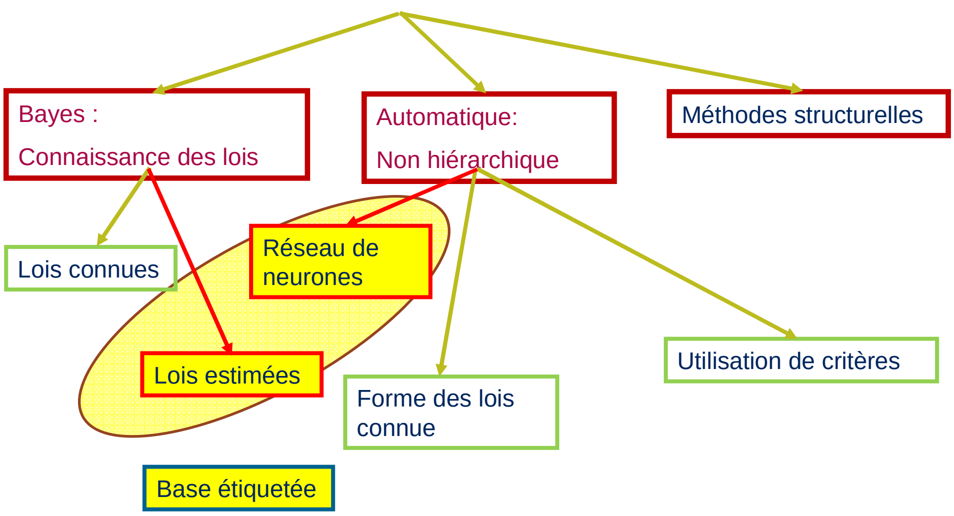
Classification automatique
Méthodes non hiérarchiques
SI 221

Jean Marie NICOLAS

Méthodes de classification



Classification bayésienne

- **N échantillons indépendants : X_i, $i \in [1,N]$**
  - Vecteurs d'état de dimension d : X_{ik}, $k \in [1,d]$
- **Classification en *c* classes**
  - Le nombre de classes *c* est connu
  - ~~On connaît les lois pour chaque classe~~
- **Pour chaque échantillon, on connaît la sortie désirée : d_i, $i \in [1,N]$**
  - Bases d'apprentissage et de test

ON ESTIME LES LOIS !!

TELECOM ParisTech

Comment estimer une loi ?

- **Méthodes paramétriques : Suppose aussi que l'on connaisse quelle est la loi suivie par les observations !!**
  - Méthode des moments
  - Méthode du maximum de vraisemblance
- **Méthodes non paramétriques :**
  - Fenêtres de Parzen
  - kn plus proches voisins

TELECOM ParisTech

La méthode des moments

- Connaître tous les moments revient à connaître la ddp (densité de probabilité)
- Dans le cas d'une loi normale (gaussienne) il suffit de connaître les deux premiers moments

$$p(x|\omega_i) = p(x|\omega_i; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)}$$

$$M = E[x]$$
$$\Sigma = E\left[(x-M)(x-M)^t\right]$$

TELECOM ParisTech

Estimation des moments

- Moments « empiriques »

$$\tilde{m}_1 = \frac{1}{N}\sum_{k=1}^{N} x_k$$
$$\tilde{m}_p = \frac{1}{N}\sum_{k=1}^{N} x_k^p$$

- Moments centrés « empiriques »

$$\tilde{M}_p = \frac{1}{N}\sum_{k=1}^{N} (x_k - \tilde{m}_1)^p$$

TELECOM ParisTech

Gaussienne 1-D

$$p(x|\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- Moment d'ordre 1 et moment centré d'ordre 2

$$E[x]=\int xp(x|\mu,\sigma)dx=\mu$$

$$E\left[(x-\mu)^2\right]=\int (x-\mu)^2\, p(x|\mu,\sigma)dx=\sigma$$

$$\tilde{m}_1=\frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\tilde{M}_2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\tilde{m}_1)^2$$

- Approche « réaliste » si N assez grand

Estimateur du maximum de vraisemblance

- A partir de N échantillons $x_k, k\in[1,N]$, on cherche les paramètres de la loi qui maximisent la vraisemblance d'avoir tiré ces N échantillons

$$\mathrm{V} = p(x_1, x_2, \cdots, x_N; \theta)$$

- On suppose que ces N échantillons $x_k, k\in[1,N]$, sont indépendants

$$\mathrm{V} = p(x_1, x_2, \cdots, x_N; \theta)=\prod_{k=1}^{N} p(x_k;\theta)$$

Vraisemblance et log vraisemblance

$$\begin{aligned}\log \mathrm{V} &= \log\left(p\left(x_1, x_2, \cdots, x_N; \theta\right)\right) \\ &= \log\left(\prod_{k=1}^{N} p\left(x_k; \theta\right)\right) \\ &= \sum_{k=1}^{N} \log\left(p\left(x_k; \theta\right)\right)\end{aligned}$$

TELECOM ParisTech

Cas de la loi normale (1 -D)

$$p\left(x \middle| \mu, \sigma\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

$$\log \mathrm{V} = \sum_{k=1}^{N} \log\left(p\left(x_k; \mu, \sigma\right)\right)$$

- **On recherche μ et σ maximisant la log-vraisemblance**
  - Dérivation selon μ
  - Dérivation selon σ

TELECOM ParisTech

Dérivation de la log-vraisemblance selon μ

$$\frac{\partial}{\partial \mu}\log \mathrm{V} = \frac{\partial}{\partial \mu}\left(\sum_{k=1}^{N}\log\left(p\left(x_k;\mu,\sigma\right)\right)\right)$$
$$= \frac{\partial}{\partial \mu}\left(\sum_{k=1}^{N}\log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right)\right)$$
$$= \sum_{k=1}^{N} -\frac{x_k-\mu}{\sigma}$$

- On obtient alors μ en annulant la dérivée

$$\boxed{\mu = \frac{1}{N}\sum_{k=1}^{N} x_k}$$

Dérivation de la log-vraisemblance selon σ

$$\frac{\partial}{\partial \mu}\log \mathrm{V} = \frac{\partial}{\partial \sigma}\left(\sum_{k=1}^{N}\log\left(p\left(x_k;\mu,\sigma\right)\right)\right)$$
$$= \frac{\partial}{\partial \sigma}\left(\sum_{k=1}^{N}\log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}\right)\right)$$
$$= \sum_{k=1}^{N}\left(-\frac{1}{\sigma}+\frac{(x_k-\mu)^2}{\sigma^3}\right)$$

- On obtient alors σ en annulant la dérivée

$$\boxed{\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k-\mu)^2}$$

Loi gaussienne : récapitulatif

- **Méthode des moments**

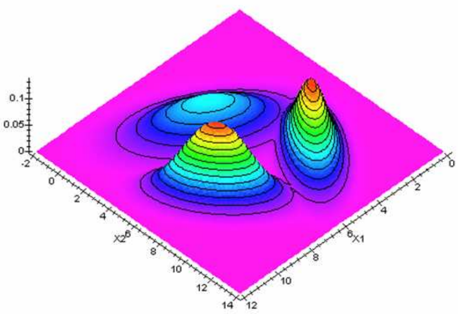
$$\boxed{\begin{aligned}\mu &= E[x]\\ \sigma &= E\left[(x-\mu)^2\right]\end{aligned}}$$

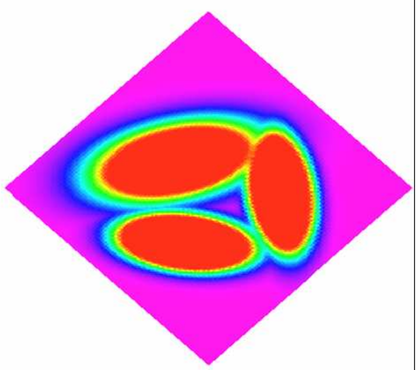
$$\boxed{\begin{aligned}\tilde{m}_1 &= \frac{1}{N}\sum_{k=1}^{N} x_k\\ \tilde{M}_2 &= \frac{1}{N}\sum_{k=1}^{N}\left(x_k-\tilde{m}_1\right)^2\end{aligned}}$$

- **Méthode du Maximum de Vraisemblance**

$$\boxed{\mu = \frac{1}{N}\sum_{k=1}^{N} x_k} \qquad \boxed{\sigma^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k-\mu)^2}$$

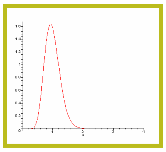
Décision bayésienne : comme avant…

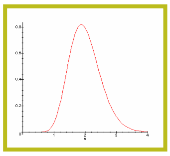




Loi Gamma : méthode des moments

$$p(x|\mu, L) = \frac{1}{\Gamma(L)} \frac{L}{\mu} \left(\frac{Lx}{\mu}\right)^{L-1} e^{-\frac{Lx}{\mu}}$$





- Les deux premiers moments

$$E[x] = \int x p(x|\mu,\sigma) dx = \mu$$

$$E\left[(x-\mu)^2\right] = \int (x-\mu)^2 p(x|\mu,\sigma) dx = \frac{\mu^2}{L}$$

$$\mu = m_1 = E[x]$$

$$L = \frac{m_1^2}{m_2 - m_1^2} = \frac{(E[x])^2}{E\left[(x-\mu)^2\right]}$$

$$\tilde{m}_1 = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\tilde{M}_2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \tilde{m}_1)^2$$

Dérivation de la log-vraisemblance selon μ

$$\frac{\partial}{\partial \mu} \log V = \frac{\partial}{\partial \mu} \left(\sum_{k=1}^{N} \log(p(x_k;\mu, L)) \right)$$

$$= \frac{\partial}{\partial \mu} \left(\sum_{k=1}^{N} \log\left(\frac{1}{\Gamma(L)} \frac{L}{\mu} \left(\frac{Lx}{\mu}\right)^{L-1} e^{-\frac{Lx}{\mu}} \right) \right)$$

$$= \sum_{k=1}^{N} - \frac{L(x_k - \mu)}{\mu^2}$$

- On obtient alors μ en annulant la dérivée

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

Dérivation de la log-vraisemblance selon L

$$\frac{\partial}{\partial L}\log V = \frac{\partial}{\partial L}\left(\sum_{k=1}^{N}\log\left(p(x_k;\mu,L)\right)\right)$$

$$= \frac{\partial}{\partial L}\left(\sum_{k=1}^{N}\log\left(\frac{1}{\Gamma(L)}\frac{L}{\mu}\left(\frac{Lx}{\mu}\right)^{L-1} e^{-\frac{Lx}{\mu}}\right)\right)$$

$$= \sum_{k=1}^{N}\left(\frac{\mu - x}{\mu} + \log\frac{x}{\mu} + \log L - \Psi(L)\right)$$

- On obtient alors L en annulant la dérivée, mais l'expression est implicite. D'où, connaissant $\hat{\mu}$

$$\boxed{\log L - \Psi(L) = \log\hat{\mu} - \frac{1}{N}\sum_{k=1}^{N}\log(x_k)}$$

TELECOM ParisTech

Loi Gamma: récapitulatif

- Méthode des moments

$$\boxed{\begin{aligned}\mu &= m_1 = E[x]\\ L &= \frac{m_1^2}{m_2 - m_1^2} = \frac{(E[x])^2}{E\left[(x-\mu)^2\right]}\end{aligned}}$$

$$\boxed{\begin{aligned}\tilde{m}_1 &= \frac{1}{N}\sum_{k=1}^{N}x_k\\ \tilde{M}_2 &= \frac{1}{N}\sum_{k=1}^{N}(x_k - \tilde{m}_1)^2\end{aligned}}$$

- Méthode du Maximum de Vraisemblance

$$\boxed{\mu = \frac{1}{N}\sum_{k=1}^{N}x_k}$$

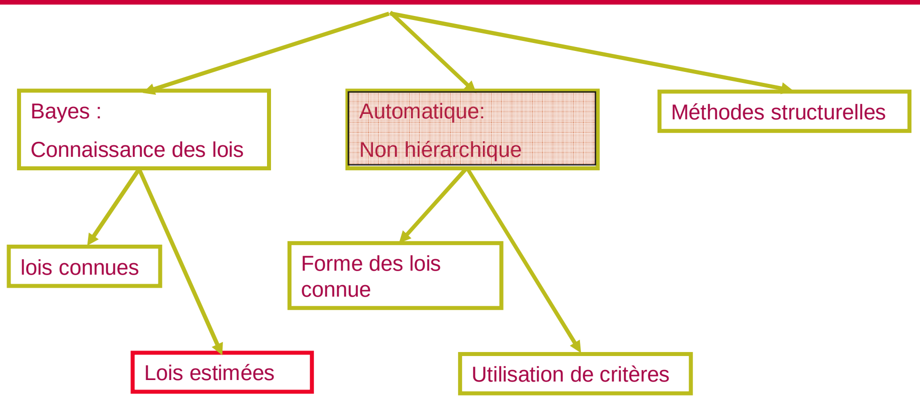
$$\boxed{\log L - \Psi(L) = \log\hat{\mu} - \frac{1}{N}\sum_{k=1}^{N}\log(x_k)}$$

TELECOM ParisTech

Un premier bilan des méthodes

- **Le maximum de vraisemblance est la méthode qui donne en général la plus petite variance des estimateurs (bornes de Cramer-Rao)**
  - Parfois expressions implicites
- **La méthode des moments peut toujours être mise en œuvre, mais :**
  - La variance des estimateurs peut être grande
  - Certaines lois (« à queue lourde ») ne peuvent être estimées par cette méthode.
- **N doit être « assez » grand**

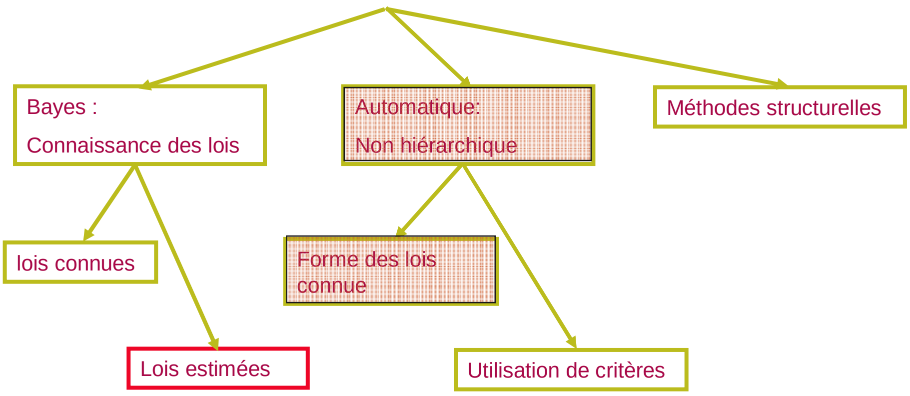
Méthodes de classification



Méthodes non hiérarchique

- **On recherche une description des classes par leur densité de probabilité :**
  - On connaît la forme
  - On ne connaît pas les **paramètres**
- **On recherche simplement un partitionnement de l'espace :**
  - Regroupement des individus
  - **Critère** de regroupement

Méthodes de classification



De Bayes aux critères

- **Formulation de type Bayes**
  - Cf classification supervisée
- **Simplification de la règle de Bayes**
  - ISODATA de base
- **Notion de critère de classification**
  - Algorithme des k-moyennes

Les données du problème

- **N échantillons indépendants : $x_k, k \in [1,N]$, Vecteurs d'état de dimension d**
- **Classification en *c* classes**
  - Le nombre de classes *c* est connu

Structures probabilistes connues

- **Pour chaque classe ω_i, on connaît la probabilité a priori** $$P(\omega_i),\ i \in [1, c]$$
- **Pour chaque classe ω_i, on connaît la forme de la densité de probabilité :**
  - Description par un vecteur de paramètre θ_i

$$p(x | \omega_i; \theta_i)$$

- **On recherche le vecteur θ = (θ_1,…, θ_q)**

Mélange de lois

- **On suppose que la densité de probabilité globale vérifie une loi de mélange linéaire**

$$p(x;\theta) = \sum_{i=1}^{c} P(\omega_i)\, p(x | \omega_i; \theta_i)$$

- **Décomposition « identifiable » si et seulement si**

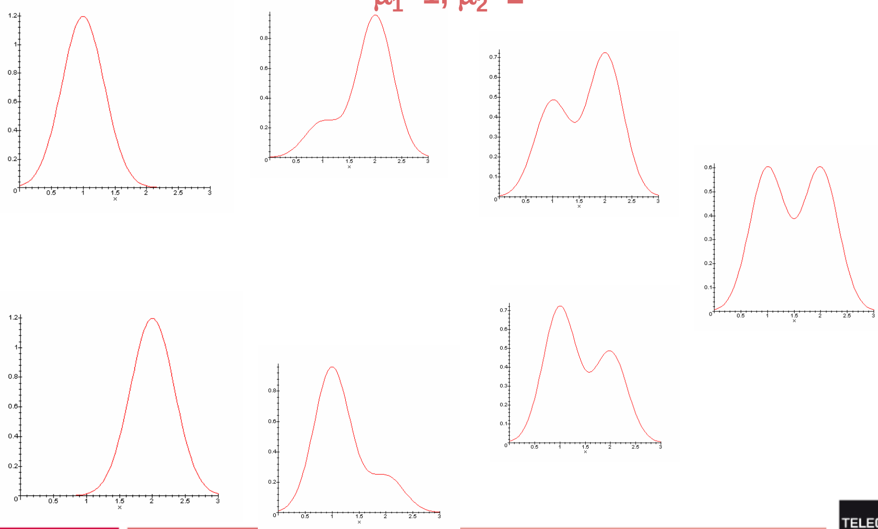
$$\theta \neq \theta' \Rightarrow \exists x\ tq\ \ p(x;\theta) \neq p(x;\theta')$$

Exemple sur la loi normale

- **On choisit la loi normale. Pour chaque classe, on a deux paramètres :**
  - µi : moyenne
  - Σ : matrice de variance-covariance

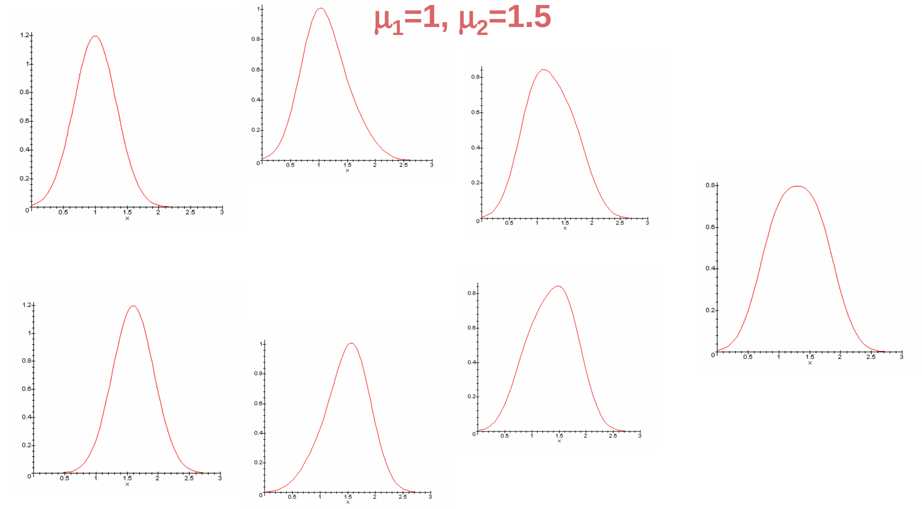
$$p(x|\omega_i;\theta_i) = p(x|\omega_i;\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)^t}$$

Cas 1-D : Mélange de 2 gaussiennes $\mu_1=1$, $\mu_2=2$



Cas 1-D : Mélange de 2 gaussiennes

$\mu_1=1$, $\mu_2=1.5$



Estimateur du maximum de vraisemblance

- A partir de n échantillons $x_k, k \in [1,N]$, on cherche le vecteur θ qui maximise la vraisemblance

$$V = p(x_1, x_2, \cdots, x_N; \theta)$$

- On suppose que ces n échantillons $x_i, i \in [1,n]$, sont indépendants

$$V = p(x_1, x_2, \cdots, x_N; \theta) = \prod_{k=1}^{N} p(x_k; \theta)$$

On calcule la log vraisemblance

$$\log \mathrm{V} = \log\left(p\left(x_1, x_2, \cdots, x_N; \theta\right)\right)$$
$$= \log\left(\prod_{k=1}^{N} p\left(x_k; \theta\right)\right)$$
$$= \sum_{k=1}^{N} \log\left(p\left(x_k; \theta\right)\right)$$

TELECOM ParisTech

Maximum de la vraisemblance

- V est maximum pour $\underline{\theta}$
- Dérivation par rapport à θ_i

$$\frac{\partial}{\partial \theta_i} \log \mathrm{V}\Bigg|_{\theta=\underline{\theta}} = \frac{\partial}{\partial \theta_i}\left(\sum_{k=1}^{N} \log\left(p\left(x_k; \theta\right)\right)\right)\Bigg|_{\theta=\underline{\theta}} = 0 \quad \forall \theta_i$$

TELECOM ParisTech

On dérive donc la log-vraisemblance

$$\frac{\partial}{\partial\theta_i}\log V = \frac{\partial}{\partial\theta_i}\left(\sum_{k=1}^{N}\log(p(x_k;\theta))\right)$$

$$= \sum_{k=1}^{N}\frac{1}{p(x_k;\theta)}\frac{\partial}{\partial\theta_i}p(x_k;\theta)$$

$$= \sum_{k=1}^{N}\frac{1}{p(x_k;\theta)}\frac{\partial}{\partial\theta_i}\left[\sum_{j=1}^{c}P(\omega_j)p(x_k|\omega_j;\theta_j)\right]$$

$$= \sum_{k=1}^{N}\frac{1}{p(x_k;\theta)}\frac{\partial}{\partial\theta_i}[P(\omega_i)p(x_k|\omega_i;\theta_i)]$$

TELECOM ParisTech

Suite…

- On applique la règle de Bayes $P(\omega_i|x_k;\theta) = \frac{P(\omega_i)p(x_k|\omega_i;\theta_i)}{p(x_k;\theta)}$

$$\frac{\partial}{\partial\theta_i}\log V = \sum_{k=1}^{N}\frac{1}{p(x_k;\theta)}\frac{\partial}{\partial\theta_i}[P(\omega_i)p(x_k|\omega_i;\theta_i)]$$

$$= \sum_{k=1}^{N}\frac{P(\omega_i|x_k;\theta)}{p(x_k|\omega_i;\theta_i)}\frac{\partial}{\partial\theta_i}[p(x_k|\omega_i;\theta_i)]$$

$$= \sum_{k=1}^{N}P(\omega_i|x_k;\theta)\frac{\partial}{\partial\theta_i}[\log(p(x_k|\omega_i;\theta_i))]$$

TELECOM ParisTech

Maximum de vraisemblance

- Pour les paramètres $\underline{\theta}_i$ vérifiant

$$\boxed{\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\theta}) \frac{\partial}{\partial \theta_i} \left[\log\left(p\left(x_k | \omega_i ; \underline{\theta}_i \right) \right) \right] = 0}$$

- Equation implicite

Cas de la loi normale

$$\boxed{p(x|\omega_i;\theta_i) = p(x|\omega_i;\mu_i,\Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)^t}}$$

On suppose les Σ_i connus $\frac{\partial}{\partial \mu_i} \left[\log\left(p(x|\omega_i;\mu_i) \right) \right] = \Sigma_i^{-1}(x-\mu_i)$

$$\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu}) \frac{\partial}{\partial \mu_i} \left[\log\left(p\left(x_k | \omega_i ; \underline{\mu}_i \right) \right) \right] =$$

$$\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu}) \Sigma_i^{-1} (x_k - \underline{\mu}_i) = 0$$

$$\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu}) \Sigma_i^{-1} \mu_i = \sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu}) \Sigma_i^{-1} x_k$$

Solution pour la loi normale

$$\underline{\mu_i} = \frac{\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu})\, x_k}{\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu})}$$

- **Très satisfaisant pour l'esprit : on a une moyenne pondérée par la probabilité a posteriori**
- **Mais équation implicite puisque**

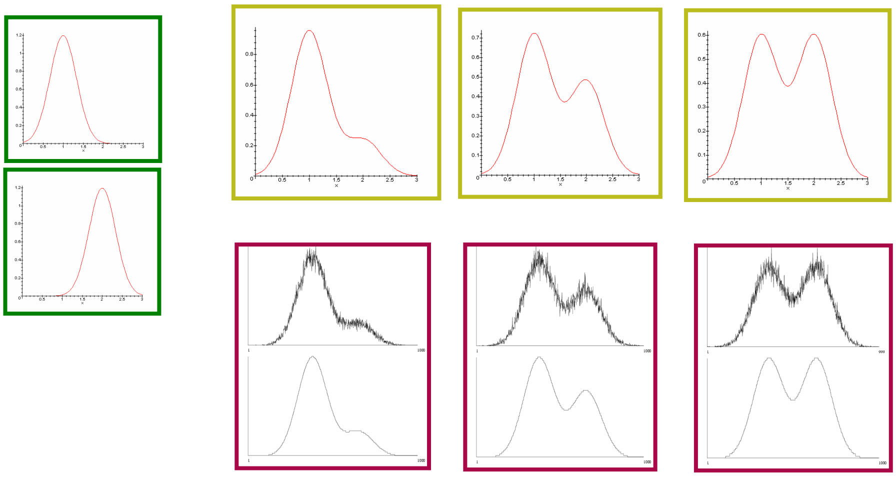
$$\boxed{P(\omega_i | x_k ; \mu) = \frac{P(\omega_i)\, p(x_k | \omega_i ; \mu_i)}{p(x_k ; \mu)}}$$

Schéma itératif

- **Etape 0 : on initialise les $\underline{\mu}_i$ par $\underline{\mu(0)}$**
  - On sait alors calculer $p(x_k | \omega_i ; \underline{\mu_i(0)})$
  - On en déduit $P(\omega_i | x_k ; \underline{\mu_i(0)})$
  - D'où

$$\underline{\mu_i(j+1)} = \frac{\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu_i(j)})\, x_k}{\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\mu(j)})}$$

Exemple : 2 gaussiennes, μ_1 =1, μ_2 =2



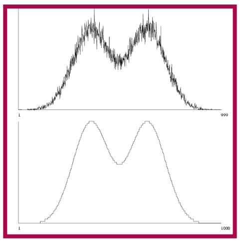
0.2 et 0.8 — 0.4 et 0.6 — 0.5 et 0.5

Exemple des deux gaussiennes (μ_1=50, μ_2=75)

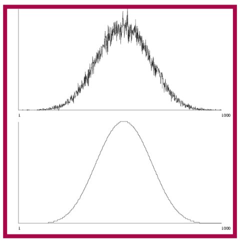
- Même variance (σ=8), $P(\omega_i)$ variable

$P(\omega_i)$	μ_1	μ_2	Nombre itérations
0.50	50,01	75,1	3
0.60	49,9	74,9	5
0.70	49,9	74,8	6
0.80	49,9	74,7	6

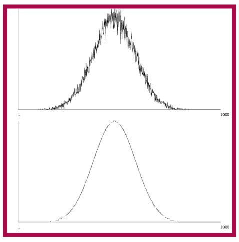
Deux gaussiennes, $P(\omega_1) = P(\omega_2) = 0.5$



$\mu1 = 50$, $\mu2 = 75$



$\mu1 = 50$, $\mu2 = 60$



$\mu1 = 50$, $\mu2 = 55$

Exemple des deux gaussiennes

- **Même variance (σ=8), $P(\omega_i)$=0,5, μ_2 varie**

μ_2	μ_1	μ_2	Nombre itérations
75	50,01	75,1	3
60	49,96	59,94	7
55	49,95	55,10	16

Cas de la loi Gamma

$$p\left(x|\omega_i;\theta_i\right)=p\left(x|\omega_i;\mu_i,L_i\right)=\frac{1}{\Gamma(L_i)}\frac{L_i}{\mu_i}\left(\frac{L_i x}{\mu_i}\right)^{L_i-1}e^{-\frac{L_i x}{\mu_i}}$$

On suppose *L* connu

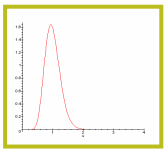
$$\frac{\partial}{\partial\mu_i}\left[\log\left(p\left(x|\omega_i;\mu_i\right)\right)\right]=\frac{L_i\left(x-\mu_i\right)}{\mu_i^2}$$

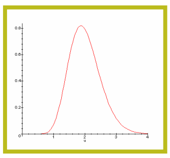
$$\sum_{k=1}^{N}P\left(\omega_i|x_k;\underline{\mu}\right)\frac{\partial}{\partial\mu_i}\left[\log\left(p\left(x_k|\omega_i;\underline{\mu_i}\right)\right)\right]=$$

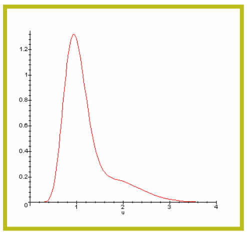
$$\sum_{k=1}^{N}P\left(\omega_i|x_k;\underline{\mu}\right)\frac{x_k-\underline{\mu_i}}{\underline{\mu_i^2}}=0$$

$$\sum_{k=1}^{N}P\left(\omega_i|x_k;\underline{\mu}\right)\frac{1}{\underline{\mu_i}}=\sum_{k=1}^{N}\frac{P\left(\omega_i|x_k;\underline{\mu}\right)}{\underline{\mu_i}^2}x_k$$

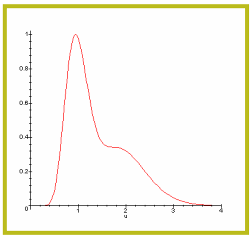
Exemple : tirages de lois Gamma (L=16, μ_1=1, μ_2=2)



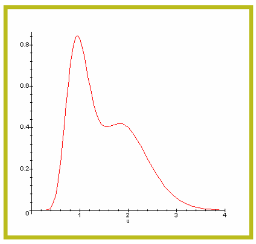




0.2 et 0.8



0.4 et 0.6



0.5 et 0.5

Exemple de deux lois Gamma (μ_1=50, μ_2=75)

- Même nombre de vues (L=16), $P(\omega_i)$ variable

$P(\omega_i)$	μ_1	μ_2	Nombre itérations
0.50	50,07	75,06	4
0.60	49,98	74,92	6
0.70	49,99	74,94	7
0.80	49,95	75,15	10

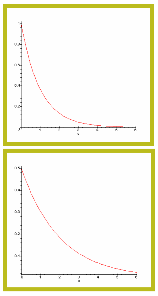
- Biais

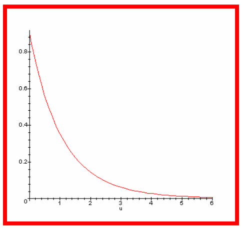
Exemple de deux lois Gamma $P(\omega_1)=P(\omega_2)=0.5$

- Même nombre de vues (L=16), $P(\omega_i)$=0,5, μ_2 varie

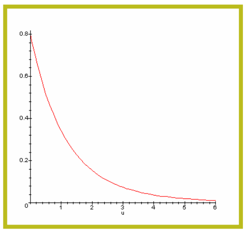
μ_2	μ_1	μ_2	Nombre itérations
75	50,07	75,06	4
60	49,92	60,02	7
55	49,88	55,00	23

Exemple : tirages de lois Gamma ($L=1$, $\mu_1=1$, $\mu_2=2$)

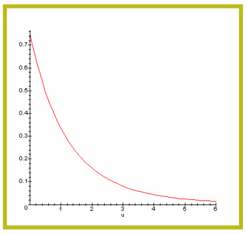




0.2 et 0.8



0.4 et 0.6



0.5 et 0.5

Exemple de deux lois Gamma $P(\omega_1)=P(\omega_2)=0.5$

- **Même nombre de vues ($L=1$), $P(\omega_i)=0,5$, μ_2 varie**

μ_2	μ_1	μ_2	Nombre itérations
75	50,02	75,19	26
60	49,59	60,20	33
55	49,39	55,58	49

Simplification du modèle

$$\sum_{k=1}^{N} P(\omega_i | x_k ; \underline{\theta}) \frac{\partial}{\partial \theta_i} \left[\log\left(p\left(x_k | \omega_i ; \underline{\theta}_i \right) \right) \right] = 0$$

- Que peut-on approximer ?

TELECOM ParisTech

Cas gaussien : une première approximation

- Le terme de « probabilité a posteriori » est inversement proportionnel à la distance de Mahalanobis

$$P(\omega_i | x_k ; \underline{\theta}) \propto \frac{1}{\left(x_k - \underline{\mu}_i \right)^t \Sigma_i^{-1} \left(x_k - \underline{\mu}_i \right)}$$

- Remplacer la distance de Mahalanobis par la distance euclidienne

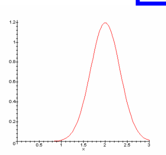
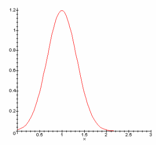
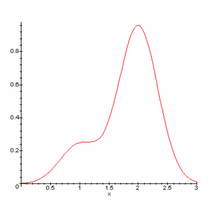
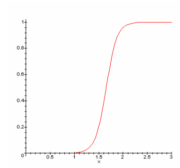
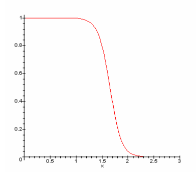
$$\left(x_k - \underline{\mu}_i \right)^t \Sigma_i^{-1} \left(x_k - \underline{\mu}_i \right) \rightarrow \left\| x_k - \underline{\mu}_i \right\|^2$$

TELECOM ParisTech

Une seconde approximation

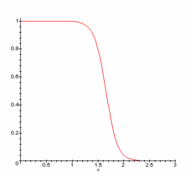
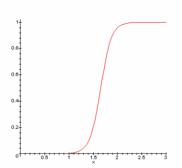
- Considérons le terme de probabilité a posteriori :

$$\boxed{P(\omega_i|x_k;\mu)=\frac{P(\omega_i)p(x_k|\omega_i;\mu_i)}{\sum_{j=1}^{c}P(\omega_j)p(x_k|\omega_j;\mu_j)}}$$

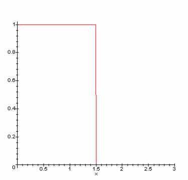
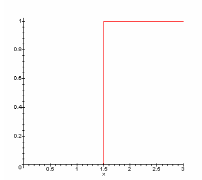






Approximation de la probabilité a posteriori

- Fonction « escalier »

- Binarisation possible : fonction d’Heaviside

Binarisation de la probabilité a posteriori

- Considérons le terme de probabilité a posteriori :

$$\boxed{P(\omega_i|x_k;\mu)=\frac{P(\omega_i)p(x_k|\omega_i;\mu_i)}{\sum_{j=1}^{c}P(\omega_j)p(x_k|\omega_j;\mu_j)}}$$

- Egal à 1 si et seulement si le point appartient à une classe unique donnée, égal à 0 sinon:

$$\boxed{P(\omega_i|x_k;\mu)=\begin{cases}1 & \text{si} \quad i=m\\ 0 & \text{sinon}\end{cases}}$$

Binarisation de la probabilité a posteriori

- On remplace donc le terme de probabilité a posteriori par une fonction d'appartenance de type ensembliste :

$$\boxed{P(\omega_i|x_k;\mu)=\begin{cases}1 & \text{si} \quad i=m\\ 0 & \text{sinon}\end{cases}}$$

- Cas de l'exemple des gaussiennes

$$\underline{\mu_i(j+1)}=\frac{\sum_{k=1}^{N_i}x_{i,k}}{N_i} \quad \text{avec} \quad x_{i,k}\in\omega_i$$

Algorithme ISODATA « de base »

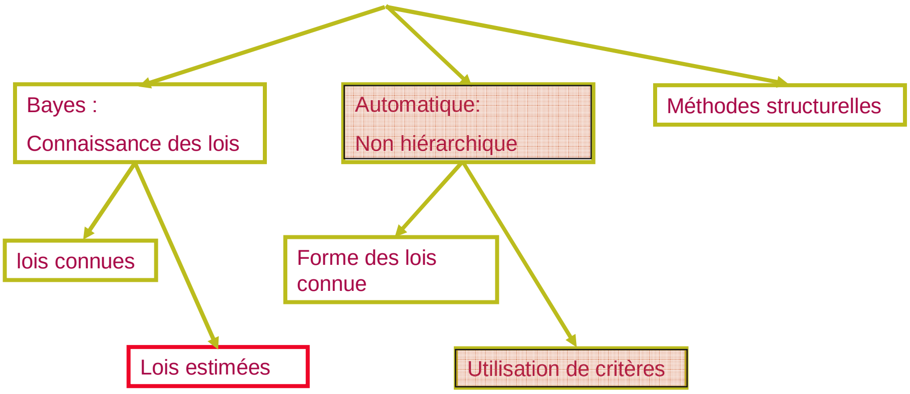
- **Choix du** critère **d'appartenance d'un échantillon à une classe**
- **ISODATA « de base » :**
  - Initialisation des moyennes
  - Recherche de la moyenne la plus proche
  - Classer les échantillons
  - Recalculer les moyennes

Problèmes divers

- **Les variances σ_i sont inconnues**
- **Les probabilités a priori $P(\omega_i)$ sont inconnues**

- **Solutions singulières**
- **Calculs lourds, simplifiables si la matrice de covariance est diagonale**

Méthodes de classification



Notion de critère : moyenne

- n_i échantillons dans la classe i
- On peut calculer la moyenne m_i

$$m_i = \frac{1}{n_i} \sum_{x \in \omega_i} x$$

- On peut calculer la moyenne générale m

$$m = \frac{1}{n} \sum_{j=1}^{n} x_j = \frac{1}{n} \sum_{i=1}^{c} \left(\sum_{x \in \omega_i} x \right) = \frac{1}{n} \sum_{i=1}^{c} n_i m_i$$

- Tout se passe comme si on avait c objets de masse n_i en m_i

Notion de critère : dispersion

- Pour une classe *i* donnée, la dispersion est donnée par :

$$S_i = \frac{1}{2n_i} \sum_{x \in \omega_i} \sum_{y \in \omega_i} \|x - y\|^2$$

- Si tous les individus d'une même classe *i* sont identiques :

$$S_i = 0$$

Théorème de Huyghens (cas 1-D)

- Introduire la moyenne dans la dispersion :

$$s_i = \frac{1}{2n_i} \sum_{x \in \omega_i} \sum_{y \in \omega_i} (x - y)^2 = \frac{1}{2n_i} \sum_{x \in \omega_i} \sum_{y \in \omega_i} (x - m_i + m_i - y)^2$$

$$= \frac{1}{2n_i} \sum_{x \in \omega_i} \sum_{y \in \omega_i} (x - m_i)^2 + \frac{1}{2n_i} \sum_{x \in \omega_i} \sum_{y \in \omega_i} (y - m_i)^2 + \frac{1}{n_i} \sum_{x \in \omega_i} \sum_{y \in \omega_i} (x - m_i)(m_i - y)$$

$$= \sum_{x \in \omega_i} (x - m_i)^2 + \frac{1}{n_i} \sum_{x \in \omega_i} (x - m_i) \sum_{y \in \omega_i} (m_i - y)$$

$$\boxed{s_i = \sum_{x \in \omega_i} (x - m_i)^2}$$

- Tout se passe comme si on se focalisait sur la moyenne (le centre de gravité)

Dispersion intraclasse

- La matrice de dispersion de la classe i s'écrit :

$$S_i = \sum_{x \in \omega_i} (x - m_i)(x - m_i)^t$$

- Dispersion Intraclasse
- Analogue aux matrices d'inertie en mécanique
- Plus il y a d'inertie, moins il y a de localisation des masses

Dispersion totale

- Dispersion totale S_T

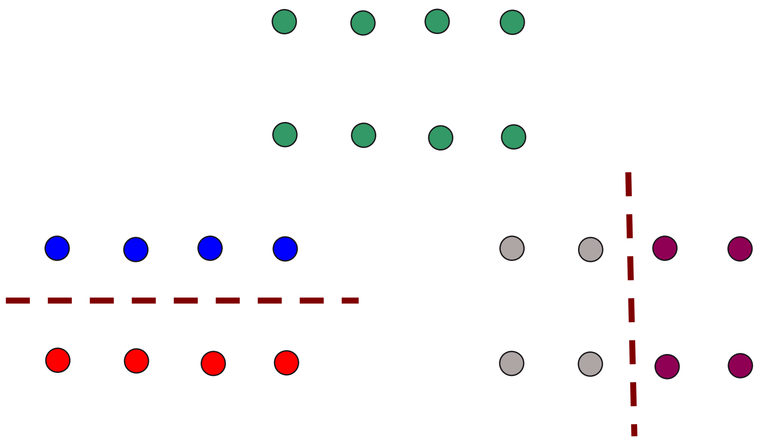
$$\begin{aligned} S_T &= \sum_{p=1}^{n} (x_p - m)(x_p - m)^t \\ &= \sum_{i=1}^{c} \sum_{x \in \omega_i} (x - m)(x - m)^t \\ &= \sum_{i=1}^{c} \sum_{x \in \omega_i} (x - m_i + m_i - m)(x - m_i + m_i - m)^t \\ &= \sum_{i=1}^{c} \sum_{x \in \omega_i} (x - m_i)(x - m_i)^t + \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^t \\ &= S_I + S_B \end{aligned}$$

Critère de classification

$$\boxed{S_T = S_I + S_B}$$

- **La dispersion totale S_T est une constante du problème**
- **Le choix des moyennes modifie :**
  - La dispersion intraclasse S_I
  - La dispersion interclasse S_B
- **Minimiser $S_I \Leftrightarrow$ maximiser S_B**

Homogénéité versus séparabilité



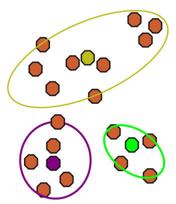
Minimiser $S_I \Leftrightarrow$ maximiser S_B

$$\boxed{S_I + S_B = \text{constante}}$$

- **On recherche des regroupements d'individus qui se ressemblent : S_I minimum**
- **On recherche des groupes d'individus qui soient le plus différent possible : S_B maximum**
- **Les deux objectifs sont** identiques

Classification par partition (K-means)

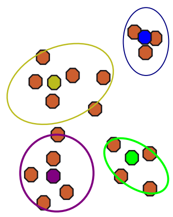
- **Exemple à 3 classes**



- **Faire des classes homogènes**
- **Ecarter les centres de gravité**

Classification par partition (K-means)

- **Exemple à 4 classes**



- **Faire des classes homogènes**
- **Ecarter les centres de gravité**

Optimisation itérative

- **Selon les cas :**
  - Il est plus facile de calculer l'intraclasse
  - Il est plus facile de calculer l'interclasse
- **Possibilité de raisonner au niveau de l'échantillon : optimisation itérative**

Optimisation itérative

- **On considère deux classes ω_i et ω_j:**
  - n_i échantillons, n_j échantillons
  - Dispersions S_i et S_j
- **On fait passer un échantillon x de la classe ω_i à la classe ω_j :**
  - Nouvelles dispersions $\underline{S_i}$ et $\underline{S_j}$
  - On compare J_i+J_j à $\underline{J_i+J_j}$

TELECOM ParisTech

Optimisation itérative

- **On montre que**

$$\underline{S_i} = S_i - \frac{n_i}{n_i - 1}\left\|x - m_i\right\|^2$$

$$\underline{S_j} = S_j - \frac{n_j}{n_j + 1}\left\|x - m_j\right\|^2$$

TELECOM ParisTech

Un premier algorithme

1. **Choix d'une partition initiale : calcul des dispersion et des centres de gravités**
2. **Prendre un échantillon x de la classe ω_i**
   - Si ni=1 aller en 5
   - Sinon calculer les écarts de dispersion
3. **Changer de classe si le test est vérifié**
4. **Calcul des centres de gravité et des dispersion**
5. **Si changement retour en 2**

TELECOM ParisTech

Réflexions sur le centre de gravité

$$S_i = \frac{1}{2n_i} \sum_{x\in\omega_i} \sum_{y\in\omega_i} \|x-y\|^2$$

$$s_i = \sum_{x\in\omega_i} (x-m_i)^2$$

- **Cas 1D :** $\boxed{(x-m)^2 = x^2 + m^2 - 2xm}$
- **Cas spécial :**
  - x normalisé
  - m normalisé
  - -2xm doit être minimisé
- **Cas n-D : étudier <x|m>**

TELECOM ParisTech

K moyennes (K means)
Méthode de « centres mobiles »

$$S_i = \sum_{x \in \omega_i} \|x - m_i\|^2$$

- **Choix du centre minimisant la dispersion (l'erreur quadratique):**

$$\frac{\partial S_i}{\partial m_i} = -2 \sum_{x \in \omega_i} (x - m_i) = 0$$

$$\Leftrightarrow$$

$$m_i = \frac{1}{n_i} \sum_{x \in \omega_i} x$$

Algorithme des K moyennes

1. **Choix de K classes**
2. **Initialisation : choix de K centres de gravité, j=0**
3. **Itération j : x est affecté à $\omega_i(j+1)$ si**

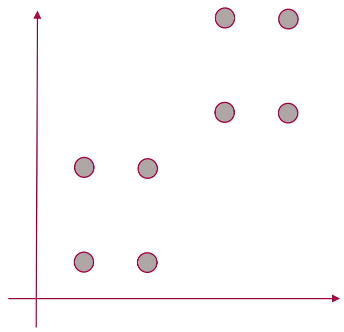
$$\|x - m_i(j)\| = \min_{l=1}^{K} \|x - m_l(j)\|$$

4. **Mise à jour des centres de gravité**

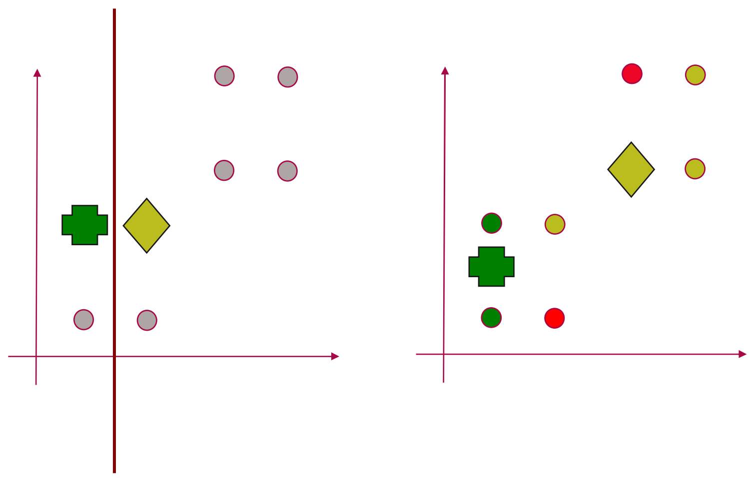
$$m_i(j+1) = \frac{1}{n_i} \sum_{x \in \omega_i(j+1)} x$$

5. **Test de convergence :**
   - Si $\forall i$, $m_i(j+1)=m_i$, alors FIN
   - Sinon, j=j+1 et retour à l'étape 3

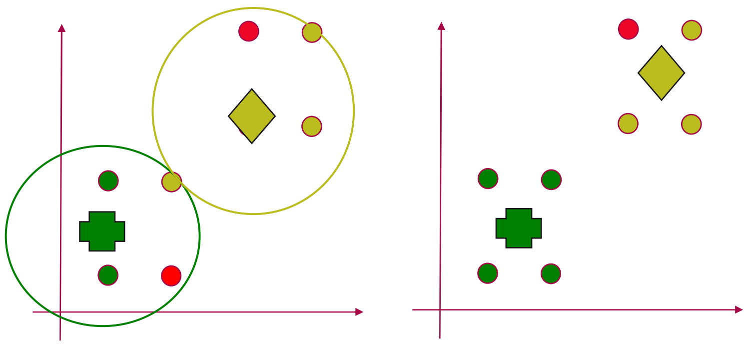
Cas d’école



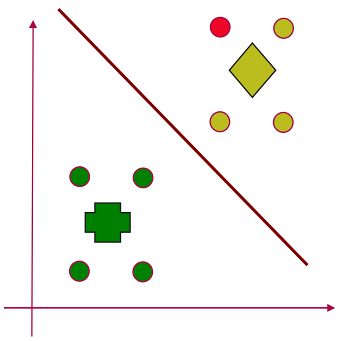
Cas d’école



Cas d'école

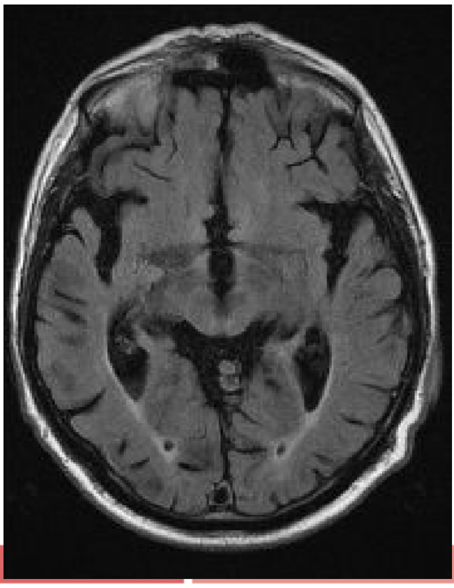


Cas d'école



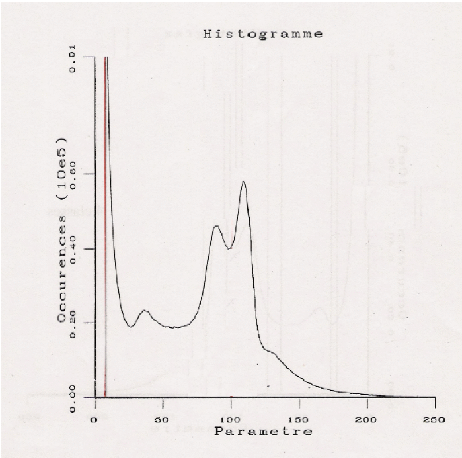
- Plus rien ne change
- Fin de l'algorithme

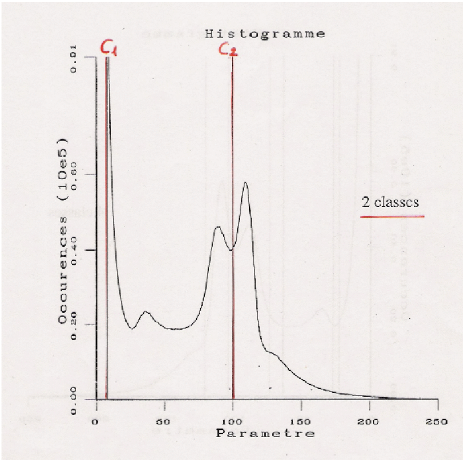
Exemple sur un IRM de cerveau

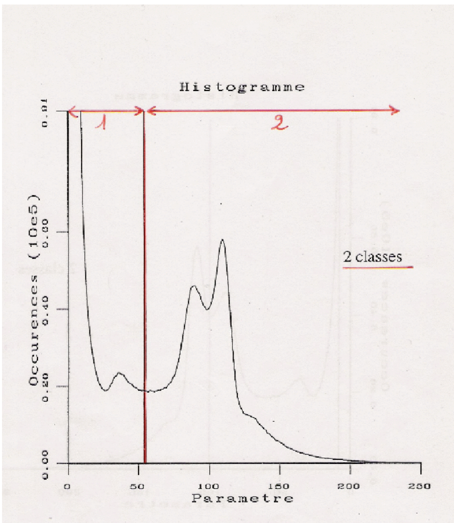


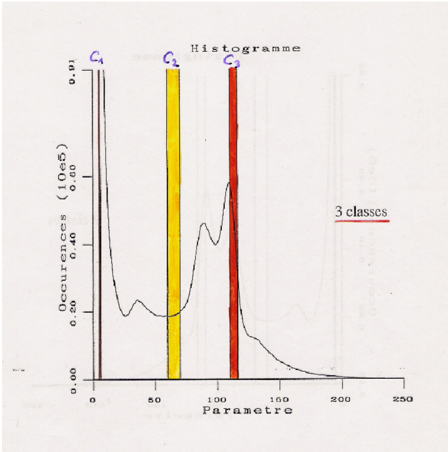
Analyse sur histogramme

- Combien de classes ?
  - 2 classes ?
  - 3 classes ?
  - 4 classes ?

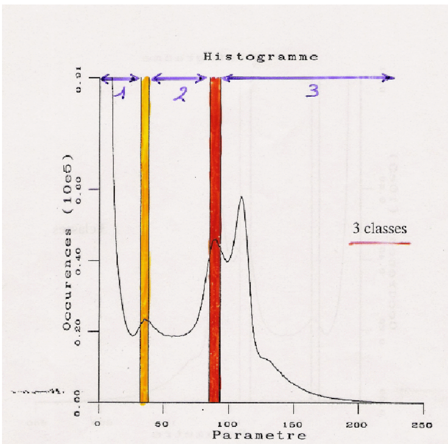



Histogramme
C_1
C_2
2 classes
Occurences (10e5)
0.00
0.20
0.40
0.60
0.81
0
50
100
150
200
250
Parametre

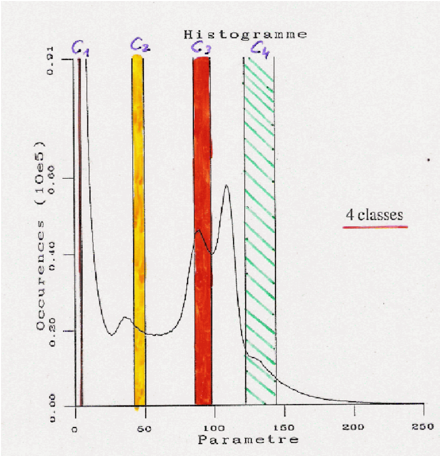

Histogramme
1
2
2 classes
Occurences (10e5)
0.00
0.20
0.40
0.60
0.81
0
50
100
150
200
250
Parametre



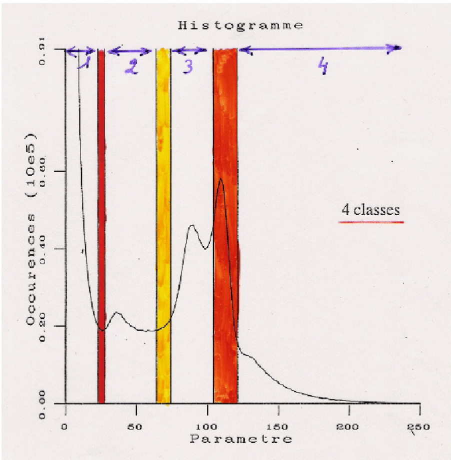
TELECOM ParisTech



TELECOM ParisTech



Histogramme
C_1 C_2 C_3 C_4
Occurences (10e5)
0.00 0.20 0.40 0.60 0.91
0 50 100 150 200 250
Parametre
4 classes



Histogramme
1 2 3 4
Occurences (10e5)
0.00 0.20 0.40 0.60 0.91
0 50 100 150 200 250
Parametre
4 classes

Autres techniques par partition

- Boules optimisées
- Fuzzy C-means
- Nuées dynamiques
- …..

Boules optimisées

Analogue aux k-moyennes, MAIS :

- Chaque classe doit être incluse dans une boule de rayon R
- Les boules ont un rayon fixé R
- Le nombre de classes n'est pas fixé

K moyennes

- Minimise l'Erreur Quadratique Moyenne (EQM)
- Sensible au nombre de classes
- Facile à mettre en œuvre

Fuzzy C-means

- *C* classes
- Degré d'appartenance à la classe k :

$$\begin{cases} \mu_k \in [0;1] \; \forall k \\ \sum_{k=1}^{C} \mu_k = 1 \end{cases}$$

- *C* prototypes
- Critère : distance euclidienne aux prototypes

Fuzzy C means

- Prototypes b_i
- On définit la fonctionnelle J

$$J(B,U,X)=\sum_{i=1}^{c}\sum_{k=1}^{n}\mu_{i,k}\,d(x_k,b_i)$$

- La fonction d'appartenance est définie par

$$\boxed{\mu_{i,k}=\frac{1}{\sum_{j=1}^{c}\left(\frac{\|x_k-b_i\|}{\|x_k-b_j\|}\right)^{\frac{2}{m-1}}}}$$

m est le « facteur de flou »

Fuzzy C means

- Le prototype b_i est alors donné par

$$b_i=\frac{\sum_{j=1}^{n}\mu_{j,i}^{m}x_j}{\sum_{j=1}^{n}\mu_{j,i}^{m}}$$

Crisp Cmeans : les k-moyennes

- **Chaque élément appartient à une classe et une seule**

$$b_i = \frac{\sum_{j=1}^{n} \mu_{j,i}^m x_j}{\sum_{j=1}^{n} \mu_{j,i}^m} = \frac{\sum_{x_j \in \omega_i}^{n} x_j}{n_i}$$

Algorithme ISODATA

- **Paramètres d'entrées**
  - Nombre d'aggrégats
  - Nombre minimum d'éléments par aggrégat
  - Distance minimale entre chaque aggrégat
  - Paramètre de contrôle des subdivisions d'aggrégat
  - Nombre d'itérations dans la première phase de l'algorithme
  - Nombre maximum de regroupements par itération
  - Nombre d'itérations maximum
- **Etapes différentes selon indice de boucle d'itération**

Nuées dynamiques (Diday)

- Les classes sont décrites par un noyau, par exemple leurs centres mi
- La « distance » entre un échantillon et un centre est décrite par une « mesure de dissemblance » :

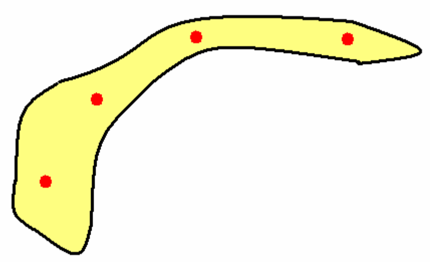
$$f(x,\omega_i)$$

- La partition P optimale vérifie le minimum de

$$\sum_{i=1}^{c}\sum_{x\in\omega_i} f(x,\omega_i)$$

Exemples de fonctions de dissemblances

- Le noyau des classes est défini par K centres $m_{i,l}$



$$\boxed{f(x,\omega_i) = \min_{l=1}^{K_i} \left\| x - m_{i,l} \right\|}$$

- Ou bien :

$$\boxed{f(x,\omega_i) = \sum_{l=1}^{K_i} \left\| x - m_{i,l} \right\|}$$

- Définition par rapport à un axe d'inertie I_i

$$\boxed{f(x,\omega_i) = \min_{l=1}^{K_i} d(x, I_i)}$$

Algorithmes des nuées dynamiques (Diday)

- **Schéma identique à celui des k-moyennes**
- **Problèmes liés :**
  - À l'initialisation des noyaux
  - Au nombre de classes

Conclusions et limites

- **En général, il faut définir :**
  - Le nombre de classes
  - la métrique (distance euclidienne, ,…)
- **Vérifier que les résultats concordent avec l'attente des opérationnels**