Homework 1 PGM

Sharone Dayan, Michael Sutton

October 2017

1 Learning in Discrete Graphical Model

Considérons le modèle suivant :
Soit Z et X deux variables discrètes qui prennent respectivement M et K valeurs tel que :

$$\begin{aligned} p(Z=m) &= \pi_m \qquad \forall m \in [\![1,\dots,M]\!] \\ p(X=k|Z=m) &= \theta_{mk} \qquad \forall k \in [\![1,\dots,K]\!], \forall m \in [\![1,\dots,M]\!] \end{aligned}$$

Données : On a un échantillon $(z^{(1)},x^{(1)}),\dots,(z^{(N)},x^{(N)})$ d'observations i.i.d du couple (Z,X).

Posons Y $=(Y_1,\dots,Y_M)$ avec $Y_m = \mathbb{1}_{Z=m}$ on a :
L'évènement $\{Z=m\}$ correspond à $\{Y_m = 1 \; et \; Y_l = 0 \;\; \forall l \neq k\}$, $Y \in \{0,1\}^K$.

$$p(Z=m) = p(Y_m=1) = \pi_m \qquad \forall m \in [\![1,\dots,M]\!]$$

$$\boxed{p(Y=y) = \prod_{m=1}^{M} \pi_m^{y_m}}$$

Posons $B = \begin{pmatrix} B_1^T \\ \vdots \\ B_M^T \end{pmatrix}$ avec $(B_m)_k = \mathbb{1}_{X=k|Z=m}$.

L'évènement $\{X=k|Z=m\}$ correspond à $\{B_{mk}=1 \; et \; B_{ij}=0 \;\; \forall (i,j)\neq(m,k)\}$, $B \in \{0,1\}^{M\times K}$, ainsi concernant B et Z on a les relations :

$$p(X=k|Z=m) = p(B_{mk}=1) = \theta_{mk} \qquad \forall k \in [\![1,\dots,K]\!], \forall m \in [\![1,\dots,M]\!]$$

$$\boxed{p(B=b) = \prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{mk}^{b_{mk}}}$$

Ecrivons la fonction de vraisemblance :

$$\begin{aligned}
L((\pi,\theta)) &= p_{(\pi,\theta)}((Z^{(1)},X^{(1)})=(z^{(1)},x^{(1)}),\dots,(Z^{(N)},X^{(N)})=(z^{(N)},x^{(N)})) \\
&= \prod_{n=1}^{N} p_{(\pi,\theta)}((Z^{(n)},X^{(n)})=(z^{(n)},x^{(n)})) \qquad \text{par indépendance} \\
&= \prod_{n=1}^{N} p(X^{(n)}=x^{(n)}|Z^{(n)}=z^{(n)})\, p(Z^{(n)}=z^{(n)}) \\
&= \prod_{n=1}^{N} p(B^{(n)}=b^{(n)})\, p(Y^{(n)}=y^{(n)}) \\
&= \prod_{n=1}^{N} \Big(\prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{mk}^{b_{mk}^{(n)}} \Big) \Big(\prod_{m\prime=1}^{M} \pi_{m\prime}^{y_{m\prime}^{(n)}} \Big) \\
&= \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{mk}^{b_{mk}^{(n)}} \pi_m^{y_m^{(n)}}
\end{aligned}$$

En remarquant que les cas ou $\pi_i = 0$ (resp. $\theta_{mk} = 0$) entraînerai que $y_i = 0$ sur toutes les observations (resp. $B_{mk} = 0$ sur toutes les observations), les valeurs correspondantes dans le produit valent 1 et "disparaissent". On en conclu donc que la fonction de log vraisemblance est bien définie et vaut :

$$\begin{aligned}
l((\pi,\theta)) &= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} (b_{mk}^{(n)} \log(\theta_{mk}) + y_m^{(n)} \log(\pi_m)) \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n=1}^{N} (b_{mk}^{(n)} \log(\theta_{mk}) + y_m^{(n)} \log(\pi_m))
\end{aligned}$$

En notant $n_m = \sum_{n=1}^{N} y_m^{(n)}$, qui correspond au nombre d'observations de Z prenant la valeur j, et en notant $n_{mk} = \sum_{n=1}^{N} b_{mk}^{(n)}$ qui correspond au nombre d'observations de (Z, X) prenant la valeur (m, k) on a :

$$l((\pi,\theta)) = \sum_{m=1}^{M}\sum_{k=1}^{K} (n_{mk} \log(\theta_{mk}) + n_m \log(\pi_m))$$

L'objectif est donc de maximiser la fonction de log vraisemblance $l((\pi,\theta))$ sous les contraintes que $\sum_{m=1}^{M} \pi_m = 1$ et $\forall m \ \sum_{k=1}^{K} \theta_{mk} = 1$, ou en d'autres termes :

$$\min_{\substack{\pi_m \geq 0 \\ \theta_{mk} \geq 0}} f((\pi,\theta)) = -l((\pi,\theta)) = \sum_{m=1}^{M}\sum_{k=1}^{K} (-n_{mk} \log(\theta_{mk}) - n_m \log(\pi_m)) \qquad s.c \quad 1^T\pi = 1 \in \mathbb{R} \ \ and \ \ 1^T\theta = 1 \in \mathbb{R}^M$$

Le Lagrangien de ce problème donne :

$$\mathcal{L}(\pi,\theta,\lambda) = -l((\pi,\theta)) + \lambda_0(\sum_{m=1}^{M} \pi_m - 1) + \sum_{m=1}^{M} (\lambda_m(\sum_{k=1}^{K} \theta_{mk} - 1))$$

Il est évident que $n_i \geq 0$ car $y_i \geq 0$ donc f est convexe comme somme de fonction convexe. De plus l'ensemble $\{\pi_m \geq 0, \theta_{mk} \geq 0 \ , \forall m \in [\![1,\dots,M \ \forall k \in [\![1,\dots,K]\!]\}$ est convexe, il s'agit d'un problème d'optimisation convexe. Les contraintes sont linéaires, et il existe $\pi_1, \pi_2, ..., \pi_M \ \ tq \ \ \sum_{i=1}^{M} \pi_i = 1$ (par exemple $\pi_i = 1/M$), donc d'après la qualification de contraintes de Slater, le problème a la propriété de forte dualité. Ainsi :

$$\min_{\substack{\pi_m \geq 0 \\ \theta_{mk} \geq 0}} f((\pi,\theta)) = \max_{\lambda} \min_{\substack{\pi_m \geq 0 \\ \theta_{mk} \geq 0}} \mathcal{L}(\pi,\theta,\lambda)$$

Comme $\mathcal{L}(\pi, \theta, \lambda)$ est convexe par rapport à π, le minimum se trouve en annulant la dérivée de $\mathcal{L}(\pi, \lambda)$ par rapport à π.
Ainsi, on obtient :

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = -\frac{K\, n_i}{\pi_i} + \lambda_0 = 0 \qquad \forall i \in [\![1, \dots, M]\!]$$

Donc :

$$\pi_i \lambda_0 = K n_i \qquad \forall i \in [\![1, \dots, M]\!]$$

En appliquant la contrainte $\sum_{i=1}^{M} \pi_i = 1$, on obtient $\lambda_0 = K \sum_{i=1}^{M} n_i$, d'où $\lambda_0 = KN$ ($\neq$ 0) avec N le nombre d'observations.
On obtient finalement :

$$\widehat{\pi_i} = \frac{K n_i}{\lambda_0} = \frac{\sum_{j=1}^{n} y_i^{(j)}}{N} = \frac{n_i}{N} \qquad \forall i \in [\![1, \dots, M]\!]$$

Pour le maximum de vraisemblance en θ, comme $\mathcal{L}(\pi, \theta, \lambda)$ est convexe par rapport à θ, le minimum se trouve en annulant la dérivée de $\mathcal{L}(\pi, \theta, \lambda)$ par rapport à θ.
Ainsi, on obtient :

$$\frac{\partial \mathcal{L}}{\partial \theta_{mk}} = -\frac{n_{mk}}{\theta_{mk}} + \lambda_m = 0 \qquad \forall m \in [\![1, \dots, M]\!], \forall k \in [\![1, \dots, K]\!]$$

En appliquant la contrainte $\sum_{k=1}^{K} \theta_{mk} = 1$, on obtient :

$$\lambda_m = \sum_{k=1}^{K} n_{mk} = \sum_{k=1}^{K} \sum_{n=1}^{N} b_{mk}^{(n)} = \sum_{n=1}^{N} \sum_{k=1}^{K} b_{mk}^{(n)} = \sum_{n=1}^{N} y_m^{(n)} = n_m.$$

Il faut faire attention au cas ou $n_m = 0$ qui correspond au cas ou aucun $Z^{(i)}$ de nos observations ne prend la valeur m. Comme il existe au moins un $n_{m0} \neq 0$, on peut poser alors pour tous les m tel que $n_m = 0$ $\theta_{mk} = 0$ $\forall k$, les autres valeurs de θ permettrons de satisfaire la contrainte. Intuitivement on se dit que si on n'a pas observer la valeur m alors la probabilité d'observer l'évènement $\{X = m | Z = k\}$ est nulle.
Finalement :

$$\widehat{\theta_{mk}} = \frac{n_{mk}}{\lambda_m} = \frac{n_{mk}}{n_m} \qquad \forall m \in [\![1, \dots, M]\!], \forall k \in [\![1, \dots, K]\!]$$

En conclusion, l'estimateur du maximum de vraisemblance de ce modèle est
$(\widehat{\pi}^{MLE}, \widehat{\theta}^{MLE})$ avec :

$$\boxed{\begin{array}{ll} \widehat{\pi}_m^{MLE} = \dfrac{n_m}{n} = \dfrac{\sum_{n=1}^{N} y_m^{(n)}}{n} & \forall i \in [\![1, \dots, M]\!] \\ \widehat{\theta}_{mk}^{MLE} = \dfrac{n_{mk}}{n_m} = \dfrac{\sum_{n=1}^{N} b_{mk}^{(n)}}{\sum_{n=1}^{N} y_m^{(n)}} & \forall m \in [\![1, \dots, M]\!], \forall k \in [\![1, \dots, K]\!] \end{array}}$$

2 Linear classification

1. **Generative model (LDA)**.
   Renommons le paramètre π en θ pour ne pas qu'il y ait de confusion avec le nombre π apparaissant dans la densité de la loi normale. On a donc le modèle suivant :

$$Y \sim Bernoulli(\theta), \quad X | \{Y = i\} \sim \mathcal{N}(\mu_i, \Sigma)$$

(a) Maximum de vraissemblance :
On suppose qu'on a un echantillon i.i.d de taille N d'observations du couple (Y, Z), la fonction de vraissemblance s'écrit :

$$\begin{aligned} L((\theta,\mu,\Sigma)) &= p_{(\theta,\mu,\Sigma)}(((Y^{(1)},X^{(1)}) = (y^{(1)},x^{(1)}),\ldots,(Y^{(N)},X^{(N)}) = (y^{(N)},x^{(N)})) \\ &= \prod_{n=1}^{N} p_{(\theta,\mu,\Sigma)}((Y^{(n)},X^{(n)}) = (y^{(n)},x^{(n)})) \qquad \text{par independance} \\ &= \prod_{n=1}^{N} p(Y^{(n)} = y^{(n)})\, p(X^{(n)} = x^{(n)} | Y^{(n)} = y^{(n)}) \end{aligned}$$

$$\begin{aligned} = \prod_{n=1}^{N} \theta^{y^{(n)}} (1-\theta)^{(1-y^{(n)})} &\left(\frac{1}{2\pi^{d/2}\sqrt{|\det(\Sigma)|}} \exp(-\frac{1}{2}((x^{(n)}-\mu_0)^T\Sigma^{-1}(x^{(n)}-\mu_0))) \right)^{(1-y^{(n)})} \\ &\times \left(\frac{1}{2\pi^{d/2}\sqrt{|\det(\Sigma)|}} \exp(-\frac{1}{2}((x^{(n)}-\mu_1)^T\Sigma^{-1}(x^{(n)}-\mu_1))) \right)^{y^{(n)}} \end{aligned}$$

En remarquant que si θ vaut 0 (resp. 1), les observations $y^{(i)}$ sont alors tous egaux à 0 (resp. 1) et les termes $\theta^{y^{(n)}}$ (resp. $(1-\theta)^{(1-y^{(n)})}$) valent 1 et "disparaissent" du produit. On a donc la log vraissemblance qui est bien définie et qui vaut :

$$\begin{aligned} l((\theta,\mu_0,\mu_1,\Sigma)) = \sum_{n=1}^{N} & y^{(n)}\log(\theta) \;+\; (1-y^{(n)})\log(1-\theta) \;-\; \frac{1}{2}\log(|\det(\Sigma)|) \\ &+ \left((1-y^{(n)})(-\frac{1}{2}((x^{(n)}-\mu_0)^T\Sigma^{-1}(x^{(n)}-\mu_0))) \right) \\ &+ \left(y^{(n)}(-\frac{1}{2}((x^{(n)}-\mu_1)^T\Sigma^{-1}(x^{(n)}-\mu_1))) \right) + cste) \end{aligned}$$

La fonction de vraissemblance est concave en θ on trouve le maximum en annulant la dérivée. En dérivant la fonction de log vraisemblance par rapport à θ et en annulant la dérivée, on obtient :

$$\begin{aligned} \frac{\partial l}{\partial \theta} = \sum_{n=1}^{N} (\frac{y^{(n)}}{\theta} - \frac{(1-y^{(n)})}{1-\theta}) = 0 \\ \sum_{n=1}^{N} ((1-\theta) * y^{(n)} - \theta * (1-y^{(n)})) = 0 \\ \sum_{n=1}^{N} (y^{(n)} - \theta) = 0 \end{aligned}$$

Finalement :

$$\boxed{\widehat{\theta}^{MLE} = \frac{\sum_{n=1}^{N} y^{(n)}}{N}}$$

Puisque Σ est symétrique définie positive alors Σ^{-1} l'est aussi et donc la fonction de vraisemblance est concave en μ_1. En dérivant la fonction de log vraisemblance par rapport à μ_1 et en annulant la dérivée, on obtient :

$$\frac{\partial l}{\partial \mu_1} = \sum_{n=1}^{N} \frac{-y^{(n)}}{2} * (-2)\Sigma^{-1}(x^{(n)}-\mu_1) = \Sigma^{-1}\sum_{n=1}^{N} y^{(n)}(x^{(n)}-\mu_1) = 0$$

Finalement :

$$\boxed{\widehat{\mu_1}^{MLE} = \frac{\sum_{n=1}^{N} y^{(n)} x^{(n)}}{\sum_{n=1}^{N} y^{(n)}}}$$

Par un calcul similaire, on a :

$$\boxed{\widehat{\mu_0}^{MLE} = \frac{\sum_{n=1}^{N} (1-y^{(n)}) x^{(n)}}{\sum_{n=1}^{N} (1-y^{(n)})}}$$

Enfin, nous voudrions calculer l'estimateur du maximum de vraisemblance de Σ. Reprenons la formulation de la log vraisemblance $l((\theta, \mu, \Sigma))$:

$$l((\theta, \mu_0, \mu_1, \Sigma)) = g(\theta, \mu_0, \mu_1) - \frac{N}{2}\log(|\det(\Sigma)|) - \frac{1}{2}\sum_{n=1}^{N}\left((1-y^{(n)})(x^{(n)} - \mu_0)^T \Sigma^{-1}(x^{(n)} - \mu_0)\right) - \frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)}(x^{(n)} - \mu_1)^T \Sigma^{-1}(x^{(n)} - \mu_1)\right)$$

Les deux derniers termes sont réels, ils sont donc égaux à leur trace.

$$l((\theta, \mu_0, \mu_1, \Sigma)) = g(\theta, \mu_0, \mu_1) - \frac{N}{2}\log(|\det(\Sigma)|)$$
$$-\text{Trace}\left(\frac{1}{2}\sum_{n=1}^{N}\left((1-y^{(n)})(x^{(n)} - \mu_0)^T \Sigma^{-1}(x^{(n)} - \mu_0)\right) - \frac{1}{2}\sum_{n=1}^{N}\left(y^{(n)}(x^{(n)} - \mu_1)^T \Sigma^{-1}(x^{(n)} - \mu_1)\right)\right)$$

On peut donc s'intéresser uniquement aux termes diagonaux des produits matriciels qui sont dans la trace.

$$l((\theta, \mu_0, \mu_1, \Sigma)) = g(\theta, \mu_0, \mu_1) - \frac{N}{2}\log(|\det(\Sigma)|)$$
$$-\frac{1}{2}\text{Trace}\left(\Sigma^{-1}\left(\sum_{n=1}^{N}(1-y^{(n)})(x^{(n)} - \mu_0)(x^{(n)} - \mu_0)^T + \sum_{n=1}^{N} y^{(n)}(x^{(n)} - \mu_1)(x^{(n)} - \mu_1)^T\right)\right)$$

Posons :

$$A = \Sigma^{-1}$$
$$\widetilde{\Sigma_0} = \sum_{n=1}^{N}(1-y^{(n)})(x^{(n)} - \mu_0)(x^{(n)} - \mu_0)^T$$
$$\widetilde{\Sigma_1} = \sum_{n=1}^{N} y^{(n)}(x^{(n)} - \mu_1)(x^{(n)} - \mu_1)^T$$

Considérons :

$$f(A) = \frac{1}{2}\text{Trace}(A(\widetilde{\Sigma_0} + \widetilde{\Sigma_1}))$$
$$f(A+H) - f(A) = \frac{1}{2}\text{Trace}(H(\widetilde{\Sigma_0} + \widetilde{\Sigma_1}))$$

Ainsi le gradient de f vaut :

$$\nabla f(A) = \frac{(\widetilde{\Sigma_0} + \widetilde{\Sigma_1})}{2}$$

et

$$df_A(H) = \text{Trace}(\nabla f(A)^T H)$$

En ce qui concerne le terme en log :
On a :

$$\begin{aligned}\log(det(A+H)) &= \log(|A^{\frac{1}{2}}(I + A^{-\frac{1}{2}}HA^{-\frac{1}{2}})A^{-\frac{1}{2}}|) \\ &= \log(|A|) + \log(\det(I + \widetilde{H}))\end{aligned}$$

En posant $\widetilde{H} = A^{-\frac{1}{2}}HA^{-\frac{1}{2}}$
Considérons :

$$g(A) = \log(|\det(A)|) \text{ avec } A = I + \widetilde{H}$$

On a donc :

$$\begin{aligned}\log(\det(I + \widetilde{H})) - \log(\det(I)) &= \sum_{i=j}^{d} \log(1 + \lambda_j) \\ &\simeq \sum_{i=j}^{d} \lambda_j + o(||\widetilde{H}||) \\ &= \text{Trace}(\widetilde{H}) + o(||\widetilde{H}||)\end{aligned}$$

Comme $\widetilde{H}$ est symmétrique, elle peut s'écrire de la sorte :

$$\widetilde{H} = U\Lambda U^T$$

où U est une matrice orthogonale et $\Lambda = \text{diag}(\lambda_1, ..., \lambda_d)$

$$d(\log(\det{}_A(H))) = \text{Trace}(A^{-\frac{1}{2}}HA^{-\frac{1}{2}}) = \text{Trace}(HA^{-1})$$

$$\nabla(\log(\det(A))) = A^{-1}$$

Finalement, le gradient de la fonction de log vraisemblance par rapport à $A = \Sigma^{-1}$ s'écrit :

$$\nabla_A(l) = \frac{1}{2}A^{-1} - \frac{N}{2}(\widetilde{\Sigma_0} + \widetilde{\Sigma_1})$$

car $\log(\det(A)) = -\log(\det(\Sigma))$

$$\nabla_A(l) = 0 \qquad \text{ssi} \qquad \widehat{\Sigma}^{MLE} = \frac{(\widetilde{\Sigma_0} + \widetilde{\Sigma_1})}{N}$$

avec :

$$\widetilde{\Sigma_0} = \sum_{n=1}^{N}(1-y^{(n)})(x^{(n)}-\mu_0)(x^{(n)}-\mu_0)^T$$
$$\widetilde{\Sigma_1} = \sum_{n=1}^{N}y^{(n)}(x^{(n)}-\mu_1)(x^{(n)}-\mu_1)^T$$

$$\boxed{\widehat{\Sigma}^{MLE} = \frac{1}{N}\left(\sum_{n=1}^{N}(1-y^{(n)})(x^{(n)}-\mu_0)(x^{(n)}-\mu_0)^T) + \sum_{n=1}^{N}y^{(n)}(x^{(n)}-\mu_1)(x^{(n)}-\mu_1)^T\right)}$$

(b) Distribution conditionnelle :

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

$$= \frac{\theta N(x|\mu_1,\Sigma)}{(1-\theta)N(x|\mu_0,\Sigma)+\theta N(x|\mu_1,\Sigma)}$$

$$= \frac{1}{1+\frac{1-\theta}{\theta}\frac{N(x|\mu_0,\Sigma)}{N(x|\mu_1,\Sigma)}}$$

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

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

Ainsi, en faisant le parallèle avec la forme de la régression logistique, on a :

$$p(y=1|x) = \frac{1}{1+\exp(-[\log(\frac{\theta}{1-\theta})-\frac{1}{2}(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1+\mu_0)+(\mu_1-\mu_0)^T\Sigma^{-1}x^{(n)}])}$$

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

avec la fonction sigmoid :

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

et :

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

Dans le cas de la régression logistique, on retrouve que $p(y=1|x)$ a la même forme c'est a dire σ(fonction linéaire). Néanmoins dans le cadre de la régression logistique on ne suppose pas a priori que $Y|X$ suit une loi normale. Cette supposition entraîne dans le cas LDA une forme particulière de la fonction linéaire dépendant des μ et Σ.

(c) Implementation :

En implémentant les estimateurs du maximum de vraisemblance calculés mathématiquement plus haut ; et en les appliquant aux sets de données A, B et C, on trouve les valeurs numériques suivantes (on prend les 4 premières décimales) :

	θ	μ_0	μ_1	Σ	a	b
A	0.3333	$\begin{pmatrix} 2.8997 \\ -0.8938 \end{pmatrix}$	$\begin{pmatrix} -2.6923 \\ 0.8660 \end{pmatrix}$	$\begin{pmatrix} 2.4419 & -1.1319 \\ -1.1319 & 0.6137 \end{pmatrix}$	$\begin{pmatrix} -6.6224 \\ -9.3462 \end{pmatrix}$	-0.1364
B	0.5	$\begin{pmatrix} 3.3406 \\ -0.8354 \end{pmatrix}$	$\begin{pmatrix} -3.2167 \\ 1.0830 \end{pmatrix}$	$\begin{pmatrix} 3.3462 & -0.1351 \\ -0.1351 & 1.7380 \end{pmatrix}$	$\begin{pmatrix} -1.9210 \\ 0.9544 \end{pmatrix}$	0.0009
C	0.625	$\begin{pmatrix} 2.7930 \\ -0.8383 \end{pmatrix}$	$\begin{pmatrix} -2.9423 \\ -0.9578 \end{pmatrix}$	$\begin{pmatrix} 2.8803 & -0.6340 \\ -0.6340 & 5.1995 \end{pmatrix}$	$\begin{pmatrix} -2.0512 \\ -0.2731 \end{pmatrix}$	0.1124

Table 1 – LDA Learnt parameters

Représentons graphiquement des données de train et de test, ainsi que la ligne définie par $p(y = 1|x) = 0.5$:

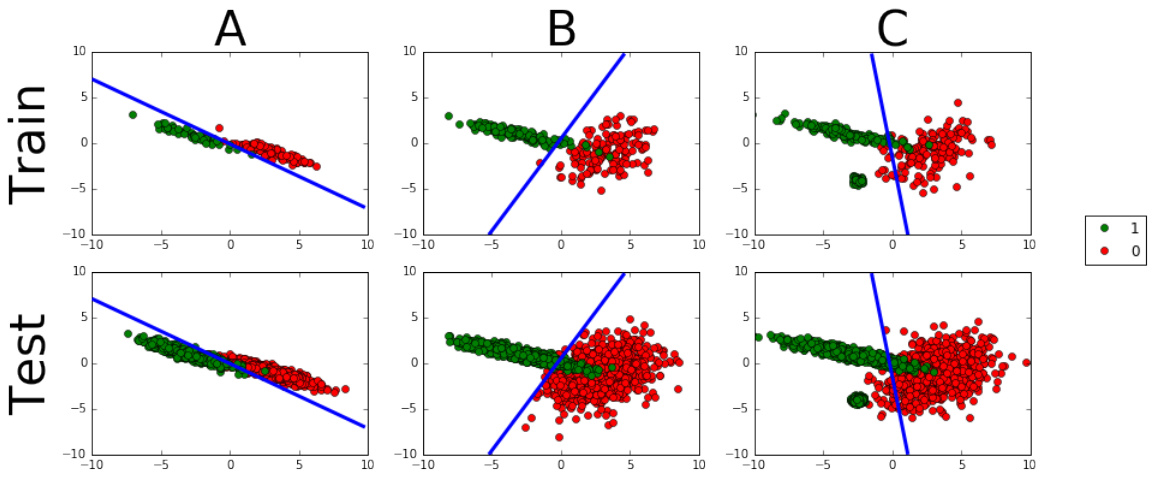

Figure 1 – LDA results

Les graphes de cette figure représentent les vrai valeurs de y les 1 en vert et 0 en rouge. La ligne bleu déterminée par les données de training de chaque sous ensemble est tracée également sur le graphe du test correspondant. Elle représente l'ensemble des points ou le critère ne sait quelle valeur attribuer a y.

2. **Régression logistique**
En reprenant les calculs fait en cours on obtient la formule itérative pour ω :

$$\omega^{(t+1)} = w^{(t)} + (X^T D_{\eta^{(t)}} X) X^T (y - \eta^{(t)})$$

Néanmoins cette méthode laisse ω prendre des valeurs assez élevées. Pour palier à ce problème, on ajoute un terme de pénalisation dans la log vraisemblance ($-\frac{\lambda}{2}\|\omega\|$) cela donne la relation suivante :

$$\omega^{(t+1)} = w^{(t)} + (X^T D_{\eta^{(t)}} X - \lambda I)(X^T (y - \eta^{(t)}) - \lambda \omega^{(t)})$$

En appliquant cette méthode de descente, on obtient numériquement :

	ω	b
A	$\begin{pmatrix} -8.33972193 \\ -13.57796665 \end{pmatrix}$	-1.49528825
B	$\begin{pmatrix} -1.7036036 \\ 1.02265864 \end{pmatrix}$	1.34676814
C	$\begin{pmatrix} -2.20181115 \\ 0.70815684 \end{pmatrix}$	0.95742125

Table 2 – Linear regression learnt parameters

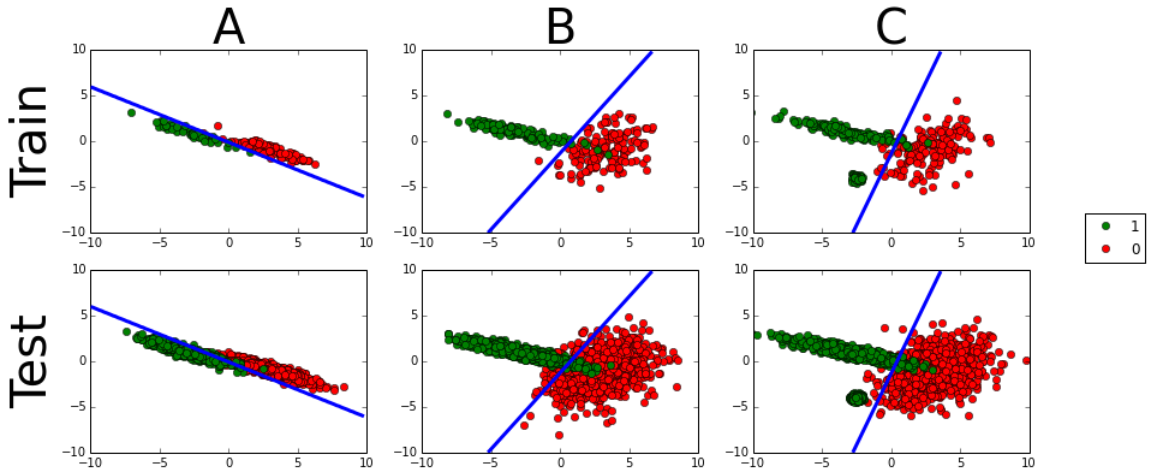

Figure 2 – Logistic regression results

Les graphes de cette figure représentent les vrai valeurs de y les 1 en vert et 0 en rouge. La ligne bleu déterminée par les données de training de chaque sous ensemble est tracée également sur le graphe du test correspondant. Elle représente l'ensemble des points ou le critère ne sait quelle valeur attribuer a y.

3. **Régression linéaire**

(a) En reprenant le cours sur la régression linéaire dans le cas $b = 0$, l'objectif est de résoudre la "normal equation" :

$$X^T X\omega - X^T y = 0$$

Ainsi, si $X^T X$ est inversible, on obtient :

$$\boxed{\widehat{\omega}^{MLE} = (X^T X)^{-1} X^T y}$$

(b) Pour l'implémentions, on souhaite un $b \neq 0$, pour ce faire on l'inclut dans ω comme ceci.

$$\begin{aligned} Y|X &\sim \mathcal{N}(\omega^T X + b, \sigma^2) \\ Y &= \omega^T X + b + \epsilon \qquad \text{avec} \qquad \epsilon \sim \mathcal{N}(0, \sigma^2) \\ Y &= \widetilde{\omega}^T \begin{pmatrix} x \\ 1 \end{pmatrix} + \epsilon \end{aligned}$$

On s'est alors ramené a un cas ou le $\widehat{b} = 0$, on peut donc appliquer la formule, puis on extrait la dernière coordonné de $\widetilde{\omega}$ pour avoir les ω et b cherché. Numériquement, on obtient :
On a trouver ω et b pour que $y \simeq \omega^T x + b$, donc pour trouver $p(y = 1|x_0) = 0.5$ on évalue $\simeq \omega^T x_0 + b$ et on applique une stratégie de seuil. Si cette valeur est plus grande que 0.5 on attribue a y la valeur 1 sinon 0. On obtient les graphiques des données de train et de test, suivant :

	ω	b
A	$\begin{pmatrix} -0.2640 \\ -0.3725 \end{pmatrix}$	0.4922
B	$\begin{pmatrix} -0.1042 \\ 0.0517 \end{pmatrix}$	0.5001
C	$\begin{pmatrix} -0.1276 \\ -0.0170 \end{pmatrix}$	0.5083

Table 3 – Linear regression learnt parameters

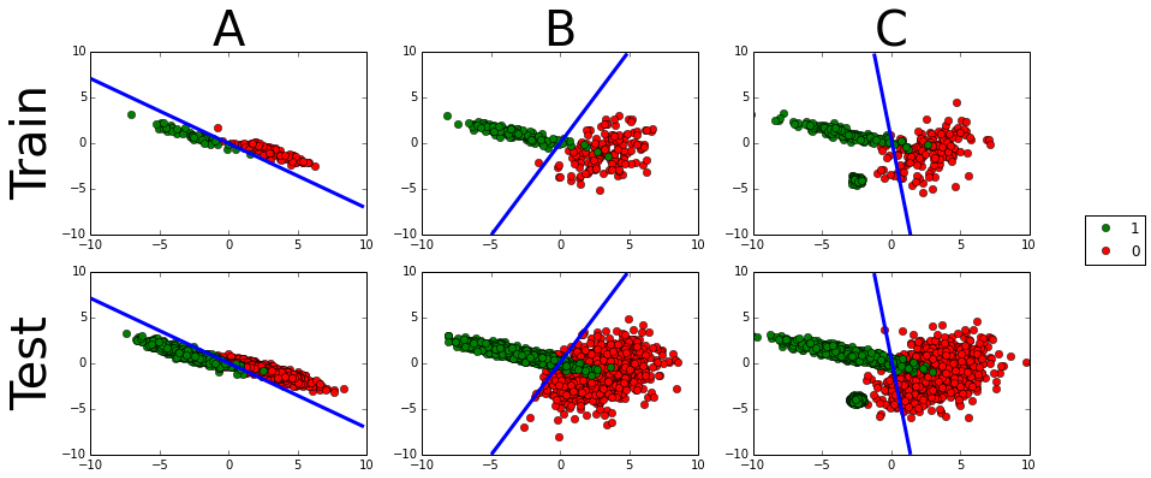

Figure 3 – Linear regression results

Les graphes de cette figure représentent les vrai valeurs de y les 1 en vert et 0 en rouge. La ligne bleu déterminée par les données de training de chaque sous ensemble est tracée également sur le graphe du test correspondant. Elle représente l'ensemble des points ou le critère ne sait quelle valeur attribuer a y.

4. **Commentaires et interprétations**

(a) Calculons l'erreur de misclassification, i.e la fraction de données mal classifiées pour les données de train et de test :

	A˙Train	A˙Test	B˙Train	B˙Test	C˙Train	C˙Test
LDA	1.33%	2.00%	3.0%	4.15%	5.5%	4.23%
Regression logistique	0.67%	2.53%	2.0%	4.3%	4.0%	2.27%
Regression linéaire	1.33%	2.07%	3.0%	4.15%	5.5%	4.23%

(b) Les trois méthodes présentées sont des méthodes permettant de faire de la classification avec un critère linéaire. Cela marche bien lorsque l'on a des données qui peuvent être séparé linéairement. On observe cela car ces méthodes performent mieux sur le dataset A que sur le B et le C.
Remarquons que l' erreur de misclassification dans le Dataset C est plus importante pour le set de Train que pour celui de Test ce qui est surprenant à premier abord. Cela nous assure au moins qu'on "overfit" pas. Quand on regarde la forme de nos data pour ce Dataset , on remarque que les deux classes sont assez melangé sur une partie et de plus il y a groupement de points, séparés des autres, représentés par les y=1 (en vert), ce qui rend les méthodes de séparation linéaire peu performantes (les erreurs de misclassification pour le dataset C sont les plus élevées).

Le fait d'utiliser la régression linéaire pour un problème de classification binaire est assez intuitif car on modélise notre variable discrète par une variable continue puis on applique un seuil.

Néanmoins, elle fournit des résultats extrêmement similaires à la LDA. On peut comprendre cela dataset par dataset :

- Le A est linéairement séparable et semble vérifier les hypothèse de la LDA
- Pour les deux autres datasets, ces hypothèses semblent être invalidés

La régression logistique est celle qui performe globalement le mieux. Les hypothèses qu'elle suppose sur les données sont moins strictes que les autres méthode et c'est pourquoi elle a de meilleur résultats.

5. **QDA model**

(a) Renomons le parametre π en θ pour ne pas qu'il y ait de confusion avec le nombre π apparaissant dans la densité de la loi normale. On a donc le modèle suivant :

$$Y \sim Bernoulli(\theta), \quad X|\{Y=i\} \sim \mathcal{N}(\mu_i, \Sigma_i)$$

En procédant exactement comme pour la question 2 sur la LDA, la log vraisemblance s'écrit :

$$\begin{aligned} l((\theta, \mu_0, \mu_1, \Sigma_0, \Sigma_1)) = \sum_{n=1}^{N} (y^{(n)} \log(\theta) \ + \ (1-y^{(n)}) \log(1-\theta) \ - \ \frac{1}{2}(1-y^{(n)}) \log(|\det(\Sigma_0)|) \\ - \ \frac{1}{2} y^{(n)} \log(|\det(\Sigma_1)|) \\ + \ \left((1-y^{(n)})(-\frac{1}{2}((x^{(n)}-\mu_0)^T \Sigma_0^{-1}(x^{(n)}-\mu_0))) \right) \\ + \ \left(y^{(n)}(-\frac{1}{2}((x^{(n)}-\mu_1)^T \Sigma_1^{-1}(x^{(n)}-\mu_1))) \right) + cste) \end{aligned}$$

Ainsi, en dérivant $l((\theta, \mu_0, \mu_1, \Sigma_0, \Sigma_1))$ par rapport à $\theta, \mu_0, \mu_1, \Sigma_0, \Sigma_1$ et en annulant les dérivées, on trouve :

$$\boxed{\begin{aligned} \widehat{\theta}^{MLE} &= \frac{\sum_{n=1}^{N} y^{(n)}}{N} \\ \widehat{\mu_1}^{MLE} &= \frac{\sum_{n=1}^{N} y^{(n)} x^{(n)}}{\sum_{n=1}^{N} y^{(n)}} \end{aligned}}$$

$$\widehat{\mu_0}^{MLE} = \frac{\sum_{n=1}^{N} (1-y^{(n)}) x^{(n)}}{\sum_{n=1}^{N} (1-y^{(n)})}$$

En raisonnant de manière similaire que pour la LDA, la dérivée de la fonction de log vraisemblance par rapport à Σ_0 et Σ_1 donne :

$$\nabla_{\Sigma_0}(l) = \frac{1}{2} \sum_{n=1}^{N} (1-y^{(n)}) \Sigma_0 - \frac{1}{2} \sum_{n=1}^{N} (1-y^{(n)})(x^{(n)}-\mu_0)^T (x^{(n)}-\mu_0)$$

$$\nabla_{\Sigma_1}(l) = \frac{1}{2} \sum_{n=1}^{N} y^{(n)} \Sigma_1 - \frac{1}{2} \sum_{n=1}^{N} y^{(n)} (x^{(n)}-\mu_1)^T (x^{(n)}-\mu_1)$$

En annulant les dérivées, on trouve les expressions de $\widehat{\Sigma_0}^{MLE}$ et $\widehat{\Sigma_1}^{MLE}$:

$$\boxed{\begin{aligned} \widehat{\Sigma_0}^{MLE} &= \frac{\sum_{n=1}^{N} (1-y^{(n)})(x^{(n)}-\mu_0)^T (x^{(n)}-\mu_0)}{\sum_{n=1}^{N} (1-y^{(n)})} \\ \widehat{\Sigma_1}^{MLE} &= \frac{\sum_{n=1}^{N} y^{(n)} (x^{(n)}-\mu_1)^T (x^{(n)}-\mu_1)}{\sum_{n=1}^{N} y^{(n)}} \end{aligned}}$$

L'implémentation des estimateurs du maximum de vraisemblance donnent les résultats numériques suivants pour A, B et C :

	θ	μ_0	μ_1	Σ_0	Σ_1	c	d	M
A	0.3333	$\begin{pmatrix}2.8997\\-0.8938\end{pmatrix}$	$\begin{pmatrix}-2.6923\\0.8660\end{pmatrix}$	$\begin{pmatrix}2.3106 & -1.0474\\-1.0474 & 0.5757\end{pmatrix}$	$\begin{pmatrix}2.7044 & -1.3008\\-1.3008 & 0.6896\end{pmatrix}$	$\begin{pmatrix}-7.3652\\-10.8733\end{pmatrix}$	-0.6262	$\begin{pmatrix}-1.5174 & -3.0272\\-3.0272 & -5.7233\end{pmatrix}$
B	0.5	$\begin{pmatrix}3.3406\\-0.8354\end{pmatrix}$	$\begin{pmatrix}-3.2167\\1.0830\end{pmatrix}$	$\begin{pmatrix}2.5388 & 1.0642\\1.0642 & 2.9600\end{pmatrix}$	$\begin{pmatrix}4.1536 & -1.3345\\-1.3345 & 0.5160\end{pmatrix}$	$\begin{pmatrix}-2.2806\\1.4570\end{pmatrix}$	3.3665	$\begin{pmatrix}-0.9596 & -3.8476\\-3.8476 & -11.0586\end{pmatrix}$
C	0.625	$\begin{pmatrix}2.7930\\-0.8383\end{pmatrix}$	$\begin{pmatrix}-2.9423\\-0.9578\end{pmatrix}$	$\begin{pmatrix}2.8991 & 1.2458\\1.2458 & 2.9247\end{pmatrix}$	$\begin{pmatrix}2.8691 & -1.7619\\-1.7619 & 6.5643\end{pmatrix}$	$\begin{pmatrix}-2.6652\\0.3488\end{pmatrix}$	0.1100	$\begin{pmatrix}0.0048 & -0.2918\\-0.2918 & 0.2361\end{pmatrix}$

Table 4 – QDA learnt parameters

(b) En ce qui concerne l'expression de la probabilité conditionnelle $p(y=1|x)$, par un calcul similaire à la question 2 sur la LDA, on a :

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

$$= \frac{\theta N(x|\mu_1,\Sigma_1)}{(1-\theta)N(x|\mu_0,\Sigma_0)+\theta N(x|\mu_1,\Sigma_1)}$$

$$= \frac{1}{1+\frac{1-\theta}{\theta}\frac{N(x|\mu_0,\Sigma_0)}{N(x|\mu_1,\Sigma_1)}}$$

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

Ainsi, on effectue les transformations suivantes pour faire apparaitre la fonction sigmoide :

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

$$= \sigma(\frac{1}{2}x^{(n)T}Mx^{(n)}+c^Tx^{(n)}+d)$$

avec σ la fonction sigmoide, et :

$$M = \Sigma_0^{-1}-\Sigma_1^{-1}$$

$$c = \mu_1^T\Sigma_1^{-1}-\mu_0^T\Sigma_0^{-1} \qquad d = \log(\frac{\theta}{1-\theta})+\frac{1}{2}\log(\frac{|\det(\Sigma_0)|}{|\det(\Sigma_1)|})+\frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0-\mu_1^T\Sigma_1^{-1}\mu_1)$$

La figure 4 est la représentation graphique des données de train et de test. Dans ce cas, la frontière de décision définie par $p(y=1|x)=0.5$ est une elliptique car c'est la solution d'une quadratique $=0$:

(c) Calculons l'erreur de misclassification pour les données de train et de test :

	A˙Train	A˙Test	B˙Train	B˙Test	C˙Train	C˙Test
LDA	1.33%	2.00%	3.0%	4.15%	5.5%	4.23%
Regression logistique	0.67%	2.53%	2.0%	4.3%	4.0%	2.27%
Regression linéaire	1.33%	2.07%	3.0%	4.15%	5.5%	4.23%
QDA	0.67%	2.00%	1.33%	2.0%	5.25%	3.83%

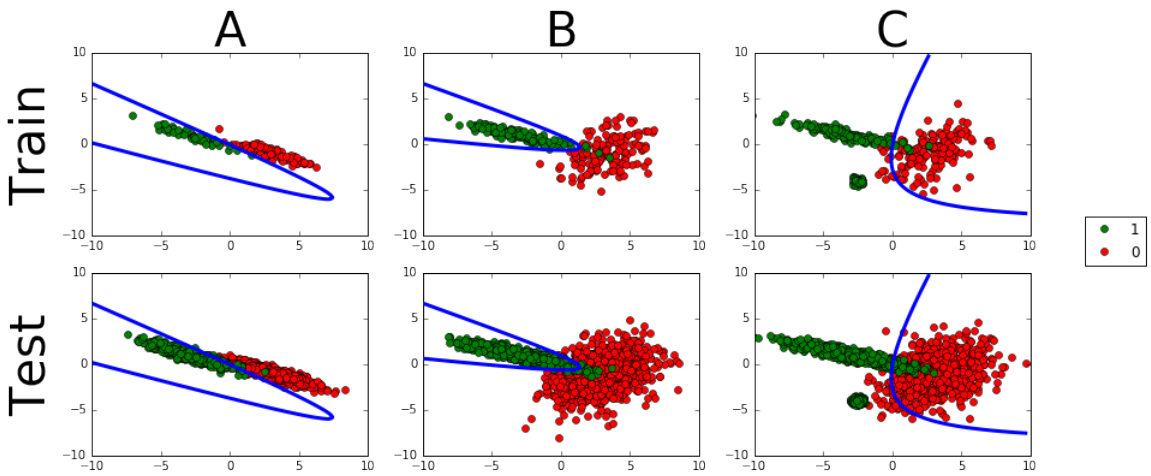

Figure 4 – QDA results

Les graphes de cette figure représentent les vrais valeurs de y les 1 en vert et 0 en rouge. La ligne bleu déterminée par les données de training de chaque sous ensemble est tracée également sur le graphe du test correspondant. Elle représente l'ensemble des points ou le critère ne sait quelle valeur attribuer a y.

(d) Pour le dataset A, on voit que la LDA fait aussi bien que les premières méthode. En effet il semble que l'hypothèse que $\Sigma_1 = \Sigma_2$ soit bonne donc la LDA est suffisante.

Pour le dataset B, la frontière de décision définie par $p(y = 1|x) = 0.5$ est une conique ce qui garantit que la QDA - qui est un modèle plus complexe que la LDA - performe beaucoup mieux que tous les autres modèles (l'erreur de misclassification est la plus basse). La QDA semble être le modèle adapté pour la séparation des données de ce dataset.

Pour le dataset C, on note une amélioration de l'erreur, la QDA permet de mieux séparer les points proche de la zone ou les deux classes se chevauchent. Mais on arrive aux mêmes conclusions que précédemment, à savoir que le le groupement de points vert séparés des autres nous laisse penser que l'hypothèse que $X|Y$ suit une loi normale est fausse. C'est pourquoi l'erreur ne décroît pas autant que sur le dataset B. On remarque d'ailleurs que a cause du groupement un peu a l'écart des point 1, dans le cas C la décision essaye d'exclure les point 0, alors que dans le cas B ou il n'y a pas cette singularité la courbe essaye d'inclure les 1.