Rapport MVA : Graphical models - HWK 1

Plassier Vincent

1 Learning in discrete graphical models

Notons $(x_i, z_i)_{1\leq i\leq n}$ les n observations. Nous avons la série d'égalités suivantes :

$$\begin{aligned} l(\pi,\theta) &= \sum_{i=1}^{n} \ln(p(x_i,z_i)) \\ &= \sum_{i=1}^{n} \ln(p(z_i)p(x_i|z_i)) \\ &= \sum_{i=1}^{n} [\ln(\pi(z_i)) + \ln(\theta_{z_i,x_i})] \\ &= \sum_{i=1}^{n}\sum_{m=1}^{M}\left[\mathbb{1}_{z_i=m}\ln(\pi_m) + \sum_{k=1}^{K}\mathbb{1}_{x_i=k}\ln(\theta_{mk})\right] \end{aligned}$$

Introduisons le lagrangien du problème :

$$\begin{aligned} L(\pi,\theta,\alpha,\lambda) = &\sum_{i=1}^{n}\sum_{m=1}^{M}\mathbb{1}_{z_i=m}\ln(\pi_m) + \sum_{i=1}^{n}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}\ln(\theta_{m,k}) \\ &+ \alpha\left(1-\sum_{m=1}^{M}\pi_m\right) + \sum_{m=1}^{M}\lambda_m\left(1-\sum_{k=1}^{K}\theta_{mk}\right) \end{aligned}$$

Les contraintes sont affines, et en prenant $(\pi_m)_m = (1/M)_m \in]0,1[^M$, $(\theta_{mk})_k = (1/K)_k \in]0,1[^K$ on a $\sum_{m=1}^{M}\pi_m = 1$,$\sum_{k=1}^{K}\theta_{mk} = 1$ .D'après le théorème de Slater nous avons la dualité forte.
On en déduit que :

$$\max_{(\alpha,\lambda)}\min_{(\pi,\theta)} L(\pi,\theta,\alpha,\lambda) = \min_{(\pi,\theta)} l(\pi,\theta)$$

De plus,

$$\partial_{\pi_m} L(\pi,\theta,\alpha,\lambda) = \sum_{i=1}^{n}\frac{\mathbb{1}_{z_i=m}}{\pi_m} - \alpha$$

Donc

$$\partial_{\pi_m} L(\pi,\theta,\alpha,\lambda) = 0 \iff \sum_{i=1}^{n}\frac{\mathbb{1}_{z_i=m}}{\pi_m} - \alpha = 0$$

En sommant sur m, on en déduit que $\alpha = n$, d'où :

$$\partial_{\pi_m} L(\pi,\theta,\alpha,\lambda) = 0 \iff \pi_m = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{z_i=m}$$

De manière analogue,

$$\partial_{\theta_{m,k}} L(\pi, \theta, \alpha, \lambda) = \sum_{i=1}^{n} \frac{\mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}}{\theta_{m,k}} - \lambda_m$$

D'où

$$\partial_{\theta_{m,k}} L(\pi, \theta, \alpha, \lambda) = 0 \iff \sum_{i=1}^{n} \frac{\mathbb{1}_{z_i=m}\mathbb{1}_{x_i=k}}{\theta_{m,k}} - \lambda_m = 0$$

En sommant sur k, on en déduit que

$$\lambda_m = |\{i : z_i = m\}|$$

$$\partial_{\theta_{m,k}} L(\pi, \theta, \alpha, \lambda) = 0 \iff \theta_{m,k} = \frac{1}{|\{i : z_i = m\}|} \sum_{i=1}^{n} \mathbb{1}_{z_i=m, x_i=k}$$

avec la convention $\frac{0}{0} = 0$. D'après le théorème de Karush-Kuhn-Tucker, les scalaires π et θ trouvés correspondent aux points minimisant le log-likelihood.

2 Linear classification

2.1 LDA formules

Notons $n_1 := \sum_{i=1}^{n} y_i$ et $n_0 = \sum_{i=1}^{n}(1 - y_i)$. Par hypothèse $x|\{y = i\} \sim \mathcal{N}(\mu_i, \Sigma)$, donc

$$-\ln(p(x_i|y_i = 1)) = \frac{(x_i - \mu_i)^T \Sigma^{-1}(x_i - \mu_i)}{2} + \frac{1}{2}\ln(|\Sigma|) + \ln(2\pi)$$

En calculant le log-likelihood, on obtient les égalités suivantes :

$$\begin{aligned}
l(\pi, \mu, \Sigma) &= \sum_{i=1}^{n} \ln(p(x_i, y_i)) \\
&= \sum_{i=1}^{n} \ln(p(y_i)p(x_i|y_i)) \\
&= \sum_{i=1}^{n} \left[y_i \left(\ln(\pi) + \ln(p(x_i|y_i = 1)) \right) + (1 - y_i) \left(\ln(1 - \pi) + \ln(p(x_i|y_i = 0)) \right) \right] \\
&= \text{constante} - \frac{n \ln(\Sigma)}{2} + \sum_{i=1}^{n} y_i \left[\ln(\pi) - \frac{1}{2}(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) \right] \\
&\qquad + \sum_{i=1}^{n} (1 - y_i) \left[\ln(1 - \pi) - \frac{1}{2}(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0) \right]
\end{aligned}$$

En dérivant par rapport à π, on trouve $\pi = \frac{n_1}{n}$. On vérifie qu'il s'agit d'un maximum.

De plus, $(\mu, \Sigma) \longmapsto l(\pi, \mu, \Sigma)$ est concave. D'après le cours, nous avons

$$\begin{aligned}
\nabla_{\mu_1} l(\pi, \mu, \Sigma) &= \sum_{i=1}^{n} y_i \Sigma^{-1}(\mu_1 - x_i) \\
&= \Sigma^{-1} \left(n_1 \mu_1 - n_1 \sum_{i=1}^{n} y_i x_i \right)
\end{aligned}$$

Par concavité, les points critiques sont des extremums. On en déduit que $\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i$ est un extremum et on vérifie qu'il s'agit d'un maximum. De manière analogue, nous avons :

$$\begin{aligned}\nabla_{\mu_0} l(\pi,\mu,\Sigma) &= \sum_{i=1}^{n}(1-y_i)\Sigma^{-1}(\mu_0 - x_i) \\ &= \Sigma^{-1}\left(n_1\mu_0 - n_1\sum_{i=1}^{n}(1-y_i)x_i\right)\end{aligned}$$

D'où $\mu_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)x_i$ Maximiser $\Sigma \longmapsto l(\pi,\mu,\Sigma)$ revient à minimiser

$$f : A \longmapsto \ln(\det(A)) - \sum_{i=1}^{n}\left[y_i(x_i-\mu_1)^T A(x_i-\mu_1) + (1-y_i)(x_i-\mu_0)^T A(x_i-\mu_0)\right]$$

D'après un raisonnement analogue à celui du cours $\nabla_A \ln(\det(A)) = A^{-1}$ et

$$\nabla_A\left[\sum_{i=1}^{n}\left(y_i(x_i-\mu_1)^T A(x_i-\mu_1) + (1-y_i)(x_i-\mu_0)^T A(x_i-\mu_0)\right)\right] = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i + \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)x_i$$

On en déduit que $\nabla_A f(A) = 0$ si et seulement si $A = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i + \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)x_i$. Toujours par concavité, on en déduit que le Σ maximisant le log-likelihood l est

$$\Sigma = \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T(x_i-\mu_0) + \frac{1}{n_1}\sum_{i=1}^{n} y_i(x_i-\mu_1)^T(x_i-\mu_1)$$

Désormais calculons $p(y=1|x)$. Nous avons :

$$\begin{aligned}p(y=1|x) &= \frac{p(y=1)p(x|y=1)}{p(y=1)p(x|y=1) + p(y=0)p(x|y=0)} \\ &= \frac{1}{1+\exp(-w^T x - b)}\end{aligned}$$

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

2.2 QDA formules

$$\begin{aligned}l(\pi,\mu,\Sigma) &= \sum_{i=1}^{n}\ln(p(x_i,y_i)) \\ &= \sum_{i=1}^{n}\ln(p(y_i)p(x_i|y_i)) \\ &= \sum_{i=1}^{n}\left[y_i\left(\ln(\pi) + \ln(p(x_i|y_i=1))\right) + (1-y_i)\left(\ln(1-\pi) + \ln(p(x_i|y_i=0))\right)\right] \\ &= \text{constante} - \frac{n_0\ln(\Sigma_0)}{2} - \frac{n_1\ln(\Sigma_1)}{2} + \sum_{i=1}^{n}\left[y_i\ln(\pi) + (1-y_i)\ln(1-\pi)\right] \\ &\quad - \frac{1}{2}\sum_{i=1}^{n} y_i(x_i-\mu_1)^T\Sigma_1^{-1}(x_i-\mu_1) - \frac{1}{2}\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T\Sigma_0^{-1}(x_i-\mu_0)\end{aligned}$$

Nous avons donc une somme de fonctions dé-corrélées à maximiser. Cela se fait en maximisant indépendamment chaque quantité. D'abord on maximise en π, on trouve : $\pi = \frac{n_1}{n}$. Comme dans la question précédente, on trouve

$$\begin{aligned}\nabla_{\mu_1} l(\mu, \Sigma) &= \sum_{i=1}^{n} y_i \Sigma_1^{-1}(\mu_1 - x_i) \\ &= \Sigma_1^{-1}\left(n_1\mu_1 - n_1 \sum_{i=1}^{n} y_i x_i\right)\end{aligned}$$

Et de même

$$\begin{aligned}\nabla_{\mu_0} l(\mu, \Sigma) &= \sum_{i=1}^{n} (1-y_i) \Sigma_0^{-1}(\mu_0 - x_i) \\ &= \Sigma_0^{-1}\left(n_0\mu_0 - n_0 \sum_{i=1}^{n} (1-y_i) x_i\right)\end{aligned}$$

Par concavité on en déduit que $\mu_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)x_i$ et $\mu_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i$ maximisent le log-likelihood.
Désormais, intéressons-nous aux Σ_0, Σ_1 qui minimisent le log-likelihood. Reprenons la démonstration du cours, définissons :

$$\begin{cases} f_0 : A_0 \longmapsto \ln(\det(A_0)) - \sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T A_0 (x_i - \mu_0) \\ f_1 : A_1 \longmapsto \ln(\det(A_1)) - \sum_{i=1}^{n} y_i (x_i-\mu_1)^T A_1 (x_i - \mu_1) \end{cases}$$

D'après le raisonnement du cours $\nabla_{A_0} \ln(\det(A_0)) = A_0^{-1}$ et

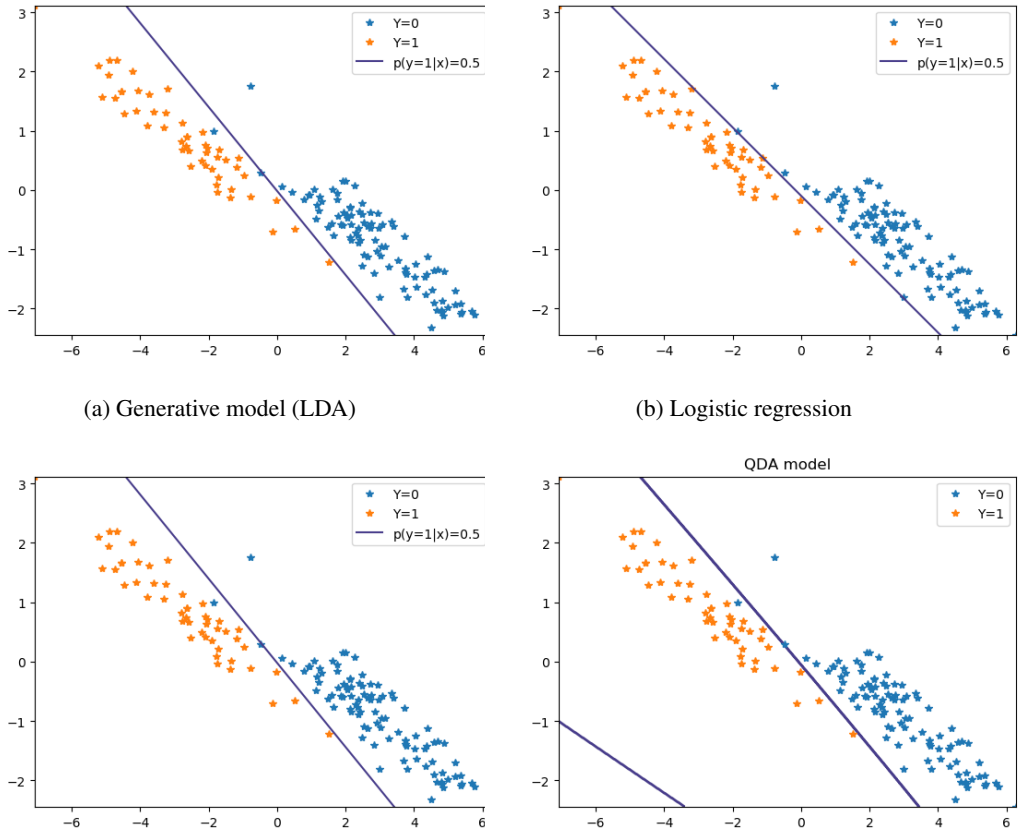
$$\nabla_{A_0}\left[\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T A_0 (x_i-\mu_0)\right] = \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)x_i$$

On en déduit que $\nabla_{A_0} f_0(A_0) = 0$ si et seulement si $A_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)x_i$. Un raisonnement identique montre que $\nabla_{A_1} f_1(A_1) = 0 \iff A_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i$.

Par concavité, on en déduit que $\Sigma_0 = \frac{1}{n_0}\sum_{i=1}^{n}(1-y_i)(x_i-\mu_0)^T(x_i-\mu_0)$ et $\Sigma_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i (x_i - \mu_1)^T(x_i-\mu_1)$ minimisent le log-likelihood.

3 Données A

fraction des données mal classifiées	Entraînement	Test
Generative model (LDA)	0.013	0.02
Logistic regression	0.0	0.034
Linear regression	0.013	0.02
QDA model	0.006	0.02



(a) Generative model (LDA)

(b) Logistic regression

(c) Linear regression

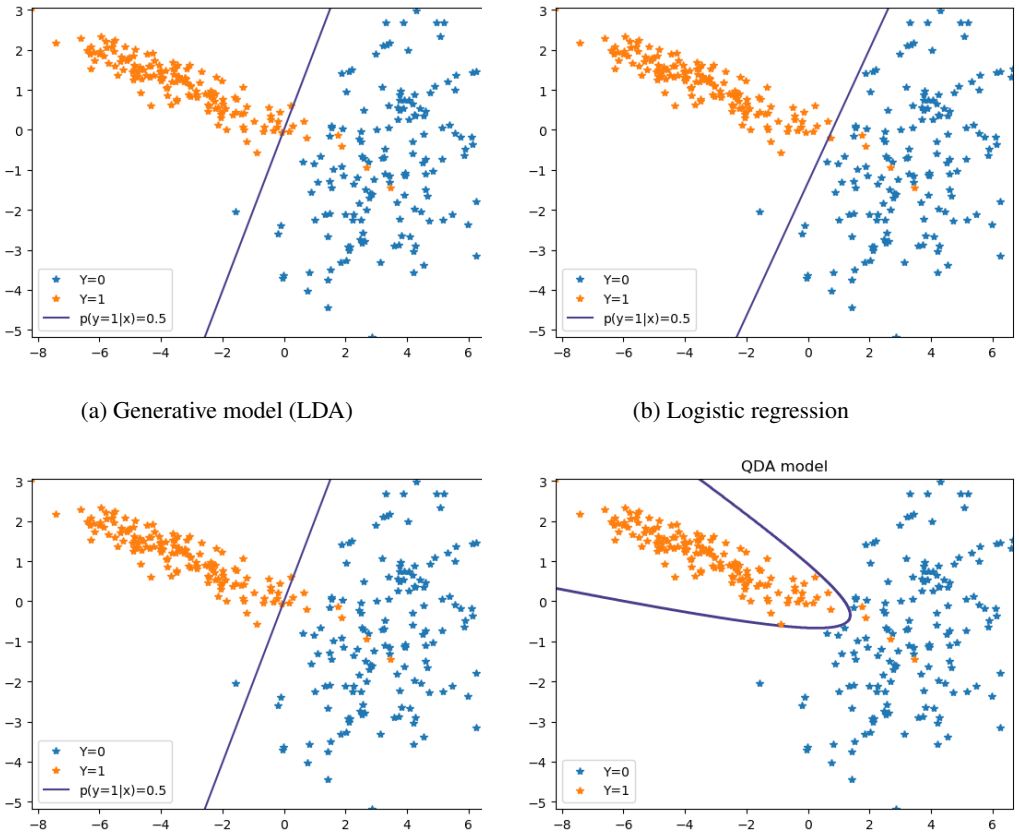
(d) QDA model

FIGURE 1: Représentation des données d'entraînement A

Pour l'ensemble (X_i, Y_i), nous pouvons constater que les données (X_i) sont réparties le long de 2 droites parallèles selon le fait que $Y_i = 0$ ou $Y_i = 1$. Cela permet aux classifieurs LDA, logistic regression, et linear regression d'être efficace en séparant les données par une droite $p(y = 1|x) = 1/2$. Nous pouvons constater que la méthode QDA possède une erreur sur les données tests bien plus importante que sur les données d'entraînement. On peut en déduire qu'il y a de l'overfitting à cause d'un nombre trop important de paramètres. De plus, nous constatons que les modèles LDA et linéaires ont des résultats identiques, en effet, on peut démontrer qu'ils apprennent les mêmes paramètres.

4 Données B

Fraction des données mal classifiées	Entraînement	Test
Generative model (LDA)	0.03	0.041
Logistic regression	0.02	0.043
Linear regression	0.03	0.041
QDA model	0.013	0.02



(a) Generative model (LDA)

(b) Logistic regression

(c) Linear regression

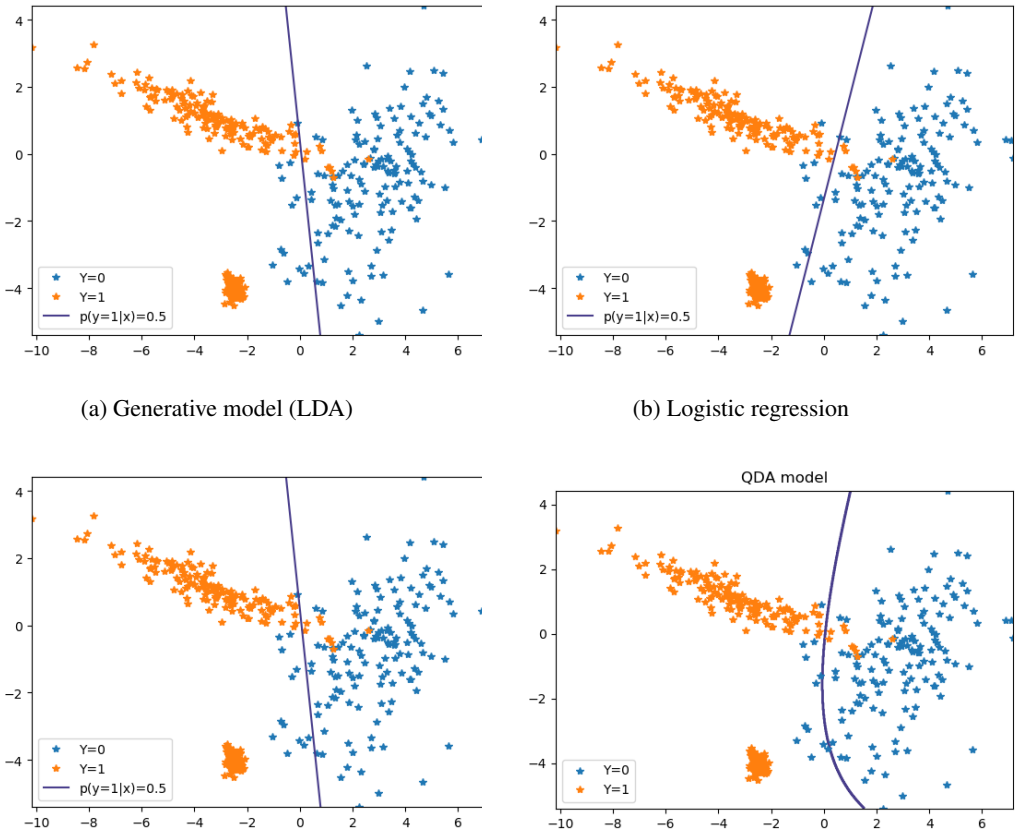
(d) QDA model

FIGURE 2: Représentation des données d'entraînement B

Cette fois-ci, les données (X_i) sont soient réparties le long d'une demi-droite si $Y_i = 1$ ou bien dispersées dans un demi-espace lorsque $Y_i = 0$. Il est désormais plus délicat pour les classifieurs linéaires de séparer les 2 jeux de données, c'est pourquoi la méthode QDA a de meilleures performances.

5 Données C

Fraction des données mal classifiées	Entraînement	Test
Generative model (LDA)	0.055	0.042
Logistic regression	0.04	0.022
Linear regression	0.055	0.042
QDA model	0.052	0.038



(a) Generative model (LDA)

(b) Logistic regression

(c) Linear regression

(d) QDA model

FIGURE 3: Représentation des données d'entraînement C

Désormais, pour $Y_i = 1$ les données (X_i) sont concentrées sur deux groupes. Cette configuration favorise la méthode QDA qui permet de séparer efficacement les données.