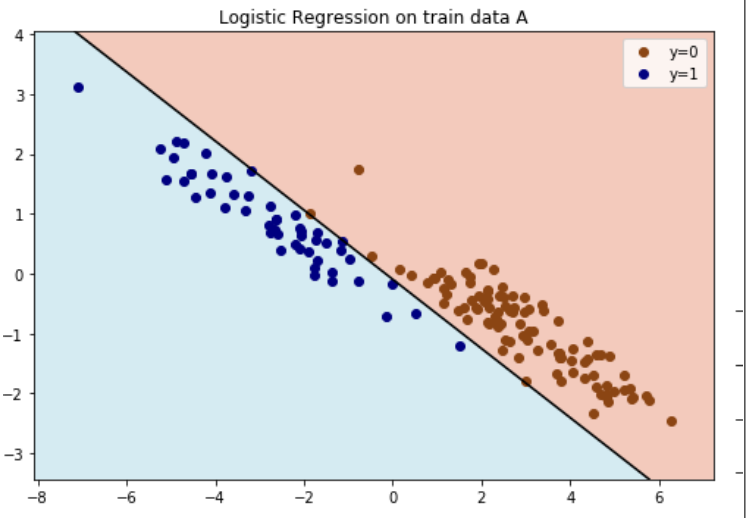
$$\nabla_{\mu_1}\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n}(1-y_i)(2\Sigma^{-1}\mu_1 - 2\Sigma^{-1}x_i) \qquad \hat{\mu_1} = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i}$$

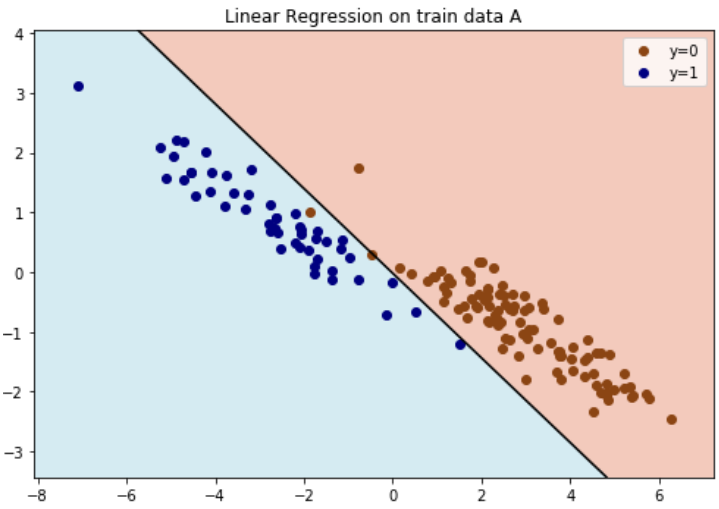
$$\nabla_{\Sigma_0^{-1}}\mathcal{L} = \sum_{i=1}^{n} \frac{1-y_i}{2}\Sigma_0 - \frac{1-y_i}{2}(x_i-\mu_0)(x_i-\mu_0)^T \qquad \hat{\Sigma_0} = \frac{\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T}{\sum_{i=1}^{n}(1-y_i)}$$

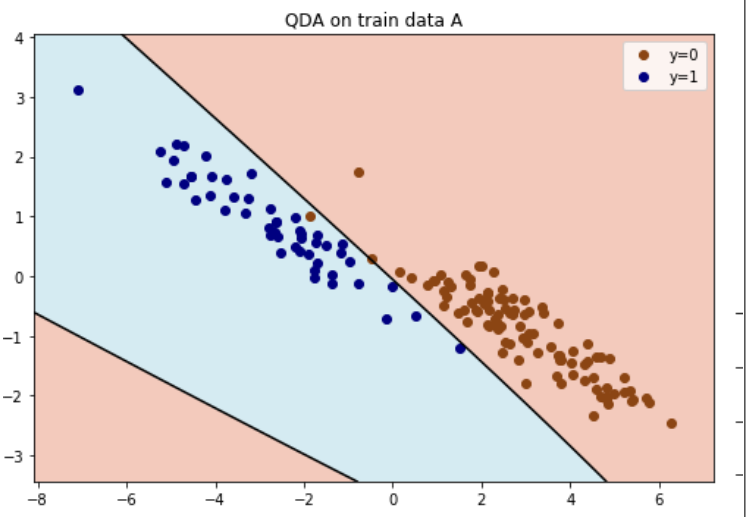
$$\nabla_{\Sigma_1^{-1}}\mathcal{L} = \sum_{i=1}^{n} \frac{y_i}{2}\Sigma_1 - \frac{y_i}{2}(x_i-\mu_1)(x_i-\mu_1)^T \qquad \hat{\Sigma_1} = \frac{\sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T}{\sum_{i=1}^{n} y_i}$$

Dataset A







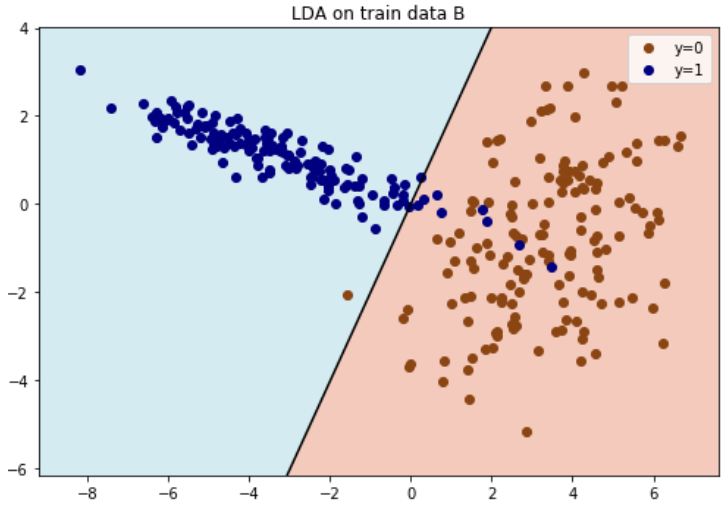


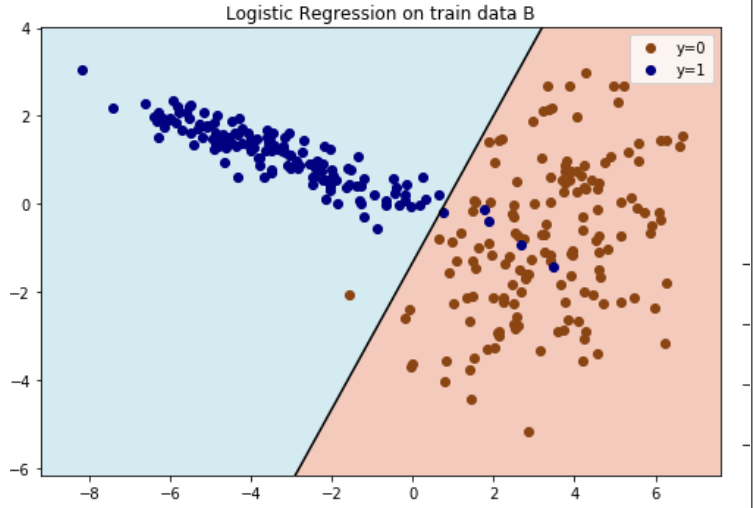
Misclassification errors

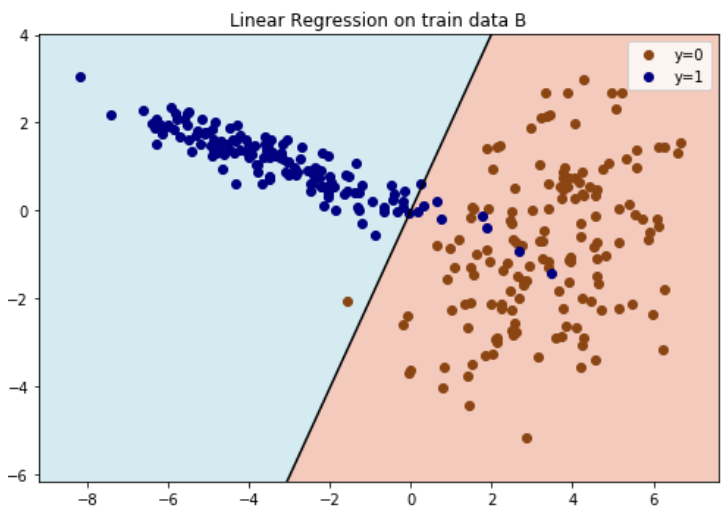
	Train	Test
LDA	1.33%	2%
Lin R	1.33%	2.07%
Log R	0%	3.4%
QDA	0.67%	2%

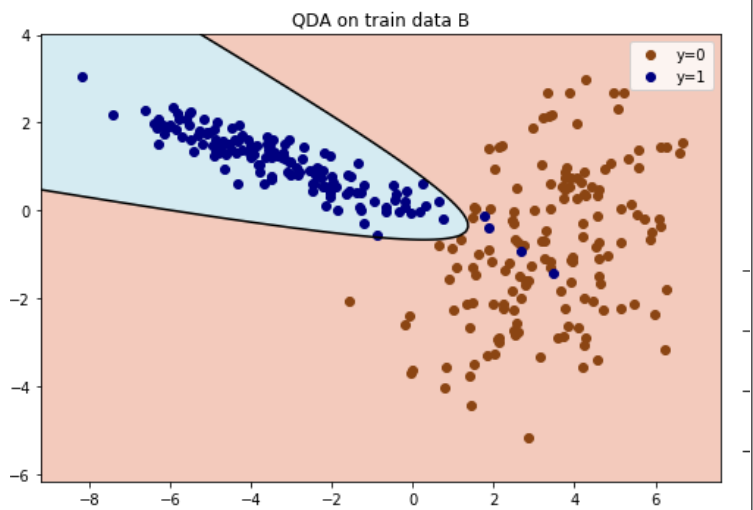
- All the misclassification errors are larger on the test data as the training data is too small to generalize well (size 150).
- The models have approximately same misclassification errors on train and test data. Noting that the logistic regression yields a perfect score on the training set,because the two distributions are easily separable by a line, but still gets the poorest score on the testing set. We can explain that by the fact that the logistic regression overfits on the training set, and can't generalize because of the small training dataset.
- We notice that the dataset A is composed by two multivariate normal distributions with almost similar covariances. Logically, LDA should perform perfectly with its covariance equality assumption, which is the case in here, and should rank second right after the QDA model with almost same performances since the small dimensionality of our problem.

Dataset B







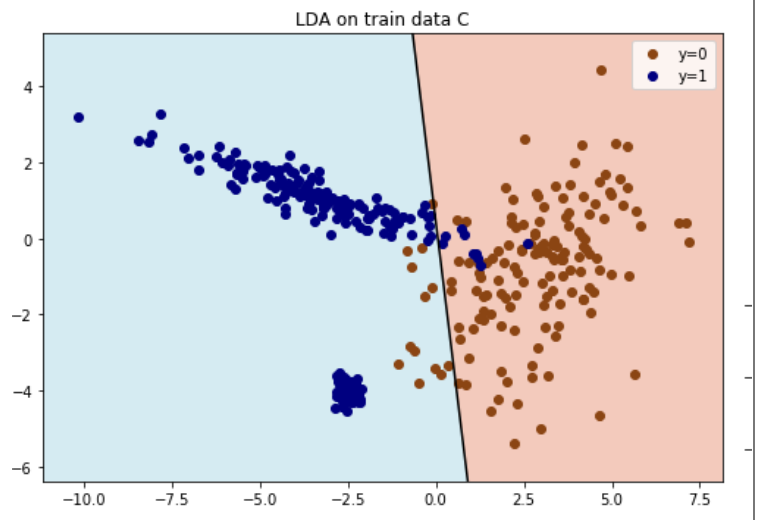


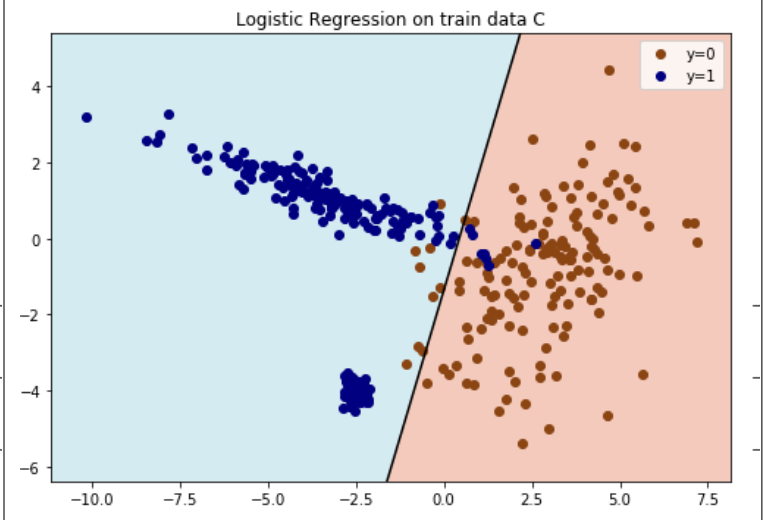
Misclassification errors

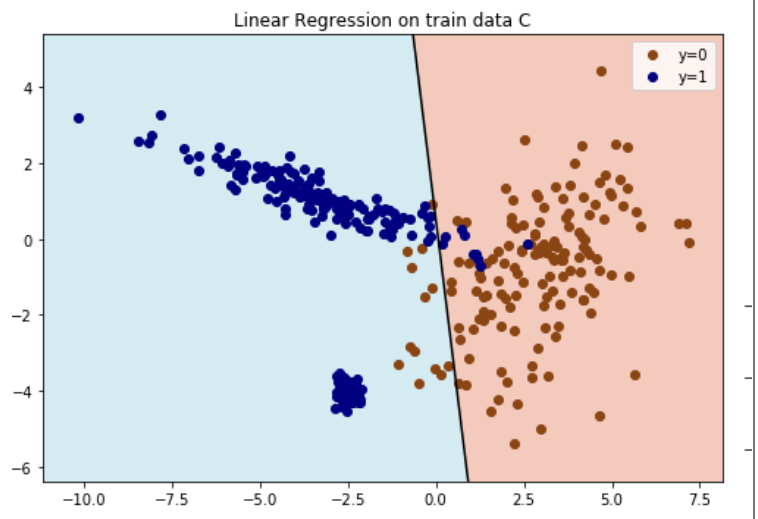
	Train	Test
LDA	3%	4.15%
Lin R	3%	4.15%
Log R	2%	4.3%
QDA	1.33%	2%

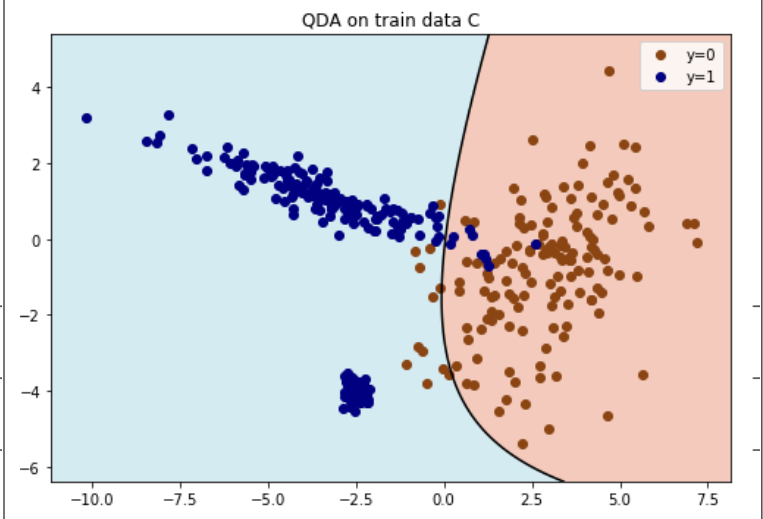
- All the misclassification errors are larger on the test data as the training data is too small to generalize well (size 300).
- The Dataset B is composed by two multivariate gaussians with completely different covariance matrices. This data fits exactly the QDA model, so we would expect it to give the best performances, which is the case in here.
- LDA and Linear Regression give the exact same performances on the training and test datasets. The logistic regression gives approximately same results, since the two class spaces are not easily separable.

Dataset C









Misclassification errors

	Train	Test
LDA	5.5%	4.23%
Lin R	5.5%	4.23%
Log R	4%	2.27%
QDA	5.25%	3.83%

- All the misclassification errors are smaller on the test data as the training data is finally getting big enough to start generalizing enough (400).
- Dataset C is composed from three different multivariate gaussians, (two for the positive class and one for the negative class). Therefore the assumptions taken in LDA and QDA are wrong, which makes them being outperformed by the logistic regression.
- The logistic regression outperforms all the other algorithms because it's the only one without any strong assumptions on a Gaussian model.

Detailed proofs

Exercise 2.1 : LDA

The assumption we consider in our model :

- $X \in \mathcal{R}^2$ and $Y \in \mathcal{R}$
- $P(X/Y=1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $P(X/Y=0) \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $P(Y) \sim \mathcal{B}(\pi)$
- Fisher's assumption : $\Sigma_0 = \Sigma_1 = \Sigma$

Computing Likelihood

$$\begin{aligned}
L(\mu_0, \mu_1, \pi, \Sigma) &= \prod_{i=1}^{n} p(x_i, y_i) = \prod_{i=1}^{n} p(y_i) p(x_i/y_i) \\
&= \prod_{i=1}^{n} \pi_i^y (1-\pi)^{1-y_i} \frac{1}{2\pi\sqrt{det\Sigma}} exp(-\frac{y_i}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) - \frac{1-y_i}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)) \\
&= \prod_{i=1}^{n} exp(y_i log(\pi) + (1-y_i) log(1-\pi) - log(2\pi) - \frac{log(det\Sigma)}{2} - \frac{y_i}{2}(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) \\
&\quad - \frac{1-y_i}{2}(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)) \\
&= \prod_{i=1}^{n} exp(y_i log(\pi) + (1-y_i) log(1-\pi) - log(2\pi) - \frac{log(det\Sigma)}{2} \\
&\quad - \frac{y_i}{2}(x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1) - \frac{1-y_i}{2}((x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma_1^{-1}\mu_0)) \\
\mathcal{L}(\mu_0, \mu_1, \pi, \Sigma) &= log(L(\mu_0, \mu_1, \pi, \Sigma)) \\
&= \sum_{i=1}^{n} y_i log(\pi) + (1-y_i) log(1-\pi) - log(2\pi) - \frac{1}{2} log(det\Sigma) \\
&\quad - \frac{y_i}{2}(x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1) - \frac{1-y_i}{2}((x_i^T\Sigma^{-1}x_i - 2x_i^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma_1^{-1}\mu_0))
\end{aligned}$$

Computing $P(Y=1/X)$

$$\begin{aligned}
p(y|x) &\propto p(x|y)p(y) \\
&\propto \pi^y(1-\pi)^{1-y} \frac{1}{2\pi\sqrt{det\Sigma}} exp\left(-\frac{y}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1-y}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
&\propto exp\left(y log(\pi) + (1-y) log(1-\pi) - log(2\pi) - \frac{log(det\Sigma)}{2} - \frac{y}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1-y}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
&\propto exp(y log(\pi) + (1-y) log(1-\pi) - log(2\pi) - \frac{log(det\Sigma)}{2} - \frac{y}{2}(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_1 + \mu_1^T\Sigma^{-1}\mu_1) \\
&\quad - \frac{1-y}{2}(x^T\Sigma^{-1}x - 2x^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}\mu_0)) \\
&\propto exp\left(y(log(\frac{\pi}{1-\pi}) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1)) + y(\mu_1^T - \mu_0^T)\Sigma^{-1}x\right)
\end{aligned}$$

We removed all expressions that are independent from y because of the constraint of normalization.
We get then :

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

with : $\begin{cases} \text{a} & = log(\frac{\pi}{1-\pi}) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) \\ b & = \Sigma^{-1}(\mu_1 - \mu_0) \end{cases}$

Exercise 2.5 : QDA

The assumption we consider in our model :

- $X \in \mathcal{R}^2$ and $Y \in \mathcal{R}$
- $P(X/Y=1) \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $P(X/Y=0) \sim \mathcal{N}(\mu_0, \Sigma_0)$
- $P(Y) \sim \mathcal{B}(\pi)$

Computing Likelihood

$$\begin{aligned}
L(\mu_0,\mu_1,\pi,\Sigma_0,\Sigma_1) &= \prod_{i=1}^{n} p(x_i,y_i) = \prod_{i=1}^{n} p(y_i)p(x_i/y_i) \\
&= \prod_{i=1}^{n} \pi_i^y (1-\pi)^{1-y_i} \frac{1}{2\pi\sqrt{det\Sigma_y}} exp(-\frac{y_i}{2}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1) - \frac{1-y_i}{2}(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)) \\
&= \prod_{i=1}^{n} exp(y_i log(\pi) + (1-y_i)log(1-\pi) - log(2\pi) - \frac{y_i}{2}log(det\Sigma_1) - \frac{1-y_i}{2}log(det\Sigma_0) - \frac{y_i}{2}(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1) \\
&\quad - \frac{1-y_i}{2}(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)) \\
&= \prod_{i=1}^{n} exp(y_i log(\pi) + (1-y_i)log(1-\pi) - log(2\pi) - \frac{y_i}{2}log(det\Sigma_1) - \frac{1-y_i}{2}log(det\Sigma_0) \\
&\quad - \frac{y_i}{2}(x_i^T\Sigma_1^{-1}x_i - 2x_i^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mu_1) - \frac{1-y_i}{2}((x_i^T\Sigma_0^{-1}x_i - 2x_i^T\Sigma_0^{-1}\mu_0 + \mu_0^T\Sigma_1^{-1}\mu_0)) \\
\mathcal{L}(\mu_0,\mu_1,\pi,\Sigma_0,\Sigma_1) &= log(L(\mu_0,\mu_1,\pi,\Sigma_0,\Sigma_1)) \\
&= \sum_{i=1}^{n} y_i log(\pi) + (1-y_i)log(1-\pi) - log(2\pi) - \frac{y_i}{2}log(det\Sigma_1) - \frac{1-y_i}{2}log(det\Sigma_0) \\
&\quad - \frac{y_i}{2}(x_i^T\Sigma_1^{-1}x_i - 2x_i^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mu_1) - \frac{1-y_i}{2}((x_i^T\Sigma_0^{-1}x_i - 2x_i^T\Sigma_0^{-1}\mu_0 + \mu_0^T\Sigma_1^{-1}\mu_0))
\end{aligned}$$

Computing $P(Y=1/X)$

$$\begin{aligned}
p(y|x) &\propto p(x|y)p(y) \\
&\propto \pi^y(1-\pi)^{1-y}(\frac{1}{2\pi\sqrt{det\Sigma_1}})^y(\frac{1}{2\pi\sqrt{det\Sigma_0}})^{1-y} exp\left(-\frac{y}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1-y}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right) \\
&\propto exp(ylog(\pi) + (1-y)log(1-\pi) - log(2\pi) - \frac{ylog(det\Sigma_1)}{2} - \frac{(1-y)log(det\Sigma_0)}{2} - \frac{y}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) \\
&\quad - \frac{1-y}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)) \\
&\propto exp(ylog(\pi) + (1-y)log(1-\pi) - log(2\pi) - \frac{ylog(det\Sigma_1)}{2} - \frac{(1-y)log(det\Sigma_0)}{2} - \frac{y}{2}(x^T\Sigma_1^{-1}x - 2x^T\Sigma_1^{-1}\mu_1 + \mu_1^T\Sigma_1^{-1}\mu_1) \\
&\quad - \frac{1-y}{2}(x^T\Sigma_0^{-1}x - 2x^T\Sigma_0^{-1}\mu_0 + \mu_0^T\Sigma_0^{-1}\mu_0)) \\
&\propto exp\left(y(log(\frac{\pi}{1-\pi}) + \frac{1}{2}log(\frac{det(\Sigma_0)}{det(\Sigma_1)}) + \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1)) + y(\mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1})x + \frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x\right)
\end{aligned}$$

We removed all expressions that are independent from y because of the constraint of normalization.
We get then :

$$\begin{aligned}
p(y=1|x) &= \frac{exp\left(y(log(\frac{\pi}{1-\pi}) + \frac{1}{2}log(\frac{det(\Sigma_0)}{det(\Sigma_1)}) + \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1)) + y(\mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1})x + \frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x\right)}{1 + exp\left(y(log(\frac{\pi}{1-\pi}) + \frac{1}{2}log(\frac{det(\Sigma_0)}{det(\Sigma_1)}) + \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1)) + y(\mu_1^T\Sigma_1^{-1} - \mu_0^T\Sigma_0^{-1})x + \frac{1}{2}x^T(\Sigma_0^{-1} - \Sigma_1^{-1})x\right)} \\
&= \sigma\left(a + b^Tx + x^Tcx\right)
\end{aligned}$$

$$\text{with : } \begin{cases} \text{a} & = log(\frac{\pi}{1-\pi}) + \frac{1}{2}log(\frac{det(\Sigma_0)}{det(\Sigma_1)}) + \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1) \\ b & = (\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) \\ c & = \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}) \end{cases}$$