

App. - Non
Supervi

D)

Chapitre 2

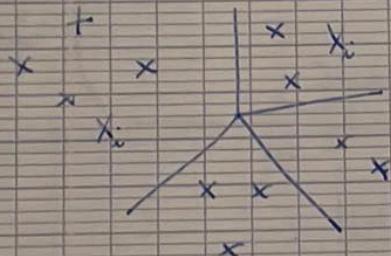
Clustering / Partitionnement

Classification: (x, y) à valeurs dans $\mathbb{R}^d \times [l, k]$ (k classes).

But: pedirle Y dachant nos X.

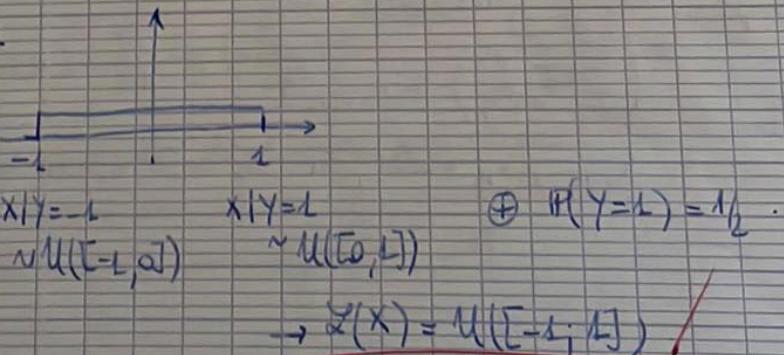
Observation: Seulement X (on n'a plus nos (Y_i) , d'entraînement)
 " (X_1, \dots, X_n) iid.

On suppose que les latents γ_i existent, et le but sera de trouver la permutation σ t.q. $f(x_i) \simeq \sigma(\gamma_i)$.

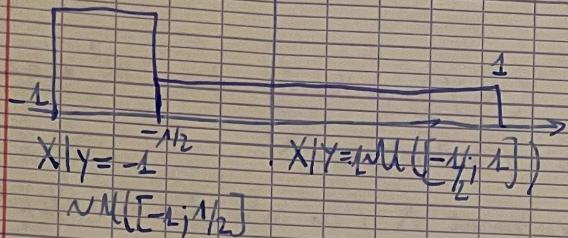


Le pb est mal posé: • K inconnu (nb de classes) $\xrightarrow{\text{choix}} K$.
• $X \rightarrow$ information partiellement privée,
ne $\propto (Y|X)$.

Fol.



(2)



$$\text{P}(Y=1) = 3/4$$

$$\text{Alors, } \mathcal{Z}(X) = N(-1; 1).$$

Clusters: rassemblement de pts "communs":

→ similarité INTRA classe.

→ dissimilarité INTER classe.

Les modes de la loi de \mathcal{Z} .

région de faible densité

) définition et contexte (implique).

= Modèle donné un $\mathcal{Z}(X)$!

Modèles de mélange:

$\mathcal{P} = \{\lambda_i f_{\mu_i}, \lambda_i \in \Delta\}$, $\forall \mu_i$ un ensemble de loi.

$$\left(\sum_{i=1}^k \pi_i f_{\mu_i}\right) \cdot \mu$$

$\lambda_1, \dots, \lambda_k \in \Delta$, et $\pi_i \in [0, 1]$ tq $\Delta^T \pi = \sum_i \pi_i = 1$

③

→ Classification:

variable latente qui permet de partitionner l'espace.

$$\left\{ \begin{array}{l} \Pr(Y=j) = \pi_j \\ X|Y=j \sim P_{yj} \in \mathcal{P} \end{array} \right.$$

$$\left\{ \begin{array}{l} X \sim \sum_{j=1}^k \pi_j P_{yj} \\ j \end{array} \right.$$

$$\Pr(Y=j|X=x) = \pi_j f_{yj}(x) = \frac{\pi_j f_{yj}(x)}{\sum_{l=1}^k \pi_l f_{yl}(x)} \quad (= \frac{\Pr(Y=j) \cdot \Pr(X=x|Y=j)}{\Pr(X=x)})$$

→ Ceci définit la loi de (X, Y) .

$\forall x \in \mathbb{R}^d$, argmax $\Pr(Y=j|X=x) : = g(x)$ va permettre de

classifier / ordonner l'espace / les données.

► Considérons $X \sim \sum_{j=1}^k \pi_j P_{yj}$ où $P_{yj} \in \mathcal{P}_m = \{M_\theta, \theta = (\pi, \lambda)\}$ tq

$$\sum_j \pi_j = 1 \quad \text{et } \lambda_{j1}, \dots, \lambda_{jm} \in \Lambda.$$

$$\log\text{-vrais.} \Rightarrow \ln(\theta) = \sum_{i=1}^n \log \left(\sum_{j=1}^k \pi_j f_{yj}(x_i) \right)$$

Exemple: $K=2$. $P_{yj} = \mathcal{U}(\mu_j, 1)$ $\theta = (\pi_1, \mu_1, \mu_2) \in \mathbb{R}^3$.

$$\ln(\theta) = \sum_{i=1}^n \log \left(\frac{\pi_1}{\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x_i - \mu_1)^2}{1}} + \frac{\pi_2}{\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(x_i - \mu_2)^2}{1}} \right)$$

→ Impossible de trouver explicitement les paramètres qui maximisent notre log-vraisemblance.

On va passer par l'algorithme EM

(4)

Supposons que l'on arrive à observer les étiquettes y_i pour des z_i :

$$z_1, \dots, z_n \sim y_1, \dots, y_m$$

$\mathcal{L}(x, y)$ est définie par la densité ($\lambda_{IR} \times \delta_{y_i=x_i} \cdot \pi_x$)

$$f_\theta(x, z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\mu_z - x)^2} \cdot \pi_x$$

$$f_\theta(x, -1) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\mu_z - x)^2} (\lambda_{IR})$$

$$\text{Puis } \mathcal{P}_g = \{ \mathcal{L}(x, y), \theta \in \Theta \}.$$

$$\text{Par suite: } l_{(x, z)_1^n}(\theta) = \sum_{i=1}^n \left(\log(\pi_x) + \log\left(\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_z)^2}{2}}\right) \cdot \mathbb{I}_{z_i=1} \right. \\ \left. + \left(\dots \right) \mathbb{I}_{z_i=-1} \cdot \mathbb{E}[\mathbb{I}_{z_i=-1}|x] \right)$$

On cherche $\hat{\theta} \in \arg\max_{\theta} l_{(x, z)_1^n}(\theta)$

$$\text{Où, } \hat{\theta} = (\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_2), \text{ avec } N = \sum_{i=1}^n \mathbb{I}_{z_i=1}$$

$$\text{Puis, } \hat{\pi}_1 = \frac{1}{N}$$

$$\hat{\mu}_1 = \frac{1}{N} \sum_{i=1}^n x_i \cdot \mathbb{I}_{z_i=1}$$

$$\hat{\mu}_2 = \frac{1}{N} \sum_{i=1}^n x_i \cdot \mathbb{I}_{z_i=-1}$$

Si $\hat{\theta}_0$ candidat estimateur:

$$z_1, \dots, z_n \text{ tq } \mathcal{L}(z_i | x_i^n) = \mathcal{L}(z_i | \hat{\theta}_0, x_i)$$

$$\Rightarrow P(z_i=1 | x_i) = \hat{\pi}_1^0 \cdot f_{\mu_1^0}(x_i)$$

$$\begin{cases} (x, y) \sim \mathcal{G}_\theta \\ \text{Si on a: } \mathcal{L}(Y) \rightarrow \mathcal{L}(Y|X) \\ \mathcal{L}(X|Y) \end{cases}$$

proviennent de $\hat{\theta}_0$.

$$\text{Alors, } P(Y=1 | X=x) = \frac{\pi_1 f_{\mu_1}(x)}{\pi_1 f_{\mu_1}(x) + (1-\pi_1) f_{\mu_2}(x)}$$

5

$\underset{\theta \in \Theta}{\operatorname{argmax}}$

$$\mathbb{E}[\ell_{(X, Z)_1}(\theta) | X_1^n] \\ = \sum_{i=1}^n p_i + (1-p_i).$$

Principe de l'algorithme EM:

- $\hat{\theta}_0$ estimateur initial à $t=0$.

- Itérations à t :

• Expectation: $Z_{1,1}^t, \dots, Z_{n,n}^t$

$$\left\{ \begin{array}{l} \text{tg } ((X_1, Z_1^t | \hat{\theta}_t), \dots, (X_n, Z_n^t | \hat{\theta}_t) \text{ sont iid (comme les } X_i) \\ Z_i^t | \hat{\theta}_t, X_i \end{array} \right.$$

$$F(\theta | \hat{\theta}_t) = \mathbb{E}[\ell_{(X, Z)_1}(\theta) | X_1^n]$$

• Maximization: $\hat{\theta}_{t+1} \in \underset{\theta \in \Theta}{\operatorname{argmax}} F(\theta | \hat{\theta}_t)$.

Recap: Obs $(X_i, Y_i) \rightarrow$ on ne connaît pas Y_i .

\rightarrow log vraisemblance loi jointe

$$\rightarrow \hat{\pi}_{11}, \hat{\pi}_1, \hat{\pi}_2.$$

$$\rightarrow Z_i, \mathbb{E}[\log \text{vrais}] = \sum_{Z_i | X_i} \mathbb{E}[Z_i | \hat{\theta}_t, X_i]$$

— Fin de l'exemple —

Example 1.4.1. $\mathbb{E}_{\text{mix}}(x') = x^T x'$ and $\mathbb{E}_{\text{mix}}(x^T x) = \mathbb{E}_{\text{mix}}(x)^T \mathbb{E}_{\text{mix}}(x)$

and feature space are not unique.
 $x \in \mathbb{R}^d$ and $\phi_2 : x \mapsto (\frac{x_1}{\sqrt{2}}, \frac{x_2}{\sqrt{2}})$
 $x \in \mathbb{R}^d$ and $G_2 = \mathbb{R}^{2d}$.

6

X à valeurs dans \mathbb{R}^d .

$X \sim \sum_{k=1}^K \pi_k \mathcal{N}(\mu_k, \Sigma_k)$.

$$\Theta = \{\pi_1, (\mu_1, \Sigma_1), \dots, (\mu_K, \Sigma_K)\}$$

$F(\theta | \hat{\theta}_t)$ soft K-means.

Maximum a posteriori:

$$\forall x \in \mathbb{R}^d, P(Y=j | X=x) = \frac{\pi_j \cdot f_j}{\sum_{k=1}^K \pi_k f_k}$$

Algorithme EM pour les mélanges gaussiens

(X, Y) à valeurs dans $\mathbb{R}^d \times \{1; K\}$.

$$P(Y=j) = \pi_j > 0 \quad (\sum_j \pi_j = 1) \text{ (de proba)}$$

et $X | Y=j \sim \mathcal{N}(\mu_j, \Sigma_j)$

(ceci définit la loi de (X, Y)).

On pose $\mathcal{P}_g = \{G_\theta = \mathcal{L}(X, Y), \theta \in \Theta\}$

$$\Theta = \{\pi_1, (\mu_1, \Sigma_1), \dots, (\mu_K, \Sigma_K)\}$$

où $\pi_i \in [0, 1]$, et $\sum_i \pi_i = 1$ et Σ_k symétrique définie pos.

$\forall k \in \{1; K\}$.

On considère les obs. x_1, \dots, x_n iid $\sim \mathcal{L}(X)$

$$\mathcal{P}_m = \{M_\theta = \mathcal{L}(X), (X, Y) \sim G_\theta\} \text{ (2^e modèle statistique)}$$

$$\mathcal{P}_c = \{Q_\theta = \mathcal{L}(Y | X), (X, Y) \sim G_\theta\}$$

Cas général:

• X a pour densité:
 $\mathcal{N}(\mu, \Sigma)$ $P_{\mu, \Sigma}(x) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det(\Sigma)}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)}$

• G_θ a pour densité:

$$P(X, Y) = P(X | Y) \cdot P(Y) \quad g_\theta(x, y) = \pi_y \cdot P_{\mu_y, \Sigma_y}(x)$$

(7)

$$\text{Mo a pour densité: } m_{\theta}(x) = \sum_{j=1}^k g_{\theta}(x, j) \quad \begin{matrix} \uparrow \\ \text{IP}(Y|X) \end{matrix}$$

$$\text{Q}_{\theta, x} a pour densité: \quad g_{\theta, x}(y) = \frac{g_{\theta}(x, y)}{m_{\theta}(x)} \quad \begin{matrix} \uparrow \\ \text{IP}(Y|X) \end{matrix} \quad \begin{matrix} \leftarrow \\ \text{IP}(X) \end{matrix}$$

(1) Expectation Step:

Iteration t: $\hat{\theta}_t$.

z_1^t, \dots, z_n^t (variables latentes qui vont estimer les Y labels)

- $\{(x_i, z_i^t) | \theta_t, \dots, (x_n, z_n^t) | \theta_t\}$ iid

- $z_i^t | x_i \approx z_i^t | \hat{\theta}_t, x_i \sim Q_{\hat{\theta}_t, x_i}$

et $P(z_i^t = j | x_i) = \hat{g}_{\hat{\theta}_t, x_i}(j) = \frac{\hat{g}_{\hat{\theta}_t}(x_i, j)}{m_{\hat{\theta}_t}(x_i)} := p_{ij}^t$.
 = $P(Y|X)$ $\text{pu} \star$

On peut ensuite définir notre fonction "posti":

$$\forall \theta: F(\theta, \hat{\theta}_t) = \mathbb{E}_{\theta} \left[\underbrace{\sum_{i=1}^n \log g_{\theta}(x_i, z_i^t)}_{=\log \text{vraisemblance}} \mid x_i^n \right]$$

→ on veut ensuite maximiser l'espérance de la log. vraisemblance
 et on va trouver un $\hat{\theta}_{t+1}$ qui le fait.

On poursuit le calcul et:

$$F(\theta, \hat{\theta}_t) = \sum_{i=1}^n \mathbb{E} [\log(g_{\theta}(x_i, z_i^t)) \mid x_i^n] \\ = \sum_{i=1}^n \sum_{j=1}^k p_{ij}^t \times \log(g_{\theta}(x_i, j))$$

par \star : si $\log(g_{\theta}(x_i, j)) = \log(\pi_j) - \frac{1}{2}(x_i - \mu_j)^T \Sigma^{-1} (x_i - \mu_j) - \frac{1}{2} \log(\det(\Sigma_j))$

$$P(Y|X) = \frac{P_{x,y}}{\sum_{j=1}^m P_{x,y_j}} \quad \leftarrow P(X \cap Y)$$

Maximization Step:

On obtient: $\hat{\theta}_{t+1} \in \text{argmax}_{\theta} F(\theta | \hat{\theta}_t)$

$$\text{Let } \theta \in \Theta, \text{ then } \hat{\beta}_{t+L} = \left(\hat{\pi}_{t+L}^1, \hat{\gamma}_1^{t+L}, \dots, \hat{\gamma}_k^{t+L}, \hat{\mu}_1^{t+L}, \dots, \hat{\mu}_k^{t+L} \right).$$

Méthodologie
par rapport à μ
et Σ

$$\text{avec : } \tilde{f}_{ij}^{t+1} = \frac{1}{m} \sum_{i=1}^m p_i^t g_i$$

$$\hat{\mu}_j^{t+1} = \frac{\sum_{i=1}^m p_{ij}^t X_{ii}}{\sum_{i=1}^m p_{ij}^t}$$

$$\sum_{ij}^{t+1} = \sum_{i=1}^m p_{ij}^{(t)} \left[(x_i - \hat{\mu}_j^{(t)}) (x_i - \hat{\mu}_j^{(t)})^\top \right]$$

Maximisation
en rapport à T;

$$\text{Maximise}_{T \in [0,1]^K} \sum_{j=1}^k \left(\frac{m}{L-1} \frac{x_j}{p_j} \right) \log(T_j)$$

$$\mathcal{P}^T \mathcal{P} = I$$

la quantité

(=)

Mapimide

$$\left\{ \sum_j \frac{1}{n} \left(\sum_i p_{ij}^{-k} \right) \log \left(\frac{1}{n} \sum_i p_{ij}^{-k} \right) - \sum_j \frac{1}{n} \left(\sum_i p_{ij}^{-k} \right) \log \left(\pi_j^{-k} \right) \right\}$$

$$= \sum_j \underbrace{\left(\frac{1}{n} \sum_i p_{ij}^{-k} \right)}_{\square_j} \log \left(\frac{\square_j}{\pi_j^{-k}} \right)$$

9

EM: $\mathcal{U}(-) \rightarrow \text{Soft K-means}$

Itérer :

$$\begin{pmatrix} p_{\theta, j}^t \\ f_j \\ z_j \end{pmatrix} \simeq \mathbb{P}(Y_i=j | X_i)$$

Général cas : Dans le cas général de l'EM: (plus la gaussianité)

$$(X_i, Y) \sim G_\theta = g_\theta \cdot (\mu \times \sigma)$$

↑ loi dominante

$$X = (X_1, \dots, X_n)$$

$$Y = (Y_1, \dots, Y_m)$$

$$\text{Alors, on a } X \sim M_\theta = M_\theta \times \mu.$$

$$Y|X \sim Q_{\theta, X} = q_{\theta, X} \times \sigma$$

$\hat{\Theta}_t$ estimateur de Θ à l'échellement t .

$$\begin{aligned} F(\theta | \hat{\Theta}_t) &= \mathbb{E}[\log(g_\theta(x, z_x)) | X] \\ Z_t | X &\sim Q_{\hat{\Theta}_t, X} \end{aligned}$$

Lemme :

Soit $Q_x = q_x \cdot \sigma$ et $Z|X=x \sim Q_x$.

On a que, $\forall \theta \in \Theta$,

$$\begin{aligned} \log(M_\theta(x)) &= \mathbb{E}[\log(g_\theta(x, z)) | X] + \\ &\quad \text{loi jointe} \\ &\quad \text{loi de } X \\ &\quad + D_{KL}(Q_x \| Q_{\theta, X}) + H(Q_x) \end{aligned}$$

où la divergence KL est : $D_{KL}(Q_x \| Q_{\theta, X}) = \mathbb{E}[\log\left(\frac{q_x(z)}{q_{\theta, X}(z)}\right) | X]$

et l'entropie $H(Q_x) = -\mathbb{E}[\log q_x(z) | X]$

10

$$\text{En particulier: } Q_x = Q_{\theta' | x}$$

$$\log(m_\theta(x)) = F(\theta | \theta') + D_{KL}(\theta' || \theta) + H(\theta').$$

Preuve:

Par la formule de Bayes:

$$g_\theta(x, z) = q_{\theta, x}(z) \cdot m_\theta(x). \quad (= \frac{g_\theta}{q_{\theta, x}})$$

$$\begin{aligned} \text{alors: } \log(m_\theta(x)) &= \mathbb{E}[\log m_\theta(x) | x] \\ &= \mathbb{E}[\underbrace{\log(g_\theta(x, z))}_{\text{Coup du } \pm 1} | x] \\ &\quad - \mathbb{E}[\log(q_{\theta, x}(z)) | x] \\ &= \mathbb{E}[\cancel{A} | x] + \mathbb{E}[\log(\frac{q_x(z)}{q_{\theta, x}(z)}) | x] - \mathbb{E}[\log(q_x(z)) | x] \\ &\quad := D_{KL}(Q_x || Q_{\theta, x}) \end{aligned}$$

Proposition:

$P = p \mu.$ et $Q = q \mu.$ (la loi abs continue pour densité)

$$\text{Alors, } 0 \leq D_{KL}(P || Q) \in \mathbb{R}.$$

De plus, si $Q \ll P,$

$$D_{KL}(P || Q) = 0 \Leftrightarrow P = q, P \text{-presque partout.}$$

Preuve:

$$D_{KL}(P || Q) = \mathbb{E}\left[\frac{\log(P(z))}{q(z)}\right] \text{ avec } z \sim P.$$

(14)

Suite parrale, si $P \ll Q$ alors $D_{KL}(P||Q) = +\infty$.

Si $P \leq Q$:

$$\text{alors } D_{KL}(P||Q) = \mathbb{E} \left[-\log \left(\frac{q(z)}{p(z)} \right) \right] \quad p(z) > 0 \text{ p.s.} \\ \geq -\log \left(\mathbb{E} \left[\frac{q(z)}{p(z)} \right] \right) \quad (\text{car } z \sim P) \\ \text{Tensur.} \\ = 0.$$

Converse
 $(-\log = -\text{concave} = \text{convexe})$

Par suite, on a bien: $0 \leq D_{KL}(P||Q) \leq +\infty$ ($D_{KL}(P||Q) \in \mathbb{R}$).

Par stricte convexité de $-\log$, alors, on a:

$$D_{KL}(P||Q) \Leftrightarrow \frac{q(z)}{p(z)} = \text{cte p.s.}$$

$$\text{i.e. } \frac{q}{p} = \text{cte} \begin{cases} P - \text{p.s.} \\ Q - \text{p.s.} \end{cases} \quad (\text{car } Q \ll P).$$

$$\text{Donc } \int \frac{q}{p} = \text{cte} \int \frac{p}{p} \text{ et cte} = \underline{\underline{1}} \\ \text{i.e. } \boxed{P = Q} \quad P - \text{p.s.}$$

□

Proposition:

$$\forall \theta, \theta' \quad \log(m_\theta(x)) \geq F(\theta|\theta') + H(\theta')$$

$$(\text{et } \log(m_\theta(x)) = F(\theta|\theta) + H(\theta) \text{ pour } \theta = \theta')$$

$$\text{car } D_{KL}(\theta||\theta) = 0 \\ = \mathbb{E} [\log(1)] \\ = 0$$

$$\log(m_\theta(x)).$$

$$\theta \mapsto \log(m_\theta(x))$$

le bon θ'

$$\theta \mapsto F(\theta|\theta_x) + H(\theta_x)$$

$$\theta \mapsto F(\theta|\theta'_x) + H(\theta'_x)$$

[par θ'_1, θ'_2]
à droite.

$$\theta \mapsto F(\theta, \theta'_1) + H(\theta'_1)$$

~~Exercice~~

12

E-step: $\hat{\theta}'_t \in \arg\max_{\theta} F(\hat{\theta}_t | \theta) + H(\theta)$.

$\rightarrow \hat{\theta}'_t = \hat{\theta}_t$ \nwarrow celui qui maximise (en rouge sur le dessin).

M-step: $\hat{\theta}_{t+1} \in \arg\max_{\theta} F(\theta | \hat{\theta}_t) + H(\theta)$.

Théorème: (Croissance selon $(\hat{\theta}_t)_t$).

$\hat{\theta}_{t+1} \in \arg\max_{\theta} F(\theta | \hat{\theta}_t)$, alors la suite

$\log(m_{\hat{\theta}_t}(x))$ est ↑.

Prouve: $\log(m_{\hat{\theta}_t}(x)) = F(\hat{\theta}_t | \hat{\theta}_t) + H(\hat{\theta}_t)$

(par propriété d'avant)

car $\hat{\theta}_{t+1}$
vient maximiser
 $F(\theta | \hat{\theta}_t) + H(\theta)$

puis

$\hat{\theta}_{t+1}$ minimisant

par rapport au M-step

(proposition 5.2 page 68).

$\log(m_{\hat{\theta}_t}(x)) > F(\hat{\theta}'_t | \hat{\theta}_t) + H(\hat{\theta}'_t)$

$\leq \log(m_{\hat{\theta}_{t+1}}(x))$.