Homework 1

Romain Vial

romain.vial@ens-paris-saclay.fr

1 Learning in discrete graphical models

Considérons le modèle suivant : z et x sont des variables aléatoires prenant respectivement M et K valeurs différentes. On a de plus $p(z = m) = \pi_m$ et $p(x = k|z = m) = \theta_{mk}$.

En utilisant un encodage one-hot, on peut écrire:

$$p(z) = \prod_{m=1}^{M} \pi_m^{z_m}$$
$$p(x|z = m) = \prod_{k=1}^{K} \theta_{mk}^{x_k \times z_m}$$

où $(x_k)_{k=1...K} = (1_{\{x=k\}})_{k=1...K}$ et $(z_m)_{m=1...M} = (1_{\{z=m\}})_{m=1...M}$

On en déduit la loi jointe du couple (x, z) avec $\theta = \{(\theta_{mk})_{m=1...M,k=1...K}\}$ et $\pi = \{(\pi_m)_{m=1...M}\}$:

$$\begin{aligned} p(x, z; \theta, \pi) &= p(x|z) \times p(z) \\ &= \prod_{m=1...M} \pi_m^{z_m} \prod_{k=1...K} \theta_{mk}^{x_k \times z_m} \end{aligned}$$

Considérons à présent la log-vraisemblance de N observations i.i.d. du couple (x, z) :

$$\begin{aligned} l(\theta, \pi) &= \sum_{i=1}^{N} \log p(x^{(i)}, z^{(i)}; \theta) \\ &= \sum_{i=1}^{N} \left[\sum_{m=1}^{M} \left[z_m^{(i)} \log \pi_m + \sum_{k=1}^{K} x_k^{(i)} \times z_m^{(i)} \log \theta_{mk} \right] \right] \\ &= \sum_{m=1}^{M} \left[N_m \log \pi_m + \sum_{k=1}^{K} N_{mk} \log \theta_{mk} \right] \end{aligned}$$

où $N_m = \sum_{i=1}^{N} z_m^{(i)}$ et $N_{mk} = \sum_{n=1}^{N} x_k^{(i)} \times z_m^{(i)}$. N_m est le nombre d'observations de la classe m et N_{mk} est le nombre d'observations du couple $(x = k, z = m)$.

Les contraintes du problème d'optimisation sont les suivantes :

$$\sum_{m=1}^{M} \pi_m = 1$$
$$\forall m \in [1 \dots M] \ \sum_{k=1}^{K} \theta_{mk} = 1$$

On en déduit que le Lagrangien du problème, définit sur ${\mathbb{R}_+^*}^K \times {\mathbb{R}_+^*}^M \times {\mathbb{R}_+^*}^{M+1}$, s'écrit :

$$L(\theta, \pi, \lambda) = -l(\theta, \pi) + \lambda_1 \left(\sum_{m=1}^{M} \pi_m - 1 \right) + \sum_{m=1}^{M} \lambda_{m+1} \left(\sum_{k=1}^{K} \theta_{mk} - 1 \right)$$

La fonction $-l$ est convexe comme somme à coefficients positifs ($N_m \geq 0$ et $N_{mk} \geq 0$) de logarithmes. De plus, la condition de Slater est vérifiée car il existe trivialement $\pi_1 \dots \pi_M$ vérifiant $\forall m, \ \pi_m > 0$ et $\sum_{m=1}^{M} \pi_m = 1$ (de même pour $\theta_{m1} ... \theta_{mK}, \ \forall m$). On en conclut ainsi que le problème vérifie la propriété de dualité forte :

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

Comme $L(\theta, \pi, \lambda)$ est convexe par rapport à (θ, π), il suffit de regarder les dérivées partielles par rapport à chaque paramètre pour obtenir $\min_{\theta,\pi} L(\theta, \pi, \lambda)$. On obtient :

$$\frac{\partial L}{\pi_m} = -\frac{N_m}{\pi_m} + \lambda_1 = 0, \ \forall m \in [1 \dots M]$$
$$\frac{\partial L}{\partial \theta_{mk}} = -\frac{N_{mk}}{\theta_{mk}} + \lambda_{m+1} = 0, \ \forall m \in [1 \dots M], \forall k \in [1 \dots K]$$

On en déduit que $\pi_m \propto N_m$ et $\theta_{mk} \propto N_{mk}$. Afin que les contraintes soient respectées, on conclut que :

$$\hat{\pi_m} = \frac{N_m}{\sum_{m=1}^{M} N_m} = \frac{N_m}{N}$$
$$\hat{\theta_{mk}} = \frac{N_{mk}}{\sum_{k=1}^{K} N_{mk}}$$

On peut faire deux remarques sur le résultat obtenu :

- $\hat{\pi_m}$ correspond à la moyenne du nombre d'observations de la classe m alors que $\hat{\theta_{mk}}$ correspond à la moyenne du nombre d'observations de la classe k dans la classe m.
- On observe que par construction du modèle graphique, les deux paramètres π et θ ont été optimisés indépendamment et correspondent aux estimateurs du maximum de vraisemblance de $p(z)$ et $p(x|z)$ respectivement.

2 Linear Classification

1. Linear Discriminant Analysis

(a) On suppose que les distributions de y et $x|y=i$ suivent les lois suivantes:

$$y \sim \text{Bernoulli}(\pi), \quad x|\{y=i\} \sim \mathcal{N}(\mu_i, \Sigma)$$

On en déduit la loi jointe du couple (x, y):

$$\begin{aligned} p(x, y; \pi, \mu_0, \mu_1, \Sigma) &= p(x|y)p(y) \\ &= \pi^y \times (1-\pi)^{1-y} \times \mathcal{N}(x, \mu_y, \Sigma) \end{aligned}$$

Considérons à présent la log-vraisemblance de n observations i.i.d. du couple (x, y) :

$$\begin{aligned} l(\pi, \mu_0, \mu_1, \Sigma) =& \sum_{i=1}^{n} \left[y^{(i)} \log \pi + (1-y^{(i)}) \log(1-\pi) + \log \mathcal{N}(x^{(i)}, \mu_{y^{(i)}}, \Sigma) \right] \\ =& \sum_{i=1}^{n} \left[y^{(i)} \log \pi + (1-y^{(i)}) \log(1-\pi) \right] \\ &+ \sum_{i \in C_0} \log \mathcal{N}(x^{(i)}, \mu_0, \Sigma) + \sum_{i \in C_1} \log \mathcal{N}(x^{(i)}, \mu_1, \Sigma) \end{aligned}$$

où $C_i = \{j, y^{(j)} = i\}$ et $n_i = |C_i|$.

On peut à présent calculer le gradient de l par rapport à chaque composante. On retrouve facilement que :

$$\begin{aligned} \hat{\pi} &= \frac{n_1}{n} \\ \hat{\mu_0} &= \frac{1}{n_0} \sum_{i \in C_0} x^{(i)} \\ \hat{\mu_1} &= \frac{1}{n_1} \sum_{i \in C_1} x^{(i)} \end{aligned}$$

On observe que $\hat{\pi}$ est la moyenne du nombre d'observations de la classe 1 et $\hat{\mu_0}$ et $\hat{\mu_1}$ sont les moyennes des observations des classes 0 et 1 respectivement.

A présent, détaillons davantage le calcul de $\hat{\Sigma}$. On a tout d'abord :

$$\begin{aligned} \sum_{i \in C_i} \log \mathcal{N}(x^{(i)}, \mu_i, \Sigma) &= -\frac{n_i d}{2} \log(2\pi) - \frac{n_i}{2} \log \det \Sigma - \frac{1}{2} \sum_{i=1}^{n_i} (x^{(i)} - \mu_i)^\intercal \Sigma^{-1} (x^{(i)} - \mu_i) \\ &= -\frac{n_i d}{2} \log(2\pi) + \frac{n_i}{2} \log \det A - \frac{n_i}{2} \text{Tr}(A \tilde{\Sigma_i}) \end{aligned}$$

en posant $A = \Sigma^{-1}$ et $\tilde{\Sigma_i} = \frac{1}{n_i} \sum_{i \in C_i} (x^{(i)} - \mu_i)(x^{(i)} - \mu_i)^\intercal$.

En s'inspirant de la preuve du maximum de vraisemblance de la distribution gaussienne, on en déduit que :

$$\nabla_A l(A) = \frac{n_1}{2} A^{-1} - \frac{n_1}{2} \tilde{\Sigma_1} + \frac{n_0}{2} A^{-1} - \frac{n_0}{2} \tilde{\Sigma_0} = 0$$

La condition d'optimalité nous donne finalement :

$$\hat{\Sigma} = \frac{n_1}{n}\tilde{\Sigma_1} + \frac{n_0}{n}\tilde{\Sigma_0}$$

On observe que $\hat{\Sigma}$ est simplement la moyenne des matrices empiriques de covariance pondérées par la proportion observée de la classe correspondante.

(b) La distribution conditionnelle $p(y = 1|x)$ s'écrit sous la forme suivante :

$$\begin{aligned} p(y = 1|x) &= \frac{p(x|y = 1) \times p(y = 1)}{p(x|y = 1) \times p(y = 1) + p(x|y = 0) \times p(y = 0)} \\ &= \frac{1}{1 + \frac{p(x|y=0)\times p(y=0)}{p(x|y=1)\times p(y=1)}} \end{aligned}$$

En posant $s_i(x) = \log(p(x|y = i) \times p(y = i))$, on obtient :

$$\begin{aligned} p(y = 1|x) &= \frac{1}{1 + \exp(-(s_1(x) - s_0(x)))} \\ &= \sigma(s_1(x) - s_0(x)) \end{aligned}$$

Or,

$$s_1(x) - s_0(x) = \log\frac{\pi}{1-\pi} - \frac{1}{2}\mu_1^\intercal\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^\intercal\Sigma^{-1}\mu_0 + x^\intercal\Sigma^{-1}(\mu_1 - \mu_0)$$

On a donc :

$$p(y = 1|x) = \sigma(\tilde{w}^\intercal x + \tilde{b})$$

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

On observe que la distribution conditionnelle peut se mettre sous une forme similaire à celle de la régression logistique. Néanmoins, lors d'une régression logistique, on ne fait pas d'hypothèse sur la forme de la distribution des classes. Ainsi, les résultats obtenus avec les deux méthodes vont largement différer selon la distribution réelle des classes.

(c) A présent, on cherche la droite de classification correspondant à $p(y = 1|x) = 0.5$ ou encore $s_1(x) = s_0(x)$. L'équation de la droite est la suivante :

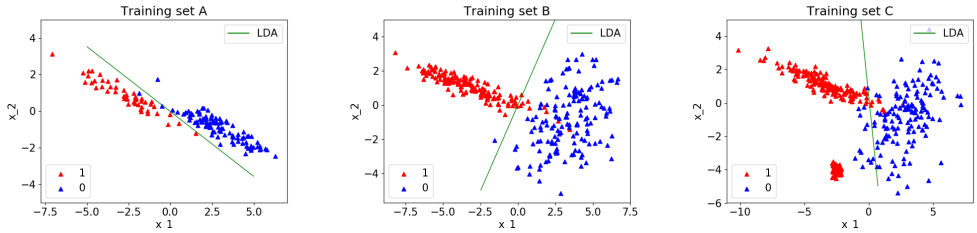
$$\log\frac{\pi}{1-\pi} - \frac{1}{2}\mu_1^\intercal\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_0^\intercal\Sigma^{-1}\mu_0 + x^\intercal\Sigma^{-1}(\mu_1 - \mu_0) = 0$$

Sur la Figure 1, on peut observer les droites de classification par la méthode LDA sur les trois jeux de données.

2. Logistic Regression

On considère le modèle de régression logistique où $y = \sigma\left(w^\intercal x + b\right) = \sigma\left((w_1 w_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + b\right)$.

(a) Les paramètres appris par le modèle après 5 itérations de l'algorithme IRLS sur chaque jeu de données sont les suivant :



(a) Dataset A (b) Dataset B (c) Dataset C

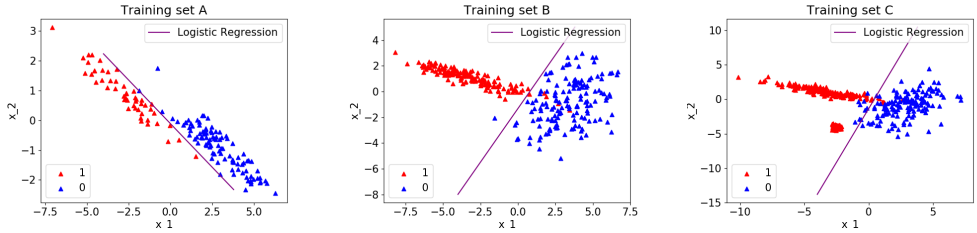
Figure 1: Droites de classification par méthode LDA sur les jeux de données d'entraînement

	Dataset A	Dataset B	Dataset C
w	$[-5.5489, -9.0189]^\intercal$	$[-1.5603, 0.9244]^\intercal$	$[-2.0276, 0.5900]^\intercal$
b	-0.7176	1.1535	0.8215

(b) A présent, on cherche la droite de classification correspondant à $y = 0.5$, c'est à dire $w^\intercal x + b = 0$. Cela correspond à l'équation suivante :

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2}$$

Sur la Figure 2, on peut observer les droites de régression logistique sur les trois jeux de données.



(a) Dataset A (b) Dataset B (c) Dataset C

Figure 2: Droites de régression logistique sur les jeux de données d'entraînement

3. Linear Regression

On considère le modèle de régression linéaire où $y = w^\intercal x + b + \epsilon = (w_1 w_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + b + \epsilon$ où $\epsilon = \mathcal{N}(0, \sigma^2)$.

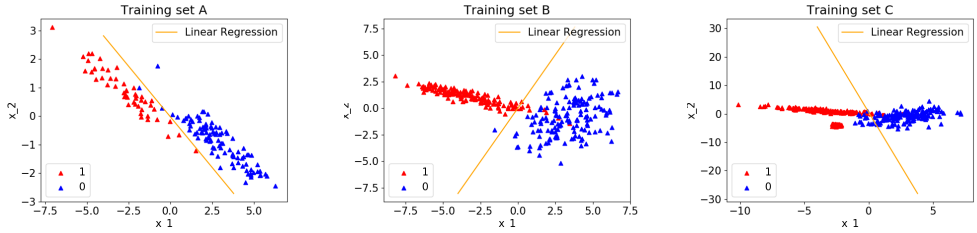
(a) Les paramètres appris par le modèle sur chaque jeu de données sont les suivant:

	Dataset A	Dataset B	Dataset C
w	$[-0.2640, -0.3726]^\intercal$	$[-0.1042, 0.0518]^\intercal$	$[-0.1277, -0.0170]^\intercal$
b	0.4923	0.5001	0.5084
σ^2	0.0399	0.0543	0.0622

(b) A présent, on cherche la droite de régression correspondant à $y = 0.5$. Cela correspond à l'équation suivante :

$$x_2 = -\frac{w_1}{w_2}x_1 + \frac{y-b}{w_2}$$

Sur la Figure 3, on peut observer les droites de régression linéaire sur les trois jeux de données.



(a) Dataset A (b) Dataset B (c) Dataset C

Figure 3: Droites de régression linéaire sur les jeux de données d'entraînement

4. Comparaison des approches

(a) Les erreurs de classification pour chaque modèle dans chaque dataset sont résumées dans le tableau suivant :

Taux d'erreur (%)	Dataset A		Dataset B		Dataset C	
	Train	Test	Train	Test	Train	Test
Linear Regression	1.33	2.07	3.00	**4.15**	5.50	4.23
Logistic Regression	**0.67**	2.47	**2.00**	4.25	**4.00**	**2.37**
LDA	1.33	**2.00**	3.00	**4.15**	5.50	4.23

(b) De façon générale, on observe que les performances sur les jeux d'entraînement sont meilleurs que sur les jeux de test. Ce résultat est cohérent car, par construction, on optimise une fonction de coût sur le jeu d'entraînement. La réduction de l'erreur sur le jeu de test est seulement traitée de façon indirecte, en supposant que sa distribution est la même que sur le jeu d'entraînement.

Pour autant, sur le dataset C, les méthodes obtiennent de meilleurs résultats sur le jeu de données de test. Cela peut éventuellement s'expliquer par la non unimodalité de la distribution de la classe 1, faussant ainsi les hypothèses de séparabilité linéaire (régression linéaire et logistique) et de distribution normale des classes (LDA), amenant, par chance, à une meilleure performance sur le jeu de test.

On remarque de plus que les méthodes de régression linéaire et LDA obtiennent des résultats quasi-identiques, alors que la régression logistique diffère notablement. C'est notamment cette dernière que l'on peut considérer comme la plus robuste sur ces trois jeux de données, ayant le meilleur résultat sur le dataset C et des résultats très similaires aux meilleurs sur les dataset A et B.

Analysons plus en détails chacun des dataset. On observe que le dataset A est linéairement séparable avec deux distributions normales de variances similaires. Les

trois méthodes linéaires sont donc adaptées, et obtiennent des résultats similaires et relativement bas avec 2% d'erreurs. Le dataset B est composé de deux distributions normales linéairement séparables avec des variances différentes (hypothèse LDA non vérifiée). Les erreurs sont plus élevées, autour de 4%. Enfin pour le dataset C, on observe une classe distribuée normalement (classe 0) et une classe distribuée de façon bimodale (classe 0). Les trois méthodes ne sont pas adaptées et obtiennent de mauvais résultats dès le jeu d'entraînement (environ 5% d'erreurs), néanmoins, comme précisé précédemment, les performances sur le jeu de test sont légèrement supérieures.

5. Quadratic Discriminant Analysis

A présent, on relâche l'hypothèse d'égalité des matrices de covariance pour obtenir un modèle d'analyse discriminante quadratique. L'estimateur du maximum de vraisemblance pour les matrices Σ_0 et Σ_1 devient simplement les matrices empiriques de covariance associées à chacune des classes 0 et 1.

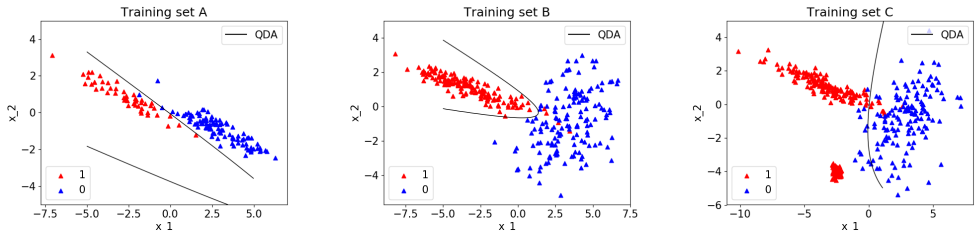
(a) Les paramètres appris par le modèle sur chaque jeu de données sont les suivant:

	Dataset A	Dataset B	Dataset C
μ_0	$[2.8997, -0.8939]^\intercal$	$[3.3407, -0.8355]^\intercal$	$[2.7930, -0.8384]^\intercal$
Σ_0	$\begin{pmatrix} 2.3107 & -1.0475 \\ -1.0475 & 0.5758 \end{pmatrix}$	$\begin{pmatrix} 2.5389 & 1.0642 \\ 1.0642 & 2.9601 \end{pmatrix}$	$\begin{pmatrix} 2.8991 & 1.2458 \\ 1.2458 & 2.9248 \end{pmatrix}$
μ_1	$[-2.6923, 0.8660]^\intercal$	$[-3.2167, 1.0831]^\intercal$	$[-2.9423, -0.9578]^\intercal$
Σ_1	$\begin{pmatrix} 2.7044 & -1.3008 \\ -1.3008 & 0.6897 \end{pmatrix}$	$\begin{pmatrix} 4.1536 & -1.3345 \\ -1.3345 & 0.5161 \end{pmatrix}$	$\begin{pmatrix} 2.8691 & -1.7620 \\ -1.7620 & 6.5644 \end{pmatrix}$
π	$\frac{1}{3}$	$\frac{1}{2}$	0.625

(b) De manière similaire au modèle LDA, la conique de classification vérifie l'équation :

$$\log \frac{\pi}{1-\pi} + \frac{1}{2} \log \frac{\det \Sigma_1}{\det \Sigma_0} - \frac{1}{2}(x-\mu_1)^\intercal \Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_0)^\intercal \Sigma_0^{-1}(x-\mu_0) = 0$$

Sur la Figure 4, on peut observer les droites de classification par méthode QDA sur les trois jeux de données.



(a) Dataset A (b) Dataset B (c) Dataset C

Figure 4: Droites de classification par méthode QDA sur les jeux de données d'entraînement

(c) On obtient les taux d'erreurs suivant pour le modèle QDA :

Taux d'erreur (%)	Dataset A		Dataset B		Dataset C	
	Train	Test	Train	Test	Train	Test
QDA	**0.67**	**2.00**	**1.33**	**2.00**	5.25	3.83

(d) On observe à présent que la méthode QDA obtient les meilleurs résultats sur les dataset A et B. Sur le dataset A, elle obtient le même résultat que la LDA. Ce résultat est cohérent car les variances des distributions des deux classes semblent similaires (on peut notamment observer que $\Sigma_0 \approx \Sigma_1$) et la méthode QDA revient à pratiquer une LDA.

Sur le dataset B, la méthode QDA prend tout son sens et obtient des résultats bien meilleurs que les 3 méthodes linéaires avec 2% d'erreurs contre 4% auparavant. En effet, dans ce cas, les deux classes sont distribuées normalement avec des variances différentes.

Pour finir, la méthode QDA obtient des résultats similaires aux méthodes linéaires sur le dataset C. Sur ce jeu de données, l'hypothèse de distribution normale des deux classes n'est pas respectée.