
COMPTE-RENDU DU DEVOIR 1

Vincent Matthys

1 Learning in discrete graphical models

Etant donné que z et x sont des variables à valeurs discrètes prenant respectivement M et K valeurs, on peut procéder au *one hot encoding*, notant respectivement \mathbf{Z} et \mathbf{X} leur encodage sur, respectivement \mathbb{R}^M et \mathbb{R}^K . Ainsi, on peut écrire :

$$p(z = m) = P(Z_m = 1) = \pi_m$$

$$p(x = k|z = m) = p(X_k = 1|Z_m = 1) = \theta_{mk}$$

En supposant que l'on ait un échantillon de n observations de (x, z) , on peut exprimer la probabilité jointe d'une observation i :

$$p(x^{(i)}, z^{(i)}; \boldsymbol{\pi}, \boldsymbol{\theta}) = p(z^{(i)}; \boldsymbol{\pi})p(x^{(i)}|z^{(i)}; \boldsymbol{\theta})$$

$$p(x^{(i)}, z^{(i)}; \boldsymbol{\pi}, \boldsymbol{\theta}) = \prod_{m=1}^M \prod_{k=1}^K \theta_{mk}^{X_k^{(i)} Z_m^{(i)}} \prod_{l=1}^M \pi_l^{Z_l^{(i)}} \quad (1)$$

On peut alors écrire la log-vraisemblance de l'échantillon dans le modèle $(\boldsymbol{\pi}, \boldsymbol{\theta})$, composé d'observations i.i.d., en utilisant (1) :

$$\begin{aligned} \ell(\boldsymbol{\pi}, \boldsymbol{\theta}) &= \sum_{i=1}^n \ln(p(x^{(i)}, z^{(i)}; \boldsymbol{\pi}, \boldsymbol{\theta})) \\ &= \sum_{i=1}^n \left(\ln \left(\prod_{m=1}^M \prod_{k=1}^K \theta_{mk}^{X_k^{(i)} Z_m^{(i)}} \prod_{l=1}^M \pi_l^{Z_l^{(i)}} \right) \right) \\ &= \sum_{i=1}^n \left(\sum_{m=1}^M \sum_{k=1}^K \ln \left(\theta_{mk}^{X_k^{(i)} Z_m^{(i)}} \right) + \sum_{l=1}^M \ln \left(\pi_l^{Z_l^{(i)}} \right) \right) \\ &= \sum_{i=1}^n \left(\sum_{m=1}^M \sum_{k=1}^K X_k^{(i)} Z_m^{(i)} \ln(\theta_{mk}) + \sum_{l=1}^M Z_l^{(i)} \ln(\pi_l) \right) \\ &= \sum_{m=1}^M \sum_{k=1}^K \left(\sum_{i=1}^n X_k^{(i)} Z_m^{(i)} \right) \ln(\theta_{mk}) + \sum_{l=1}^M \left(\sum_{i=1}^n Z_l^{(i)} \right) \ln(\pi_l) \\ &= \sum_{m=1}^M \sum_{k=1}^K \alpha_{mk} \ln(\theta_{mk}) + \sum_{l=1}^M \beta_l \ln(\pi_l) \end{aligned} \quad (2)$$

avec

$$\alpha_{mk} = \sum_{i=1}^n X_k^{(i)} Z_m^{(i)}$$

$$\beta_l = \sum_{i=1}^n Z_l^{(i)}$$

D'après (2), la log-vraisemblance est donc strictement concave en chaque composantes de $\boldsymbol{\pi}$ et $\boldsymbol{\theta}$, par combinaison linéaire de logarithmes. D'autre part, on a les contraintes linéaires suivantes :

$$\sum_{m=1}^M \pi_m - 1 = 0$$

$$\sum_{k=1}^K \theta_{mk} - 1 = 0, \forall m : 1..M \quad (3)$$

On peut donc écrire le Lagrangien, en utilisant les multiplicateurs de Lagrange correspondants, associé au problème de maximisation de la log-vraisemblance (2) :

$$\mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda, \boldsymbol{\gamma}) = \sum_{m=1}^M \sum_{k=1}^K \alpha_{mk} \ln(\theta_{mk}) + \sum_{l=1}^M \beta_l \ln(\pi_l) - \lambda \left(\sum_{m=1}^M \pi_m - 1 \right) - \boldsymbol{\gamma}^\top \begin{pmatrix} \vdots \\ \sum_{k=1}^K \theta_{mk} - 1 \\ \vdots \end{pmatrix} \quad (4)$$

Le problème admet une solution unique, notée $(\widehat{\boldsymbol{\pi}_{MLE}}, \widehat{\boldsymbol{\theta}_{MLE}})$ que l'on trouve par dérivation de (4), puisque cette expression est différentiable.

1.1 Par rapport à π_m

$$\frac{\partial \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda, \boldsymbol{\gamma})}{\partial \pi_m} = \frac{\beta_m}{\pi_m} - \lambda = 0 \implies \pi_m \propto \beta_m \implies \pi_m = \frac{\beta_m}{\sum_{m=1}^M \beta_m} = \frac{\sum_{i=1}^n Z_m^{(i)}}{\sum_{m=1}^M \sum_{i=1}^n Z_m^{(i)}}$$

D'où finalement :

$$(\widehat{\boldsymbol{\pi}_{MLE}})_m = \frac{1}{n} \sum_{i=1}^n Z_m^{(i)} \quad (5)$$

1.2 Par rapport à θ_{mk}

$$\frac{\partial \mathcal{L}(\boldsymbol{\pi}, \boldsymbol{\theta}, \lambda, \boldsymbol{\gamma})}{\partial \theta_{mk}} = \frac{\alpha_{mk}}{\theta_{mk}} - \gamma_m = 0 \implies \theta_{mk} \propto \alpha_{mk} \implies \theta_{mk} = \frac{\alpha_{mk}}{\sum_{k=1}^K \alpha_{mk}} = \frac{\sum_{i=1}^n X_k^{(i)} Z_m^{(i)}}{\sum_{i=1}^n \left(\sum_{k=1}^K X_k^{(i)} \right) Z_m^{(i)}}$$

D'où finalement :

$$(\widehat{\boldsymbol{\theta}_{MLE}})_{mk} = \frac{\sum_{i=1}^n X_k^{(i)} Z_m^{(i)}}{\sum_{i=1}^n Z_m^{(i)}} \quad (6)$$

On remarque que si $\sum_{i=1}^n Z_m^{(i)}$ est nulle, alors $\left(\widehat{\theta}_{MLE}\right)_{mk}$ n'est pas défini, et ce pour tout k . Mais, dans de telles conditions, cela signifie que la $m^{\text{ième}}$ valeur de z n'a pas été observée, et donc que l'on peut réduire $\boldsymbol{\theta}$ d'une ligne entière, puisqu'alors la probabilité d'observer une quelconque valeur de x sachant que z prend la valeur m est nulle.

En définitif, on peut s'affranchir de l'hypothèse implicite qu'aucune composante de $\boldsymbol{\pi}$ ni de $\boldsymbol{\theta}$ n'est nulle. En effet, si tel est le cas, alors la probabilité d'observer une telle observation est nulle.

1.3 Conclusion

On remarque donc que l'estimateur de $\boldsymbol{\pi}$ n'est rien d'autre que la moyenne empirique des observations de z , et que l'estimateur de $\boldsymbol{\theta}$ est également la moyenne des observations de x pour une valeur de z donnée.

2 Linear classification

2.1 Generative model : Linear Discriminant Analysis

Supposons un échantillon de N observations de x, y , notées, (x_n, y_n) , pour n variant de 1 à N , où $y_n \in \{0, 1\}$. Etant donné le modèle suivant :

$$y \sim \text{Bernoulli}(\boldsymbol{\pi})$$

$$p(x|y = i) = \frac{1}{2\pi\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x - \mu_i)^\top \Sigma^{-1}(x - \mu_i)\right)$$

paramétré par :

$$\Theta = (\mu_0, \mu_1, \Sigma, \boldsymbol{\pi})$$

on peut réécrire la probabilité d'observer (x_n, y_n) sous la forme comprenant déjà la contrainte sur $\boldsymbol{\pi}$:

$$p(x_n, y_n) = (\boldsymbol{\pi}p(x_n|y_n = 1))^{y_n} ((1 - \boldsymbol{\pi})p(x_n|y_n = 0))^{1-y_n} \quad (7)$$

ce qui permet de déterminer la log-vraisemblance de l'échantillon composé de N observations i.i.d dans le modèle Θ en utilisant (7) :

$$\begin{aligned} \ell(\Theta) &= \sum_{n=1}^N \ln(p(x_n, y_n; \Theta)) \\ \ell(\Theta) &= \sum_{n=1}^N \ln((\boldsymbol{\pi}p(x_n|y_n = 1))^{y_n} ((1 - \boldsymbol{\pi})p(x_n|y_n = 0))^{1-y_n}) \\ \ell(\Theta) &= \sum_{n=1}^N \left(y_n \left(\ln\left(\frac{\boldsymbol{\pi}}{2\pi\sqrt{\det \Sigma}}\right) - \frac{1}{2}(x_n - \mu_1)^\top \Sigma^{-1}(x_n - \mu_1) \right) \right. \\ &\quad \left. + \sum_{n=1}^N \left((1 - y_n) \left(\ln\left(\frac{1 - \boldsymbol{\pi}}{2\pi\sqrt{\det \Sigma}}\right) - \frac{1}{2}(x_n - \mu_0)^\top \Sigma^{-1}(x_n - \mu_0) \right) \right) \right) \\ \ell(\Theta) &= \sum_{n=1}^N \left(y_n \left(\ln \boldsymbol{\pi} - \ln(2\pi) - \frac{1}{2} \ln \det \Sigma - \frac{1}{2}(x_n - \mu_1)^\top \Sigma^{-1}(x_n - \mu_1) \right) \right) \\ &\quad + \sum_{n=1}^N \left((1 - y_n) \left(\ln(1 - \boldsymbol{\pi}) - \ln(2\pi) - \frac{1}{2} \ln \det \Sigma - \frac{1}{2}(x_n - \mu_0)^\top \Sigma^{-1}(x_n - \mu_0) \right) \right) \end{aligned} \quad (8)$$

$\ell(\Theta)$ est donc une fonction strictement concave en π par stricte concavité du logarithme, en μ_0 et μ_1 par stricte concavité de l'opposé d'une forme quadratique (avec les valeurs propres de Σ strictement positives), et en Σ par convexité du logdet. De plus, $\ell(\Theta)$ est différentiable. Le problème de maximisation $\ell(\Theta)$ admet une solution unique, notée $(\widehat{\mu_{0,MLE}}, \widehat{\mu_{1,MLE}}, \widehat{\Sigma_{MLE}}, \widehat{\pi_{MLE}})$ que l'on trouve par dérivation de (8).

2.1.1 Calcul du MLE

Par rapport à π

$$\frac{\partial \ell(\mu_0, \mu_1, \Sigma, \pi)}{\partial \pi} = \sum_{n=1}^N \left(\frac{y_n}{\pi} - \frac{1-y_n}{1-\pi} \right) = 0 \implies \frac{1-\pi}{\pi} = \frac{\sum_{n=1}^N (1-y_n)}{\sum_{n=1}^N y_n} \implies \pi = \frac{\sum_{n=1}^N y_n}{\sum_{n=1}^N 1}$$

D'où finalement :

$$(\widehat{\pi_{MLE}}) = \frac{1}{N} \sum_{n=1}^N y_n \quad (9)$$

qui est la moyenne empirique des classes observées.

Par rapport à μ_0 et μ_1

$$\frac{\partial \ell(\mu_0, \mu_1, \Sigma, \pi)}{\partial \mu_0} = \sum_{n=1}^N ((1-y_n)(-\Sigma^{-1}(x_n - \mu_0))) = 0 \implies \mu_0 = \frac{\sum_{n=1}^N (1-y_n)x_n}{\sum_{n=1}^N (1-y_n)}$$

En introduisant $A = \sum_{n=1}^N y_n$, on a donc :

$$(\widehat{\mu_{0,MLE}}) = \frac{1}{N-A} \sum_{y_n=0} x_n \quad (10)$$

Et de la même façon :

$$(\widehat{\mu_{1,MLE}}) = \frac{1}{A} \sum_{y_n=1} x_n \quad (11)$$

Par rapport à Σ En introduisant temporairement $\Lambda = \Sigma^{-1}$, la log-vraisemblance se réécrit :

$$\begin{aligned} \ell(\Theta) = & \sum_{n=1}^N \left(y_n \left(\ln \pi - \ln(2\pi) + \frac{1}{2} \ln \det \Lambda - \frac{1}{2} (x_n - \mu_1)^\top \Lambda (x_n - \mu_1) \right) \right) \\ & + \sum_{n=1}^N \left((1-y_n) \left(\ln(1-\pi) - \ln(2\pi) + \frac{1}{2} \ln \det \Lambda - \frac{1}{2} (x_n - \mu_0)^\top \Lambda (x_n - \mu_0) \right) \right) \end{aligned} \quad (12)$$

En introduisant aussi :

$$\begin{aligned} \widetilde{\Sigma}_0 &= \frac{1}{N-A} \sum_{y_n=0} (x_n - \mu_0)^\top (x_n - \mu_0) \\ \widetilde{\Sigma}_1 &= \frac{1}{A} \sum_{y_n=1} (x_n - \mu_1)^\top (x_n - \mu_1) \end{aligned} \quad (13)$$

On peut réécrire (12) sous la forme suivante :

$$\ell(\Theta) = -N \ln(2\pi) + A \ln \pi + (N - A) \ln(1 - \pi) + \frac{N}{2} \ln \det \Lambda - \frac{N - A}{2} \text{tr}(\tilde{\Sigma}_0 \Lambda) - \frac{A}{2} \text{tr}(\tilde{\Sigma}_1 \Lambda) \quad (14)$$

Dérivant (14) par rapport à Λ :

$$\frac{\partial \ell(\mu_0, \mu_1, \Lambda, \pi)}{\partial \Lambda} = \frac{N}{2} \Lambda^{-1} - \frac{N - A}{2} \tilde{\Sigma}_0 - \frac{A}{2} \tilde{\Sigma}_1 = 0 \implies \Lambda^{-1} = \frac{1}{N} \left((N - A) \tilde{\Sigma}_0 + A \tilde{\Sigma}_1 \right)$$

D'où finalement :

$$\widehat{\Sigma_{MLE}} = \frac{1}{N} \left((N - A) \tilde{\Sigma}_0 + A \tilde{\Sigma}_1 \right) \quad (15)$$

2.1.2 Comparaison avec la régression logistique

Le modèle d'une régression logistique repose sur un modèle (ω, \mathbf{b}) tel que :

$$p(y = 1|x) = \sigma(\omega^\top x + \mathbf{b}) \quad (16)$$

Dans le cas du *LDA*, on peut également calculer cette probabilité comme suit :

$$\begin{aligned} p(y = 1|x) &= \frac{p(x|y = 1)p(y = 1)}{p(x)} \\ &= \frac{p(x|y = 1)p(y = 1)}{p(x|y = 0)p(y = 0) + p(x|y = 1)p(y = 1)} \\ &= 1 / \left(1 + \frac{p(x|y = 0)p(y = 0)}{p(x|y = 1)p(y = 1)} \right) \\ &= 1 / \left(1 + \frac{1 - \pi}{\pi} \exp \left(-\frac{1}{2}(x - \mu_0)^\top \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^\top \Sigma^{-1}(x - \mu_1) \right) \right) \\ &= \sigma \left(-\ln \left(\frac{1 - \pi}{\pi} \right) - \frac{1}{2} Q(\mu_0, \mu_1, \Sigma^{-1}, x) \right) \end{aligned} \quad (17)$$

Où :

$$\begin{aligned} Q(\mu_0, \mu_1, \Sigma^{-1}, x) &= -(x - \mu_0)^\top \Sigma^{-1}(x - \mu_0) + (x - \mu_1)^\top \Sigma^{-1}(x - \mu_1) \\ Q(\mu_0, \mu_1, \Sigma^{-1}, x) &= -x^\top \Sigma^{-1}x + \mu_0^\top \Sigma^{-1}x + x^\top \Sigma^{-1}\mu_0 - \mu_0^\top \Sigma^{-1}\mu_0 \\ &\quad + x^\top \Sigma^{-1}x - \mu_1^\top \Sigma^{-1}x - x^\top \Sigma^{-1}\mu_1 + \mu_1^\top \Sigma^{-1}\mu_1 \\ Q(\mu_0, \mu_1, \Sigma^{-1}, x) &= 2(\mu_0 - \mu_1)^\top \Sigma^{-1}x + \mu_1^\top \Sigma^{-1}\mu_1 - \mu_0^\top \Sigma^{-1}\mu_0 \end{aligned}$$

Ainsi, le terme quadratique disparaît sous l'hypothèse de l'égalité des matrices de covariance entre les deux gaussiennes, d'où le L de LDA, et on peut écrire (17) sous la forme :

$$\begin{aligned} p(y = 1|x) &= \sigma \left((\mu_1 - \mu_0)^\top \Sigma^{-1}x + \mu_0^\top \Sigma^{-1}\mu_0 - \mu_1^\top \Sigma^{-1}\mu_1 - \ln \left(\frac{1 - \pi}{\pi} \right) \right) \\ p(y = 1|x) &= \sigma(\omega_{LDA}^\top x + \mathbf{b}_{LDA}) \end{aligned} \quad (18)$$

Où :

$$\begin{aligned} \omega_{LDA} &= \Sigma^{-1}(\mu_1 - \mu_0) \\ \mathbf{b}_{LDA} &= \mu_0^\top \Sigma^{-1}\mu_0 - \mu_1^\top \Sigma^{-1}\mu_1 - \ln \left(\frac{1 - \pi}{\pi} \right) \end{aligned} \quad (19)$$

La LDA est donc équivalent à une régression logistique de paramètres $(\omega_{LDA}, \mathbf{b}_{LDA})$

2.1.3 Résultats numériques

Jeu de données d'entraînement	$\widehat{\pi}_{MLE}$	$\widehat{\mu_{0,MLE}}$	$\widehat{\mu_{1,MLE}}$	$\widehat{\Sigma}_{MLE}$
A	0.333	$\begin{pmatrix} 2.89 \\ -0.894 \end{pmatrix}$	$\begin{pmatrix} -2.69 \\ 0.866 \end{pmatrix}$	$\begin{pmatrix} 2.44 & -1.13 \\ -1.13 & 0.614 \end{pmatrix}$
B	0.500	$\begin{pmatrix} 3.34 \\ -0.835 \end{pmatrix}$	$\begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix}$	$\begin{pmatrix} 3.35 & -0.135 \\ -0.135 & 1.74 \end{pmatrix}$
C	0.625	$\begin{pmatrix} 2.79 \\ -0.838 \end{pmatrix}$	$\begin{pmatrix} -2.94 \\ -0.958 \end{pmatrix}$	$\begin{pmatrix} 2.88 & -0.634 \\ -0.634 & 5.20 \end{pmatrix}$

TABLE 1 – Valeurs numériques des estimateurs MLE du modèle LDA

Les résultats numériques obtenus pour les estimateurs pour le modèle LDA sont présentés en Table 1, Graphiquement, on représente également en Figure 1 les zones de \mathbb{R}^2 en deux couleurs suivant la classe attribuée par le modèle LDA déterminé.

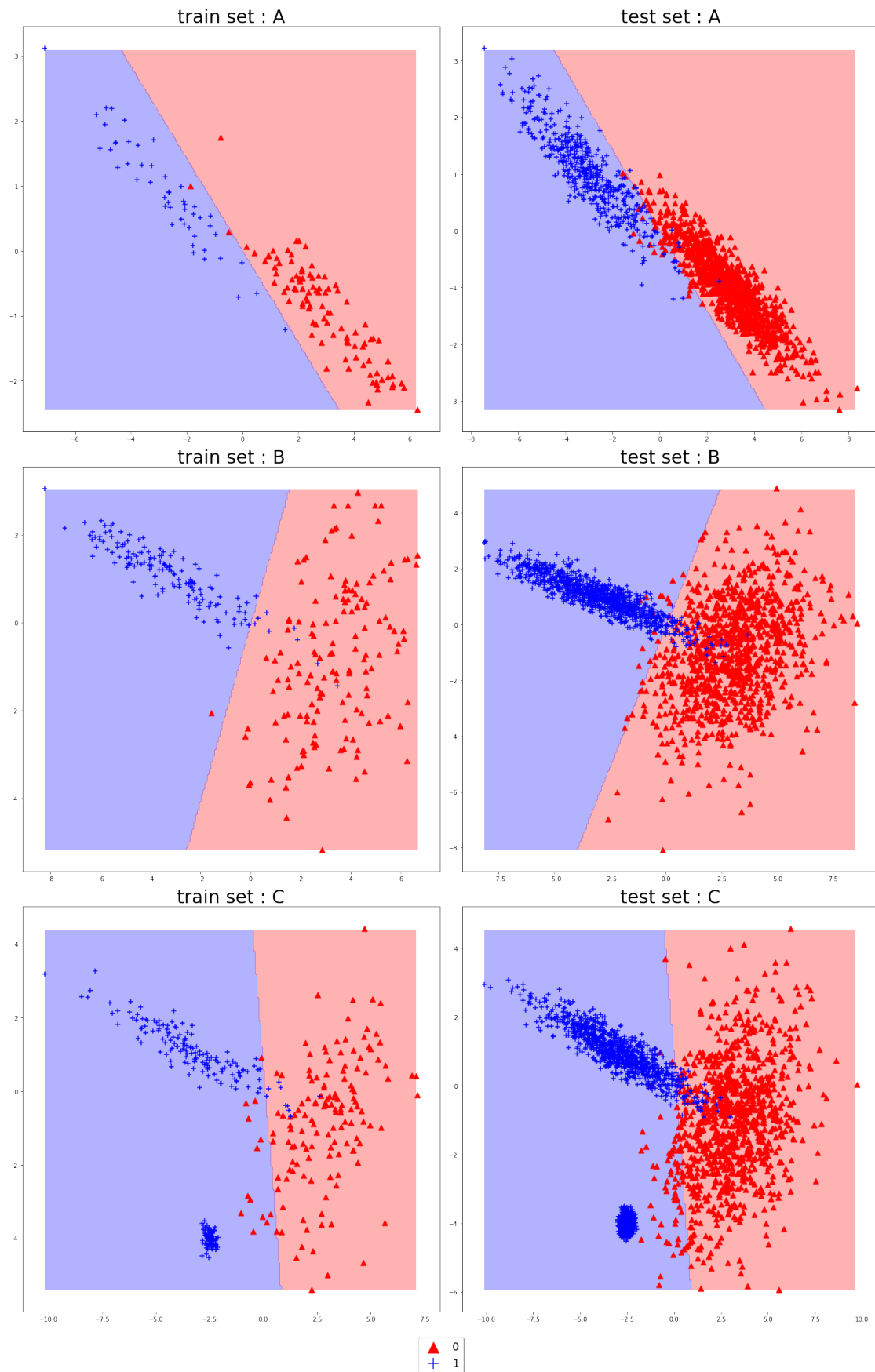


FIGURE 1 – Représentation graphique obtenue pour le modèle LDA sur les 3 jeux de données A, B, C respectivement de en haut en bas. Sur la colonne de gauche, les sous-ensembles d'entraînement. Sur la colonne de droite les sous-ensembles de test. La courbe de transition de la classe bleue à la classe rouge est définie par l'équation $p(y = 1|x) = 0.5$

2.2 Logistic regression

Jeu de données d'entraînement	$\widehat{\omega}_{MLE}$	\widehat{b}_{MLE}
A	$\begin{pmatrix} -16.2 \\ -27.2 \end{pmatrix}$	-3.09
B	$\begin{pmatrix} -1.71 \\ 1.02 \end{pmatrix}$	1.35
C	$\begin{pmatrix} -2.20 \\ 0.709 \end{pmatrix}$	0.959

TABLE 2 – Valeurs numériques des estimateurs MLE obtenus par IRLS du modèle de régression logistique

Les résultats numériques obtenus pour les estimateurs pour le modèle de régression logistique sont présentés en Table 2, Graphiquement, on représente également en Figure 2 les zones de \mathbb{R}^2 en deux couleurs suivant la classe attribuée par le modèle de régression logistique déterminé.

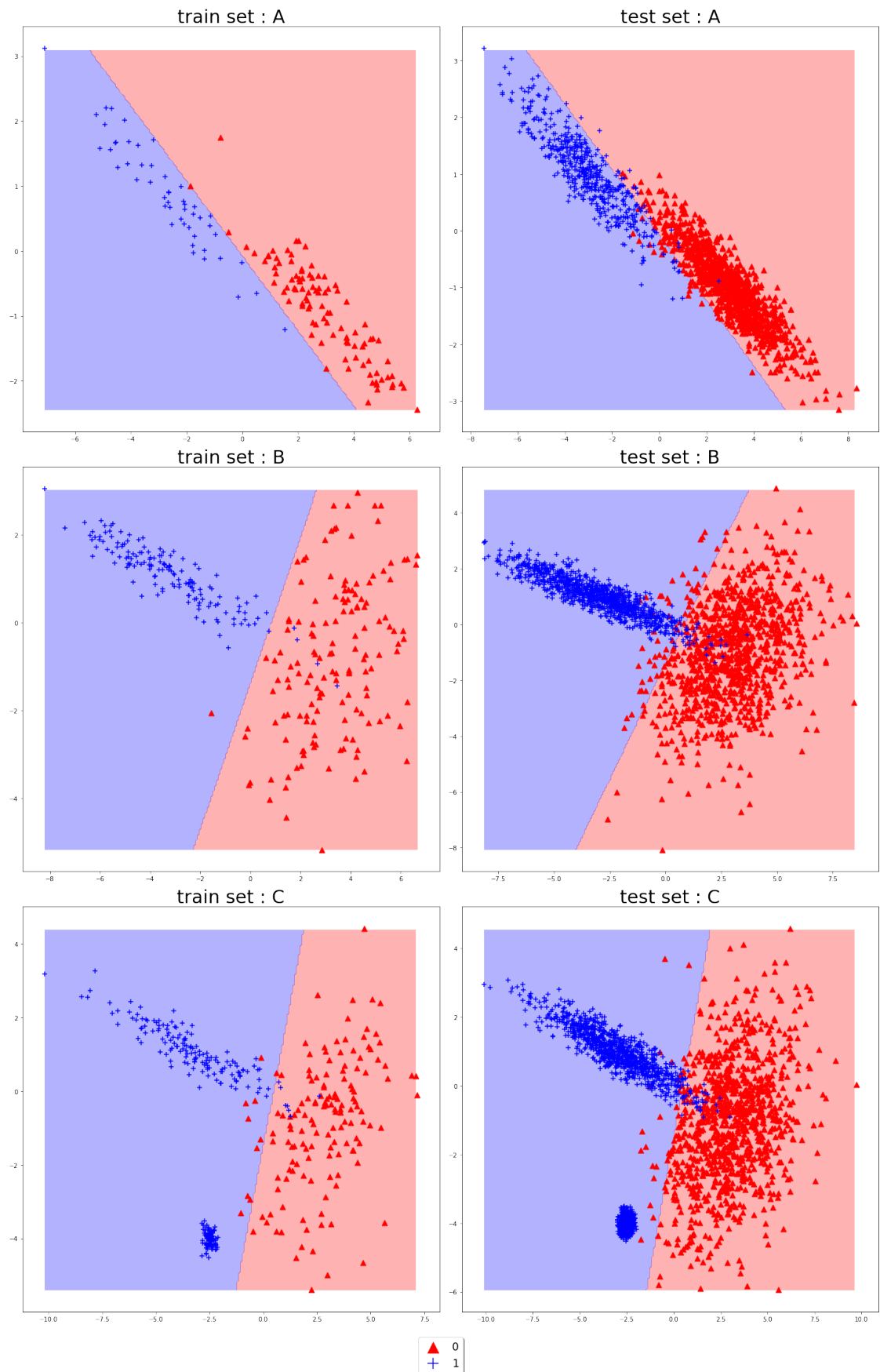


FIGURE 2 – Représentation graphique obtenue pour le modèle de régression logistique sur les 3 jeux de données A, B, C respectivement de en haut en bas. Sur la colonne de gauche, les sous-ensembles d’entraînement. Sur la colonne de droite les sous-ensembles de test. La courbe de transition de la classe bleue à la classe rouge est définie par l’équation $p(y = 1|x) = 0.5$

2.3 Linear regression

Jeu de données d'entraînement	$\widehat{\omega}_{MLE}$	\widehat{b}_{MLE}	$\widehat{\sigma}_{MLE}^2$
A	$\begin{pmatrix} -0.264 \\ -0.373 \end{pmatrix}$	0.492	0.0399
B	$\begin{pmatrix} -0.104 \\ 0.0518 \end{pmatrix}$	0.500	0.0543
C	$\begin{pmatrix} -0.128 \\ -0.0170 \end{pmatrix}$	0.508	0.0622

TABLE 3 – Valeurs numériques des estimateurs MLE obtenus par les équations normales du modèle de régression linéaire

Les résultats numériques obtenus pour les estimateurs pour le modèle de régression linéaire sont présentés en Table 3, Graphiquement, on représente également en Figure 3 les zones de \mathbb{R}^2 en deux couleurs suivant la classe attribuée par le modèle de régression logit déterminé. La courbe de séparation pour $p(y = 1|x) = 0.5$ n'est plus linéaire comme précédemment du fait de la forme quadratique qui intervient dans cette probabilité qui suit une loi de distribution gaussienne $\sim N(\omega^\top X + b, \sigma^2)$

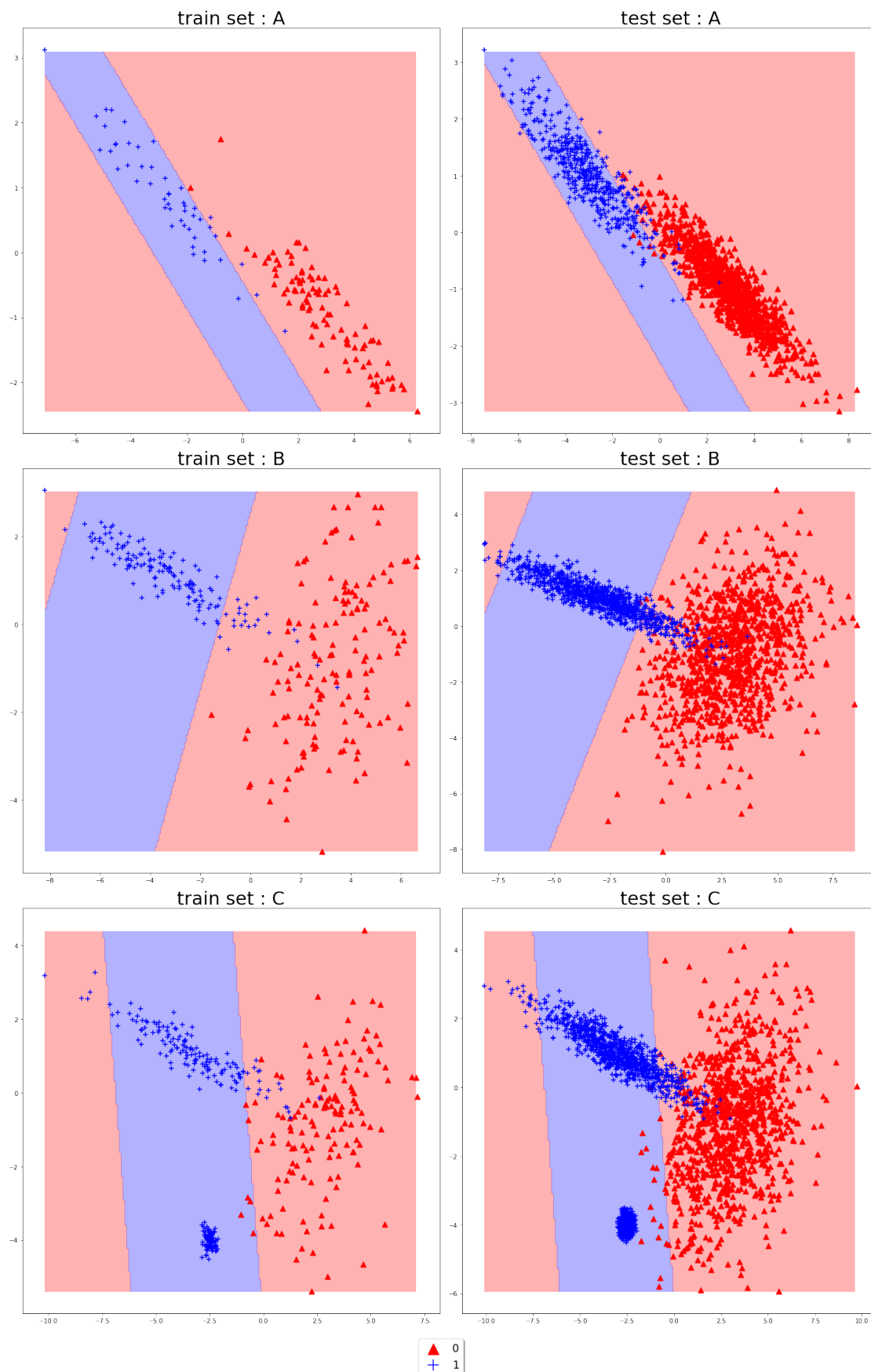


FIGURE 3 – Représentation graphique obtenue pour le modèle de régression linéaire sur les 3 jeux de données A, B, C respectivement de en haut en bas. Sur la colonne de gauche, les sous-ensembles d’entraînement. Sur la colonne de droite les sous-ensembles de test. La courbe de transition de la classe bleue à la classe rouge est définie par l’équation $p(y = 1|x) = 0.5$

2.4 Compare previous models

Modèle	A train	A test	B train	B test	C train	C test
LDA	0.0133	0.0200	0.0300	0.0415	0.0550	0.0423
Régression logistique	0.000	0.0326	0.0200	0.0430	0.0400	0.0227
Régression linéaire	0.0400	0.0473	0.0800	0.0965	0.0725	0.0610

TABLE 4 – Erreur de classification de chacun des trois modèles précédemment présentés, sur chaque jeu de données d'apprentissage (train) et de test.

L'erreur de classification en entraînement et en test a été calculée et est présentée en Table 4. D'après cette table, on constate que :

- A une exception près (qui est la régression logistique sur C test), l'erreur en apprentissage est moindre qu'en test. Cela va de soi, puisque les estimateurs sont construits de sorte à maximiser la vraisemblance des observations déjà effectuées, et non pas celles qui ne l'ont pas encore été.
- La LDA et la régression linéaire donne des erreurs similaires en test et en apprentissage, tandis que la régression logistique a tantôt une erreur nulle en apprentissage mais pas en test, tantôt une erreur similaire en apprentissage et en test, tantôt une erreur plus faible en test qu'en apprentissage. Cela s'explique par le fait que la régression logistique n'est parfaitement capable que de séparer des données linéairement séparables.
- Sur le premier jeu de données, linéairement séparables en apprentissage, la régression logistique donne une erreur nulle. La LDA donne une erreur entre 1 % en apprentissage et 2 % en test, stable avec la multiplication de données à la frontière de décision. La stabilité et la relative performance s'expliquent par la vérification de l'hypothèse de distribution des données suivant deux gaussiennes très étirées, de matrice de covariance similaire, avec peu de recouvrement.
- Sur le troisième jeu de données, en comparant le modèle LDA au modèle de régression logistique, on constate que la direction de la frontière de décision est largement différente, résultant, dans le cas du modèle LDA, dans le passage d'une partie non négligeable du premier noeud gaussien dans la zone opposée du classifieur. Cette attraction de la frontière par le 3^e noeud bleu, s'explique par la masse représentée par ce dernier noeud dans les estimateurs MLE, masse qui n'intervient pas dans l'évaluation itérative de ω en régression logistique, une fois qu'elle est passée dans le bon côté de la frontière de décision pour un ω donné.
- La régression linéaire, pour sa part, donne des résultats toujours substantiellement moins bons.
- Le jeu de données C ne peut plus raisonnablement être considéré comme issu de la distribution de deux gaussiennes. On y constate au moins 3 noeuds, et tous les modèles présentés y performant le moins bien, à part le modèle de régression logistique en test pour une raison déjà expliquée.

2.5 QDA model

Jeu de données d'entraînement	$\widehat{\pi}_{MLE}$	$\widehat{\mu}_{0,MLE}$	$\widehat{\mu}_{1,MLE}$	$\widehat{\Sigma}_{0,MLE}$	$\widehat{\Sigma}_{1,MLE}$
A	0.333	$\begin{pmatrix} 2.89 \\ -0.984 \end{pmatrix}$	$\begin{pmatrix} -2.69 \\ 0.866 \end{pmatrix}$	$\begin{pmatrix} 2.31 & -1.05 \\ -1.05 & 0.576 \end{pmatrix}$	$\begin{pmatrix} 2.70 & -1.30 \\ -1.30 & 0.690 \end{pmatrix}$
B	0.500	$\begin{pmatrix} 3.34 \\ -0.835 \end{pmatrix}$	$\begin{pmatrix} -3.22 \\ 1.08 \end{pmatrix}$	$\begin{pmatrix} 3.35 & -0.135 \\ -0.135 & 1.74 \end{pmatrix}$	$\begin{pmatrix} 4.15 & -1.33 \\ -1.33 & 0.516 \end{pmatrix}$
C	0.625	$\begin{pmatrix} 2.79 \\ -0.838 \end{pmatrix}$	$\begin{pmatrix} -2.94 \\ -0.958 \end{pmatrix}$	$\begin{pmatrix} 2.90 & 1.24 \\ 1.24 & 2.92 \end{pmatrix}$	$\begin{pmatrix} 2.87 & -1.76 \\ -1.76 & 6.56 \end{pmatrix}$

TABLE 5 – Valeurs numériques des estimateurs MLE du modèle QDA

Les résultats numériques obtenus pour les estimateurs pour le modèle de régression linéaire sont présentés en Table 5, Graphiquement, on représente également en Figure 4 les zones de \mathbb{R}^2 en deux couleurs suivant la classe attribuée par le modèle de régression logitique déterminé. La courbe de séparation pour $p(y = 1|x) = 0.5$ n'est plus linéaire comme précédemment dans le modèle LDA du fait de la forme quadratique qui intervient dans cette probabilité.

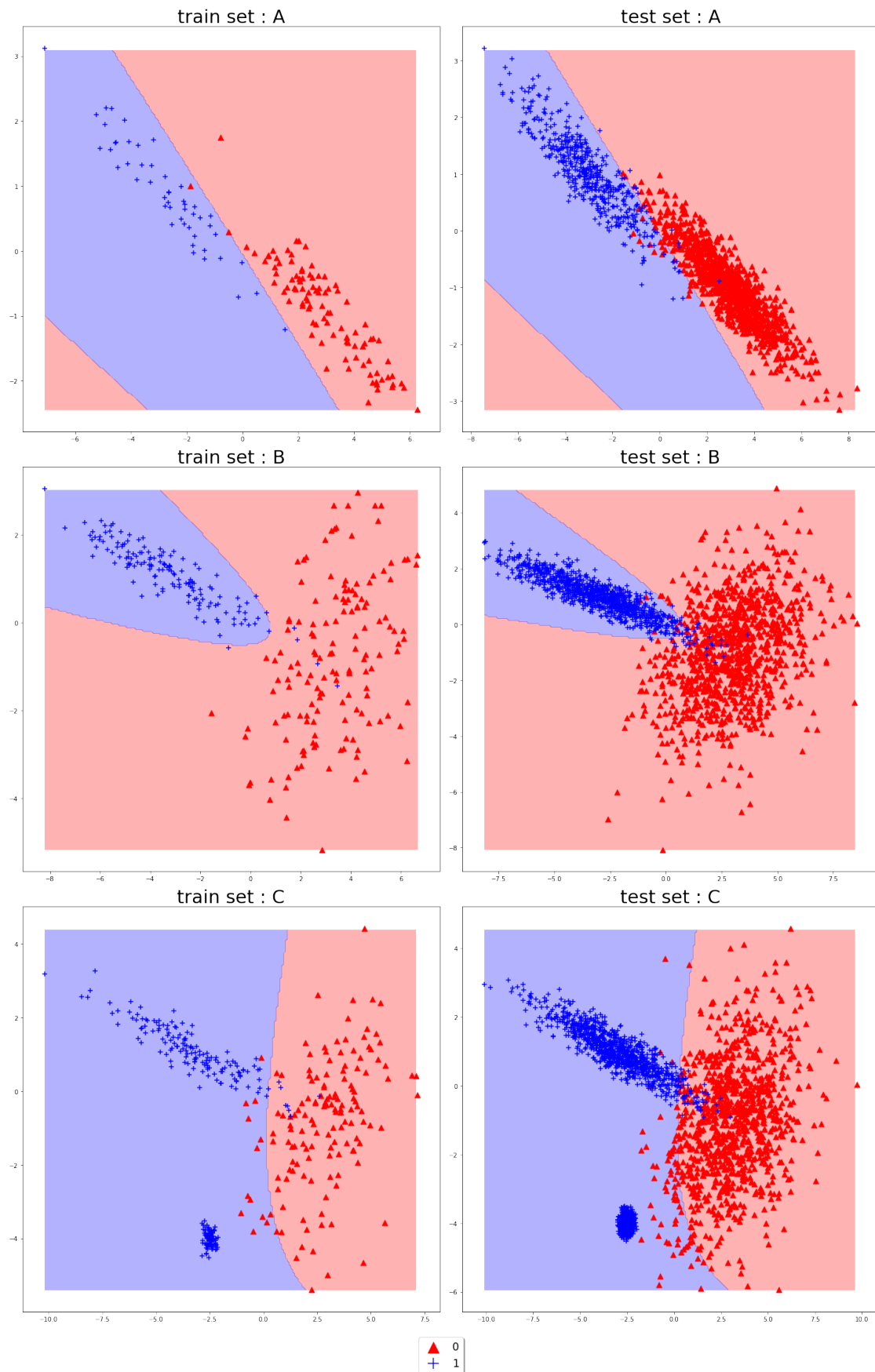


FIGURE 4 – Représentation graphique obtenue pour le modèle QDA sur les 3 jeux de données A, B, C respectivement de en haut en bas. Sur la colonne de gauche, les sous-ensembles d'entraînement. Sur la colonne de droite les sous-ensembles de test. La courbe de transition de la classe bleue à la classe rouge est définie par l'équation $p(y = 1|x) = 0.5$

Modèle	A train	A test	B train	B test	C train	C test
LDA	0.0133	0.0200	0.0300	0.0415	0.0550	0.0423
Régression logistique	0.000	0.0326	0.0200	0.0430	0.0400	0.0227
Régression linéaire	0.0400	0.0473	0.0800	0.0965	0.0725	0.0610
QDA	0.00667	0.0187	0.0233	0.0234	0.0525	0.0403

TABLE 6 – Erreur de classification de chacun des trois modèles précédemment présentés et du modèle QDA, sur chaque jeu de données d'apprentissage (train) et de test

L'erreur de classification en entraînement et en test a été calculée et est présentée et rajoutée au tableau déjà existant en Table 6. D'après cette table, on constate que : pour les jeux de données A et B, c'est le modèle QDA qui établit le plus petit taux d'erreur de classification. Le dernier jeu de set, C, pose toujours problème, puisqu'étant issus d'au moins 3 neuds gaussiens distincts. On constate également que l'erreur en apprentissage et en test est très similaire. Cela peut s'interpréter aisément étant donné que la frontière de décision à $p(y = 1|x)$, régulière et quadratique, permet d'aller chercher des points de même classe inaccessible avec des séparations linéaires tout en étant assez généralisatrice (par sa régularité), ce que le modèle LDA ne peut pas faire. D'autre part, si on compare toujours au modèle LDA, on constate que les matrices de covariances estimées par le modèle QDA en A et B sont assez proches pour chaque classe, et que le modèle LDA devrait donc être un modèle valable pour ces jeux de données A et B. Pour autant, il y a jusqu'à un facteur 2 d'erreur en moins en apprentissage et en test pour le QDA par rapport au LDA, explicable donc principalement par la forme quadratique de la frontière de décision.