Master M2 MVA 2018/2019 - Graphical models - HWK 1

Souhaib ATTAIKI

October 19, 2018

Disclaimer : Tous les résultats présentés sur cette page sont démontrés dans la section Démonstration, page 5

Exercise 1 : Learning in discrete graphical models

L'estimateur du maximum de vraisemblance pour π et θ basé sur un échantillon i.i.d. d'observations pour le modèle fourni est le suivant :

$$\hat{\pi}_m = \frac{N_m}{N} \quad \forall m \in \{1, ..., M\} \qquad \hat{\theta}_{mk} = \frac{N_{mk}}{N_m} \quad \forall k \times m \in \{0, ..., K\} \times \{0, ..., M\}$$

où : $N_m = \sum_{i=1}^{N} z_{im}$ et $N_{mk} = \sum_{i=1}^{N} z_{im} x_{ik}$

Exercise 2.1 : LDA formulas

(a) L'estimateur du maximum de vraisemblance est le suivant :

$$\hat{\pi} = \frac{n_1}{n} \qquad \hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i \qquad \hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n} (1-y_i) x_i$$

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

où : $n_1 = \sum_{i=1}^{n} y_i$ et $n_0 = n - n_1$

(b) Après calcul, on trouve que

$$p(y=1|x) = \sigma(\alpha^T x + \beta)$$

où : $\alpha = \hat{\Sigma}^{-1}(\hat{\mu_1} - \hat{\mu_0})$ et $\beta = \frac{1}{2}(\hat{\mu}_0^T \Sigma^{-1} \hat{\mu}_0 - \hat{\mu}_1^T \Sigma^{-1} \hat{\mu}_1) + \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$

Exercise 2.5.(a) : QDA formulas

De même que pour LDA, on montre que :

$$\hat{\pi} = \frac{n_1}{N} \qquad \hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i x_i \qquad \hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n} (1-y_i) x_i$$

$$\hat{\Sigma}_0 = \frac{1}{n_0}\sum_{i=1}^{n} (1-y_i)(x_i - \mu_0)(x_i - \mu_0)^T \qquad \hat{\Sigma}_1 = \frac{1}{n_1}\sum_{i=1}^{n} y_i (x_i - \mu_1)(x_i - \mu_1)^T$$

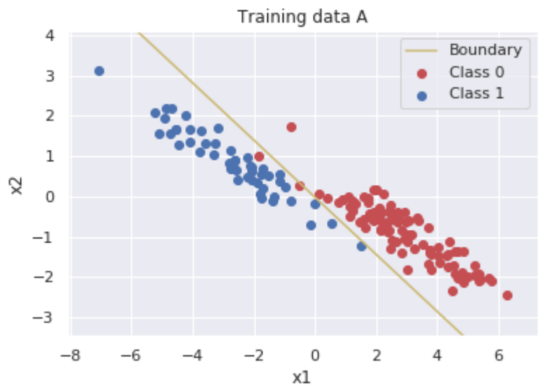
Et on a :

$$p(y=1|x) = \sigma\left(\frac{1}{2} x^T Q x + \alpha^T x + \beta\right)$$

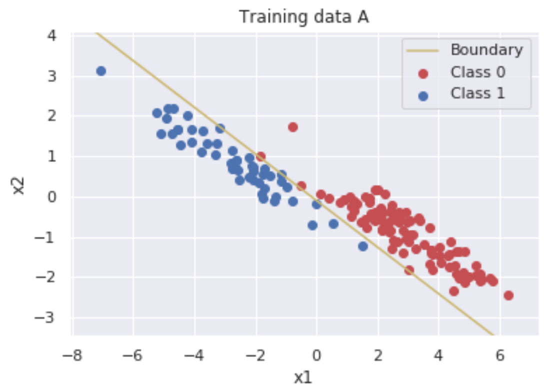
où : $Q = \hat{\Sigma}_0^{-1} - \hat{\Sigma}_1^{-1}$ et $\alpha = \hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\Sigma}_0^{-1}\hat{\mu}_0$

et $\beta = \frac{1}{2}\left(\hat{\mu}_0^T \Sigma_0^{-1} \hat{\mu}_0 - \hat{\mu}_1^T \Sigma_1^{-1} \hat{\mu}_1 + \log\left(\frac{det\hat{\Sigma}_0}{det\hat{\Sigma}_1}\right)\right) + \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$

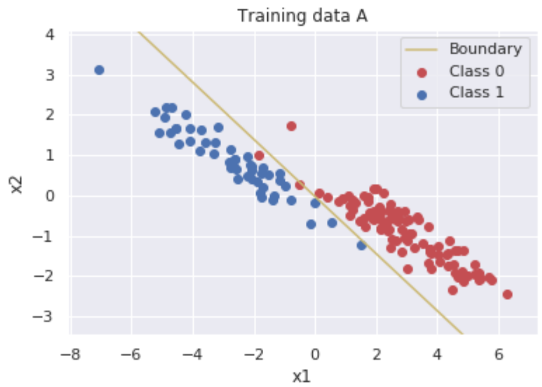
Dataset A



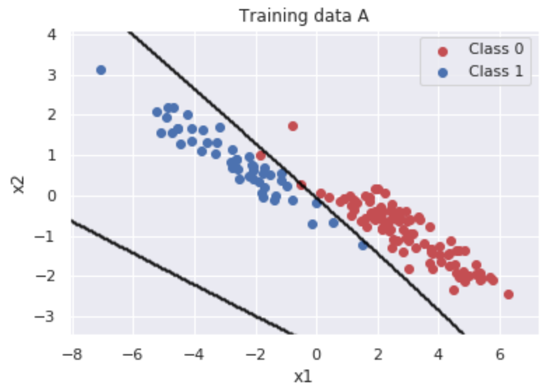
“Genrative Model”



“Logistic regression”



“Linear regression”



“QDA model”

	Train	Test
LDA	1.33	2
Logistic reg.	0	3.4
Linear reg.	1.33	2.07
QDA	0.67	2

– Nous constatons que l’erreur de classification est plus importante sur la dataset de test que sur la dataset d’apprentissage, ce qui est tout à fait prévisible

– Nous pouvons remarquer que la régression logistique atteint une précision de 100 % sur l’ensemble d’entraînement. Ceci est dû au fait que les données sont linéaires et séparables, mais cela engendre un comportement de *"overfitting"*, ce qui explique l’erreur de classification élevé pour la dataset de test

– LDA et QDA donnent les meilleures performances de test, ce qui peut s’expliquer par le fait que les données sont peut-être générées par deux Gaussiens qui ont la même matrice de covariance

Dataset B

“Genrative Model”

“Logistic regression”

“Linear regression”

“QDA model”

	Train	Test
LDA	3	4.15
Logistic reg.	2	4.3
Linear reg.	3	4.15
QDA	1.33	2

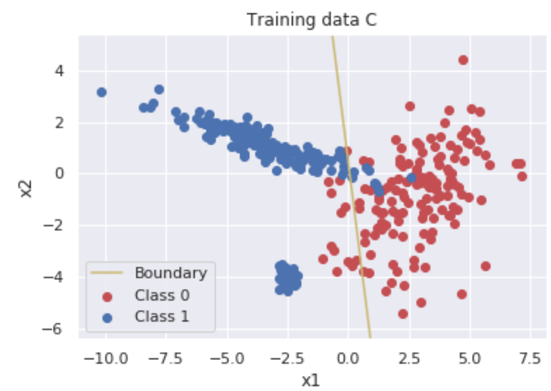
– Nous pouvons voir que la LDA et la régression linéaire ont des performances similaires, ce qui est valable pour les autres ensembles de données.

– Les données semblent être générées par deux gaussiens de matrice de covariance différente, de ce fait, les trois premiers modèles ont du mal à trouver un séparateur linéaire, et l’erreur de classification est importante, sauf pour le QDA qui a une bonne performance

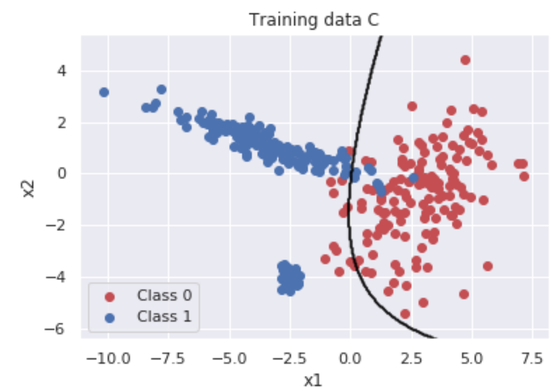
Dataset C

"Genrative Model"

"Logistic regression"



"Linear regression"



"QDA model"

	Train	Test
LDA	5.5	4.23
Logistic reg.	4	2.27
Linear reg.	5.5	4.23
QDA	5.25	3.83

– On constate que les données ne semblent pas être générées par des gaussiennes, en particulier la présence du petit ensemble de données en bas à gauche
– Tous les modèles présentent une erreur de classification élevée par rapport aux autres ensembles de données
– Contrairement aux cas précédents, la régression logistique est la plus performante
– L'erreur de classification sur l'ensemble d'entraînement est élevée par rapport à l'ensemble de test, ce qui est atypique, mais qui indique l'absence de *"overfitting"*

Exercise 1 : Learning in discrete graphical model

On cherche à estimer π_m et θ_{mk}, $\forall m \in [\![1, M]\!]$, $\forall k \in [\![1, K]\!]$.

Soient $(x_1, z_1), \ldots, (x_N, z_N)$ N observations

et on pose x_i le vecteur tel que $\begin{cases} x_{ik} = 1 & \text{si } x_i = k \\ 0 & \text{sinon} \end{cases}$

et z_i " " $\begin{cases} z_{im} = 1 & \text{si } z_i = m \\ 0 & \text{sinon} \end{cases}$

on a $P(X_i = x, Z_i = z) = P(X_i = x \mid Z_i = z) \cdot P(Z_i = z)$

et $P(X_i = x \mid Z_i = z) = \prod_{m=1}^{M} \prod_{k=1}^{K} \theta_{mk}^{z_{im} x_{ik}}$ et $P(Z_i = z) = \prod_{m=1}^{M} \pi_m^{z_{im}}$

donc sous l'hypothèse de iid, la log vraisemblance s'écrit de la façon suivante :

$$\mathcal{L}(\pi, \theta) = \sum_{i=1}^{N} \left(\sum_{m=1}^{M} z_{im} \log(\pi_m) + \sum_{m=1}^{M} \sum_{k=1}^{K} z_{im} x_{ik} \log(\theta_{mk}) \right)$$

$$= \sum_{m=1}^{M} N_m \log(\pi_m) + \sum_{m=1}^{M} \sum_{k=1}^{K} N_{mk} \log(\theta_{mk})$$

avec $N_m = \sum_{i=1}^{N} z_{im}$ et $N_{mk} = \sum_{i=1}^{N} z_{im} x_{ik}$

on a $\mathcal{L}(\pi, \theta) = \mathcal{L}_1(\pi) + \mathcal{L}_2(\theta)$

donc $\max_{\pi, \theta} \mathcal{L}(\pi, \theta) = \max_{\pi, \theta} \left(\mathcal{L}_1(\pi) + \mathcal{L}_2(\theta) \right) = \max_{\pi} \mathcal{L}_1(\pi) + \max_{\theta} \mathcal{L}_2(\theta)$

* Maximisation de $\mathcal{L}_1$

$\mathcal{L}_1$ est strictement concave, comme somme positive de fonctions concaves, donc son maximiseur est unique, et on peut réécrire le problème de la façon suivante $\begin{cases} \min_{\pi} (-\mathcal{L}_1) \\ \sum_{m=1}^{M} \pi_m = 1 \end{cases}$. Ce problème est un problème d'optimisation convexe et pour $\pi_i = \frac{1}{M}$, $\forall 1 \leq i \leq M$, les conditions de Slater sont vérifiées, on a donc la dualité forte, on peut obtenir le minimum de $(-\mathcal{L}_1)$ en dérivant le lagrangien L_1 et vérifier les conditions

$$L_1 = -\sum_{m=1}^{M} N_m \log \pi_m + \lambda \left(\sum_{m=1}^{M} \pi_m - 1 \right)$$

$$\frac{\partial L_1}{\partial \pi_i} = -\frac{N_i}{\pi_i} + \lambda = 0 \rightarrow \pi_i = \frac{-N_i}{\lambda}$$

$$\sum_{i=1}^{M} \pi_i = 1 \rightarrow \lambda = N \qquad \text{d'ou} \quad \boxed{\hat{\pi}_m = \frac{N_m}{N}}$$

* Maximisation de $\mathcal{L}_2$:

Comme avant, $\mathcal{L}_2$ est strictement concave et le problème peut s'écrire sous la forme suivante

$$\begin{cases} \min_\theta -\mathcal{L}_2(\theta) \\ \sum_{k=1}^{K} \theta_{mk} = 1 \ , \forall m \in [\![1, M]\!] \end{cases}$$

pour $\theta_{mk} = \frac{1}{K}$, $\forall m$, les conditions de Slater sont vérifiées, donc comme avant, on a

$$L_2 = -\sum_{m=1}^{M} \sum_{k=1}^{K} N_{mk} \log(\theta_{mk}) + \sum_{m=1}^{M} \lambda_m \left(\sum_{k=1}^{K} \theta_{mk} - 1 \right)$$

$$\frac{\partial L_2}{\partial \theta_{mk}} = 0 \Leftrightarrow -\frac{N_{mk}}{\theta_{mk}} + \lambda_m = 0 \Leftrightarrow \frac{N_{mk}}{\lambda_m} = \theta_{mk}$$

et $\sum_{k=1}^{K} \theta_{mk} = 1 \Leftrightarrow \lambda_m = \sum_{k=1}^{K} N_{mk} = \sum_{i=1}^{N} z_i = N_m$

d'ou $\boxed{\hat{\theta}_{mk} = \frac{N_{mk}}{N_m}}$

Exercice 2.1. (a) ; LDA

$$y \sim \mathcal{B}(\pi) \quad , \quad x \mid y = i \sim \mathcal{N}(\mu_i, \Sigma)$$

Soient $(x_1, y_1), \dots, (x_n, y_n)$ n observations. on a

$$P(x_i, y_i) = \left(\pi \mathcal{N}_i(\mu_1, \Sigma)\right)^{y_i} \left((1-\pi) \mathcal{N}_i(\mu_0, \Sigma) \right)^{1-y} \qquad \left(\text{formule de Bayes} \right)$$

donc la log vraisemblance est :

$$\mathcal{L}(\pi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{n} \Big(y_i \log \pi + (1-y_i) \log(1-\pi) + y_i \log \mathcal{N}_i(\mu_1, \Sigma) + (1-y_i) \log \mathcal{N}_i(\mu_0, \Sigma) \Big)$$

on a $\log \mathcal{N}_i(\mu_j, \Sigma) = -\log(2\pi) - \frac{1}{2}\log\det\Sigma - \frac{1}{2}(x_i-\mu_j)^T\Sigma^{-1}(x_i-\mu_j)$

et on note $n_1 = \sum_{i=1}^n y_i$ et $n_0 = n - n_1$, donc on a

$$\mathcal{L}(\pi,\mu_0,\mu_1,\Sigma) = n_1\log\pi + n_0\log(1-\pi) - \frac{n}{2}\log\det\Sigma - \frac{1}{2}\sum_{i=1}^n\Big(y_i(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) + (1-y_i)(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\Big) + Cst$$

la fonction $\mathcal{L}$ est strictement concave séparément par rapport à chacun des variables, donc ses maximiseurs par rapport aux variables sont uniques, on peut les trouver par dérivation.

* $\frac{\partial\mathcal{L}}{\partial\pi} = 0 \Leftrightarrow \frac{n_1}{\pi} + \frac{n_0}{1-\pi} = 0 \longrightarrow \boxed{\hat{\pi} = \frac{n_1}{n}}$

* $\frac{\partial\mathcal{L}}{\partial\mu_1} = 0 \Leftrightarrow \Sigma^{-1}\left(n_1\mu_1 - \sum_{i=1}^n y_i x_i\right) = 0 \Leftrightarrow \boxed{\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^n y_i x_i}$

* de même pour μ_0, on trouve $\boxed{\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^n (1-y_i)x_i}$

* la partie de $\mathcal{L}$ qui dépend de Σ est la suivante:

$$\mathcal{L}_{|\Sigma} = +\frac{n}{2}\log\det\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^n\Big(y_i(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1) + (1-y_i)(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)\Big)$$

$$\mathcal{L}_{|\Sigma} = +\frac{n}{2}\log\det\Sigma^{-1} - \frac{1}{2}\Big(tr\Big(\Sigma^{-1}\Big(\sum_{i=1}^n y_i(x_i-\mu_1)(x_i-\mu_1)^T\Big)\Big) + tr\Big(\Sigma^{-1}\Big(\sum_{i=1}^n(1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T\Big)\Big)\Big)$$

on a vu dans le cours (Lecture 1) les différentes dérivées par rapport à Σ des differents termes de $\mathcal{L}_{|\Sigma}$, d'où

$$\frac{\partial\mathcal{L}_{|\Sigma}}{\partial\Sigma} = 0 \Leftrightarrow \boxed{\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n y_i(x_i-\mu_1)(x_i-\mu_1)^T + \frac{1}{n}\sum_{i=1}^n(1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T}$$

* Calcul de $P(y=1|x)$

$$P(y=1|x) = \frac{P(y=1, x)}{P(x)} = \frac{P(x|y=1)\cdot\pi}{\pi P(x|y=1) + (1-\pi)P(x|y=0)}$$ (Probabilité totale)

$$= \frac{1}{1 + \frac{1-\pi}{\pi}\frac{P(x|y=0)}{P(x|y=1)}}$$

$$P(y=1|x) = \frac{1}{1 + \frac{1-\pi}{\pi}\exp\left(-\frac{1}{2}\left(\cancel{x^T\Sigma^{-1}x} + \mu_0^T\Sigma^{-1}\mu_0 - 2x^T\Sigma^{-1}\mu_0 - \cancel{x^T\Sigma^{-1}x} - \mu_1^T\Sigma^{-1}\mu_1 + 2x^T\Sigma^{-1}\mu_1\right)\right)}$$

$$= \sigma(\alpha^T x + \beta)$$

où $\alpha = \Sigma^{-1}(\mu_1 - \mu_0)$ et $\beta = -\log\left(\frac{1-\pi}{\pi}\right) + \frac{1}{2}\left(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right)$

Exercice 2.5 (a) : QDA

En suivant les mêmes étapes que dans l'exercice précédent, on trouve les résultats de la page 1.