DM 1 Probalistic Graphical Model

Arthur Lavergne and Tamim El Ahmad

22 novembre 2019

1 Learning in discrete Graphical Models

On a $p(z = m) = \pi_m$ et $p(x = k|z = m) = \theta_{mk}$. z peut prendre M valeurs différentes et x peut prendre K valeurs différentes.
On considère N échantillons $(x^i,\ z^i)$. Pour écrire correctement le maximum de vraisemblance du modèle, on se ramène au cas multinomial étudié dans le cours. Ainsi on a :

$$\begin{aligned} \forall i \in 1, ..., N, z^i = m \\ \iff Z^i = (Z_1^i, ..., Z_M^i)^T \\ \text{avec} \quad \{Z_m^i = 1 \quad \text{et} \quad Z_l = 0 \quad \forall l \neq m\} \end{aligned}$$

De même, on a pour (x=k|z=m) :

$$\begin{aligned} \forall i \in 1, ..., N, (x^i|z = m) = k \\ \iff (X^i|z = m) = ((X_1^i|z = m), ..., X_K^i|z = m))^T \\ \text{avec} \quad \{(X_k^i|z = m) = 1 \quad \text{et} \quad (X_l|z = m) = 0 \quad \forall l \neq k\} \end{aligned}$$

On peut maintenant écrire la vraisemblance du modèle :

$$\begin{aligned} L(\theta, \pi) &= \prod_{i=1}^{N} p(x^i, z^i|\theta, \pi) \\ &= \prod_{i=1}^{N} p(x^i|\theta, \pi, z^i)p(z^i|\theta, \pi) \end{aligned}$$

Or, on sait que :

$$p(z^i|\theta, \pi) = \prod_{m=1}^{M} \pi_m^{z_m^i}$$

et que :

$$p(x_i|\theta, z^i) = \prod_{k=1}^{K} (p(x^i = k|z^i, \theta))^{x_k^i}$$

$$\text{comme} \quad z^i \in \{1, ..., M\} \quad \text{il vient que :}$$

$$p(x_i|\theta, z^i) = \prod_{k=1}^{K} \prod_{m=1}^{M} \theta_{mk}^{x_k^i z_m^i}$$

finalement, on obtient la vraisemblance :

$$\boxed{L(\theta, \pi) = \prod_{i=1}^{N} \prod_{m=1}^{M} \pi_m^{z_m^i} \prod_{i=1}^{N} \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{mk}^{z_m^i x_k^i}} \tag{1}$$

A partir de l'équation (1), on peut calculer la log vraisemblance :

$$\mathcal{L}(\theta, \pi) = \sum_{i=1}^{N} \sum_{m=1}^{M} z_m^i log(\pi_m) + \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} z_m^i x_k^i log(\theta_{mk})$$

On peut donc poser le problème d'optimisation suivant :

$$\max_{\theta, \pi} \quad \sum_{i=1}^{N} \sum_{m=1}^{M} z_m^i log(\pi_m) + \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} z_m^i x_k^i log(\theta_{mk})$$

$$\underline{\text{sujet à :}} \quad \sum_{m=1}^{M} \pi_m = 1, \quad \forall m \in 1, ..., M, \sum_{k=1}^{K} \theta_{mk} = 1$$

On peut séparer ce problème d'optimisation en deux problèmes distinct :

$$\begin{aligned} \max_{\pi} \quad & \sum_{i=1}^{N} \sum_{m=1}^{M} z_m^i log(\pi_m) \\ \text{s.t.} \quad & \sum_{m=1}^{M} \pi_m = 1 \end{aligned} \tag{P}$$

$$\begin{aligned} \max_{\theta} \quad & \sum_{i=1}^{N} \sum_{m=1}^{M} \sum_{k=1}^{K} z_m^i x_k^i log(\theta_{mk}) \\ \text{s.t.} \quad & \forall m \in 1, ..., M, \sum_{k=1}^{K} \theta_{mk} = 1 \end{aligned} \tag{Q}$$

On peut changer l'écriture du problème (P) :

$$\text{on a :} \quad \sum_{i=1}^{N}\sum_{m=1}^{M} z_m^i log(\pi_m) = \sum_{m=1}^{M} log(\pi_m) \sum_{i=1}^{N} z_m^i$$
$$= \sum_{m=1}^{M} n_m log(\pi_m)$$
$$\text{en posant :} \quad n_m = \sum_{i=1}^{N} z_m^i$$

Le problème (P) est équivalent à :

$$\min_{\pi} - \sum_{m=1}^{M} n_m log(\pi_m)$$
$$\text{s.t.} \quad \sum_{m=1}^{M} \pi_m = 1$$

Le problème est clairement convexe, et on peut donc le minimiser via la maximisation du lagrangien du problème. Le lagrangien du problème est donné par :

$$g(\pi, \lambda) = - \sum_{m=1}^{M} n_m log(\pi_m) + \lambda(\sum_{m=1}^{M} \pi_m - 1)$$

On dérive par rapport à π_m et on a :

$$\frac{\partial g(\pi, \lambda)}{\partial \pi_m} = -\frac{n_m}{\pi_m} + \lambda = 0$$
$$\Longleftrightarrow \pi_m = \frac{n_m}{\lambda}, \forall m \in \{1, ..., M\}$$

En utilisant le contrainte du problème, il vient que :

$$\sum_{m=1}^{M} \pi_m = 1 \Longleftrightarrow \sum_{m=1}^{M} n_m = \lambda \Longrightarrow \lambda = N$$
$$\text{d'où :} \boxed{\hat{\pi}_m = \frac{n_m}{N}}, \forall m \in \{1, ..., M\}$$

En utilisant la même méthode que pour le problème (P), on pose $n_{mk} = \sum_{i=1}^{N} z_m^i x_k^i$, et le problème devient :

$$\min_{\theta} - \sum_{m=1}^{M} \sum_{k=1}^{K} n_{mk} log(\theta_{mk})$$
$$\text{s.t.} \quad \forall m \in \{1, ..., M\}, \sum_{k=1}^{K} \theta_{mk} = 1$$

La lagragien du problème s'écrit :

$$g(\theta, \lambda_1, ..., \lambda_M) = - \sum_{m=1}^{M} \sum_{k=1}^{K} n_{mk} log(\theta_{mk}) + \sum_{m=1}^{M} \lambda_m (\sum_{k=1}^{K} \theta_{mk} - 1)$$

La maximisation du lagrangien par rapport à θ_{mk} nous donne donc :

$$\frac{\partial g(\theta, \lambda_1, ..., \lambda_M)}{\partial \theta_{mk}} = - \frac{n_{mk}}{\theta_{mk}} + \lambda_m = 0$$
$$\text{d'où} \quad \theta_{mk} = \frac{n_{mk}}{\lambda_m}, \forall m \in \{1, ..., M\}$$

En utilisant la contrainte $\sum_{k=1}^{K} \theta_{mk} = 1, \forall m \in \{1, ..., M\}$, on a :

$$\sum_{k=1}^{K} \frac{n_{mk}}{\lambda_m} = 1$$
$$\text{donc,} \quad \sum_{k=1}^{K} n_{mk} = \lambda_m$$
$$\text{puis :} \quad \lambda_m = n_m = \sum_{i=1}^{N} z_m^i$$
$$\text{et enfin, on trouve :} \boxed{\hat{\theta}_{mk} = \frac{n_{mk}}{n_m}}$$

A partie de là, nous pouvons donc implémenter l'algorithme de descente de gradient, et trouver le vecteur w qui permet résoudre le problème d'optimisation.

2 Linear Classiffication

2.1 Generative Model

Question 1

Soit y $\hookrightarrow$ Bernouilli(θ) et $(x|y = i) \hookrightarrow \mathcal{N}(\mu_1, \Sigma) \in \mathbb{R}^2$.
Soient $(x^n, y^n)_{1 \leq n \leq N}$, N échantillons i.i.d.

On commence par calculer la vraisemblance :

$$
\begin{aligned}
L_{x,y}(\theta,\mu_0,\mu_1,\Sigma) &= \prod_{n=1}^{N} p(x^n, y^n|\theta,\mu_0,\mu_1,\Sigma)\\
&= \prod_{i=1}^{N} p(y^n|\theta,\mu_0,\mu_1,\Sigma)p(x^n|y^n,\theta,\mu_0,\mu_1,\Sigma)
\end{aligned}
$$

Comme y suit une loi de Bernoulli de paramètre θ, on a :

$$p(y|\theta,\mu_0,\mu_1,\Sigma) = p(y|\theta) = \theta^y(1-\theta)^{1-y}$$

De plus, comme $(x|y=i) \hookrightarrow \mathcal{N}(\mu_i,\Sigma)$, on déduit que :

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

Donc, si y=0, on a :

$$p(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)) = p(x|0,\theta,\mu_0,\mu_1,\Sigma) \tag{2}$$

De même, si y=1, on a :

$$p(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)) = p(x|1,\theta,\mu_0,\mu_1,\Sigma) \tag{3}$$

Donc, en injectant (2) et (3) dans l'expression de la vraisemblance et en passant au log pour avoir la log-vraisemblance, on a :

$$
\begin{aligned}
\mathcal{L}_{x,y}(\theta,\mu_0,\mu_1,\Sigma) = \sum_{n=1}^{N}[&y^n\log(\theta) + (1-y^n)\log(1-\theta) - \frac{d}{2}\log(2\pi)\\
&-\frac{1/2}{\log}(|\Sigma|) + y^n(-\frac{1}{2}(x^n-\mu_1)^T\Sigma^{-1}(x^n-\mu_1)\\
&+(1-y^n)(-\frac{1}{2}(x^n-\mu_0)^T\Sigma^{-1}(x^n-\mu_0))]
\end{aligned}
$$

On cherche maintenant à maximiser la log-vraisemblance en prenant le gradient par rapport à chacun des paramètre :

Paramètre θ :

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{n=1}^{N} (\frac{y^n}{\theta} - \frac{1-y^n}{1-\theta})$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = 0 \Longleftrightarrow \frac{1}{\theta}\sum_{n=1}^{N} y^n = \frac{1}{1-\theta}(N - \sum_{n=1}^{N} y^n)$$

$$\Longleftrightarrow \theta = \frac{1}{N - \sum_{n=1}^{N} y^n} \frac{(\sum_{n=1}^{N} y^n)(N - \sum_{n=1}^{N} y^n)}{N}$$

$$\Longleftrightarrow \boxed{\theta = \frac{\sum_{n=1}^{N} {}^{n}}{y} N}$$

Paramètre μ_1 :

$$\begin{aligned} \text{on pose : } \quad A_n &= (x^n - \mu_1)\Sigma^{-1}(x^n - \mu_1) \\ &= x^n \Sigma^{-1} x^n - 2\mu_1^T \Sigma^{-1} x^n + \mu_1^T \Sigma^{-1} \mu_1 \\ \text{car : } x^n \Sigma^{-1} \mu_1 &= (\Sigma^{-1}\mu_1)^T x^n = \mu_1 \Sigma^{-1} x^n \end{aligned}$$

$$\text{d'où : } \quad \frac{dA_n}{d\mu_1} = -2\Sigma^{-1}x^n + 2\Sigma^{-1}\mu_1 \quad \text{car} \quad \Sigma^{-1} \quad \text{est symétrique}$$

Ainsi, on en déduit que :

$$\frac{\partial \mathcal{L}}{\partial \mu_1} = 0 \Longleftrightarrow \Sigma^{-1}\mu_1(\sum_{n=1}^{N} y^n) = \Sigma^{-1}(\sum_{n=1}^{N} y^n x^n)$$

$$\Longleftrightarrow \boxed{\mu_1 = \frac{\sum_{n=1}^{N} y^n x^n}{\sum_{n=1}^{N} y^n}}$$

Paramètre μ_0 : On procède de manière équivalente :

$$\frac{\partial \mathcal{L}}{\partial \mu_0} = \sum_{n=1}^{N} -\frac{1-y^n}{2}(2\Sigma^{-1}x^n + 2\Sigma^{-1}\mu_0) = \sum_{n=1}^{N}(1-y^n)(\Sigma^{-1}x^n - \Sigma^{-1}\mu_0)$$

$$\text{donc : } \quad \frac{\partial \mathcal{L}}{\partial \mu_0} = 0 \Longleftrightarrow \Sigma^{-1}\mu_0(\sigma_{n=1}^{N}(1-y^n)) = \Sigma^{-1}(\sum_{n=1}^{N}(1-y^n)x^n)$$

$$\text{puis : } \quad \boxed{\mu_0 = \frac{\sum_{n=1}^{N} x^n(1-y^n)}{N - \sum_{n=1}^{N} y^n}}$$

Paramètre Σ :
On pose :

$$\begin{aligned} B_n &= y^n(x^n-\mu_1)^T\Sigma^{-1}(x^n-\mu_1) \\ C_n &= (1-y^n)(x^n-\mu_0)^T\Sigma^{-1}(x^n-\mu_0) \\ D &= -\log(|\Sigma|) = \log(\Sigma^{-1}) \\ M &= \Sigma^{-1} \end{aligned}$$

Comme $B_n = y^n(x^n-\mu_1)^T M(x^n-\mu_1) \in \mathbb{R}$. Donc :

$$\begin{aligned} B_n &= Tr(y^n(x^n-\mu_1)^T M(x^n-\mu_1)) \\ &= Tr(M(x^n-\mu_1)(x^n-\mu_1)^T) \\ &= Tr(M\tilde{\Sigma_1^n}) \quad \text{avec } \tilde{\Sigma_1^n} = (x^n-\mu_1)(x^n-\mu_1)^T \end{aligned}$$

on pose donc $f(M) = Tr(M\tilde{\Sigma_1^n})$ Pour $H \in \mathbb{R}^{d\times d}, f(M+H)-f(M) = Tr(H\tilde{\Sigma_1^n})$, d'où :

$$\boxed{\nabla f(M) = \tilde{\Sigma_1^n}}$$

Soit H$\in \mathbb{R}^{d\times d}$, définie positive :

$$\begin{aligned} \log(|M+H|) &= \log(|M^{1/2}(\mathbf{Id}+M^{-1/2}HM^{-1/2})M^{1/2}) \\ &= \log(|M|)+\log(|\mathbf{Id}+\tilde{H}|) \end{aligned}$$

$M^{1/2}$ est la matrice racine carrée de M qui existe car M est une matrice définie positive

$$\tilde{H} = M^{-1/2}HM^{-1/2} \quad \text{est définie positive}$$

Notons $(\lambda_1, \lambda_1, ..., \lambda_d)$ ses valeurs propres, et on peux exprimer le déterminant grâce à ses valeurs propres :

$$\log(|\mathbf{Id}+\tilde{H}|) = \log(\prod_{i=1}^d(1+\lambda_i)) = \sum_{i=1}^d \log(1+\lambda_i) = \sum_{i=1}^d \lambda_i + o(\|\tilde{H}\|)$$
$$\text{car} \quad \|\tilde{H}\| \to 0 \Longleftrightarrow \forall i \in \{1,...,d\}, \lambda_i \to 0$$

or, :

$$\sum_{i=1}^d \lambda_i = Tr(\tilde{H}) = Tr(M^{-1/2}H^{-1/2}) = Tr(HM^{-1/2}M^{-1/2}) = Tr(M^{-1})$$

On en déduit donc que :

$$\log(|M+H|)-\log(|M|) = tr(HM^{-1}) + o(\|H\|)$$

car : $\tilde{H} = M^{-1/2} H M^{-1/2}$, donc $\|\tilde{H}\| \to 0 \iff \|H\| \to 0$ Puis, on :

$$\boxed{\nabla \log(|M|) = M^{-1}}$$

On obtient donc :

$$\begin{aligned} \frac{\partial B_n}{\partial M} = \tilde{\Sigma_1^n} \quad \frac{\partial C_n}{\partial M} = \tilde{\Sigma_0^n} \\ \text{avec} \quad \tilde{\Sigma_0^n} = (1 - y^n)(x^n - \mu_0)(x^n -_m u_0)^T \\ \frac{\partial D}{\partial M} = \frac{\partial}{\partial M}(log(|M|)) = M^{-1} \end{aligned}$$

Donc :

$$\begin{aligned} \frac{\partial L}{\partial M} &= \sum_{n=1}^{N} \frac{\partial}{\partial M}(\frac{1}{2}C - \frac{1}{2}B_n - \frac{1}{2}C_n) \\ &= \sum_{n=1}^{N}(\frac{1}{2}M^{-1} - \frac{1}{2}\tilde{\Sigma_1^n} - \frac{1}{2}\tilde{\Sigma_0^n}) \\ &= \frac{N}{2}M^{-1} - \frac{1}{2}\sum_{n=1}^{N}(\tilde{\Sigma_1^n} + \tilde{\Sigma_0^n}) \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial M} = 0 \iff \hat{M^{-1}} = \frac{1}{N}\sum_{n=1}^{N}\tilde{\Sigma_1^n} + \frac{1}{N}\sum_{n=1}^{N}\tilde{\Sigma_0^n} \\ \boxed{\iff \hat{\Sigma} = \frac{n_1}{N}\Sigma_1 + \frac{n_0}{N}\Sigma_0} \end{aligned}$$

Avec :

$$\begin{aligned} \Sigma_1 = \frac{1}{n_1}\sum_{n=1}^{N}\tilde{\Sigma_1^n} = \frac{1}{n_1}\sum_{i=1}^{N} y^n(x^n - \mu_1)(x^n - \mu_1)^T \\ \Sigma_0 = \frac{1}{n_0}\sum_{i=1}^{N}\tilde{\Sigma_0^n} = \frac{1}{n_0}\sum_{n=1}^{N}(1 - y^n)(x^n - \mu_0)(x^n - \mu_0)^T \end{aligned}$$

Question 2

Ici, on montre en détail comment calculer p(y=1|x) dans le cadre d'un tel modèle.

$$
\begin{aligned}
p(y|x) = p(y)p(x,y) \quad \text{(Règle de Bayes)} \\
=\propto \pi^y(1-\pi)^{1-y}\exp(|\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)y - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)(1-y)) \\
=\propto \pi^y(1-\pi)^{1-y}\exp(-\frac{1}{2}y(2(\mu_0-\mu_1)^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0)) \\
=\propto \exp(y\alpha + y\beta^T x)
\end{aligned}
$$

On a posé pour l'équation précédente :

$$\boxed{\alpha = \log(\frac{\pi}{1-\pi}) - \frac{1}{2}(\mu_1\Sigma^{-1} - \mu_0\Sigma^{-1}\mu_0)} \quad \boxed{\beta = -(\Sigma^{-1})^T(\mu_0-\mu_1)}$$

Ainsi, par une normalisation pour annuler tous les termes constants, on trouve que :

$$\boxed{p(y=1|x) = \frac{\exp(\alpha+\beta^T x)}{1+\exp(\alpha+\beta^T x)} = \sigma(\alpha+\beta^T x)}$$

2.2 Modèle QDA

Y $\hookrightarrow$Bernoulli(θ), x|y=i$\hookrightarrow$ $\mathcal{N}(\mu_i, \Sigma_i)$. De même que pour le modèle LDA, on trouve :

$$
\begin{aligned}
L_{(x,y)}(\theta,\mu_0,\mu_1,\Sigma_0,\Sigma_1) = \sum_{n=1}^{N}[y^n\log(\theta) + (1-y^n)\log(1-\theta) \\
-\frac{y^n}{2}\log(|\Sigma_1|) + y^n[-\frac{1}{2}(x^n-\mu_1)^T\Sigma_1^{-1}(x^n-\mu_1)] \\
-\frac{1-y^n}{2}\log(|\Sigma_0|) + (1-y^n) \\
\times[-\frac{1}{2}(x^n-\mu_0)^T\Sigma_0^{-1}(x^n-\mu_0)]] - N\frac{d}{2}\log(2\pi)
\end{aligned}
$$

Finalement, on a que :

$$\boxed{\hat{\theta} = \frac{n_1}{N}} \quad \boxed{\hat{\mu_1} = \frac{\sum_{n=1}^N y^n x^n}{n_1}} \quad \boxed{\hat{\mu_0} = \frac{\sum_{n=1}^N (1-x^n)x^n}{n_0}}$$

<u>Paramètre Σ_1</u>

On pose $M_1 = \Sigma_0^{-1}$ et $\tilde{\Sigma_1^n} = y^n(x^n-\mu_1)(x^n-\mu_1)^T$

Comme dans le modèle LDA, on a :

$$\frac{\partial L}{\partial M_1} = \sum_{n=1}^{N} (\frac{y^n}{2} M_1^{-1} - \frac{1}{2}\tilde{\Sigma_1^n})$$

$$\text{donc} \quad \frac{\partial L}{\partial M_1} = 0 \Longleftrightarrow (\sum_{n=1}^{N} y^n) M_1^{-1} = \sum_{n=1}^{N} y^n (x^n - \mu_1)(x^n - y^n)^T$$

$$\Longleftrightarrow \hat{\Sigma_1} = \frac{1}{n_1} \sum_{n=1}^{N} y^n (x^n - \mu_1)(x^n - \mu_1)^T$$

En posant $\mu_0 = \Sigma_0^{-1}$ et $\tilde{\Sigma_0^n} = (1 - y^n)(x^n - \mu_0)(x^n - \mu_0)^T$, et on trouve :

$$\boxed{\hat{\Sigma_0} = \frac{1}{n_0} \sum_{n=1}^{N} (1 - y^n)(x^n - \mu_0)(x^n - \mu_0)^T}$$

Question 2

$$\boxed{p(y = 1|x) = \sigma(-\frac{1}{2} x^T M x + (\mu_1^T \Sigma_1^{-1} - \mu_0^T \Sigma_0^{-1})x + \log \frac{\pi \sqrt{|\Sigma_0|}}{(1-\pi)\sqrt{|\Sigma_1|}} - \frac{1}{2}(\mu_1^T \Sigma_1^{-1} \mu_1 - \mu_0^T \Sigma_0^{-1} \mu_0))}$$

2.3 Logistic Model

Dans le modèle de régression logistique, on suppose que Y suit une loi de Bernoulli de paramètre θ, avec $\theta = \sigma(w^T x)$, où σ représente la fonction sigmoïde :

$$\begin{array}{cccc} f & : & \mathbb{R} & \rightarrow &]0, 1[\\ & & z & \mapsto & \frac{1}{1+e^{-z}} \end{array}$$

Comme vu en cours, on a la vraisemblance qui est donnée par :

$$\mathcal{L}(w) = \sum_{i=1}^{N} y_i \log(\sigma(w^T x_i)) + (1 - y_i) \log(\sigma(-w^T x_i))$$

Optimiser cette vraisemblance revient à résoudre le problème d'optimisation suivant :

$$w^{t+1} = w^t + Hl(w^t)^{-1} \nabla_w l(w)$$

$$\text{avec} \quad \nabla_w l(w) = X^T (y - \eta)$$

$$\text{et} \quad Hl(w) = -X^T Diag(\eta_i(1 - \eta_i))X$$

$$\text{en posant :} \quad \eta_i = \sigma(\theta^T x_i)$$

2.4 Linear Model

Pour le modèle linéaire, on peut directement appliquer l'équation normale à savoir : $w = (X^TX)^{-1}XY$ Dans le cadre du modèle étudié, il y a un offset b, qui peut être intégré dans le vecteur w, sous condition d'avoir ajouté une colonne de 1 à la matrice X. Dans ce cas, on peut appliquer l'équation précédente. De plus, on sait que que la variance est donné par :

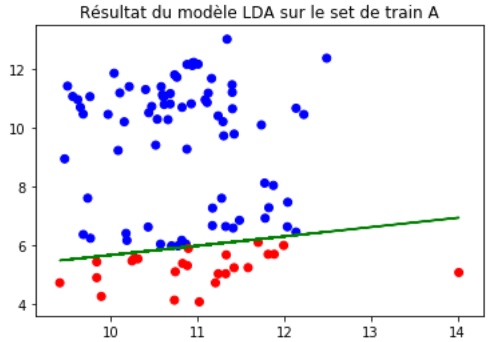
$$\hat{\sigma^2} = \frac{1}{N}(Y - Xw)^T(Y - Xw)$$

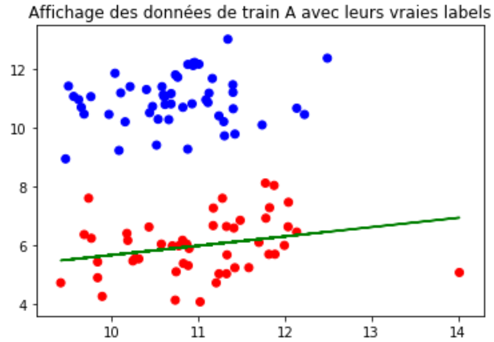
Pour tracer l'hyperplan de séparation des données, étant donné que $Y|X \sim \mathcal{N}(w^TX, \sigma^2)$, on considère que l'événement $\{Y = 1\}$ dans la classification binaire, équivaut à $\{Y > 0|X\}$. Donc, pour que cet événement soit de probabilité $\frac{1}{2}$, tout dépend de sa moyenne w^TX :

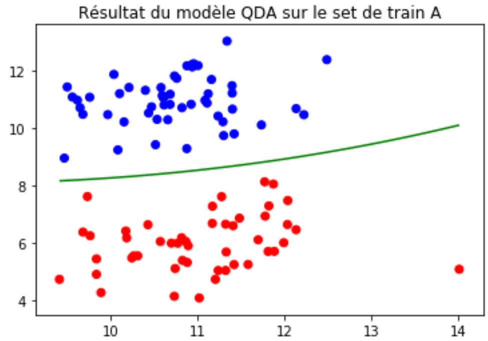
$$p(y = 1|x) = \frac{1}{2} \iff w^TX = 0$$

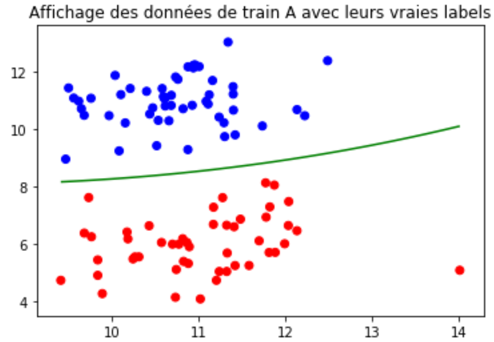
3 Applications

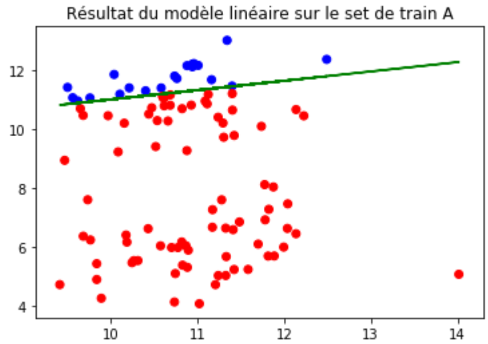
3.1 Dataset train A

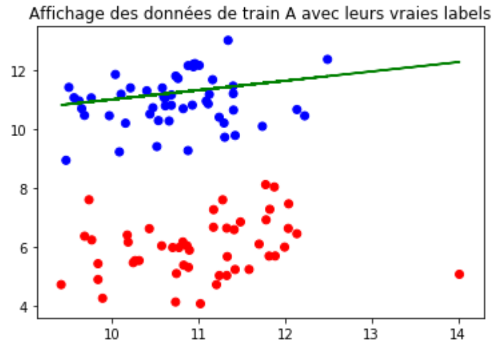


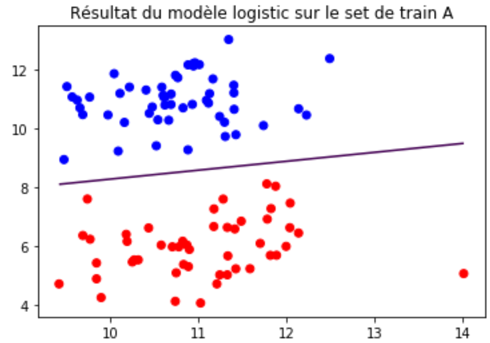


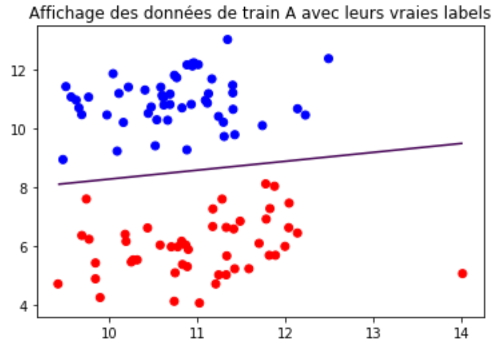






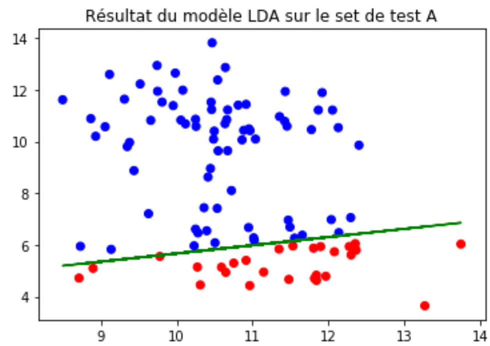


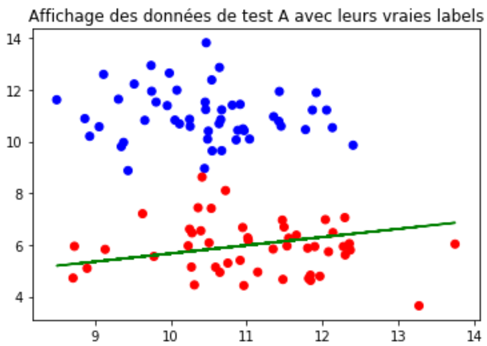


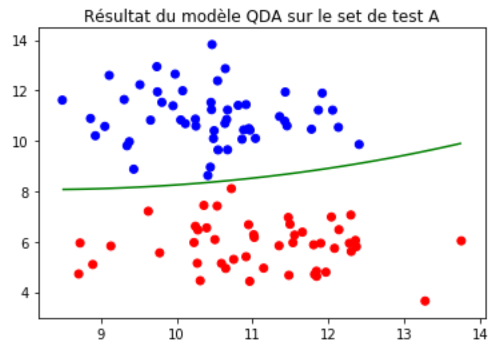


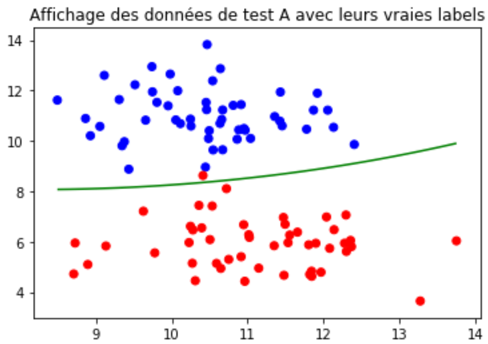
Récapitulatif des erreurs pour le train A	
modele LDA	0.24
modele QDA	0.0
modele Linéaire	0.32
modèle logistique	0.0

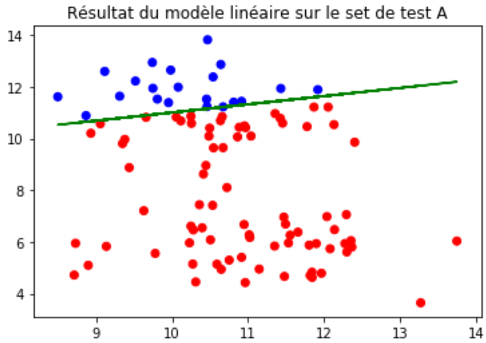
3.2 Dataset test A

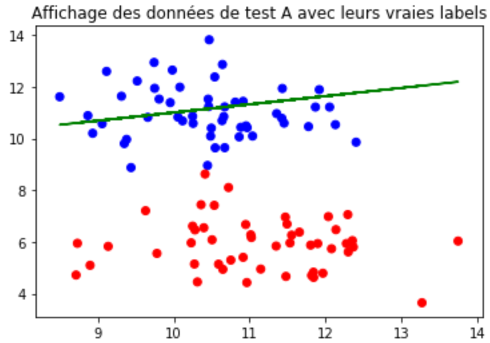


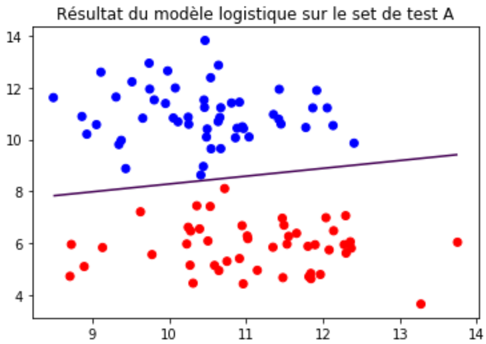


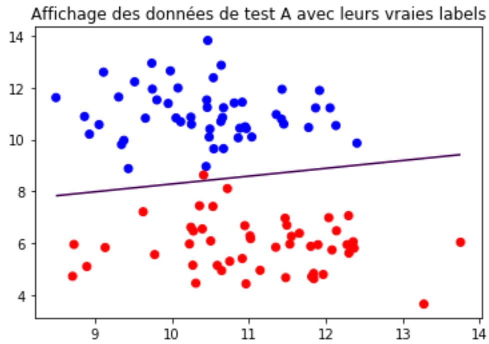






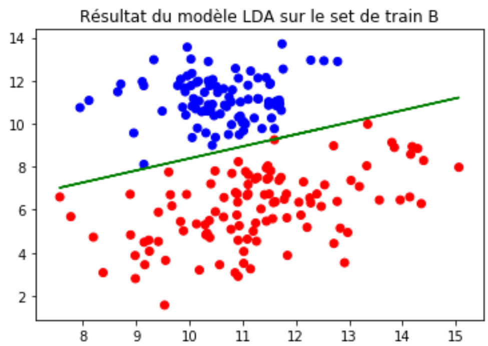


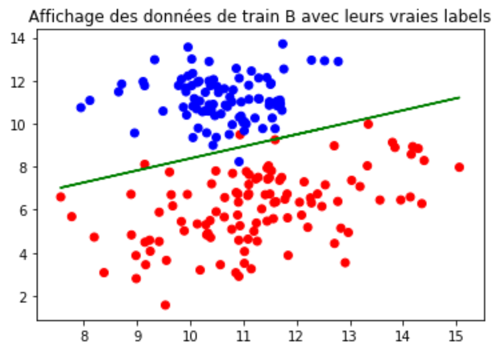


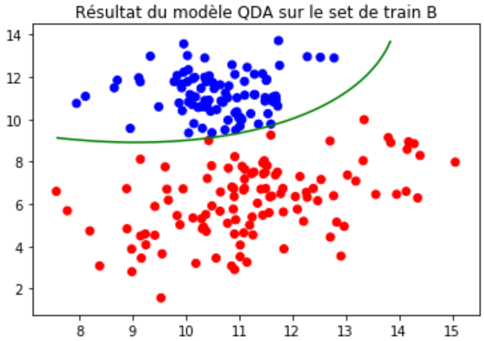


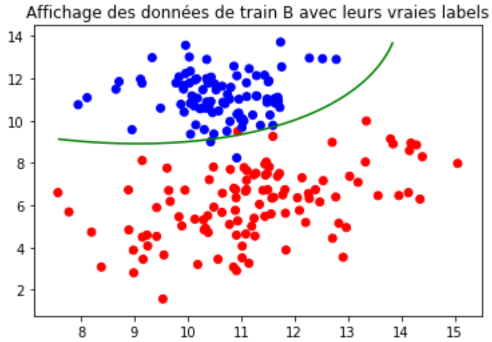
Récapitulatif des erreurs pour le test A	
modele LDA	0.22
modele QDA	0.01
modele Linéaire	0.3
modèle logistique	0.01

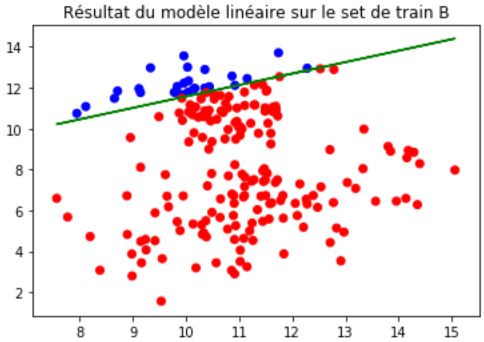
3.3 Dataset train B

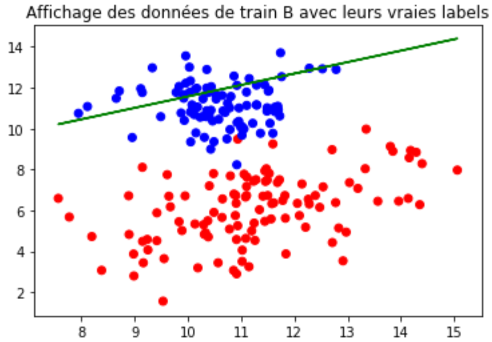


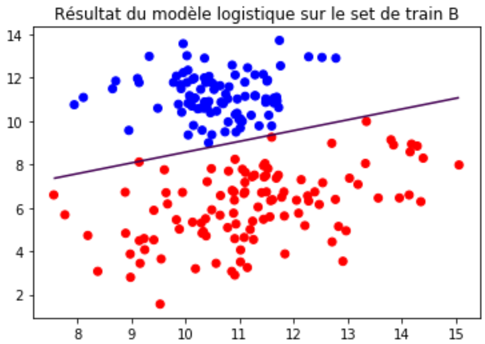


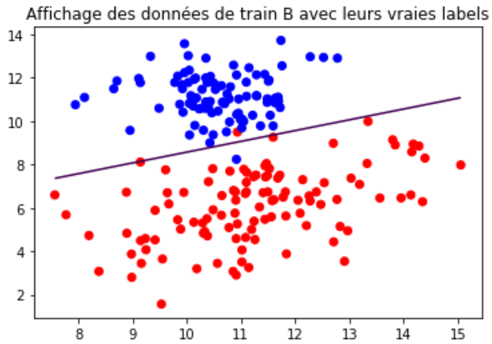






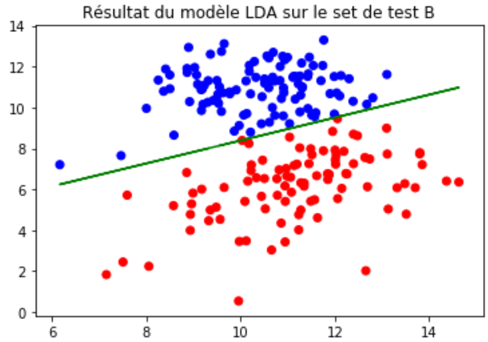


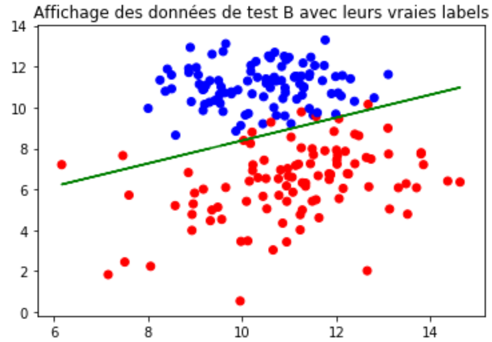


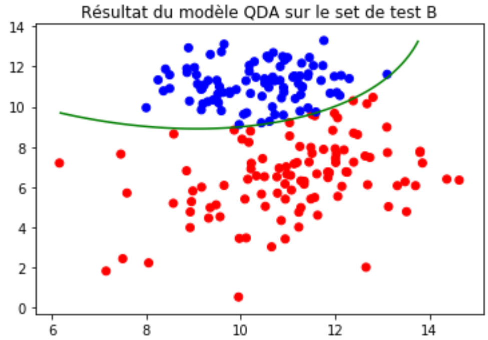


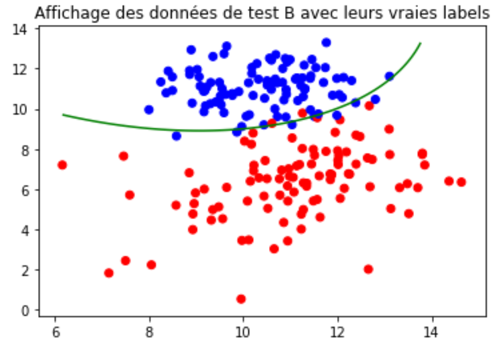
Récapitulatif des erreurs pour le train B	
modele LDA	0.015
modele QDA	0.015
modele Linéaire	0.325
modèle logistique	0.01

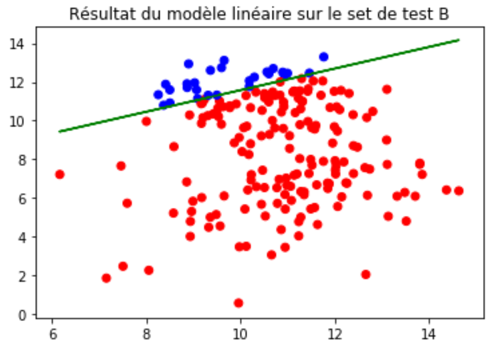
3.4 Dataset test B

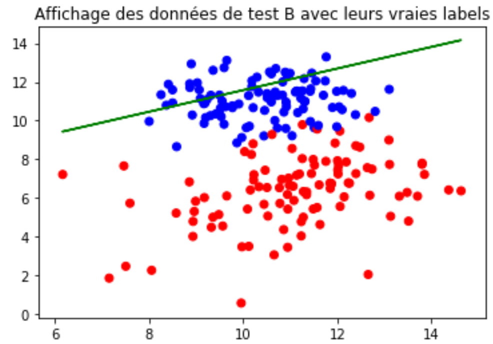

Résultat du modèle LDA sur le set de test B
0 2 4 6 8 10 12 14
6 8 10 12 14

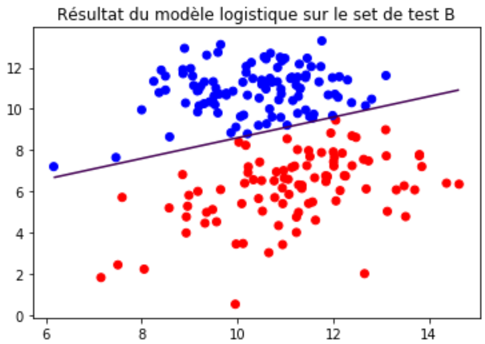

Affichage des données de test B avec leurs vraies labels
0 2 4 6 8 10 12 14
6 8 10 12 14

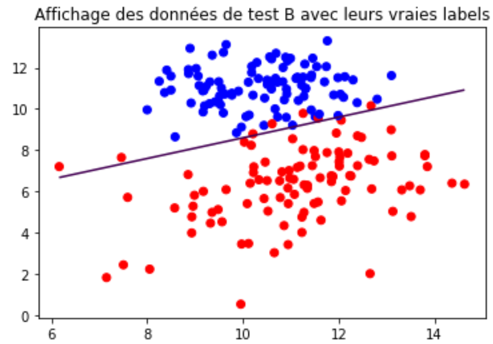

Résultat du modèle QDA sur le set de test B
0 2 4 6 8 10 12 14
6 8 10 12 14


Affichage des données de test B avec leurs vraies labels
0 2 4 6 8 10 12 14
6 8 10 12 14


Résultat du modèle linéaire sur le set de test B
0 2 4 6 8 10 12 14
6 8 10 12 14

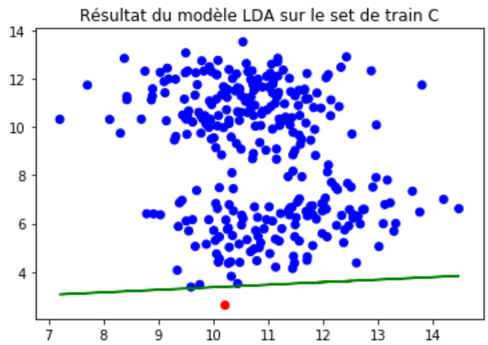

Affichage des données de test B avec leurs vraies labels
0 2 4 6 8 10 12 14
6 8 10 12 14

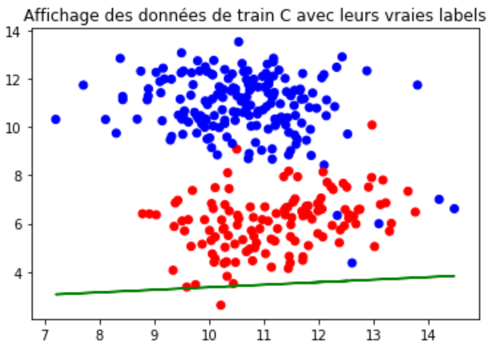

Résultat du modèle logistique sur le set de test B
0 2 4 6 8 10 12
6 8 10 12 14

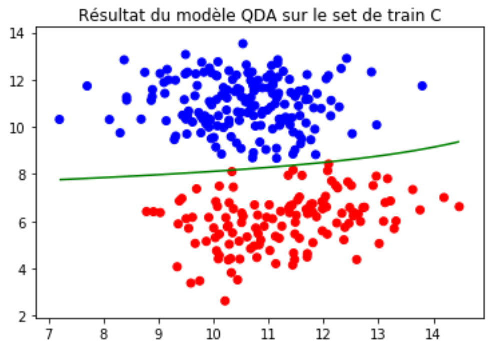

Affichage des données de test B avec leurs vraies labels
0 2 4 6 8 10 12
6 8 10 12 14

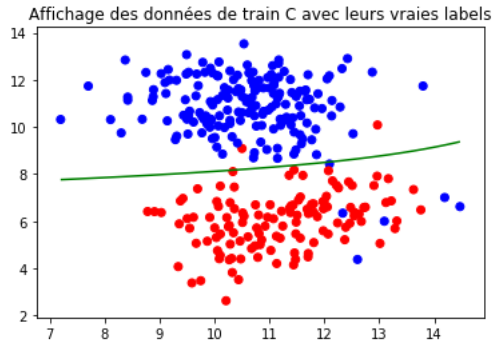
Récapitulatif des erreurs pour le test B	
modele LDA	0.035
modele QDA	0.045
modele Linéaire	0.36
modèle logistique	0.035

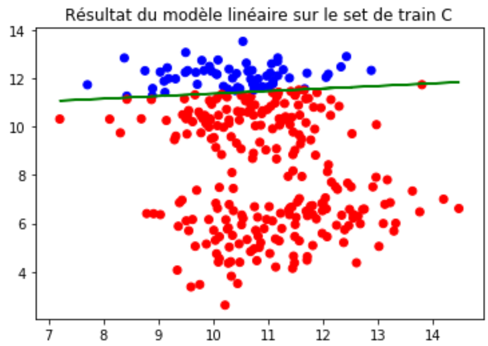
3.5 Dataset train C

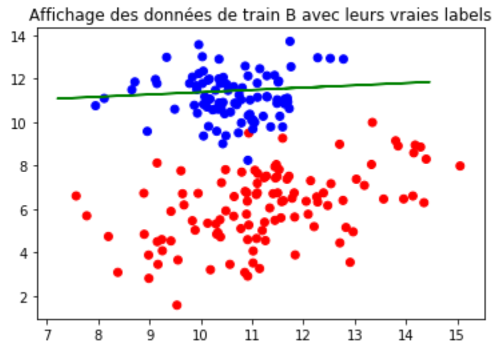


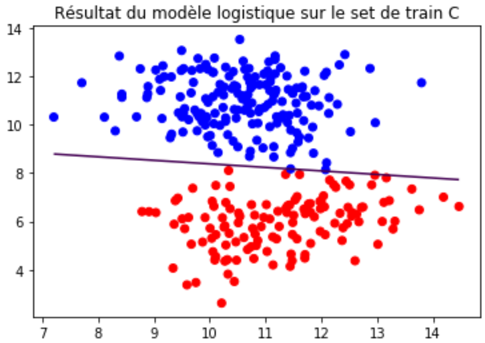


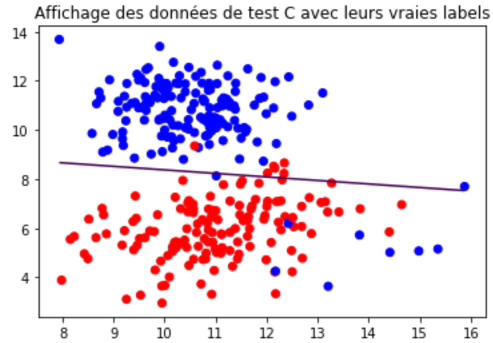






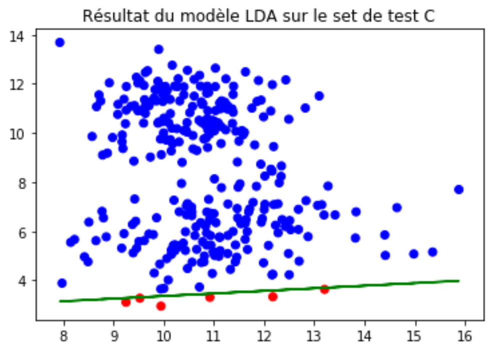


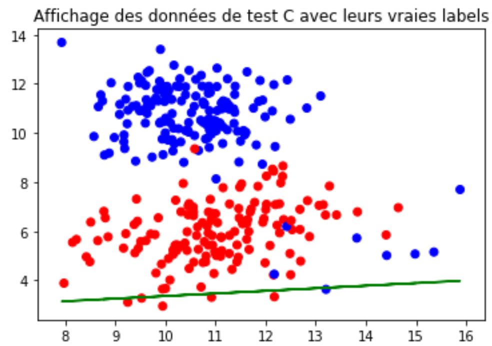


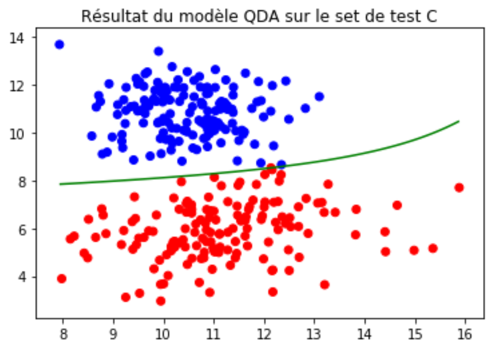


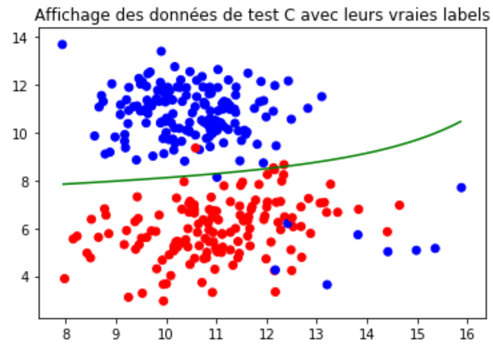
Récapitulatif des erreurs pour le train C	
modele LDA	0.413
modele QDA	0.0267
modele Linéaire	0.0383
modèle logistique	0.03

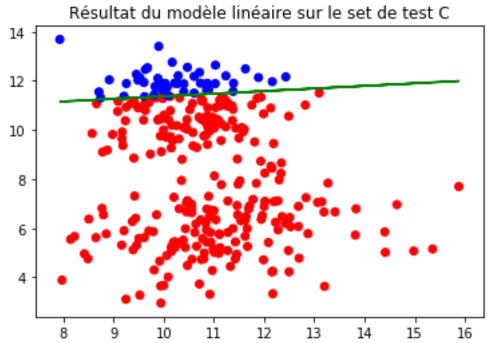
3.6 Dataset test C

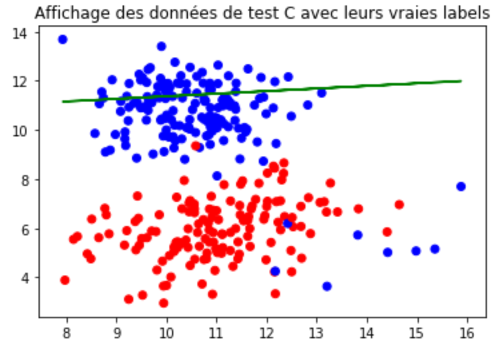


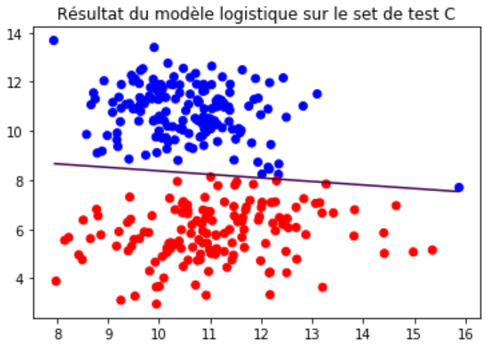


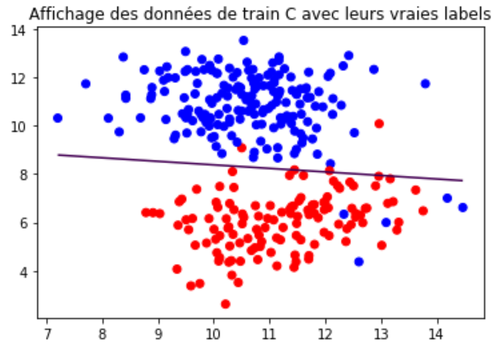












Récapitulatif des erreurs pour le test C	
modele LDA	0.46
modele QDA	0.0367
modele Linéaire	0.0367
modèle logistique	0.0467

Interprétations

Données A

erreur	train A	test A
modele LDA	0.24	0.22
modele QDA	0.0	0.01
modele Linéaire	0.32	0.3
modèle logistique	0.0	0.01

— Pour les modèles linéaire et LDA, les erreurs sont plus élevées pour les ensembles d'entraînement, et sont particulièrement hautes. Ces deux modèles peuvent donc être écartés pour classer nos données.

— Pour cet ensemble de données, mis à part le modèle logistique, les erreurs sont assez élevées. Ce problèmes est probablement attribuable au faible nombre de données disponibles pour entraîner le modèle de classification.
— Le modèle logistiques et le modèle QDA nous donne exactement les même erreurs pour les deux ensembles de données.

<u>Données B</u>

erreur	train B	test B
modele LDA	0.015	0.035
modele QDA	0.015	0.045
modele Linéaire	0.325	0.36
modèle logistique	0.01	0.035

— Les erreurs sont toujours plus élevées dans les ensembles de tests.
— Le modèle linéaire n'est clairement pas adapté pour classer ces données.
— Les modèle LDA et QDA sont très proches l'un de l'autre en terme de classification.
— Compte tenu de la différence d'erreurs entre les données d'entraînement et les données de test pour le modèle logistiques en peu raisonnablement penser qu'il y eu overfitting.

<u>Données C</u>

erreur	train C	test C
modele LDA	0.413	0.46
modele QDA	0.0267	0.0367
modele Linéaire	0.383	0.367
modèle logistique	0.03	0.047

— Les erreurs sont plus élevées dans les ensembles de test.
— Le modèle LDA ne semble pas convenir du tout pour modéliser les données, sûrement car la distribution des données supposée dans le modèle est bien plus complexe que la réalité. En effet, l'hypothèse selon laquelle les deux distributions auraient la même matrice de covariance semble erronée.
— Le modèle linéaire ne semble pas convenir non plus pour classifier cet ensemble de données.
— Le modèle logistique et le modèle QDA ont des performances comparables, et surtout des écarts d'erreurs entre l'entraînement et le test relativement faible, ce qui nous prouve qu'il n'y pas eu d'overfitting.

Paramètres des modèles

<u>Données A</u>
Modèle LDA

$$\mu_1 = \begin{pmatrix} 11.032 & 5.993 \end{pmatrix} \quad \mu_0 = \begin{pmatrix} 10.732 & 10.939 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 0.588 & 0.139 \\ 0.139 & 0.819 \end{pmatrix} \quad \beta = \begin{pmatrix} 2.019 & -6.378 \end{pmatrix} \quad \gamma = 15.999$$
$$\theta = 0.48$$

Modèle QDA

$$\mu_1 = \begin{pmatrix} 11.033 & 5.993 \end{pmatrix}$$
$$\mu_0 = \begin{pmatrix} 10.732 & 10.939 \end{pmatrix}$$
$$\Sigma_1 = \begin{pmatrix} 0.722 & 0.183 \\ 0.183 & 0.935 \end{pmatrix}$$
$$\Sigma_0 = \begin{pmatrix} 0.465 & 0.099 \\ 0.099 & 0.713 \end{pmatrix}$$
$$Q = \begin{pmatrix} 0.761 & -0.0229 \\ -0.0229 & 0.319 \end{pmatrix}$$
$$\beta = \begin{pmatrix} -6.069 & -8.900 \end{pmatrix}$$
$$\theta = 0.48$$
$$\gamma = 87.138$$

modèle linéaire

$$w = \begin{pmatrix} 0.056 & -0.176 & 1.383 \end{pmatrix} \quad \sigma^2 = 0.0276$$

Modèle logistique

$$w = \begin{pmatrix} 1.809 & -5.992 & 31.549 \end{pmatrix}$$

<u>Données B</u>
Modèle LDA

$$\mu_1 = \begin{pmatrix} 11.247 & 6.095 \end{pmatrix}$$
$$\mu_0 = \begin{pmatrix} 10.582 & 11.172 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 1.644 & 0.701 \\ 0.701 & 2.060 \end{pmatrix}$$
$$\beta = \begin{pmatrix} 1.703 & -3.043 \end{pmatrix}$$
$$\gamma = 8.491$$
$$\theta = 0.55$$

Modèle QDA

$$\mu_1 = \begin{pmatrix} 11.247 & 6.095 \end{pmatrix} \quad \mu_0 = \begin{pmatrix} 10.582 & 11.172 \end{pmatrix}$$
$$\Sigma_1 = \begin{pmatrix} 2.366 & 1.231 \\ 1.231 & 2.840 \end{pmatrix} \quad \Sigma_0 = \begin{pmatrix} 0.762 & 0.0535 \\ 0.0535 & 1.107 \end{pmatrix} \beta = \begin{pmatrix} -8.533 - 9.339 \end{pmatrix}$$
$$\gamma = 96.972 Q = \begin{pmatrix} 0.771 & 0.173 \\ 0.173 & 0.451 \end{pmatrix}$$
$$\theta = 0.55$$

modèle linéaire

$$w = \begin{pmatrix} 0.082 & -0.147 & 0.882 \end{pmatrix} \quad \sigma^2 = 0.0485$$

Modèle logistique

$$w = \begin{pmatrix} 1.657 & -3.352 & 12.090 \end{pmatrix}$$

<u>Données C</u>
Modèle LDA

$$\mu_1 = \begin{pmatrix} 11.185 & 6.0425 \end{pmatrix} \quad \mu_0 = \begin{pmatrix} 10.619 & 10.839 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 1.278 & -0.062 \\ -0.062 & 1.666 \end{pmatrix} \quad \beta = \begin{pmatrix} 0.302 & -2.868 \end{pmatrix} \quad \gamma = 6.626 \theta = 0.417$$

Modèle QDA

$$\mu_1 = \begin{pmatrix} 11.185 & 6.042 \end{pmatrix}$$
$$\mu_0 = \begin{pmatrix} 10.619 & 10.839 \end{pmatrix}$$
$$\Sigma_1 = \begin{pmatrix} 1.268 & 0.457 \\ 0.457 & 1.441 \end{pmatrix}$$
$$\Sigma_0 = \begin{pmatrix} 1.286 & -0.433 \\ -0.433 & 1.826 \end{pmatrix}$$
$$\beta = \begin{pmatrix} -2.898 & -7.010 \end{pmatrix}$$

$$\gamma = 55.109$$
$$Q = \begin{pmatrix} -0.045 & 0.483 \\ 0.483 & -0.188 \end{pmatrix}$$
$$\theta = 0.417$$

modèle linéaire

$$w = \begin{pmatrix} 0.017 & -0.159 & 1.640 \end{pmatrix} \quad \sigma^2 = 0.055$$

Modèle logistique

$$w = \begin{pmatrix} -0.280 & -1.919 & 18.882 \end{pmatrix}$$