1 Exercise 1: learning in discrete graphical models

$$\hat{\pi}_m = \frac{n_m}{n}$$

$$\hat{\theta}_{mk} = \frac{n_{mk}}{n}$$

2 Exercise 2.1: LDA Formulas

$$\hat{\mu}_0 = \frac{\sum_i x_i^T (1 - y_i)}{\sum_i (1 - y_i)}$$

donc l'estimateur du maximum de vraisemblance pour μ_0 est la moyenne empirique des x_i pour les observations où $y_i = 0$

$$\hat{\mu}_1 = \frac{\sum_i y_i x_i}{\sum_i y_i}$$

donc l'estimateur du maximum de vraisemblance pour μ_1 est la moyenne empirique des x_i pour les observations où $y_i = 1$

$$\hat{\pi} = \frac{\sum_i y_i}{n}$$

donc l'estimateur du maximum de vraisemblance de π est la moyenne empirique des y_i

$$p(y = 1|x) = 0.5 \Leftrightarrow a^T x + b = 0$$

$a = (\mu_1 - \mu_0)^T \Sigma^{-1}$ et $b = \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) + log(\frac{1-\pi}{\pi})$

3 Exercise 2.5: QDA Formulas

$$\hat{\mu}_0 = \frac{\sum_i (1 - y_i) x_i}{\sum_i (1 - y_i)}$$

$$\hat{\mu}_1 = \frac{\sum_i y_i x_i}{\sum_i y_i}$$

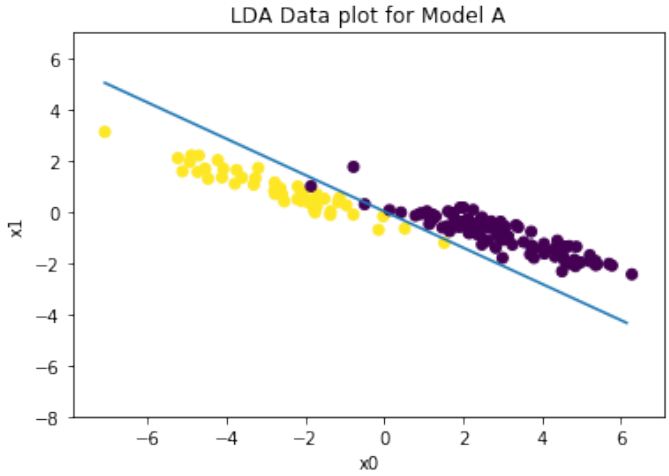
$$\hat{\pi} = \frac{\sum_i y_i}{n}$$

$$\hat{\Sigma}_0 = \frac{\sum_i (1 - y_i)(x_i - \mu_0)^T (x_i - \mu_0)}{\sum_i (1 - y_i)}$$

$$\hat{\Sigma}_1 = \frac{\sum_i y_i (x_i - \mu_1)^T (x_i - \mu_1)}{\sum_i y_i}$$

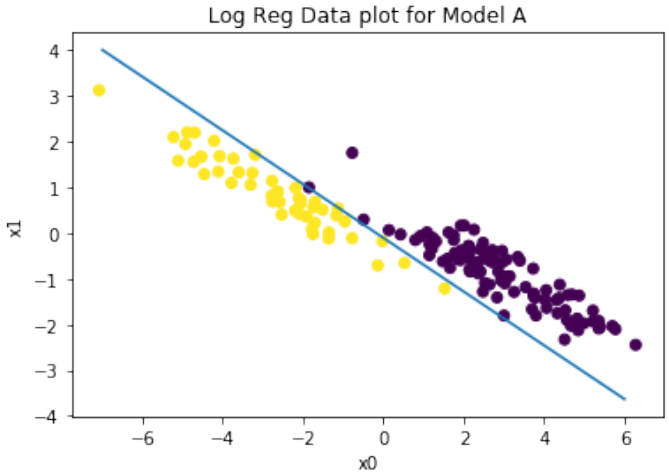
$$p(y = 1|x) = 0.5 \Leftrightarrow \frac{1}{2}(x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - \frac{1}{2}(x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) + log(\frac{1 - \pi}{\pi}) = 0$$

$$\Leftrightarrow x^T (\Sigma_1^{-1} - \Sigma_0^{-1}) x - 2x^T (\Sigma_1^{-1} \mu_1 - \Sigma_0^{-1} \mu_0) + \mu_1 \Sigma_1^{-1} \mu_1 - \mu_0 \Sigma_0^{-1} \mu_0 + 2log(\frac{1 - \pi}{\pi}) = 0$$

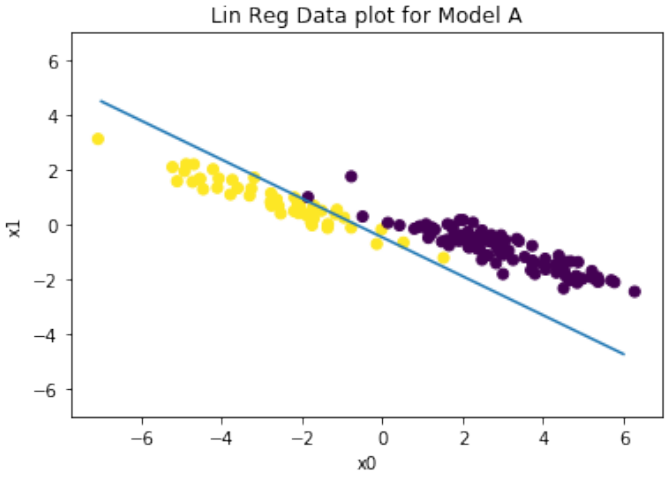


$$\hat{\pi} = 0.33 \quad \mu_0 = \begin{pmatrix} 2.90 \\ -0.89 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.69 \\ 0.87 \end{pmatrix}$$

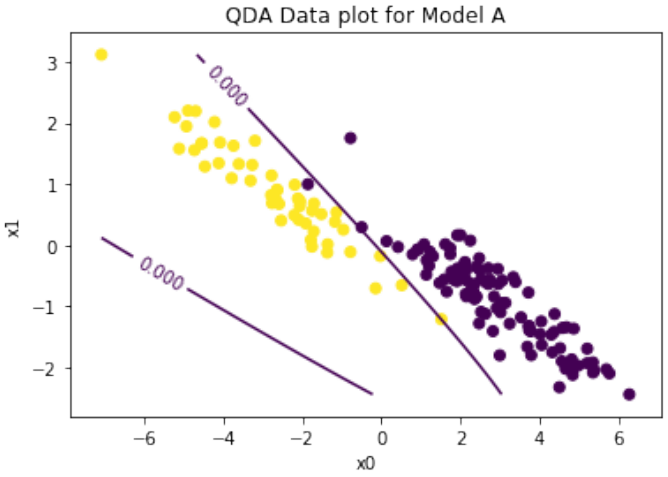
$$\Sigma = \begin{pmatrix} 2.44 & -1.13 \\ -1.13 & 0.61 \end{pmatrix}$$



$$\hat{w} = \begin{pmatrix} -27.798 \\ -47.358 \end{pmatrix} \quad b = -5.142$$



$$\hat{w} = \begin{pmatrix} -0.264 \\ -0.373 \end{pmatrix} \quad b = 0.492 \quad \sigma = 0.200$$



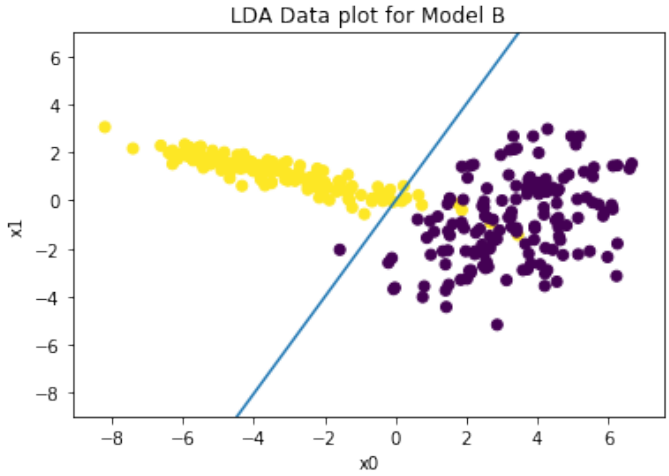
$$\hat{\pi} = 0.33 \quad \mu_0 = \begin{pmatrix} 2.90 \\ -0.89 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.69 \\ 0.87 \end{pmatrix}$$

$$\Sigma_0 = \begin{pmatrix} 2.31 & -1.05 \\ -1.05 & 0.58 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 2.70 & -1.30 \\ -1.30 & 0.69 \end{pmatrix}$$

Table 1: Table of errors (%)

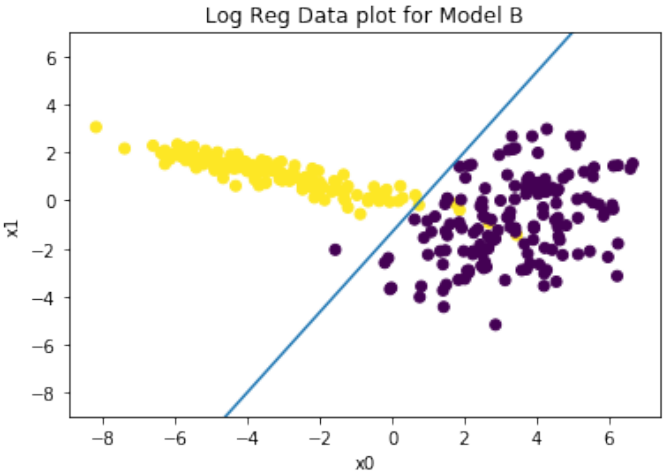
Data Type	LDA	Log Reg	Lin Reg	QDA
train	1.33	0	1.33	0.67
test	2.13	3.53	2.07	1.87

Comments: - Generative models: ici au vue des données, l'hypothèse selon laquelle les deux classes ont une covariance identique est raisonnable. D'ailleurs on voit que les valeurs de Σ pour le LDA et Σ_0 et Σ_1 pour le QDA ne sont pas très éloignées. Le LDA donne donc des performances qui sont comparables au QDA.
- Conditional models: une frontière raisonnable est apprise par les deux algos bien que la régression logistique semble être dans un cas d'overfitting.

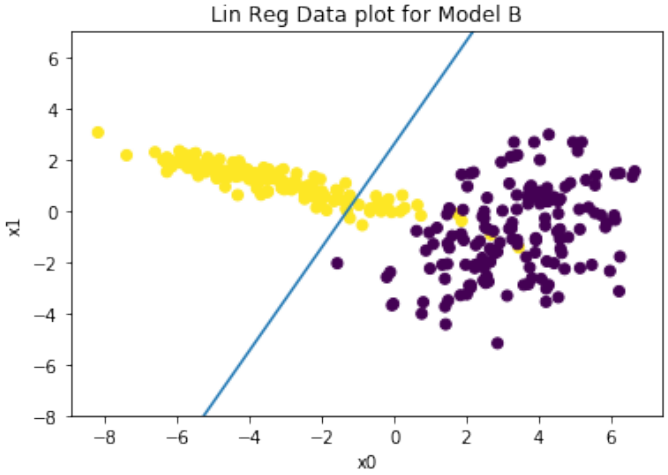


$$\hat{\pi} = 0.5 \quad \mu_0 = \begin{pmatrix} 3.34 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix}$$

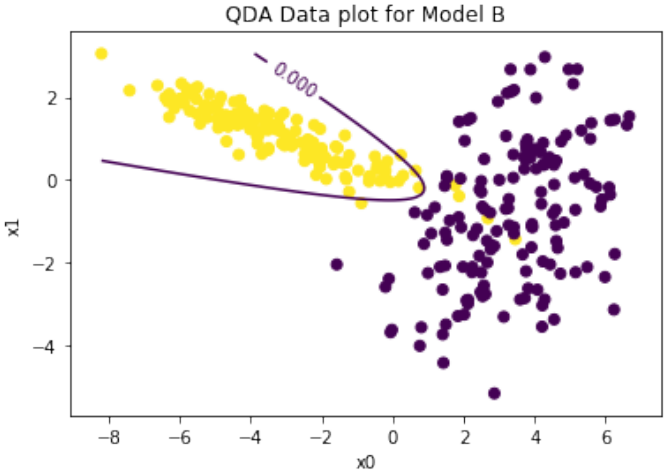
$$\Sigma = \begin{pmatrix} 3.35 & -0.14 \\ -0.14 & 1.74 \end{pmatrix}$$



$$\hat{w} = \begin{pmatrix} -1.705 \\ 1.024 \end{pmatrix} \quad b = 1.350$$



$$\hat{w} = \begin{pmatrix} -0.104 \\ -0.052 \end{pmatrix} \quad b = 0.500 \quad \sigma = 0.233$$



$$\hat{\pi} = 0.5 \quad \mu_0 = \begin{pmatrix} 3.34 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix}$$

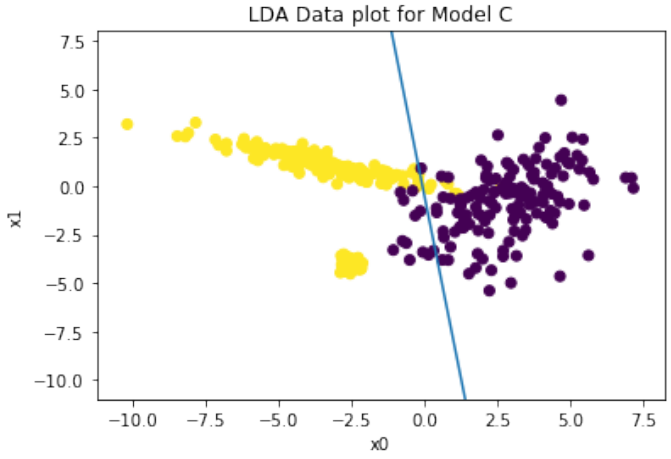
$$\Sigma_0 = \begin{pmatrix} 2.54 & 1.06 \\ 1.06 & 2.96 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 4.15 & -1.33 \\ -1.33 & 0.52 \end{pmatrix}$$

Table 2: Table of errors (%)

Data Type	LDA	Log Reg	Lin Reg	QDA
train	3.00	2.00	3.00	2.33
test	4.15	4.30	4.15	2.35

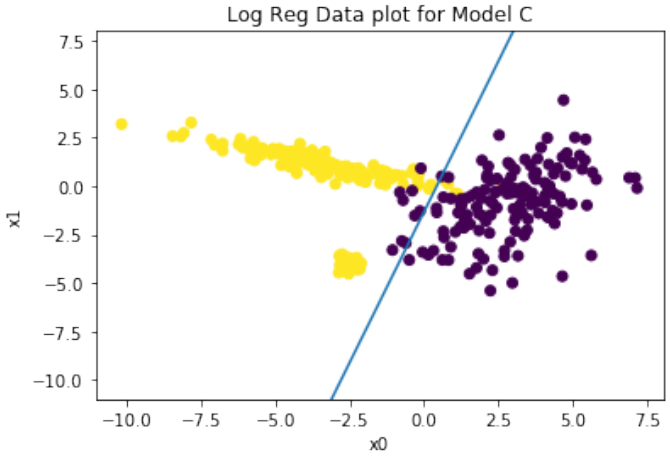
Comments: - Generative models: ici, les deux classes ont une covariance bien différente (hypothèse LDA moins acceptable). La classe $y = 0$ (violet) est plus étalée selon 2 axes alors que la classe $y = 1$ (jaune) et quasiment répartie uniquement selon un axe. On constate cela dans les valeurs de Σ_0 et Σ_1 qui sont très différentes. Sans surprise, le QDA donne donc de bien meilleurs résultats que le LDA.

- Conditional models: on constate que la régression linéaire est très sensible aux outliers. La classe y=1 (jaune) est très allongée vers la gauche ce qui décale la frontière vers la gauche du graphe. Néanmoins, sur le jeu de test la régression logistique ne performe pas mieux que la régression linéaire. La régression logistique est peut etre encore sujette à de l'overfitting.

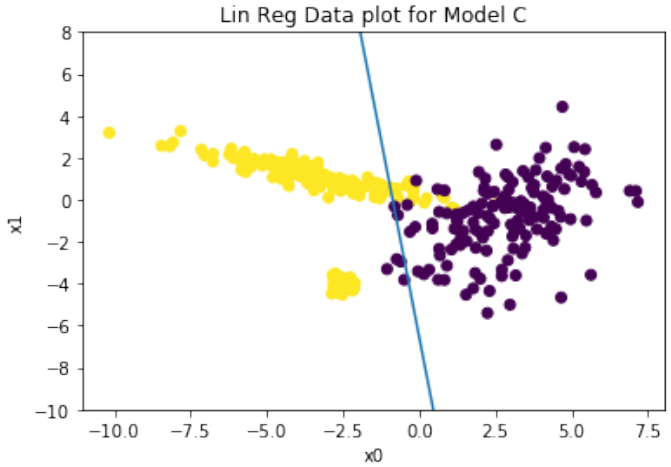


$$\hat{\pi} = 0.625 \quad \mu_0 = \begin{pmatrix} 2.79 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.94 \\ -0.96 \end{pmatrix}$$

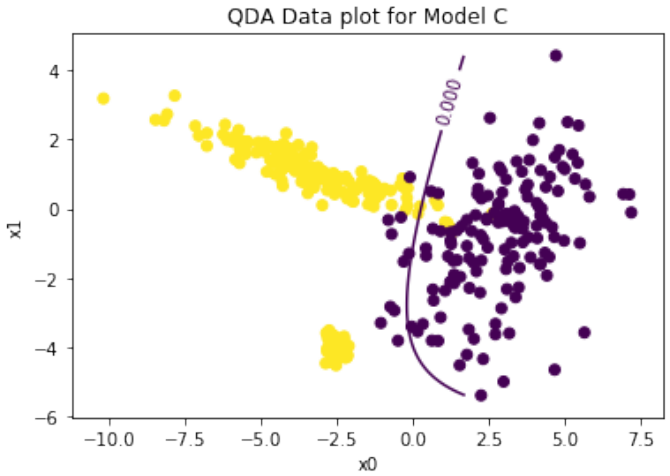
$$\Sigma = \begin{pmatrix} 2.88 & -0.63 \\ -0.63 & 5.20 \end{pmatrix}$$



$$\hat{w} = \begin{pmatrix} -2.203 \\ 0.709 \end{pmatrix} \quad b = 0.959$$



$$\hat{w} = \begin{pmatrix} -0.128 \\ -0.017 \end{pmatrix} \quad b = 0.508 \quad \sigma = 0.249$$



$$\hat{\pi} = 0.625 \quad \mu_0 = \begin{pmatrix} 2.79 \\ -0.84 \end{pmatrix} \quad \mu_1 = \begin{pmatrix} -2.94 \\ -0.96 \end{pmatrix}$$

$$\Sigma_0 = \begin{pmatrix} 2.90 & 1.25 \\ 1.25 & 2.92 \end{pmatrix} \quad \Sigma_1 = \begin{pmatrix} 2.87 & -1.76 \\ -1.76 & 6.56 \end{pmatrix}$$

Table 3: Table of errors (%)

Data Type	LDA	Log Reg	Lin Reg	QDA
train	5.25	4.00	5.50	5.25
test	4.33	2.27	4.233	4.03

Cette fois non seulement les distributions ont une covariance différente mais la classe $y = 1$ semble être la superposition de deux distributions. Aucun des modèles ne semble vraiment pertinent pour cette distribution non-triviale. En particulier, la frontière trouvée par le QDA semble relativement pertinente mais le $mu1$ calculé ne correspond pas à un barycentre pertinent (puisque qu'il y a "deux groupes").

4 PROOF Exercise 1: learning in discrete graphical models

Soit les observations $Y = (Y_1, ..., Y_n)$ avec

$$Y_i = \begin{pmatrix} X_i \\ Z_i \end{pmatrix}$$

alors

$$p(y) = \prod_i p(y_i) = \prod_i p(x_i|z_i)p(z_i) = \prod_i \theta_{z_i x_i} \pi_{z_i}$$

$$p(y) = \prod_{m \in \mathcal{M}} \pi_m^{n_m} \prod_{m \in \mathcal{M}, k \in \mathcal{K}} \theta_{mk}^{n_{mk}}$$

où $n_{mk} = \sum_i 1\{x_i = k, z_i = m\}$ = "Nombre de couples (k, m) dans les observations" et $n_m = \sum_i \sum_k n_{m,k} = \sum_i 1\{z_i = m\}$ = "Nombre d'observations où $z_i = m$" Alors, on veut maximiser

$$log(p(y)) = \sum_m n_m log(\pi_m) + \sum_{m,k} n_{mk} log(\theta_{mk})$$

sous les contraintes $\sum_m \pi_m = 1$ et $\sum_{m,k} \theta_{mk} = 1$ Posons donc le Lagrangien :

$$\mathcal{L}(\pi, \theta, \lambda, \mu) = \sum_m n_m log(\pi_m) + \sum_{m,k} n_{mk} log(\theta_{mk}) - \lambda(\sum_m \pi_m - 1) - \mu(\sum_{m,k} \theta_{mk} - 1)$$

En dérivant par rapport à π:

$$\frac{\partial \mathcal{L}}{\partial \pi_m} = \frac{n_m}{\pi_m} - \lambda = 0 \Rightarrow \pi_m = \frac{n_m}{\lambda}$$

or avec la contrainte

$$\sum_m \pi_m = 1 \Rightarrow \hat{\pi}_m = \frac{n_m}{n}$$

De même, en dérivant par rapport à θ:

$$\frac{\partial \mathcal{L}}{\partial \theta_{mk}} = \frac{n_{mk}}{\theta_{mk}} - \mu = 0 \Rightarrow \theta_{mk} = \frac{n_{mk}}{\mu}$$

or avec la contrainte

$$\sum_{m,k} \theta_{mk} = 1 \Rightarrow \hat{\theta}_{mk} = \frac{n_{mk}}{n}$$

5 PROOF Exercise 2.1: LDA Formulas

Pour une seule observation :

$$p(y|x) \propto p(x|y)p(y) = \pi^y(1-\pi)^{1-y}p(x|y)$$

$$p(y|x) \propto \pi^y(1-\pi)^{1-y}(det(\Sigma))^{-n/2}exp(-\frac{y}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1-y}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))$$

$$p(y|x) \propto \pi^y(1-\pi)^{1-y}(det(\Sigma))^{-n/2}\ exp(yx^T\Sigma^{-1}(\mu_1-\mu_0) + \frac{y}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))$$

Donc, pour n observations (X_i, Y_i) iid, on a :

$$log(p(y|x)) = log(\pi)\sum_i y_i + log(1-\pi)\sum_i (1-y_i) - \frac{n}{2}log(det(\Sigma)) + \sum_i y_i x_i^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1)\sum_i \frac{y_i}{2}$$

$$-\frac{1}{2}\sum_i (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0) + Cste$$

Comme cette expression est concave en π, μ_0, μ_1, et Σ, on cherche un maximum pour chacun des paramètre lorsque le gradient par rapport à ce paramètre vaut 0.

* Estimation de μ_0:

$$\nabla_{\mu_0}\Big(log(p(y|x))\Big) = -\sum_i y_i x_i^T\Sigma^{-1} + 2\mu_0^T\Sigma^{-1}\sum_i \frac{y_i}{2} + \sum_i (x_i-\mu_0)^T\Sigma^{-1} = 0$$

$$\Leftrightarrow \mu_0^T\sum_i y_i + \sum_i x_i^T - \mu_0^T\sum_i 1 = \sum_i y_i x_i^T$$

$$\Leftrightarrow \mu_0^T\sum_i (1-y_i) = \sum_i x_i^T(1-y_i)$$

$$\Leftrightarrow \hat{\mu}_0 = \frac{\sum_i x_i^T(1-y_i)}{\sum_i (1-y_i)}$$

donc l'estimateur du maximum de vraisemblance pour μ_0 est la moyenne empirique des x_i pour les observations où $y_i = 0$

* Estimation de μ_1:

$$\nabla_{\mu_1}\Big(log(p(y|x))\Big) = \sum_i y_i x_i^T\Sigma^{-1} - 2\mu_1^T\Sigma^{-1}\sum_i \frac{y_i}{2} = 0$$

$$\Leftrightarrow \sum_i y_i x_i = \mu_1\sum_i y_i$$

$$\Leftrightarrow \hat{\mu}_1 = \frac{\sum_i y_i x_i}{\sum_i y_i}$$

donc l'estimateur du maximum de vraisemblance pour μ_1 est la moyenne empirique des x_i pour les observations où $y_i = 1$

* Estimation de π:

$$\nabla_{\pi}\Big(log(p(y|x))\Big) = \frac{\sum_i y_i}{\pi} - \frac{\sum_i (1-y_i)}{1-\pi} = 0$$

$$\Leftrightarrow \hat{\pi} = \frac{\sum_i y_i}{n}$$

donc l'estimateur du maximum de vraisemblance de π est la moyenne empirique des y_i

* Estimation de Σ:

En posant $\Lambda = \Sigma^{-1}$ on peut réecrire, en ne gardant que les termes dépendant de Λ:

$$log(p(y|x)) = \frac{n}{2}log(det(\Lambda)) + \sum_i tr(y_i x_i^T\Lambda(\mu_1-\mu_0)) + (tr(\mu_0^T\Lambda\mu_0) - tr(\mu_1^T\Lambda\mu_1))\sum_i \frac{y_i}{2}$$

$$-\frac{1}{2}\sum_i tr((x_i-\mu_0)^T\Lambda(x_i-\mu_0)) + Cste$$

$$= \frac{n}{2}log(det(\Lambda)) + tr(\Lambda \sum_i y_i x_i (\mu_1 - \mu_0)^T) + (tr(\Lambda\, \mu_0 \mu_0^T) - tr(\Lambda \mu_1 \mu_1^T)) \sum_i \frac{y_i}{2} - \frac{1}{2} tr(\Lambda \sum_i (x_i - \mu_0)(x_i - \mu_0)^T) + Cste$$

or, pour A, B deux matrices carrées $\nabla_A tr(AB) = B^T$, et $\nabla_A \Big(log(det(A)) \Big) = A^{-1}$ d'où

$$\nabla_\Lambda \Big(log(p(y|x)) \Big) = \frac{n}{2}\Sigma + \sum_i y_i x_i^T (\mu_1 - \mu_0) + (\mu_0^T \mu_0 - \mu_1^T \mu_1) \sum_i \frac{y_i}{2} - \frac{1}{2} \sum_i (x_i - \mu_0)^T (x_i - \mu_0) = 0$$

$$\Leftrightarrow \Sigma = \frac{1}{n} (\sum_i y_i (-2 x_i^T \mu_1 + \mu_1^T \mu_1 + x_i^T x_i - x_i^T x_i) + \sum_i y_i (2 x_i^T \mu_0 - \mu_0^T \mu_0 + x_i^T x_i - x_i^T x_i) + \sum_i (x_i - \mu_0)^T (x_i - \mu_0))$$

$$\Leftrightarrow \Sigma = \frac{1}{n} (\sum_i y_i (x_i - \mu_1)^T (x_i - \mu_1) - \sum_i y_i (x_i - \mu_0)^T (x_i - \mu_0) + \sum_i (x_i - \mu_0)^T (x_i - \mu_0))$$

$$\Leftrightarrow \hat{\Sigma} = \frac{1}{n} (\sum_i y_i (x_i - \mu_1)^T (x_i - \mu_1) + \sum_i (1 - y_i) (x_i - \mu_0)^T (x_i - \mu_0))$$

* Calcul de $p(y = 1|x)$:

$$p(y = 1|x) = \frac{p(x|y = 1)p(y = 1)}{p(x|y = 1)p(y = 1) + p(x|y = 0)p(y = 0)} = \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}$$

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

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

avec $a = (\mu_1 - \mu_0)^T \Sigma^{-1}$ et $b = \frac{1}{2}(\mu_1^T \Sigma^{-1} \mu_1 - \mu_0^T \Sigma^{-1} \mu_0) + log(\frac{1-\pi}{\pi})$

Ainsi

$$p(y = 1|x) = 0.5$$
$$\Leftrightarrow a^T x + b = 0$$

6 PROOF Exercise 2.5: QDA Formulas

* Estimation de μ_0 (même calcul que le LDA):

$$\hat{\mu}_0 = \frac{\sum_i x_i^T (1 - y_i)}{\sum_i (1 - y_i)}$$

* Estimation de μ_1 (même calcul que le LDA):

$$\hat{\mu}_1 = \frac{\sum_i y_i x_i}{\sum_i y_i}$$

* Estimation de π(même calcul que le LDA):

$$\hat{\pi} = \frac{\sum_i y_i}{n}$$

* Estimation de Σ_0: On pose $\Lambda_0 = \Sigma_0^{-1}$, et par un calcul similaire à celui du LDA on a:

$$\nabla_{\Lambda_0} \Big(log(p(y|x)) \Big) = 0$$

$$\Leftrightarrow \frac{\sum_i (1 - y_i)}{2} \Sigma_0 - \sum_i \frac{(1 - y_i)}{2} (x_i - \mu_0)^T (x_i - \mu_0) = 0$$

$$\Leftrightarrow \hat{\Sigma}_0 = \frac{\sum_i (1 - y_i)(x_i - \mu_0)^T (x_i - \mu_0)}{\sum_i (1 - y_i)}$$

* Estimation de Σ_1 (idem):

$$\hat{\Sigma}_1 = \frac{\sum_i y_i (x_i - \mu_1)^T (x_i - \mu_1)}{\sum_i y_i}$$

* Calul de $p(y=1|x)$:

$$p(y=1|x) = \frac{1}{1+\frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}$$

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

Donc

$$p(y=1|x) = 0.5$$

$$\Leftrightarrow \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + log(\frac{1-\pi}{\pi}) = 0$$

$$\Leftrightarrow x^T(\Sigma_1^{-1}-\Sigma_0^{-1})x - 2x^T(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0) + \mu_1\Sigma_1^{-1}\mu_1 - \mu_0\Sigma_0^{-1}\mu_0 + 2log(\frac{1-\pi}{\pi}) = 0$$

qui est bien une forme quadratique si $\Sigma_0 \neq \Sigma_1$, ce qui est l'hypothèse du QDA. Autrement, on retombe sur le cas du LDA question 1).