SPLEX
Statistiques pour la classification et fouille de données en génomique

Classification probabiliste - EM

Pierre-Henri WUILLEMIN

DEcision, Système Intelligent et Recherche opérationnelle
LIP6
pierre-henri.wuillemin@lip6.fr
http://webia.lip6.fr/~phw/splex

Rappels : Probabilité

Soit Ω un ensemble (fini), $\mathcal{E} \subset \mathcal{P}(\Omega)$ (fermé pour $\cup$ et $\cap$).

Définition (Probabilité)

$p : \mathcal{E} \longrightarrow [0,1]$ est une loi de probabilité si et seulement si :

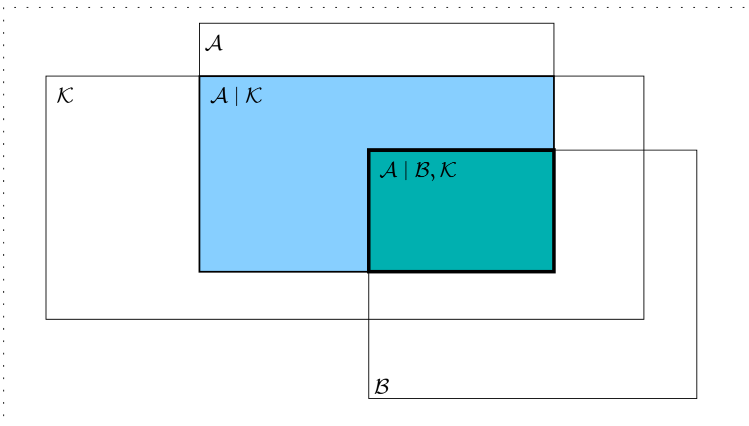
- $\forall \mathcal{A} \in \mathcal{E}, 0 \leq p(\mathcal{A}) \leq 1$
- $\forall \mathcal{A}, \mathcal{B} \in \mathcal{E}, \mathcal{A} \cap \mathcal{B} = \emptyset \Rightarrow p(\mathcal{A} \cup \mathcal{B}) = p(\mathcal{A}) + p(\mathcal{B})$
  $\mathcal{A}$ et $\mathcal{B}$ sont alors dits mutuellement exclusifs.
- $p(\Omega) = 1$.

$p(\emptyset) = 0$?
$p(\overline{\mathcal{A}}) = 1 - p(\mathcal{A})$?

À partir de maintenant : $\mathcal{E} = \mathcal{P}(\Omega)$.
Tout $\mathcal{A} \subset \Omega$ est un évènement ; tout $\omega \in \Omega$ est un évènement élémentaire

Rappels : Probabilité conditionnelle

La probabilité devrait (normalement) toujours indiquer dans quel contexte ($\mathcal{K}$) elle est calculée. Ce qu'on pourrait noter $p_{\mathcal{K}}(\mathcal{A})$ ou encore $p(\mathcal{A} \mid \mathcal{K})$.

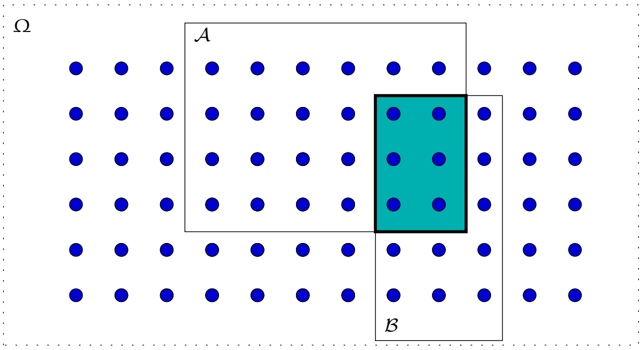


$p(\mathcal{A})$ devrait être écrit $p(\mathcal{A} \mid \Omega)$.

$p(. \mid \mathcal{X})$ et $p(. \mid \mathcal{Y})$ sont deux lois différentes !

Rappels : Calcul de probabilités : le cas fini

Avec $\Omega \neq \emptyset$ fini ; $A, B \subseteq \Omega$,



- $p(\mathcal{A} \mid \Omega) = \frac{|\mathcal{A}|}{|\Omega|}$
- $p(\mathcal{A} \mid \mathcal{B}, \Omega) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{B}|}$
- $p(\mathcal{B} \mid \mathcal{A}, \Omega) = \frac{|\mathcal{B} \cap \mathcal{A}|}{|\mathcal{A}|}$
- $\frac{|\mathcal{B} \cap \mathcal{A}|}{|\Omega|} = p(\mathcal{A} \cap \mathcal{B} \mid \Omega)$

Rappels : Variable aléatoire

Les notations ensemblistes sont un peu lourdes. Elles nécessitent de donner un nom à chaque sous-ensemble intéressant de Ω. Peut-on proposer un peu mieux ?

Définition (Variable aléatoire (v.a.))

- *Une v.a. est une fonction X définie sur $\Omega \longrightarrow \mathcal{D}_X$.*
- $\forall x \in \mathcal{D}_X, \{X = x\} = \{\omega \in \Omega, X(\omega) = x\} \subseteq \Omega$
- *La famille $(\{X = x\})_{x \in \mathcal{D}_X}$ forme une partition de Ω.*
  - $\forall x \neq x' \in \mathcal{D}_X$,
    $\{X = x\}$ *et* $\{X = x'\}$ *sont mutuellement exclusifs.*
  - $\bigcup_{x \in \mathcal{D}_X} \{X = x\} = \Omega$.

On peut considére une variable aléatoire comme un *attribut* sur chaque évènement élémentaire.
Par exemple : $\mathcal{A} = \{C = \mathit{rouge}\}$ est l' "ensemble des objets dont la **couleur** est **rouge**".

Rappels : Variables aléatoires et probabilités

Soient X et Y deux variables aléatoires sur Ω.

Définition

- *loi marginale de X*
  $p(X = x) = p(\{X = x\} \mid \Omega)$
- *loi jointe de X et Y*
  $p(X = x, Y = y) = p(\{X = x\} \cap \{Y = y\} \mid \Omega)$
- *loi conditionnelle de X sachant Y*
  $p(X = x \mid Y = y) = p(\{X = x\} \mid \{Y = y\}, \Omega)$

Propriété

- $$\sum_x p(X = x \mid Y = y) = 1$$
- ⚠ $$\sum_y p(X = x \mid Y = y) \neq 1$$
- $$\sum_x p(X = x, Y = y \mid Z = z) = p(Y = y \mid Z = z)$$

Rappels : Probabilités, fonctions — théorème de Bayes

- $p(X = x|Y = y) \in [0, 1]$ est une probabilité.
- $p(X|Y = y)$ est une distribution de probabilité (une fonction).
- $p(X|Y)$ est une famille de distributions de probabilités sur des espaces différents.

Il y a une différence de nature entre $p(X \mid Y = y)$ et $p(Y \mid X = x)$. On connaît néanmoins un rapport entre ces grandeurs :$p(X, Y) = p(X \mid Y) \cdot p(Y) = p(Y \mid X) \cdot P(X)$

Théorème (de Bayes)

$$p(X \mid Y) = \frac{p(Y \mid X) \cdot p(X)}{p(Y)}$$

ou encore : $p(X \mid Y, Z) = \frac{p(Y|X,Z) \cdot p(X|Z)}{p(Y|Z)}$

Rappels : Variables aléatoires et indépendances

Soient X, Y et Z des variables aléatoires sur Ω.

Définition (Indépendance conditionnelle)

$X \perp\!\!\!\perp Y | Z$ *selon* p *si et seulement si* $p(X|Y, Z) = p(X|Z)$

Définition (indépendance marginale)

$X \perp\!\!\!\perp Y$ *selon* p *si et seulement si* $p(X|Y) = p(X)$

$p(X|Y, Z) = p(X|Z)$ peut se lire :
Augmenter notre état de connaissance Z en apprenant Y n'influence pas la distriution de probabilité qu'on attribue à X.

À noter que cette relation est **symétrique** : $X \perp\!\!\!\perp Y | Z$ si et seulement si $Y \perp\!\!\!\perp X | Z$.

On remarque que, si $X \perp\!\!\!\perp Y$ alors $p(X, Y) = p(X \mid Y) \cdot p(Y) = p(X) \cdot p(Y)$

Approche probabiliste de la classification

Soient, à nouveau, deux v.a. X (de dimension d) discrète et Y (de dimension 1) discrète (*pas forcément binaire*).
Sur la base Π_a, on peut estimer les probabilités par des fréquences pour $P(X, Y)$.
Soit x une instantiation de X, on cherche sa classe y (valeur de Y).

1 Maximum de vraisemblance

$$y = \arg\max_{y_i} P(x \mid y_i)$$

2 Maximum a posteriori

$$y = \arg\max_{y_i} P(y_i \mid x)$$

D'après la règle de Bayes, $P(Y \mid X) \propto P(X \mid Y) \cdot P(Y)$, on comprend que l'intérêt du MAP est de prendre en compte un *a priori* sur la fréquence de chaque classe.

⚠ Il peut être difficile d'obtenir ces distributions.
Particulièrement : $P(X \mid Y)$ peut demander beaucoup d'observation ! !

Exo 1 : français ou suédois ?

On considère deux attributs pour déterminer la nationalité d'un individu. L'attribut taille qui peut prendre les valeurs grand ou petit, l'attribut couleur des cheveux qui peut prendre les valeurs brun ou blond. Les nationalités possibles sont français et suédois.
On suppose que les populations françaises et suédoises se répartissent selon le tableau suivant :

	petit, brun	petit, blond	grand, brun	grand, blond
Suédois	10	20	30	40
Français	25	25	25	25

1. Dans une assemblée comprenant 60% de suédois et 40% de français, décrire
   1. la règle de décision majoritaire
   2. la règle du maximum de vraisemblance
   3. la règle de Bayes
2. Calculez les probabilités d'erreur de chacune de ces règles
3. On suppose maintenant que l'on ne connaît plus les proportions respectives des suédois et des français. On note p la proportion des suédois ($p \in [0, 1]$). Décrire, selon les valeurs possibles de p , les règles de Bayes correspondantes.

Classifieur Bayésien Naïf

Il peut être difficile d'obtenir $P(Y), P(X \mid Y), P(Y \mid X)$.
Particulièrement : $P(X \mid Y)$ peut demander beaucoup d'observation ! !

Hypothèse du classifieur bayésien naïf

On supposera que, $\forall k \neq l, X^k \perp\!\!\!\perp X^l \mid Y$

Cette hypothèse est très forte. Elle a peu de chance de s'avouer exacte dans un cas réel. Néanmoins cette approximation donne des résultats souvent satisfaisants.

Alors, le calcul du MAP s'écrit :

3 Maximum a posteriori

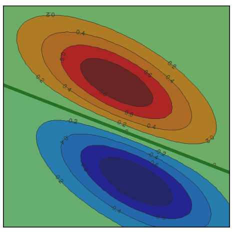
$$y = \arg\max_{y_i} \left(P(y_i) \cdot \prod_{k=1}^{d} P(x^k \mid y_i) \right)$$

Cette hypothèse permet donc de simplifier fortement les calculs nécessaires pour estimer le MAP.

Cas gaussien

Cadre

- Modèle : $\widehat{C}(x) = \sigma(g(x)) = \sigma\left(g_{\oplus}(x) - g_{\ominus}(x)\right)$
- Régions de décision :
  $\forall c \in \left\{\widehat{\ominus}, \widehat{\oplus}\right\}, R_c = \left\{x \in \mathcal{D}, \widehat{C}(x) = c\right\}$
- Frontière de décision : $F = \left\{x \in \mathcal{D}, \widehat{C}(x) = 0\right\}$
- Multinormalité : $\forall c \in \{\ominus, \oplus\}, g_c(x) = P(x \mid c) \sim \mathcal{N}(\mu_c, \Sigma_c)$



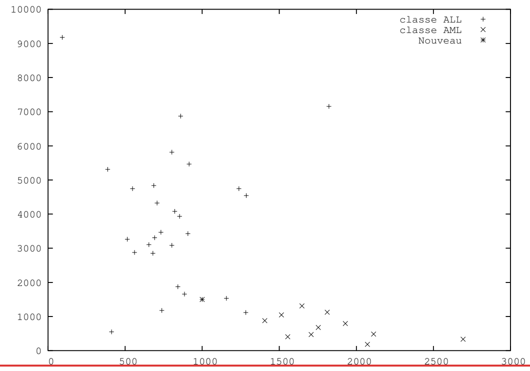
Définition (Densité normale)

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)} \text{ où}$$

- *μ le vecteur des moyennes,*
- *Σ la matrice de covariance :*
  - $\forall i, j \in \{1, \cdots, d\}, \Sigma_{[i,j]} = cov(X_i, X_j)$ *et* $\Sigma_{[i,i]} = var(X_i)$
  - *Σ semi-définie positive* ($\forall x, x^t \Sigma x \geq 0$)
  - $|\Sigma| \geq 0$ *et* Σ^{-1} *existe*

Un exemple

Les données –les couples (gènes 1, gène 2) – suivent un certain modèle : deux distributions de probabilité sur D_X (une pour chaque classe).



Le classifieur Bayésien naïf - 1

Données :

- l'ensemble d'apprentissage $X = (x_1, x_2, \cdots, x_n)$ où $x_i = (v_{i1}, v_{i2}, \cdots, v_{ip})$
- un nouvel élément $x = (v_1, \cdots, v_p)$

Problème : trouver la classe de x.
Approche probabiliste :

$$c_{MAP} = \text{argmax}_c P(c|x) = \text{argmax}_c \frac{P(x|c)P(c)}{P(x)} = \text{argmax}_c P(x|c).$$

Il faut estimer les probabilités $P(c)$ et P $(x|c)$ à partir de X.

Le classifieur Bayésien naïf - 2

- Estimation de $P(c)$

$$\begin{aligned} P(0) &= n_0/n = 27/38 \\ P(1) &= n_1/n = 11/38 \end{aligned}$$

- Le problème est plus compliqué pour $P(x|c)$

Approche naïve Bayes : les attributs sont indépendants étant donnée la classe :

$$P(x|c) = P(g_1|c)P(g_2|c)$$

Il faut estimer $P(g_1|c)$ et $p(g_2|c)$ à partir de X.

Le classifieur Bayésien naïf - 3

Hypothèse de normalité :

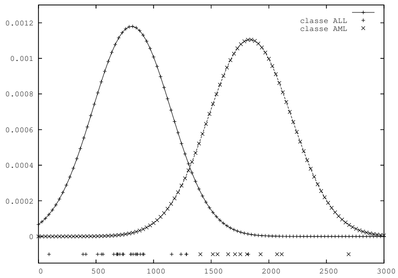
- $P(g_1|c) \rightsquigarrow \mathcal{N}(g_1, \mu_{g1,c}, \sigma^2_{g1,c})$
- $P(g_2|c) \rightsquigarrow \mathcal{N}(v_p, \mu_{g2,c}, \sigma^2_{g2,c})$

On n'a plus qu'a estimer les moyennes et écart-types à partir de X.

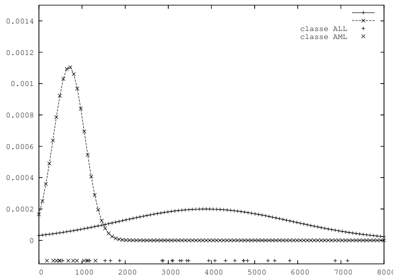
$$\mu_{g1,0} \approx 810.29 \qquad \sigma_{g1,0} \approx 338.17$$
$$\mu_{g1,1} \approx 1836.27 \qquad \sigma_{g1,1} \approx 360.88$$

$$\mu_{g1,0} \approx 3863.29 \qquad \sigma_{g1,0} \approx 1999.78$$
$$\mu_{g1,0} \approx 702.36 \qquad \sigma_{g1,0} \approx 360.34$$

Le classifieur Bayésien naïf - 4



Fumarylacetoacetate



C-myb

Le classifieur Bayésien naïf - 5

On veut déterminer la classe de $x = (1000, 1500)$:

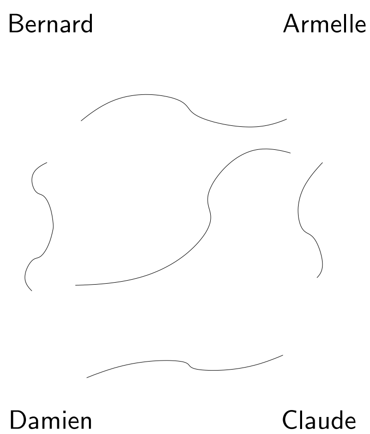
$$\begin{aligned}
P(0|x) &= P(0)\mathcal{N}\left(1500, \mu_{g1,0}, \sigma^2_{g1,0}\right)\mathcal{N}\left(4000, \mu_{g2,0}, \sigma^2_{g2,0}\right) \\
&= 0.71053 \times 1.01\ 10^{-3} \times 9.92\ 10^{-5} = 7.11\ 10^{-8} \\
P(1|x) &= P(1)\mathcal{N}\left(1500, \mu_{g1,1}, \sigma^2_{g1,1}\right)\mathcal{N}\left(4000, \mu_{g2,1}, \sigma^2_{g2,1}\right) \\
&= 0.28947 \times 7.54\ 10^{-5} \times 9.55\ 10^{-5} = 2.09\ 10^{-9} \\
\Rightarrow c_{MAP} &= 0
\end{aligned}$$

Expectation-Maximisation

1. Quelques rappels de maths
2. L'algorithme EM
3. Pourquoi fonctionne-t-il ?
4. Mixtures de Gaussiennes et EM

Motivations : Système de recommandation

- Armelle, Bernard, Claude $\Longrightarrow$ notes films (r_A, r_B, r_C)
- Problème : quel film conseiller à Damien ?
- Solution : échantillons $\langle r_A, r_B, r_C, r_D \rangle \Longrightarrow P(r_A, r_B, r_C, r_D)$
- conseiller Damien en exploitant $P(r_D | r_A, r_B, r_C)$



Motivations : Système de recommandation

Film	r_A	r_B	r_C	r_D
I robot	4	2	3	3
Forest Gump	1	3	2	4
Intouchables	2	2	3	2
Le parrain	1	3	2	1
Pulp fiction	2	4	3	4

- évaluations : $1, 2, 3, 4$
- $P(r_D | r_A, r_B, r_C)$: distribution multivariée
- $\Theta =$ paramètres de $P(r_D | r_A, r_B, r_C) \Longrightarrow 4^4 = 256$ paramètres θ_i
- $\sum_{r_D} P(r_D | r_A, r_B, r_C) = 1 \Longrightarrow 3 \times 4^3 = 192\ \theta_i$

Motivations : Système de recommandation

Vraisemblance : $L(\mathbf{x}, \Theta) = \prod_{\text{film}} \theta_{\text{film}}$

θ_{abcd} : paramètre pour un film ayant obtenu $(r_A = a, r_B = b, r_C = c, r_D = d)$

N_{abcd} : nombre de films ayant obtenu $(r_A = a, r_B = b, r_C = c, r_D = d)$

$$L(\mathbf{x}, \Theta) = \prod_{a=1}^{4}\prod_{b=1}^{4}\prod_{c=1}^{4}\left[\left(1 - \sum_{d=1}^{3}\theta_{abcd}\right)^{N_{abc4}} \prod_{d=1}^{3}\theta_{abcd}^{N_{abcd}}\right]$$

$$\log L(\mathbf{x}, \Theta) = \sum_{a=1}^{4}\sum_{b=1}^{4}\sum_{c=1}^{4}\left[N_{abc4}\log\left(1 - \sum_{d=1}^{3}\theta_{abcd}\right) + \sum_{d=1}^{3} N_{abcd}\log\theta_{abcd}\right]$$

$$\frac{\partial \log L(\mathbf{x}, \Theta)}{\partial \theta_{abcd}} = -\frac{N_{abc4}}{1 - \sum_{d'=1}^{3}\theta_{abcd'}} + \frac{N_{abcd}}{\theta_{abcd}} = 0 \quad \forall d \in \{1,2,3\}$$

$$-N_{abc4}\theta_{abcd} + N_{abcd}\left(1 - \sum_{d'=1}^{3}\theta_{abcd'}\right) = 0 \quad \forall d \in \{1,2,3\}$$

Motivations : Système de recommandation

- $-N_{abc4}\theta_{abcd} + N_{abcd}\left(1 - \sum_{d'=1}^{3}\theta_{abcd'}\right) = 0 \quad \forall d \in \{1,2,3\}$
- $\frac{N_{abc4}}{N_{abcd}}\theta_{abcd} + \sum_{d'=1}^{3}\theta_{abcd'} = 1 \quad \forall d \in \{1,2,3\}$
- $\begin{bmatrix} 1+\frac{N_{abc4}}{N_{abc1}} & 1 & 1 \\ 1 & 1+\frac{N_{abc4}}{N_{abc2}} & 1 \\ 1 & 1 & 1+\frac{N_{abc4}}{N_{abc3}} \end{bmatrix}\begin{bmatrix}\theta_{abc1}\\ \theta_{abc2}\\ \theta_{abc3}\end{bmatrix} = \begin{bmatrix}1\\1\\1\end{bmatrix}$

$$\theta_{abcd} = \frac{N_{abcd}}{\sum_{d'=1}^{4} N_{abcd'}}$$

Motivations : Système de recommandation

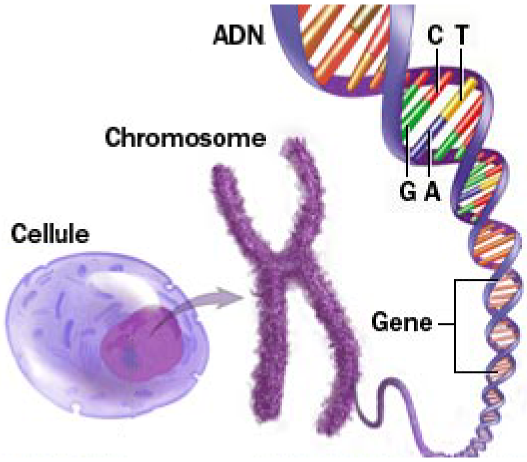
Film	r_A	r_B	r_C	r_D
I robot	4	2	3	3
Forest Gump	1	3	2	4
Intouchables	2	2	3	2
Le parrain	1	3	2	1
Pulp fiction	2	4	3	4

Film	r_A	r_B	r_C	r_D
I robot	4		3	3
Forest Gump			2	4
Intouchables	2	2	3	
Le parrain	1		2	
Pulp fiction	2	4	3	4

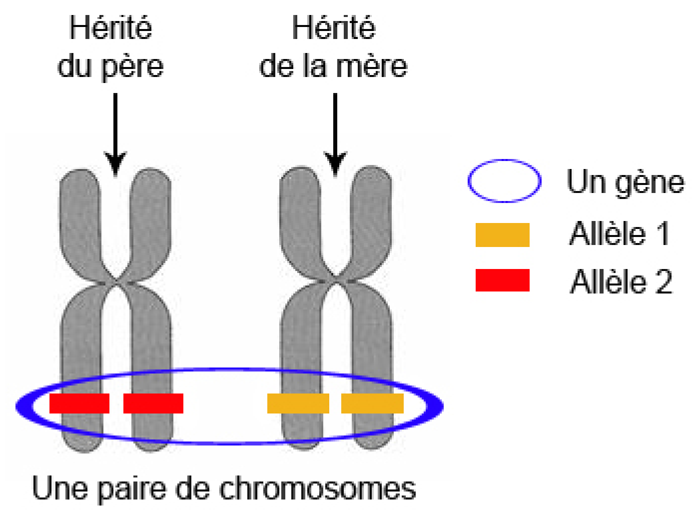
- $\sum_{r_D} P(r_D|r_A, r_B, r_C) = 1 \Longrightarrow 3 \times 4^3 = 192\ \theta_i$
- probabilités $\Longrightarrow$ au moins 2000 évaluations
- certains films n'ont pas été vus $\Longrightarrow$ données manquantes

Problème : comment estimer $P(r_D|r_A, r_B, r_C)$?

Motivations : reconstruction de génotypes



Motivations : reconstruction de génotypes



Motivations : reconstruction de génotypes

Génotype, phénotype

Génotype = paire d'allèles d'un segment d'ADN.
Phénotype = caractère observable d'un génotype.

Exemple : groupe sanguin (allèles A, B, O)

génotype	phénotype	X	Y
AA	A	1	1
AB	AB	2	3
AO	A	3	1
BB	B	4	2
BO	B	5	2
OO	O	6	4

⚠ 6 génotypes mais seulement 4 phénotypes !

Problème : Apprendre en présence de variables latentes (jamais observées)

Motivations : résumé

Problèmes étudiés dans ce cours :

1. estimation de paramètres en présence de données manquantes
2. estimation de paramètres en présence de variables latentes

Solution : l'algorithme EM

Typologies de données incomplètes

- $\mathbf{x}^o$: données observées, $\mathbf{x}^h$: données manquantes
- $\mathbf{x} = \mathbf{x}^o \cup \mathbf{x}^h$

Film	r_A	r_B	r_C	r_D
I robot	4	?	3	3
Forest Gump	?	?	2	4
Intouchables	2	2	3	?
Le parrain	1	?	2	?
Pulp fiction	2	4	3	4

- $\mathcal{M}_{ij} = P(r_i^j \in \mathbf{x}^h)$: position des données manquantes

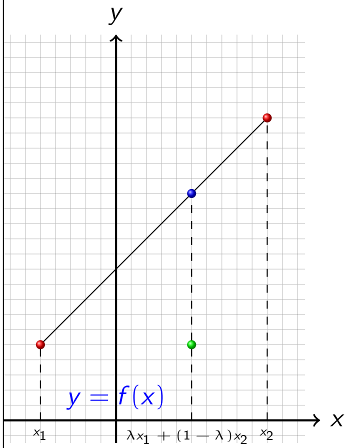
Typologies de données incomplètes

Typologies

- Missing Completely at Random (MCAR) : $P(\mathcal{M}|\mathbf{x}) = P(\mathcal{M})$ Aucune relation entre le fait qu'une donnée soit manquante ou observée
- Missing at Random (MAR) : $P(\mathcal{M}|\mathbf{x}) = P(\mathcal{M}|\mathbf{x}^o)$ données manquantes en relation avec les données observées mais pas avec les autres données manquantes
- Not Missing At Random (NMAR) : $P(\mathcal{M}|\mathbf{x})$ données manquantes en relation avec *toutes* les données

⚠ On n'étudiera que MCAR !

Rappel : fonctions convexes



Définition

f convexe $\iff \forall \lambda \in [0,1], \forall x_1, x_2$:

$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$

fonction concave

f concave $\iff -f$ convexe

Généralisation : l'inégalité de Jensen

Inégalité de Jensen

- f convexe définie sur D
- $x_1, \ldots, x_n \in D$
- $\lambda_1, \ldots, \lambda_n \geq 0, \quad \sum_{i=1}^n \lambda_i = 1$

Alors :

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i)$$

Inégalité de Jensen

- f convexe
- X : variable aléatoire à n dimensions $x_1, \ldots, x_n$
- $\lambda_1, \ldots, \lambda_n \geq 0, \quad \sum_{i=1}^n \lambda_i = 1 \Longrightarrow$ probabilité P_λ
- $f(\mathbb{E}_{P_\lambda}(X)) \leq \mathbb{E}_{P_\lambda}(f(X))$ où $\mathbb{E}_{P_\lambda}$ = espérance

Démonstration de l'inégalité de Jensen

$$f\left(\sum_{i=1}^n \lambda_i x_i\right) \leq \sum_{i=1}^n \lambda_i f(x_i)$$

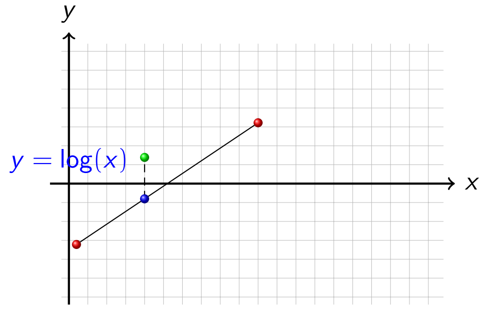
- par récurrence : si $n = 1$: trivial, si $n = 2$: convexité
- $$\begin{aligned} f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_{n+1}x_{n+1} + \sum_{i=1}^n \lambda_i x_i\right) \\ &= f\left(\lambda_{n+1}x_{n+1} + (1-\lambda_{n+1}) \sum_{i=1}^n \frac{\lambda_i}{1-\lambda_{n+1}} x_i\right) \\ &\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) f\left(\sum_{i=1}^n \frac{\lambda_i}{1-\lambda_{n+1}} x_i\right) \\ &\leq \lambda_{n+1} f(x_{n+1}) + (1-\lambda_{n+1}) \sum_{i=1}^n \frac{\lambda_i}{1-\lambda_{n+1}} f(x_i) \\ &= \lambda_{n+1} f(x_{n+1}) + \sum_{i=1}^n \lambda_i f(x_i) = \sum_{i=1}^{n+1} \lambda_i f(x_i) \end{aligned}$$

Conséquences de l'inégalité de Jensen

Inégalité de Jensen pour le logarithme

Logarithme = fonction concave :

$$\log\left(\sum_{i=1}^{n}\lambda_i x_i\right) \geq \sum_{i=1}^{n}\lambda_i \log(x_i)$$



$$\mathbb{E}(\log(X)) = \log(\mathbb{E}(X)) \Longrightarrow X = \mathbb{E}(X) = \text{constante}$$

Log-vraisemblance et données incomplètes

- Échantillon $\mathbf{x} = \{x_1, \ldots, x_n\}$ de taille n
- données complètes : $\log L(\mathbf{x}, \Theta) = \sum_{i=1}^{n} \log P(x_i|\Theta)$
- $\mathbf{x}^o$: données observées, $\mathbf{x}^h$: données manquantes
- $\log L(\mathbf{x}^o, \Theta)$ = log-vraisemblance des données observées

$$= \sum_{i=1}^{n} \log P(x_i^o|\Theta) = \sum_{i=1}^{n} \log\left(\sum_{x_i^h \in \mathbf{x}^h} P(x_i^o, x_i^h|\Theta)\right)$$

- Soit $Q_i(x_i^h)$ une loi de proba quelconque alors :

$$\log L(\mathbf{x}^o, \Theta) = \sum_{i=1}^{n} \log\left(\sum_{x_i^h \in \mathbf{x}^h} Q_i(x_i^h)\frac{P(x_i^o, x_i^h|\Theta)}{Q_i(x_i^h)}\right)$$

Log-vraisemblance et données incomplètes

- $\log L(\mathbf{x}^o, \Theta) = \sum_{i=1}^{n} \log\left(\sum_{x_i^h \in \mathbf{x}^h} Q_i(x_i^h)\frac{P(x_i^o, x_i^h|\Theta)}{Q_i(x_i^h)}\right)$

inégalité de Jensen $\Longrightarrow \log\left(\sum_{i=1}^{n}\lambda_i y_i\right) \geq \sum_{i=1}^{n}\lambda_i \log(y_i)$

$$\log L(\mathbf{x}^o, \Theta) \geq \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i(x_i^h) \log\left(\frac{P(x_i^o, x_i^h|\Theta)}{Q_i(x_i^h)}\right)$$

Jensen $\Longrightarrow$ égalité ssi $\frac{P(x_i^o, x_i^h|\Theta)}{Q_i(x_i^h)}$ = constante

$$\text{choisir } Q_i(x_i^h) \propto P(x_i^o, x_i^h|\Theta) \Longrightarrow Q_i(x_i^h) = P(x_i^h|x_i^o, \Theta)$$

Algorithme EM

Algorithme

1. choisir une valeur initiale $\Theta = \Theta^0$
2. Étape E (expectation) :
   - $Q_i^{t+1}(x_i^h) \leftarrow P(x_i^h | x_i^o, \Theta^t) \quad \forall i \in \{1, \ldots, n\}$
   - $\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left(\frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right)$
3. Étape M (maximization) :
   - $\Theta^{t+1} \leftarrow \text{argmax}_\Theta \log L^{t+1}(\mathbf{x}^o, \Theta)$
4. Tant qu'on n'a pas convergé, revenir en 2

À convergence, Θ^{t+1} = optimum local par max de vraisemblance

Algorithme EM : un exemple

- 2 variables aléatoires $A \in \{a, b\}$ et $C \in \{c, d\}$

$$P(A, C|\Theta) = \begin{bmatrix} \theta_{ac} & \theta_{ad} \\ \theta_{bc} & \theta_{bd} \end{bmatrix} \Longrightarrow \Theta = \{\theta_{ac}, \theta_{ad}, \theta_{bc}, \theta_{bd}\}$$

but : estimer Θ par EM

A	C
a	?
b	?
a	d
b	d
a	c

- A toujours observé :

  $\Longrightarrow \theta_{ac} + \theta_{ad} = \frac{3}{5}$ par max de vraisemblance

  $\theta_{bc} + \theta_{bd} = \frac{2}{5}$ par max de vraisemblance

- initialisation possible :

  $\Theta^0 = \{\theta_{ac}^0 = 0.3, \theta_{ad}^0 = 0.3, \theta_{bc}^0 = 0.2, \theta_{bd}^0 = 0.2\}$

- Étape E (expectation) : $Q_i^1(x_i^h) \leftarrow P(x_i^h | x_i^o, \Theta^0) \quad \forall i \in \{1, 2\}$

$$Q_1^1(C) = P(C|A = a, \Theta^0) = \frac{P(A=a, C|\Theta^0)}{\sum_C P(A=a, C|\Theta^0)} = [\tfrac{0.3}{0.6} \ , \ \tfrac{0.3}{0.6}] = [0.5 \ , \ 0.5]$$

$$Q_2^1(C) = P(C|A = b, \Theta^0) = \frac{P(A=b, C|\Theta^0)}{\sum_C P(A=b, C|\Theta^0)} = [\tfrac{0.2}{0.4} \ , \ \tfrac{0.2}{0.4}] = [0.5 \ , \ 0.5]$$

Algorithme EM : un exemple

$$\Theta^0 = \{\theta_{ac}^0 = 0.3, \theta_{ad}^0 = 0.3, \theta_{bc}^0 = 0.2, \theta_{bd}^0 = 0.2\}$$

$Q_1^1(C) = [0.5 \ , \ 0.5]$ $\quad Q_2^1(C) = [0.5 \ , \ 0.5]$ $\quad P(x_i^h, x_i^o|\Theta^0) = \begin{bmatrix} 0.3 & 0.3 \\ 0.2 & 0.2 \end{bmatrix}$

$$\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left(\frac{P(x_i^o, x_i^h | \Theta)}{Q_i^{t+1}(x_i^h)} \right)$$

A	B
a	?
b	?
a	d
b	d
a	c

$\Longrightarrow$

A	B	Q_i^{t+1}	P/Q_i^{t+1}	$\log(P/Q_i^{t+1})$
a	c	0.5	$\theta_{ac}/0.5$	$\log \theta_{ac} - \log 0.5$
a	d	0.5	$\theta_{ad}/0.5$	$\log \theta_{ad} - \log 0.5$
b	c	0.5	$\theta_{bc}/0.5$	$\log \theta_{bc} - \log 0.5$
b	d	0.5	$\theta_{bd}/0.5$	$\log \theta_{bd} - \log 0.5$
a	d	1	θ_{ad}	$\log \theta_{ad}$
b	d	1	θ_{bd}	$\log \theta_{bd}$
a	c	1	θ_{ac}	$\log \theta_{ac}$

⚠ $\Longrightarrow$ revient à observer l'échantillon avec poids Q_i^{t+1}

Algorithme EM

$$\log L^{t+1}(\mathbf{x}^o,\Theta) = \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left(\frac{P(x_i^o, x_i^h|\Theta)}{Q_i^{t+1}(x_i^h)} \right)$$
$$= \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \left[\log(P(x_i^o, x_i^h|\Theta)) - \log(Q_i^{t+1}(x_i^h))\right]$$

$$\Rightarrow \Theta^{t+1} = \text{argmax}_\Theta \log L^{t+1}(\mathbf{x}^o,\Theta)$$
$$= \text{argmax}_\Theta \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log(P(x_i^o, x_i^h|\Theta))$$

Principe de EM

Étape M $\Longrightarrow$ maximum de vraisemblance avec un échantillon dont chaque enregistrement x_i a un poids Q_i^{t+1}

Algorithme EM : un exemple

$$\Theta^1 = \text{argmax}_\Theta \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log \left(\frac{P(x_i^o, x_i^h|\Theta)}{Q_i^{t+1}(x_i^h)} \right)$$

A	B	Q_i^{t+1}	$\log\theta$
a	c	0.5	$\log\theta_{ac}$
a	d	0.5	$\log\theta_{ad}$
b	c	0.5	$\log\theta_{bc}$
b	d	0.5	$\log\theta_{bd}$
a	d	1	$\log\theta_{ad}$
b	d	1	$\log\theta_{bd}$
a	c	1	$\log\theta_{ac}$

$= \text{argmax}_\Theta [0.5+1]\log\theta_{ac} + [0.5+1]\log\theta_{ad} + 0.5\log\theta_{bc} + [0.5+1]\log\theta_{bd}$
Sous contrainte : $\theta_{ac}+\theta_{ad}+\theta_{bc}+\theta_{bd}=1$

$$\Theta^1 = \{\theta^1_{ac} = \tfrac{3}{10},\ \theta^1_{ad} = \tfrac{3}{10},\ \theta^1_{bc} = \tfrac{1}{10},\ \theta^1_{bd} = \tfrac{3}{10}\}$$

Algorithme EM : un exemple

$$\Theta^1 = \{\theta^1_{ac} = \tfrac{3}{10},\ \theta^1_{ad} = \tfrac{3}{10},\ \theta^1_{bc} = \tfrac{1}{10},\ \theta^1_{bd} = \tfrac{3}{10}\}$$

A	C
a	?
b	?
a	d
b	d
a	c

- Étape E (expectation) : $Q_i^2(x_i^h) \leftarrow P(x_i^h|x_i^o,\Theta^1) \quad \forall i \in \{1,2\}$

$Q_1^2(C) = P(C|A=a,\Theta^1) = \frac{P(A=a,C|\Theta^1)}{\sum_C P(A=a,C|\Theta^1)} = [\frac{0.3}{0.6}\ ,\ \frac{0.3}{0.6}] = [0.5\ ,\ 0.5]$

$Q_2^2(C) = P(C|A=b,\Theta^1) = \frac{P(A=b,C|\Theta^1)}{\sum_C P(A=b,C|\Theta^1)} = [\frac{0.1}{0.4}\ ,\ \frac{0.3}{0.4}] = [0.25\ ,\ 0.75]$

Algorithme EM : un exemple

$$\Theta^1 = \{\theta^1_{ac} = \tfrac{3}{10},\ \ \theta^1_{ad} = \tfrac{3}{10},\ \ \theta^1_{bc} = \tfrac{1}{10},\ \ \theta^1_{bd} = \tfrac{3}{10}\}$$

$Q_1^2(C) = [0.5\ ,\ 0.5]$ $\quad Q_2^2(C) = [0.25\ ,\ 0.75]$ $\quad P(x_i^h, x_i^o|\Theta^0) = \begin{bmatrix} 0.3 & 0.3 \\ 0.1 & 0.3 \end{bmatrix}$

$$\log L^{t+1}(\mathbf{x}^o, \Theta) = \sum_{i=1}^{n} \sum_{x_i^h \in \mathbf{x}^h} Q_i^{t+1}(x_i^h) \log\left(\frac{P(x_i^o, x_i^h|\Theta)}{Q_i^{t+1}(x_i^h)}\right)$$

A	B
a	?
b	?
a	d
b	d
a	c

$\Longrightarrow$

A	B	Q_i^{t+1}	P/Q_i^{t+1}	$\log(P/Q_i^{t+1})$
a	c	0.5	$\theta_{ac}/0.5$	$\log\theta_{ac} - \log 0.5$
a	d	0.5	$\theta_{ad}/0.5$	$\log\theta_{ad} - \log 0.5$
b	c	0.25	$\theta_{bc}/0.25$	$\log\theta_{bc} - \log 0.25$
b	d	0.75	$\theta_{bd}/0.75$	$\log\theta_{bd} - \log 0.75$
a	d	1	θ_{ad}	$\log\theta_{ad}$
b	d	1	θ_{bd}	$\log\theta_{bd}$
a	c	1	θ_{ac}	$\log\theta_{ac}$

$= \text{argmax}_\Theta [0.5+1]\log\theta_{ac} + [0.5+1]\log\theta_{ad} + 0.25\log\theta_{bc} + [0.75+1]\log\theta_{bd}$

Sous contrainte : $\theta_{ac} + \theta_{ad} + \theta_{bc} + \theta_{bd} = 1$

Algorithme EM : un exemple

- $\Theta^2 = \{\theta^2_{ac} = \frac{3}{10},\ \ \theta^2_{ad} = \frac{3}{10},\ \ \theta^2_{bc} = \frac{1}{20},\ \ \theta^2_{bd} = \frac{7}{20}\}$
- $\Theta^3 = \{\theta^3_{ac} = \frac{3}{10},\ \ \theta^3_{ad} = \frac{3}{10},\ \ \theta^3_{bc} = \frac{1}{40},\ \ \theta^2_{bd} = \frac{15}{40}\}$

...

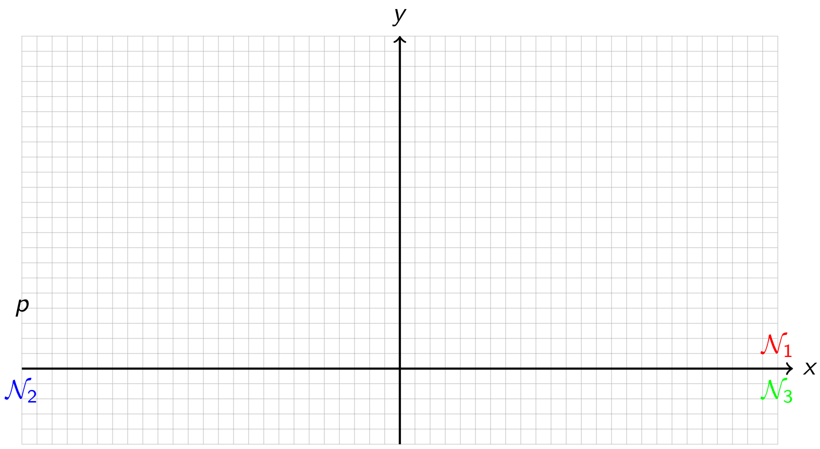
- $\theta_{ac} = \theta_{bc} = 0,3$
- $\theta_{bc} + \theta_{bd} = 0,4$ et θ_{bc} divisé par 2 à chaque étape.

$\Longrightarrow$ à convergence : $\Theta = \{\theta_{ac} = \frac{3}{10},\ \ \theta_{ad} = \frac{3}{10},\ \ \theta_{bc} = 0,\ \ \theta_{bd} = \frac{4}{10}\}$

Mixtures de Gaussiennes et EM

Mixture de gaussiennes

$$p(\cdot) = 0,3 \times \mathcal{N}(0, 2^2) + 0,4 \times \mathcal{N}(4, 3^2) + 0,3 \times \mathcal{N}(-3, 1^2)$$



Application : apprentissage de prix fonciers

postulat : prix de biens similaires dans un quartier ~ identiques

$\Longrightarrow$ prix dépendent $\begin{cases} \text{des caractéristiques du bien (e.g. nombre de pièces)} \\ \text{du quartier} \end{cases}$

$\Longrightarrow$ modélisation par une mixture de gaussiennes (ici 2 gaussiennes)

Modélisation du problème

- $\Theta = \{\mu_1, \mu_2, \sigma_1, \sigma_2, \pi_1, \pi_2\}$
- $p(x|\Theta) = \pi_1 \; \mathcal{N}(\mu_1, \sigma_1^2) + \pi_2 \; \mathcal{N}(\mu_2, \sigma_2^2)$

Apprentissage non supervisé

- échantillon $\mathbf{x} = \langle x_1, \dots, x_n \rangle$
- x_i = prix $\Longrightarrow$ on ne connaît pas la Gaussienne à laquelle le bien appartient

⚠ échantillon supposé complet (pas de données manquantes)

Application : apprentissage de prix fonciers

échantillon complet $\Longrightarrow$ estimation par max de vraisemblance

$$L(\mathbf{x}, \Theta) = \prod_{i=1}^{2} p(x_i|\Theta) = \prod_{i=1}^{2} \sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_k}{\sigma_k}\right)^2\right\}$$

$$\log L(\mathbf{x}, \Theta) = \sum_{i=1}^{2} \log \left[\sum_{k=1}^{2} \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_k}{\sigma_k}\right)^2\right\}\right]$$

⚠ trop compliqué à maximiser analytiquement !

Solution : EM

1. x_i appartient à une classe $y_{k(i)}$ non observée $\sim \mathcal{N}(\mu_{k(i)}, \sigma_{k(i)})$
2. échantillon $\mathbf{x} = \langle (x_i, y_{k(i)}) \rangle$
3. échantillon maintenant avec données manquantes $\Longrightarrow$ EM

Application : apprentissage de prix fonciers

Nouvelle modélisation du problème

$$p(X_i, Y_{k(i)}|\Theta) = p(X_i|Y_{k(i)},\Theta)P(Y_{k(i)}|\Theta) = \begin{bmatrix} \mathcal{N}(\mu_1,\sigma_1^2)\ \pi_1 \\ \mathcal{N}(\mu_2,\sigma_2^2)\ \pi_2 \end{bmatrix}$$

$\Longrightarrow$ pour x_i connu :

$$P(Y_{k(i)}|x_i,\Theta) = \frac{p(x_i, Y_{k(i)}|\Theta)}{p(x_i|\Theta)} \propto \begin{bmatrix} \frac{1}{\sqrt{2\pi}\sigma_1} exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2\right\} \times \pi_1 \\ \frac{1}{\sqrt{2\pi}\sigma_2} exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2\right\} \times \pi_2 \end{bmatrix}$$

- Initialisation d'EM : choisir une valeur $\Theta^0 = \{\mu_1^0, \mu_2^0, \sigma_1^0, \sigma_2^0, \pi_1^0, \pi_2^0\}$
- Étape E : $Q_i^1(y_k) \leftarrow P(y_k|x_i,\Theta^0)$ pour $k = 1, 2$

  $\Longrightarrow Q_i^1(\cdot)$ très facile à calculer
- Étape M :

$$\underset{\Theta}{\text{argmax}} \log L^{t+1}(\mathbf{x}^o,\Theta) = \underset{\Theta}{\text{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{2} Q_i^{t+1}(y_k)\log\left(\frac{p(x_i,y_k|\Theta)}{Q_i^{t+1}(y_k)}\right)$$

Application : apprentissage de prix fonciers

Étape M :

$$\begin{aligned}
&\text{argmax}_\Theta \log L^{t+1}(\mathbf{x}^o,\Theta) \\
&= \underset{\Theta}{\text{argmax}} \sum_{i=1}^{n}\sum_{k=1}^{2} Q_i^{t+1}(y_k)\log\left(\frac{p(x_i,y_k|\Theta)}{Q_i^{t+1}(y_k)}\right) \\
&= \underset{\Theta}{\text{argmax}} \sum_{i=1}^{n} Q_i^{t+1}(y_1)\log\left(\pi_1\frac{1}{\sqrt{2\pi\sigma_1^2}} exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2\right\}\right) + \\
&\qquad Q_i^{t+1}(y_2)\log\left(\pi_2\frac{1}{\sqrt{2\pi\sigma_2^2}} exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2\right\}\right) \\
&= \underset{\Theta}{\text{argmax}} \sum_{i=1}^{n} Q_i^{t+1}(y_1)\left[\log(\pi_1) - \frac{1}{2}\log(\sigma_1^2) - \frac{1}{2}\left(\frac{x_i-\mu_1}{\sigma_1}\right)^2\right] + \\
&\qquad Q_i^{t+1}(y_2)\left[\log(\pi_2) - \frac{1}{2}\log(\sigma_2^2) - \frac{1}{2}\left(\frac{x_i-\mu_2}{\sigma_2}\right)^2\right]
\end{aligned}$$

Argmax facile à calculer !