Modèles graphiques probabiliste - DM1

Olivier Jais-Nielsen & Jérémy Rapin

October 26, 2010

1 Apprentissage dans les modèles discret

On a pour modèle $p(Y=m)=\pi_m$ and $p(X=k|Y=m)=\theta_{m,k}$ pour tout $m=1,...,M$ et $k=1,...,K$.

On dispose de N observations i. i. d. du couple de variables aléatoires (X,Y) :

$$((x_1,y_1),...,(x_N,y_N))$$

La vraisemblance du modèle décrit par $\theta=(\pi_1,...,\pi_M,\theta_{1,1},...,\theta_{M,K})$s'écrit alors:

$$V(\theta)=p_\theta((x_1,y_1),...,(x_N,y_N))$$

Par indépendance des observations :

$$\begin{aligned}
V(\theta) &= \prod_{n=1,...,N} p_\theta(x_n,y_n) \\
&= \prod_{n=1,...,N} p_\theta(x_n|y_n)\, p_\theta(y_n) \\
&= \prod_{n=1,...,N} \left(\prod_{m=1,...,M} \pi_m^{\delta(y_n=m)} \right) \left(\prod_{\substack{m=1,...,M \\ k=1,...,K}} \theta_{m,k}^{\delta(x_n=k,y_n=m)} \right) \\
V(\theta) &= \left(\prod_{m=1,...,M} \pi_m^{N_m^y} \right) \left(\prod_{\substack{m=1,...,M \\ k=1,...,K}} \theta_{m,k}^{N_{m,k}^{xy}} \right)
\end{aligned}$$

avec $N_m^y=\sum_{n=1,...,N}\delta(y_n=m)$ le nombre de réalisations (x_n,y_n) telles que $y_n=m$ et $N_{m,k}^{xy}=\sum_{n=1,...,N}\delta(x_n=k,y_n=m)$ le nombre de réalisations (x_n,y_n) telles que $x_n=k$ et $y_n=m$.

On a alors :

$$\log(V(\theta))=\sum_{m=1,...,M} N_m^y \log(\pi_m)+\sum_{\substack{m=1,...,M \\ k=1,...,K}} N_{m,k}^{xy}\log\left(\theta_{m,k}\right)$$

On minimise donc :

$$\mathscr{L}(\theta,\lambda,\mu) = -\log(V(\theta)) + \lambda\left(\sum_{m=1,\dots,M}\pi_m - 1\right) + \sum_{m=1,\dots,M}\mu_m\left(\sum_{k=1,\dots,K}\theta_{m,k} - 1\right)$$

$\mathscr{L}$ est convexe comme somme de fonctions convexes (fonctions $-\log$ et fonctions affines). Son éventuel minimum global est donc atteint en annulant son gradient.

On cherche donc à résoudre le système suivant :

$$\begin{cases} \frac{\partial\mathscr{L}}{\partial\pi_m} = -\frac{N_m^y}{\pi_m} + \lambda = 0 & \forall m = 1,\dots,M \\ \frac{\partial\mathscr{L}}{\partial\theta_{m,k}} = -\frac{N_{m,k}^{xy}}{\theta_{m,k}} + \mu_m = 0 & \forall m = 1,\dots,M \\ & \forall k = 1,\dots,K \\ \frac{\partial\mathscr{L}}{\partial\lambda} = \sum_{m=1,\dots,M}\pi_m - 1 = 0 & \\ \frac{\partial\mathscr{L}}{\partial\mu_m} = \sum_{k=1,\dots,K}\theta_{m,k} - 1 = 0 & \forall m = 1,\dots,M \end{cases}$$

On trouve : $\sum_{m=1,\dots,M}\hat{\pi}_m = 1$ et $\sum_{k=1,\dots,K}\hat{\theta}_{m,k} = 1$ pour tout $m = 1,\dots,M$.

Et $\lambda = \sum_{m=1,\dots,M} N_m^y = N$ et donc

$$\hat{\pi}_m = \frac{N_m^y}{N}$$

pour tout $m = 1,\dots,M$.

Et $\mu_m = \sum_{k=1,\dots,K} N_{m,k}^{xy} = N_m^y$ pour tout $m = 1,\dots,M$ et donc

$$\hat{\theta}_{m,k} = \frac{N_{m,k}^{xy}}{N_m^y}$$

pour tout $m = 1,\dots,M$ et $k = 1,\dots,K$.

2 Classification linéaire

2.1 Modèles génératifs (LDA et QDA)

2.1.1 Estimation du modèle

On a pour modèle $Y \backsim \mathscr{B}(\pi)$ et $X|Y=0 \backsim \mathscr{N}(\mu_0,\Sigma_0)$ et $X|Y=1 \backsim \mathscr{N}(\mu_1,\Sigma_1)$.

On fera les calculs en considérant que les matrices de covariance sont différentes pour traiter dans un même cadre les exercices **2.1** et **2.5**. On détermine des estimateurs des paramètres du modèle en maximisant la vraisemblance, comme dans l'exercice **1.**

La vraisemblance du modèle décrit par $\theta = (\pi,\mu_0,\mu_1,\Sigma_0,\Sigma_1)$ s'écrit :

$$\begin{aligned} V(\theta) &= \prod_{n=1,\dots,N} p_\theta(x_n,y_n) = \prod_{n=1,\dots,N} p_\theta(x_n|y_n)\, p_\theta(y_n) \\ &= \pi^{N_1^Y}(1-\pi)^{N-N_1^Y} \prod_{i=\{0,1\}} \prod_{\substack{n=1,\dots,N \\ y_n=i}} \frac{1}{2\pi}\frac{1}{\sqrt{\det\Sigma_i}} e^{-\frac{1}{2}(x_n-\mu_i)^\top \Sigma_i^{-1}(x_n-\mu_i)} \end{aligned}$$

$$\begin{aligned} \log(V(\theta)) &= N_1^Y \log(\pi) + \left(N-N_1^Y\right)\log(1-\pi) \\ &\quad + \sum_{i=\{0,1\}} \sum_{\substack{n=1,\dots N \\ y_n=i}} -\log(2\pi) - \frac{1}{2}\log(\det\Sigma_i) - \frac{1}{2}(x_n-\mu_i)^\top \Sigma_i^{-1}(x_n-\mu_i) \end{aligned}$$

avec N_1^Y le nombre de réalisations (x_n,y_n) telles que $y_n=1$.

On cherche ensuite à annuler le gradient pour obtenir le maximum de vraisemblance :

$$\begin{cases} \frac{\partial \log V}{\partial \pi} = \frac{N_1^Y}{\pi} - \frac{N-N_1^Y}{1-\pi} = 0 & \\ \frac{\partial \log V}{\partial \mu_i} = \sum_{\substack{n=1,\dots N \\ y_n=i}} \Sigma_i^{-1}(x_n-\mu_i) = \Sigma_i^{-1}\left(\left(\sum_{\substack{n=1,\dots N \\ y_n=i}} x_n\right) - N_i\mu_i\right) = 0 & \forall i \in \{0,1\} \\ \frac{\partial \log V}{\partial \Lambda_i} = \frac{N_i}{2}\Lambda_i^{-1} - \frac{1}{2}\sum_{\substack{n=1,\dots N \\ y_n=i}} (x_n-\hat{\mu}_i)^\top (x_n-\hat{\mu}_i) & \forall i \in \{0,1\} \end{cases}$$

avec $\Lambda_i = \Sigma_i^{-1}$ car $\det(\Sigma) = \frac{1}{\det\left(\Sigma^{-1}\right)}$ et $\frac{\partial}{\partial\Lambda}\log(\det(\Lambda)) = \Lambda^{-1}$.

Finalement :

$$\hat{\pi} = \frac{N_1^Y}{N}$$

et

$$\hat{\Sigma}_i = \frac{1}{N_i} \sum_{\substack{n=1,\dots N \\ y_n=i}} (x_n-\hat{\mu}_i)^\top (x_n-\hat{\mu}_i)$$

avec $i \in \{0,1\}$.

2.1.2 Choix d'affectation

On a $P(y=1|x) = \dfrac{P(x,y=1)}{P(x)} = \dfrac{P(x|y=1)P(y=1)}{P(x)}$.

Calculons le Log-odds-ratio :

$$\begin{aligned}\log\frac{P(y=1|x)}{P(y=0|x)} &= \log\frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)} \\ &= \log\frac{\pi_1}{\pi_0} + \log\frac{P(x|y=1)}{P(x|y=0)} \\ \log\frac{P(y=1|x)}{P(y=0|x)} &= \log\left(\frac{\pi_1}{\pi_0}\sqrt{\frac{\Sigma_0}{\Sigma_1}}\right) - \frac{1}{2}(x-\mu_1)^\top \Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0 1)^\top \Sigma_0^{-1}(x-\mu_0)\end{aligned}$$

On développant les produits et en regroupant en suivant le degré on trouve : $\log\dfrac{P(y=1|x)}{P(y=0|x)} = \beta_0 + f(x)$ avec :

$$f(x) = \frac{1}{2}x^\top\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right)x + x^\top\left(\Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0\right)$$

$$\beta_0 = \log\left(\frac{\pi_1}{\pi_0}\sqrt{\frac{\Sigma_0}{\Sigma_1}}\right) + \frac{1}{2}\mu_0^\top\Sigma_0^{-1}\mu_0 - \frac{1}{2}\mu_1^\top\Sigma_1^{-1}\mu_1$$

Ainsi, on trouve une condition d'affectation :

$$y=1 \iff P(y=1|x) > P(y=0|x) \iff \log\frac{P(y=1|x)}{P(y=0|x)} > 0 \iff \beta_0 + f(x) > 0$$

et la frontière entre les classes correspond à l'équations : $\beta_0 + f(x) = 0$.

Dans le cas où $\Sigma_1 = \Sigma_0 = \Sigma$ (LDA), on a les simplifications suivantes :

$$f(x) = x^\top\Sigma^{-1}(\mu_1 - \mu_0)$$

$$\beta_0 = \log\frac{\pi_1}{\pi_0} + \frac{1}{2}\mu_0^\top\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^\top\Sigma^{-1}\mu_1$$

2.2 Régression logistique

2.2.1 Estimation du modèle

On a pour modèle $p(Y=1|X=x)=\sigma\left(\tilde{\theta}^T\tilde{x}\right)$ avec $\tilde{\theta}=\begin{pmatrix}\theta\\ \theta_0\end{pmatrix}$ et $\tilde{x}=\begin{pmatrix}x\\ 1\end{pmatrix}$ et $\sigma: z\mapsto\frac{1}{1+e^{-z}}$.

On maximise la log-vraisemblance conditionnelle :

$$\begin{aligned}\ell\left(\tilde{\theta}\right) &= \sum_{n=1,\dots,N}\log\left(p\left(y_n=1|x_n\right)\right)\\ &= \sum_{\substack{n=1,\dots,N\\ y_n=1}}\log\left(\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)+\sum_{\substack{n=1,\dots,N\\ y_n=0}}\log\left(1-\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)\\ \ell\left(\tilde{\theta}\right) &= \sum_{n=1,\dots,N} y_n\log\left(\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)+(1-y_n)log\left(\sigma\left(-\tilde{\theta}^T\tilde{x}_n\right)\right)\end{aligned}$$

Alors

$$\begin{aligned}\nabla_{\tilde{\theta}}\ell &= \sum_{n=1,\dots,N} y_n\tilde{x}_n\frac{\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\left(1-\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)}{\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)}-(1-y_n)\tilde{x}_n\frac{\sigma\left(-\tilde{\theta}^T\tilde{x}_n\right)\left(1-\sigma\left(-\tilde{\theta}^T\tilde{x}_n\right)\right)}{\sigma\left(-\tilde{\theta}^T\tilde{x}_n\right)}\\ &= \sum_{n=1,\dots,N} y_n\tilde{x}_n\left(1-\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)-(1-y_n)\tilde{x}_n\left(\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)\\ \nabla_{\tilde{\theta}}\ell &= \sum_{n=1,\dots,N}\tilde{x}_n\left(y_n-\sigma\left(\tilde{\theta}^T\tilde{x}_n\right)\right)\end{aligned}$$

2.2.2 Optimisation (algorithme IRLS)

On maximise la log-vraisemblance par l'algorithme IRLS.

On note $X=\begin{pmatrix}\tilde{x}_1^T\\ \dots\\ \tilde{x}_N^T\end{pmatrix}$ et $y=\begin{pmatrix}y_1\\ \dots\\ y_n\end{pmatrix}$ et $\eta=\sigma\left(\tilde{\theta}^TX\right)$ et $W=diag\left(\eta_1\left(1-\eta_1\right),\dots,\eta_N\left(1-\eta_N\right)\right)$.

On a $\nabla_{\tilde{\theta}}l=X^T(y-\eta)$.

On a également $\nabla^2_{\tilde{\theta}}l=-X^TWX$

On approche alors $\tilde{\theta}$ comme la limite $\hat{\theta}$ de la suite suivante:

$$\begin{cases}\hat{\theta}_0=0\\ \hat{\theta}_{n+1}=\hat{\theta}_n-\left(\nabla^2_{\tilde{\theta}}l\left(\hat{\theta}_n\right)\right)^{-1}.\nabla_{\tilde{\theta}}l\left(\hat{\theta}_n\right)\quad\forall n\geq 0\end{cases}$$

2.2.3 Choix d'affectation

On a

$$p(y=1|x) > \frac{1}{2} \Leftrightarrow \frac{1}{1+\exp\left(-\tilde{\theta}^T\hat{x}\right)} > \frac{1}{2}$$
$$\Leftrightarrow \exp\left(-\tilde{\theta}^T\hat{x}\right) < 1$$
$$p(y=1|x) > \frac{1}{2} \Leftrightarrow \theta^T x + \theta_0 > 0$$

On prend donc comme critère: $\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}^T x + \hat{\theta}_3 > 0$ pour assigner la classe 1 au point x et $\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}^T x + \hat{\theta}_3 < 0$ pour lui assigner la classe 0.

2.3 Régression linéaire

2.3.1 Estimation du modèle

On a pour modèle $Y|X \backsim \mathcal{N}\left(\tilde{\theta}^T \tilde{x}, \sigma^2\right)$ avec $\tilde{\theta} = \begin{pmatrix} \theta \\ \theta_0 \end{pmatrix}$ et $\tilde{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$.

On maximise la log-vraisemblance conditionnelle :

$$\begin{aligned}
\ell\left(\tilde{\theta}, \sigma\right) &= \sum_{n=1,\ldots,N} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(\tilde{\theta}^T \tilde{x}_n - y_n^2\right)}{2\sigma^2}\right)\right) \\
&= \sum_{n=1,\ldots,N} -\frac{1}{2}\log\left(2\pi\sigma\right) - \frac{1}{2\sigma^2}\left(\tilde{\theta}^T \tilde{x}_n - y_n^2\right) \\
\ell\left(\tilde{\theta}, \sigma\right) &= -\frac{N}{2}\log\left(2\pi\sigma\right) - \frac{1}{2\sigma^2}\left\|y - X\tilde{\theta}\right\|
\end{aligned}$$

Alors

$$\nabla_{\tilde{\theta}} \ell = -\frac{1}{\sigma^2} X^T \left(X\tilde{\theta} - y\right)$$

Pour $\hat{\theta}$ annulant le gradient, on trouve l'équation normale :

$$X^T X \hat{\theta} = X^T y$$

Finalement, on évalue $\tilde{\theta}$ par :

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

2.3.2 Choix d'affectation

La fonction de régression est : $f(x) = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}^T x + \hat{\theta}_3$.

La droite correspondant à:

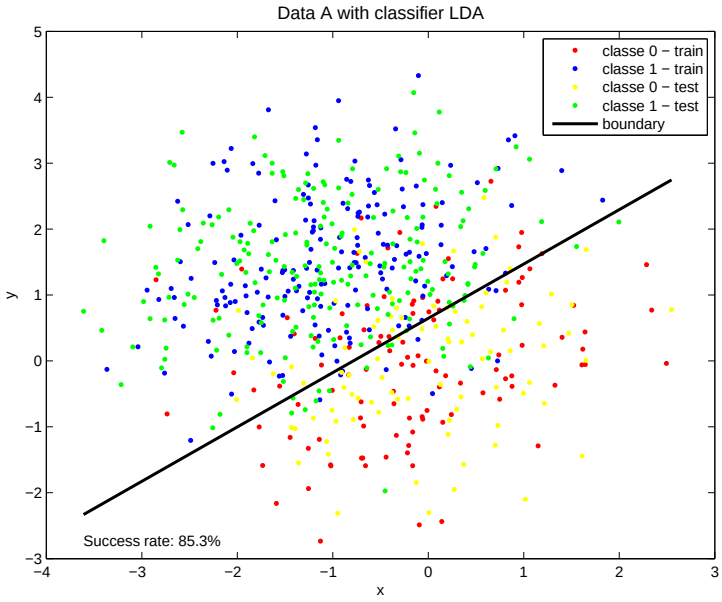
$$f(x) = \frac{1}{2} \Leftrightarrow \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}^T x + \hat{\theta}_3 - \frac{1}{2} = 0$$

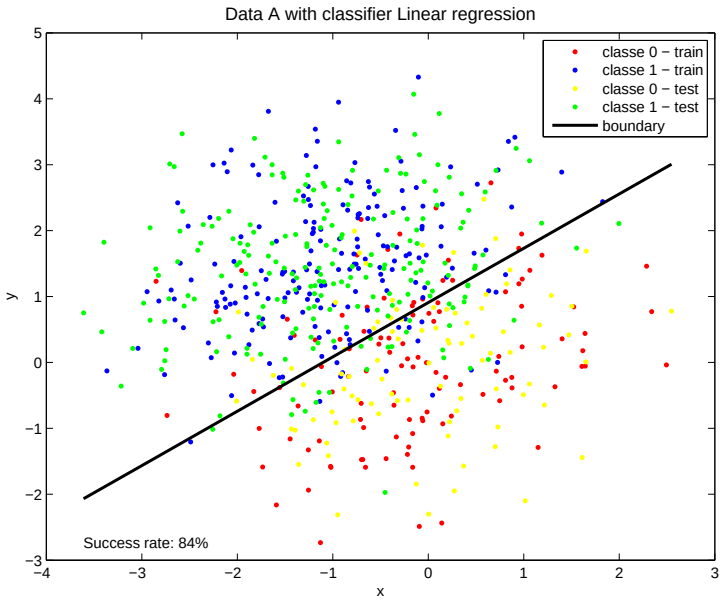
permet de délimiter les classes si on prend comme critère $E[Y|X = x] > \frac{1}{2}$ pour que le point x soit de classe 1 et de classe 0 sinon.

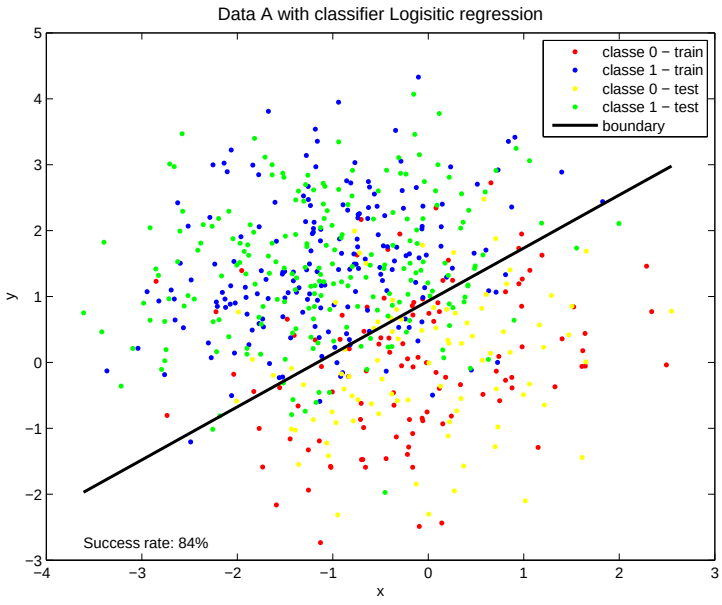
2.4 Comparaison

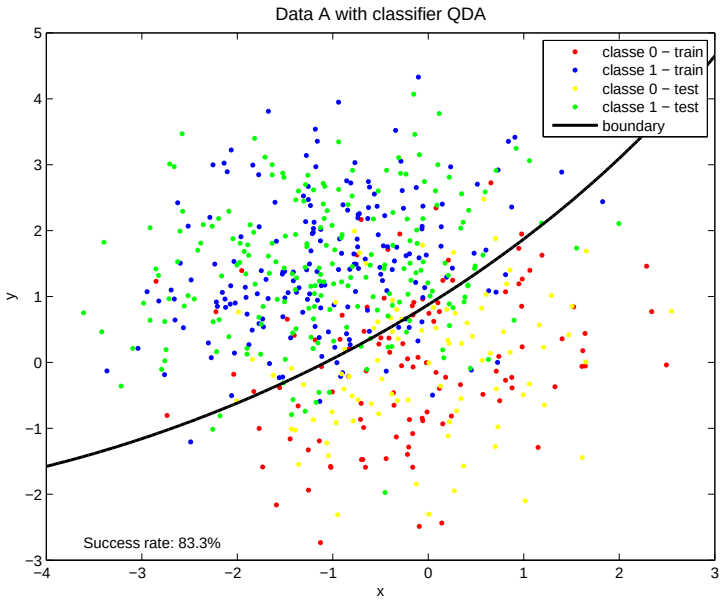
Pour chaque jeu de données et pour chaque type de modèle envisagé, on estime les paramètres du modèle sur les données d'entraînement, on représente sur un même graphe, les données d'entraînement, les données de test et le classifieur et on détermine la matrice de confusion concernant les données de test. On détermine la qualité d'un classifieur par son pourcentage de succès sur les données test.

2.4.1 Données A




Data A with classifier Linear regression
classe 0 − train
classe 1 − train
classe 0 − test
classe 1 − test
boundary
Success rate: 84%
x
y
5
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3


Data A with classifier Logisitic regression
classe 0 − train
classe 1 − train
classe 0 − test
classe 1 − test
boundary
Success rate: 84%
x
y
5
4
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3

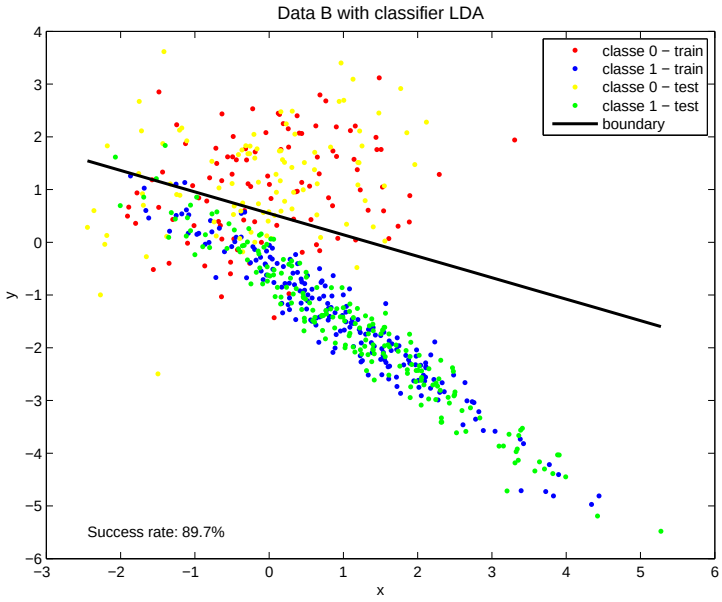


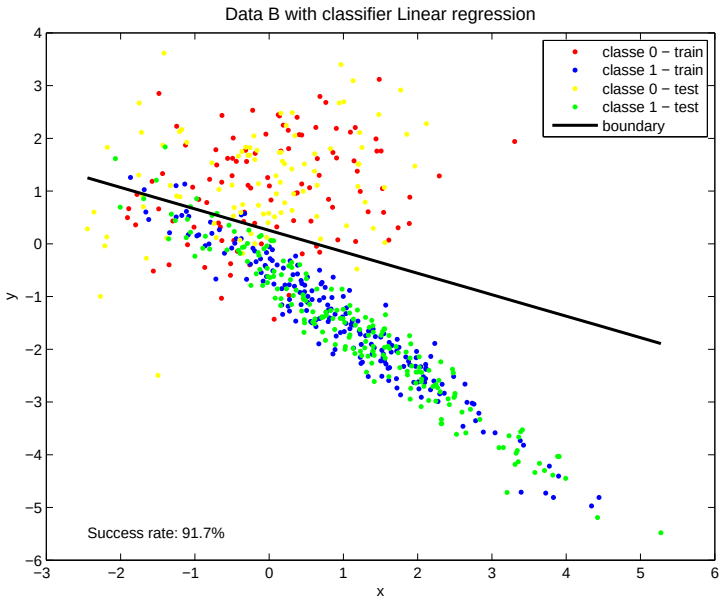
Les deux distributions sont à peu près gaussiennes de même covariance, étalées sans direction privilégiée.

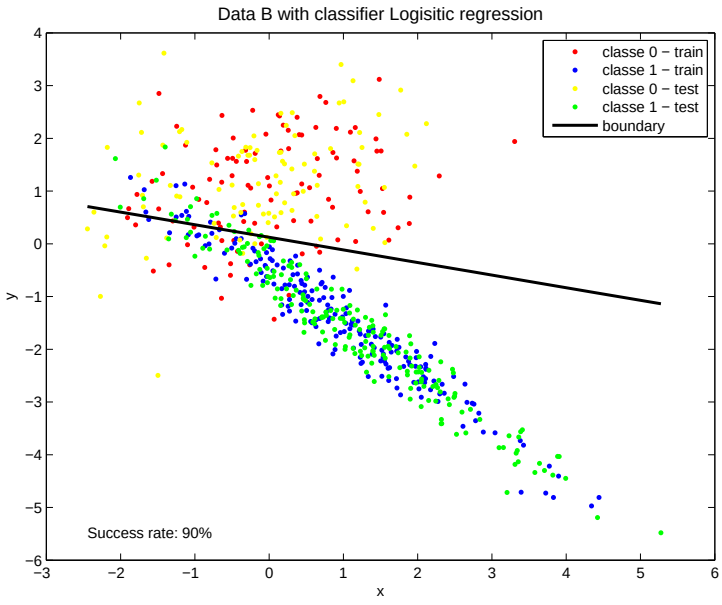
Les quatres classifieurs fonctionnent de manière identique à peu de chose près. On peut remarquer que le classifieur LDA fonctionne mieux que le QDA, alors que le QDA est sensé prendre plus d'information en compte. Cela peut être dû au fait que le classifieur LDA considère que les covariances sont identiques (ce qui semble être le cas) alors que le QDA approxime deux covariances différentes. On peut aussi voir que dans ces conditions le classifieur QDA tend à délimiter les deux zones par une droite

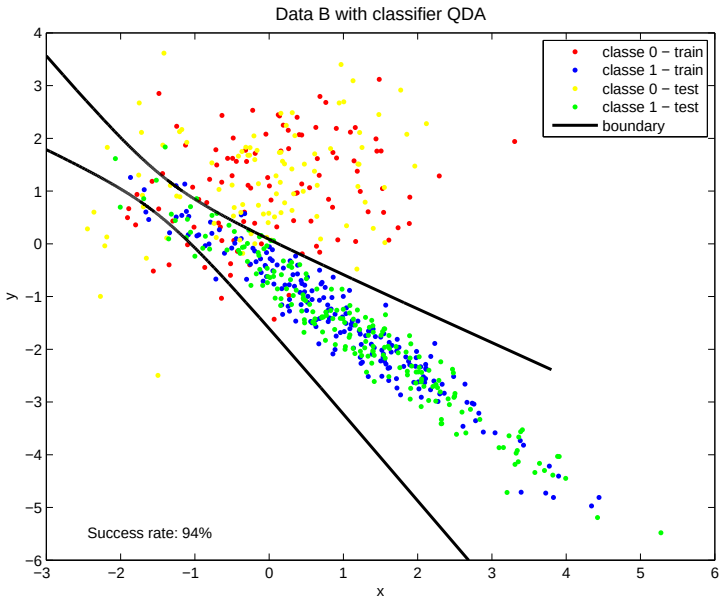
comme les autres classifieurs.

2.4.2 Données B




Data B with classifier Linear regression
classe 0 − train
classe 1 − train
classe 0 − test
classe 1 − test
boundary
Success rate: 91.7%
x
y


Data B with classifier Logisitic regression
classe 0 − train
classe 1 − train
classe 0 − test
classe 1 − test
boundary
Success rate: 90%
x
y

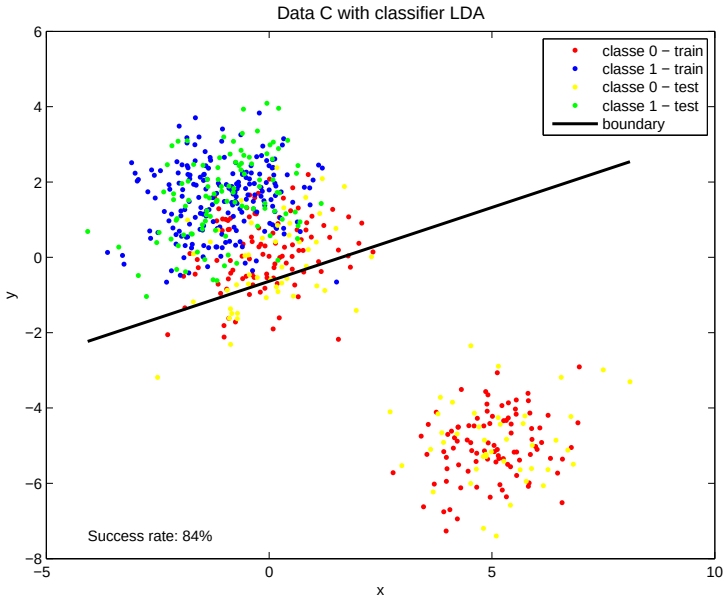


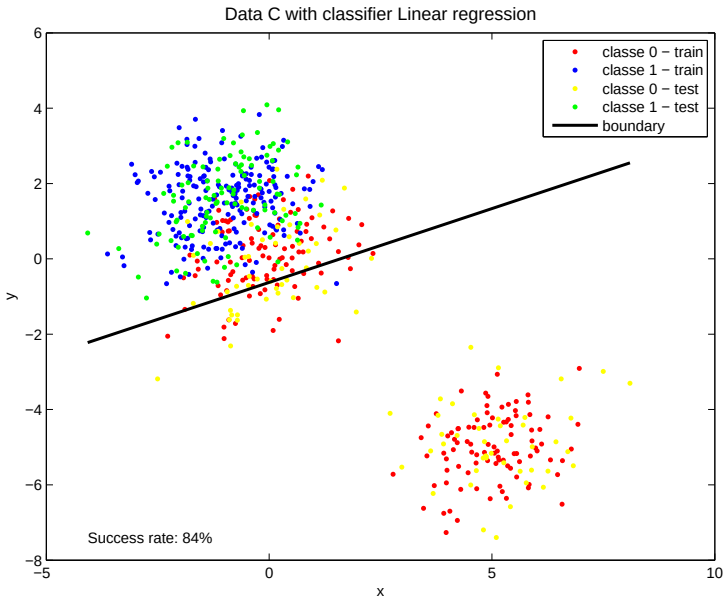
Les deux distributions sont à peu près gaussiennes avec des matrices de covariance très différentes. Une des classe est étalée sans direction particulière, alors que l'autre distribution est très allongée.

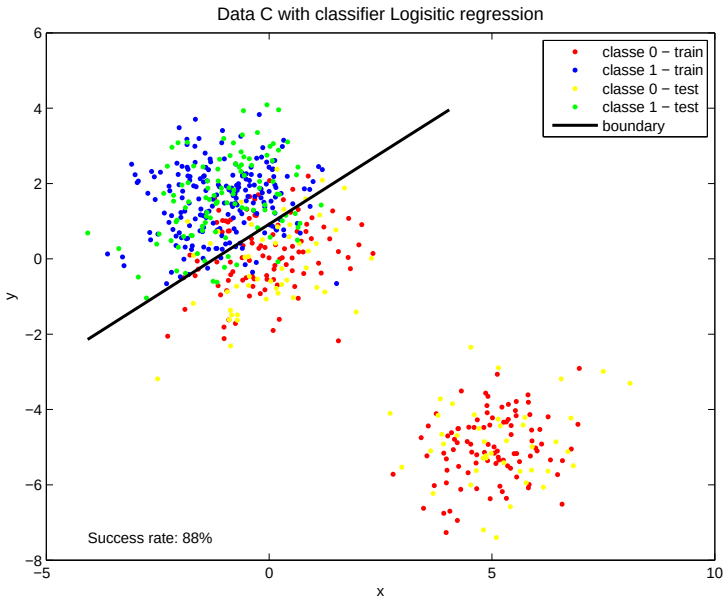
On est dans le cas précis où le classifieur QDA est le plus efficace par rapport aux autres classifieurs. Il donne ainsi un grand taux de succés (94% quand les autres ne font "que" 90%). Une grande différence du QDA avec les autres méthodes et qu'ici la limite est formée de deux branches d'hyperbole délimitant une "allée" le long de l'axe principal de la distribution allongée, et de par et d'autre de cette allée, la probabilité

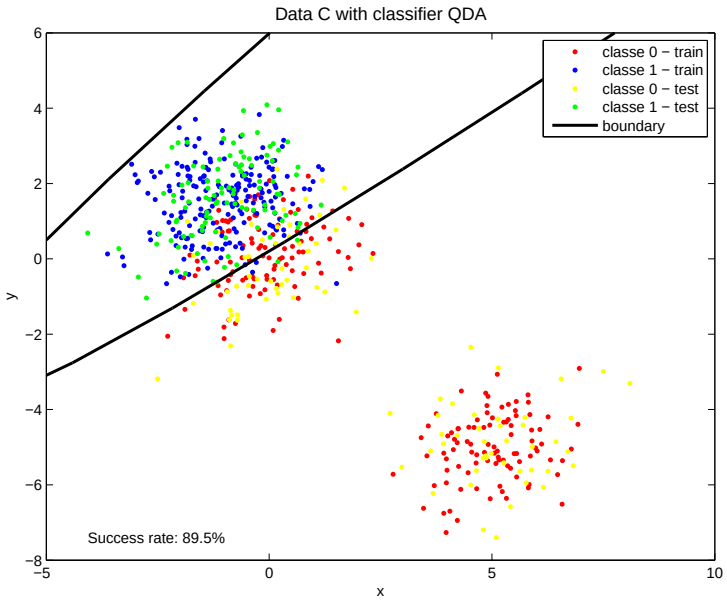
d’être dans l’autre classe est plus élevée.

2.4.3 Données C




Data C with classifier Linear regression
classe 0 − train
classe 1 − train
classe 0 − test
classe 1 − test
boundary
Success rate: 84%
x
y
6
4
2
0
−2
−4
−6
−8
−5
0
5
10


Data C with classifier Logisitic regression
classe 0 − train
classe 1 − train
classe 0 − test
classe 1 − test
boundary
Success rate: 88%
x
y
6
4
2
0
−2
−4
−6
−8
−5
0
5
10



Une des classes a une distribution à peu près gaussienne alors que l'autre absolument pas, car sa distribution est séparée en deux zones distinctes (à priori, une somme de deux gaussiennes).

Dans le cas où les distributions ne sont plus gaussiennes, les méthodes LDA et linear regression sont beaucoup moins performantes. QDA peut s'y adapter mais la régression logistique, qui ne prend pas comme à priori une distribution gaussienne et n'est donc pas biaisé par un faux à priori, a des résultat comparables.