le cnam

Apprentissage, réseaux de neurones et modèles graphiques

Cours : Michel Crucianu (PU)
Fouad Badran (PU)
MezianeYacoub (MCF)
Marin Ferecatu (MCF)
TP : Meziane Yacoub

http://idf.pleiad.net/index.php

le cnam

Objectifs de l'enseignement

- Méthodes décisionnelles basées sur l'apprentissage à partir des données : réseaux de neurones, Machines à Vecteurs Supports (SVM), méthodes (réseaux) graphiques ; leur utilisation dans des applications réelles
  - Suite de l'UE RCP208 « Reconnaissance des formes et réseaux de neurones »

→ Fouille de données (*data mining*) : recherche de régularités ou de relations inconnues *a priori* dans de grands volumes de données

→ Reconnaissance des formes (*pattern recognition*) = (sens strict) identifier à quelle catégorie appartient une « forme » décrite par des données brutes

le cnam

Contenu de l'ensemble du cours

- Estimation de fonctions de densité (2, MC)
- Cartes auto-organisatrices appliquées aux données quantitatives, catégorielles et mixtes (2, MY)
- Perceptrons multicouches : fonctions d'erreur, maximum de vraisemblance, modèle multi-expert (1, MY)
- Apprentissage, généralisation, régularisation (1 MY, 1 MC)
- Machines à vecteurs supports (SVM), ingénierie des noyaux (2, MF)
- Chaînes de Markov cachées (2, FB)
- Introduction aux réseaux bayesiens (2, FB)

le cnam

Travaux pratiques (TP)

- Contenu
  - Estimation de fonctions de densité
  - Cartes auto-organisatrices appliquées aux données quantitatives, catégorielles et mixtes
  - Perceptrons multicouches
  - Machines à vecteurs supports, ingénierie des noyaux
  - *Travail suivi sur le mini-projet*
- Logiciel privilégié : Matlab (http://www.mathworks.fr) ou, comme alternative gratuite, Octave (http://www.gnu.org/software/octave/)
  - D'autres solutions sont envisageables, suivant les préférences et les disponibilités des auditeurs
    - Weka (Java) : http://sourceforge.net/projects/weka/
    - Divers logiciels statistiques

le cnam

Évaluation

- Note finale : moyenne entre
  - Examen final
  - Mini-projets réalisés en partie durant les TP (juin + septembre)
    - Sujets proposés courant avril, choix arrêtés fin avril
    - 3 étapes
      - Compréhension du problème (fin mai)
      - Analyse des données (fin juin)
      - Finalisation du projet et envoi d'un bref mémoire (septembre)

le cnam

Bibliographie générale

Dreyfus, G., J. Martinez, M. Samuelides, M. Gordon, F. Badran, S. Thiria, *Apprentissage statistique : Réseaux de neurones - Cartes topologiques - Machines à vecteurs supports*. Éditions Eyrolles, 2008.

Hand D. J., H. Mannila, P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

Hastie T., R. Tibshirani, J. H. Friedman. *The Elements of Statistical Learning*. Springer, 2003.

Naïm P., P.-H. Wuillemin, Ph. Leray, O. Pourret, A. Becker. *Réseaux bayesiens*. Eyrolles, 2004.

Schölkopf B., A. Smola. *Learning with Kernels*. MIT Press, 2002.

le cnam

Apprentissage, réseaux de neurones et modèles graphiques

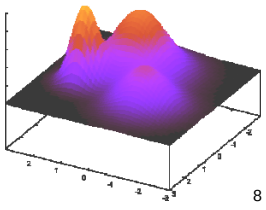
Chapitre 1 : Estimation de fonctions de densité

Michel Crucianu
(avec contributions de Jean-Philippe Tarel)
http://cedric.cnam.fr/~crucianm/ml.html

le cnam

Estimation de fonctions de densité

- Objectif : à partir d'observations, estimer la loi qui les a générées
  - Soit $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ un ensemble d'observations dans $\mathcal{X}$
  - L'observation $\mathbf{x}_i$ est une réalisation de la variable aléatoire X_i
  - On considère que les variables X_i, $i = 1, \ldots, n$, sont indépendantes et identiquement distribuées (i.i.d.) suivant la densité de probabilité $f(X)$ et on cherche à estimer cette densité à partir des observations
  - Notation : la valeur de $f(X)$ dans $\mathbf{x} \in \mathcal{X}$ sera notée $p(\mathbf{x})$
- A quoi sert l'estimation de fonctions de densité ?
  - Mettre en évidence des régularités présentes dans les données
  - Décision bayesienne – approche générative : modéliser $p(\mathbf{x} | c_j)$ et, avec $P(c_j)$, obtenir les probabilités *a posteriori* $P(c_j | \mathbf{x})$



le cnam

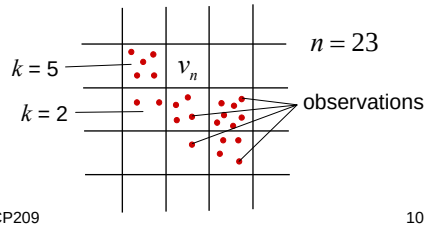
Méthodes d'estimation de densités

- Non paramétriques : absence d'hypothèses sur les lois
  1. Estimation par histogramme
  2. Méthode des noyaux de Parzen
  3. Méthode des k_n plus proches voisins (non abordée dans la suite)
- Paramétriques : hypothèses sur la nature des lois (appartenance à une famille paramétrée) → estimation des paramètres de ces lois
  - Maximisation de la vraisemblance
  - Maximisation de l'*a posteriori*
  - Modèles de mélanges
  - Algorithme *Expectation-Maximization* (EM)
  - EM appliqué aux mélanges gaussiens

le cnam

Estimation par histogramme

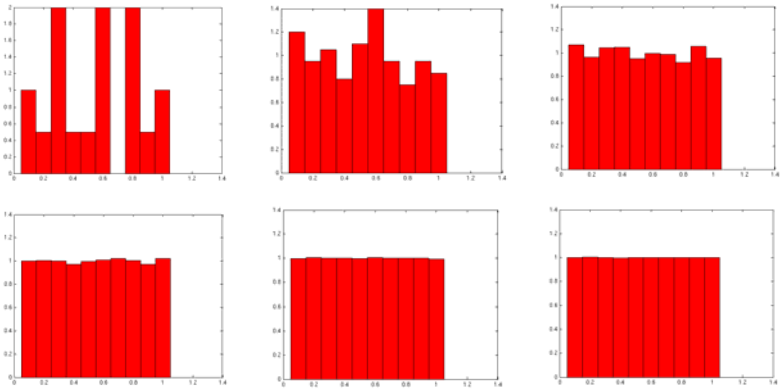
- Sans hypothèses sur les lois qui régissent $p(\mathbf{x})$
- Principe de l'estimation : $\hat{f}(\mathbf{x}) \cong \frac{k/n}{v}$ (n obs. au total, k dans v)
  - Division de l'espace en volumes v, comptage des observations
  - v fixé, n augmente : l'estimation s'améliore (la variance diminue), mais représente une moyenne sur v
  - n fixé, v diminue : la précision par rapport à $\mathbf{x}$ s'améliore, mais la qualité d'estimation se dégrade (la variance augmente)
- Conditions nécessaires :
  - $\lim_{n\to\infty} v_n = 0$ (précision)
  - $\lim_{n\to\infty} k_n = \infty$ (estimation)
  - $\lim_{n\to\infty} k_n/n = 0$ (estimation)



le cnam

Exemple : loi uniforme, v = 1/10

n = 20 n = 200 n = 2 000

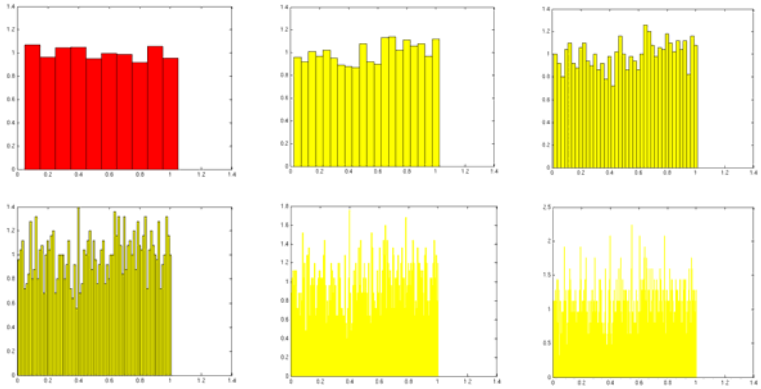


n = 20 000 n = 200 000 n = 2 000 000

le cnam

Exemple : loi uniforme, n = 2 000

v = 1/10 v = 1/20 v = 1/40



v = 1/80 v = 1/160 v = 1/320

le cnam

Exemple d'application

- Détection de la chaussée

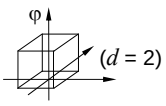
le cnam

Méthode des fenêtres de Parzen

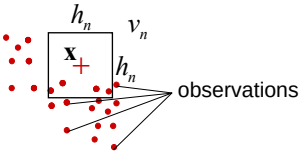
- Considérons d'abord des (hyper)cubes de coté h_n (et dimension d)
- Définissons une première fonction-fenêtre (ou noyau)

$$\varphi(\mathbf{u}) = \begin{cases} 1 & \text{si } \forall j,\ |u_j| < 1/2 \\ 0 & \text{sinon} \end{cases}$$

(u_j étant la composante j de $\mathbf{u}$)



- Densité autour de $\mathbf{x}$?



→ Pour l'hypercube $\varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right) = 1$

- Le volume est $v_n = h_n^d$
- Le nombre d'observations situées à l'intérieur est $k_n = \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_n}\right)$

le cnam

Méthode des fenêtres de Parzen (2)

→ L'estimateur sera alors $\left(\hat{f}_n(\mathbf{x}) = \frac{k_n/n}{v_n} \right)$

$$\hat{f}_n(\mathbf{x}) = \frac{1}{n h_n^d} \sum_{i=1}^{n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n} \right)$$

→ résultat global : somme des fenêtres centrées sur les observations !

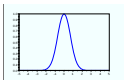
qui possède h_n (« largeur » de la fenêtre) comme seul paramètre

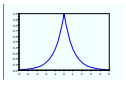
25/09/2013 RCP209 15

le cnam

Méthode des fenêtres de Parzen (3)

- Conditions suffisantes pour que l'estimateur soit lui-même une densité :
  - $\varphi(\mathbf{u}) \geq 0, \forall \mathbf{u} \in \mathbb{R}^d$
  - $\int_{\mathbb{R}^d} \varphi(\mathbf{u}) d\mathbf{u} = 1$
- L'approche peut être généralisée en considérant d'autres noyaux :
  - Normal (gaussien) : $\varphi(\mathbf{u}) = \frac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\mathbf{u}^T\mathbf{u}}$
  - Cauchy : $\varphi(\mathbf{u}) = \frac{1}{\pi(1+\mathbf{u}^T\mathbf{u})}$
  - Exponentiel : $\varphi(\mathbf{u}) = (1/2) e^{-\|\mathbf{u}\|}$
  - Triangulaire : $\varphi(\mathbf{u}) = \max\{0, 1-\|\mathbf{u}\|\}$



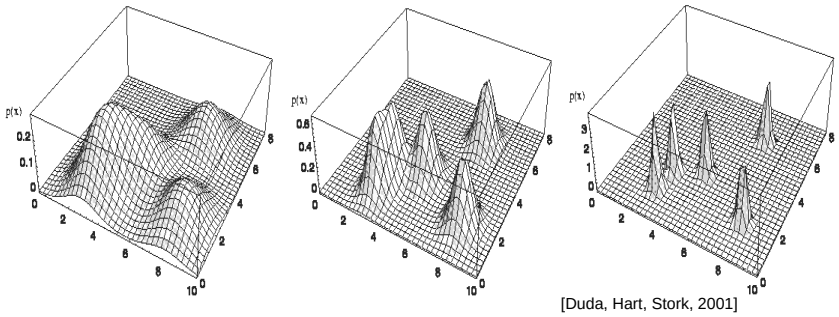


25/09/2013 RCP209 16

le cnam

Fenêtres de Parzen : exemple

- Densités obtenues avec un échantillon de 5 vecteurs de $\mathbb{R}^2$ et des noyaux gaussiens d'écart-types 2 (gauche), 1 (centre) et 0,5 (droite)



[Duda, Hart, Stork, 2001]

- La « largeur » du noyau a plus d'impact que le type de noyau !
- Choix de la largeur : nombreuses méthodes, dont la validation croisée

le cnam

Fenêtres de Parzen : convergence

- Conditions de convergence :
  - $p(\mathbf{x})$ continue dans $\mathbf{x}$
  - $\sup_{\mathbf{u}} \varphi(\mathbf{u}) < \infty$ ($\varphi(\mathbf{u})$ est bornée)
  - $\lim_{\|\mathbf{u}\| \to \infty} \|\mathbf{u}\|^d \varphi(\mathbf{u}) = 0$
  - $\lim_{n \to \infty} h_n = 0$ et $\lim_{n \to \infty} n h_n^d = 0$ (par ex. $h_n = h_0 / \sqrt{n}$, $h_n = h_0 / \ln n$)
- Nature de la convergence :
  - Moyenne : $\lim_{n \to \infty} \overline{p}_n(\mathbf{x}) = p(\mathbf{x})$
  - Variance : $\lim_{n \to \infty} \sigma_n^2(\mathbf{x}) = 0$

le cnam

Estimation paramétrique

- On considère que la densité recherchée appartient à une famille paramétrée par des vecteurs $\boldsymbol{\theta} \in \Omega$ et on indique cette dépendance en écrivant $p(\mathbf{x}|\boldsymbol{\theta})$ au lieu de $p(\mathbf{x})$
- L'échantillon $\mathcal{D}$ étant issu de variables i.i.d., nous pouvons écrire

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})$$

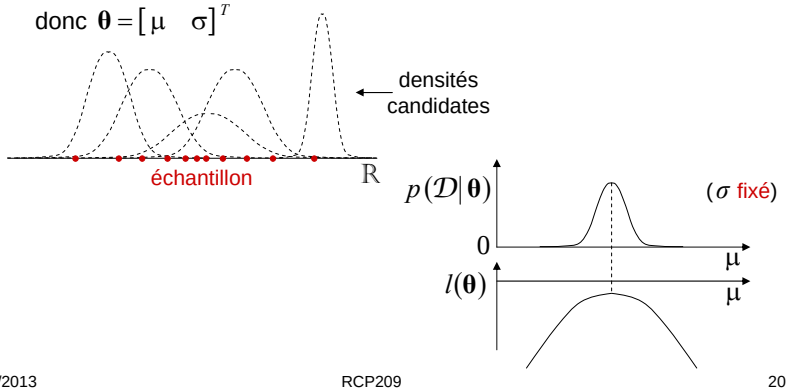
- Comme fonction de $\boldsymbol{\theta}$, $p(\mathcal{D}|\boldsymbol{\theta})$ est la vraisemblance (*likelihood*) de $\boldsymbol{\theta}$ par rapport à l'échantillon $\mathcal{D}$
- Il est en général plus facile de travailler avec le logarithme de la vraisemblance (*log-likelihood*)

$$l(\boldsymbol{\theta}) \equiv \ln[\, p(\mathcal{D}|\boldsymbol{\theta})\,] = \sum_{i=1}^{n} \ln[\, p(\mathbf{x}_i|\boldsymbol{\theta})\,]$$

le cnam

Un cas simple : loi normale 1D

- Considérons le cas suivant :
  - $\mathcal{X} = \mathrm{R}$
  - La famille paramétrée candidate est celle des lois normales $\mathcal{N}(\mu,\sigma)$, donc $\boldsymbol{\theta} = [\,\mu \quad \sigma\,]^T$



le cnam

Maximum de vraisemblance (MV)

- L'estimation $\hat{\boldsymbol{\theta}}$ la plus en accord avec les observations $\mathcal{D}$ est celle qui correspond au maximum de la vraisemblance $p(\mathcal{D}|\boldsymbol{\theta})$
- Le logarithme en base e étant une fonction monotone croissante, le maximum de la vraisemblance est atteint pour le même $\hat{\boldsymbol{\theta}}$ que le maximum de $l(\boldsymbol{\theta})$
- Exemple : pour le cas normal unidimensionnel considéré,

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln[\, p(x_i|\boldsymbol{\theta})\,] = \sum_{i=1}^{n} \left[-\ln \sigma\sqrt{2\pi} - \frac{1}{2\sigma^2}(x_i-\mu)^2 \right]$$

donc en dérivant on obtient

$$\frac{\partial \sum_{i=1}^{n} \ln\left[p\left(x_i|\boldsymbol{\theta}\right)\right]}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$

$$\frac{\partial \sum_{i=1}^{n} \ln\left[p\left(x_i|\boldsymbol{\theta}\right)\right]}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

le cnam

Maximum de vraisemblance (2)

- Les dérivées partielles sont nulles

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu}) = 0 \quad \text{et} \quad -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = 0$$

pour

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \;\; (=\bar{x}) \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$$

(on vérifie que la solution correspond bien à un maximum)

- Remarques :
  - L'estimation pour μ est non biaisée (son espérance sur tous les échantillons de taille *n* est égale à la vraie valeur de μ)
  - L'estimation de σ^2 est en revanche biaisée (mais asymptotiquement non biaisée) ; une estimation non biaisée est $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\hat{\mu})^2$

le cnam

Explication du biais

- La moyenne d'un échantillon $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est une variable aléatoire d'espérance $E(\bar{x}) = E(X)$ et de variance $V(\bar{x}) = \frac{1}{n} V(X)$
- La variance d'un échantillon $v = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$ est une variable aléatoire d'espérance

$$E(v) = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i{}^2\right) - E\left(\bar{x}^2\right) = E\left(x_i^2\right) - E\left(\bar{x}^2\right)$$

or $V(x_i) = E\left(x_i^2\right) - E(x_i)^2$, donc $E\left(x_i^2\right) = E(X)^2 + V(X)$

et $E\left(\bar{x}^2\right) = E(\bar{x})^2 + V(\bar{x}) = E(X)^2 + (1/n)V(X)$

par conséquent,

$$E(v) = E(X)^2 + V(X) - E(X)^2 - \frac{1}{n}V(X) = \frac{n-1}{n}V(X) \quad (\neq V(X))$$

le cnam

Maximum de vraisemblance (3)

- Exemple 2 – lois normales multidimensionnelles :
  - $\mathcal{X} = \mathbb{R}^d$
  - La famille paramétrée candidate est celle des lois normales $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(x|\theta) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

- Dans ce cas, le maximum de la vraisemblance est donné par

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \quad \text{et} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

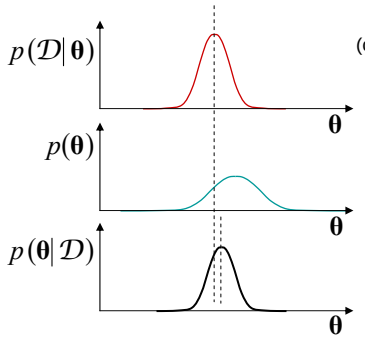
- Ici encore, l'estimation de $\boldsymbol{\mu}$ est non biaisée mais celle de la matrice de variances-covariances $\boldsymbol{\Sigma}$ est biaisée (simplement asymptotiquement non biaisée) ; une estimation non biaisée est

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$$

le cnam

Maximum *a posteriori* (MAP)

- Si on connaît la densité de probabilité *a priori* pour $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, il est préférable de choisir la solution $\hat{\boldsymbol{\theta}}$ qui maximise la probabilité *a posteriori* $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})$



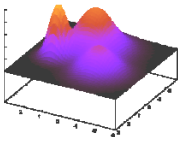
(comme fonction de $\boldsymbol{\theta}$, non normalisée)

le cnam

Modèles de mélange

- On s'intéresse à des densités de probabilité qui sont des mélanges additifs de plusieurs densités décrites par des lois élémentaires (par exemple, lois normales multidimensionnelles)

$$p(\mathbf{x}|\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{j=1}^{m} \alpha_j\, p_j(\mathbf{x}|\boldsymbol{\theta}_j)$$



avec

m : nombre de composantes du mélange

$p_j(\mathbf{x}|\boldsymbol{\theta}_j)$: densité qui définit une composante (supposée appartenir à une famille paramétrée par $\boldsymbol{\theta}_j$)

α_j : coefficients de mélange tels que $\sum_{j=1}^{m} \alpha_j = 1$

$\boldsymbol{\alpha}$: représentation vectorielle des coefficients de mélange

$\boldsymbol{\theta}$: vecteur qui représente tous les $\boldsymbol{\theta}_j$

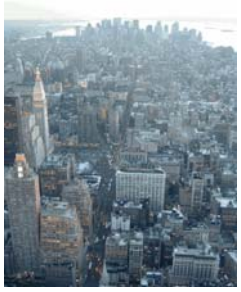
le cnam

Exemple d'application : compression

256 couleurs

128 couleurs

64 couleurs



(source : J.-Ph. Tarel, LCPC)

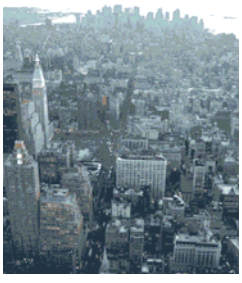
le cnam

Exemple d'application : compression

32 couleurs

16 couleurs

8 couleurs



(source : J.-Ph. Tarel, LCPC)

le cnam

Modèles de mélange : estimation MV

- On considère $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ un ensemble d'observations de $\mathbb{R}^d$ issues de variables i.i.d. suivant la densité de probabilité $p(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\theta})$
- On cherche à trouver $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$ qui maximisent la vraisemblance

$$p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\theta})$$

- À partir de l'expression de $p(\mathbf{x}|\boldsymbol{\alpha}, \boldsymbol{\theta})$, on obtient l'expression à maximiser (ici, le logarithme de la vraisemblance)

$$\ln p(\mathcal{D}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left[\sum_{j=1}^{m} \alpha_j \, p_j(\mathbf{x}_i|\boldsymbol{\theta}_j) \right] \text{ sous la contrainte } \sum_{j=1}^{m} \alpha_j = 1$$

- En raison de la présence d'une somme sous le logarithme, on ne peut pas obtenir de solution analytique à ce problème de maximisation, il faut donc avoir recours à des méthodes d'optimisation itérative pour déterminer $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}})$

le cnam

Expectation-Maximization

- Algorithme itératif, introduit dans [DLR77]
- Une solution au problème général de l'estimation MV des paramètres d'une distribution à partir de données incomplètes ou présentant des valeurs manquantes
- Pré-requis : considérons $\mathcal{D} = \{\mathcal{D}_o, \mathcal{D}_m\}$
  - $\mathcal{D}_o$ est l'ensemble des données observées
  - $\mathcal{D}_m$ correspond aux données manquantes
  - $\mathbf{z}$ sera associé à $\mathcal{D}$, $\mathbf{x}$ à $\mathcal{D}_o$ et $\mathbf{y}$ à $\mathcal{D}_m$
  - Données suivant la densité $p(\mathbf{z}|\boldsymbol{\theta}) = p(\mathbf{x}, \mathbf{y}|\boldsymbol{\theta}) \propto p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}|\boldsymbol{\theta})$
    $\propto p(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta})$

le cnam

Expectation-Maximization (2)

- On cherche le paramètre $\boldsymbol{\theta}^*$ qui maximise la vraisemblance $p(\mathcal{D}_o|\boldsymbol{\theta})$ ou son logarithme $\ln p(\mathcal{D}_o|\boldsymbol{\theta})$
- Question : comment maximiser la vraisemblance $p(\mathcal{D}_o|\boldsymbol{\theta})$ quand les données $\mathcal{D}_m$ sont manquantes ?
- Si Y suit une densité $q(\mathbf{y})$, $\int_{\mathbf{y}} q(\mathbf{y})\,d\mathbf{y} = 1$ et on peut écrire $(\forall \boldsymbol{\theta}, \forall q)$

$$\ln p(\mathcal{D}_o|\boldsymbol{\theta}) = \ln p(\mathcal{D}_o|\boldsymbol{\theta}) \int_{\mathbf{y}} q(\mathbf{y})\,d\mathbf{y} = \int_{\mathbf{y}} q(\mathbf{y}) \ln p(\mathcal{D}_o|\boldsymbol{\theta})\,d\mathbf{y}$$

$p(\mathcal{D}_o,\mathbf{y}|\boldsymbol{\theta}) \propto p(\mathbf{y}|\mathcal{D}_o,\boldsymbol{\theta})\,p(\mathcal{D}_o|\boldsymbol{\theta})$

$$\propto \int_{\mathbf{y}} q(\mathbf{y}) [\ln p(\mathcal{D}_o,\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\mathcal{D}_o,\boldsymbol{\theta}) - \ln q(\mathbf{y}) + \ln q(\mathbf{y})]\,d\mathbf{y}$$

$$= \underbrace{\int_{\mathbf{y}} q(\mathbf{y}) \ln \frac{p(\mathcal{D}_o,\mathbf{y}|\boldsymbol{\theta})}{q(\mathbf{y})}\,d\mathbf{y}}_{l(q,\boldsymbol{\theta})} + \underbrace{\int_{\mathbf{y}} q(\mathbf{y}) \ln \frac{q(\mathbf{y})}{p(\mathbf{y}|\mathcal{D}_o,\boldsymbol{\theta})}\,d\mathbf{y}}_{\mathrm{KL}(q(\mathbf{y}) \| p(\mathbf{y}|\mathcal{D}_o,\boldsymbol{\theta}))}$$

le cnam

Expectation-Maximization (3)

- Comme $\mathrm{KL}(q(\mathbf{y}) \| p(\mathbf{y}|\mathcal{D}_o,\boldsymbol{\theta})) \geq 0$, $\ln p(\mathcal{D}_o|\boldsymbol{\theta}) \geq l(q,\boldsymbol{\theta})$

→ Algorithme EM :

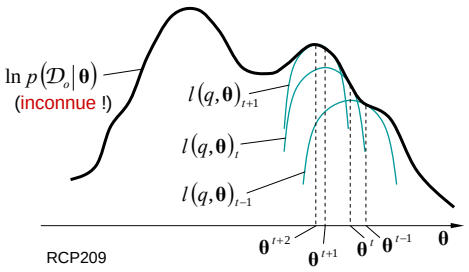
Initialisation : choix de $\boldsymbol{\theta}^0$

Itérer

Étape **E** : calculer $l(q,\boldsymbol{\theta})_t$ en considérant $q(\mathbf{y}) \approx p(\mathbf{y}|\mathcal{D}_o,\boldsymbol{\theta}^t)$

Étape **M** : identifier $\boldsymbol{\theta}^{t+1}$ maximisant $l(q,\boldsymbol{\theta})_t$

jusqu'à la convergence



le cnam

Expectation-Maximization (4)

- Déterminer $l(q,\boldsymbol{\theta})$ avec $q(\mathbf{y})$ fixé à $p\left(\mathbf{y}\middle|\mathcal{D}_o,\boldsymbol{\theta}^t\right)$ revient à calculer

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^t\right) = E_{\mathcal{D}_m}\left[\ln p(\mathcal{D}_o,\mathcal{D}_m|\boldsymbol{\theta}) \,|\, \mathcal{D}_o,\boldsymbol{\theta}^t\right] = \int_{\mathbf{y}}\left[\ln p\left(\mathcal{D}_o,\mathbf{y}\middle|\boldsymbol{\theta}\right)\right] p\left(\mathbf{y}\middle|\mathcal{D}_o,\boldsymbol{\theta}^t\right) d\,\mathbf{y}$$

- Difficulté : présence en général d'extrema locaux vers lesquels l'algorithme peut converger (plutôt que vers un maximum global)
- Types d'utilisations de EM :
  - Valeurs et/ou variables manquantes dans les données (exemple : difficultés d'observation)
  - Introduction de variables cachées, dont les valeurs manquent, permettant de simplifier l'expression de la vraisemblance ($l(q,\boldsymbol{\theta})$ au lieu de $\ln p\left(\mathcal{D}_o\middle|\boldsymbol{\theta}\right)$) et donc la détermination d'un maximum

le cnam

EM pour modèles de mélange

- Données manquantes $\{Y_i\}_{i=1}^n : y \in \{1,\ldots,m\}$, $y_i = j$ si l'observation $\mathbf{x}_i$ a été générée par la composante j du mélange
- Le vecteur des paramètres à identifier est dans ce cas $\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\theta}\end{bmatrix}$
- Le caractère i.i.d. des variables dont sont issues les observations implique $\ln p(\mathcal{D}_o,\mathcal{D}_m|\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{i=1}^n \ln p(\mathbf{x}_i,y_i|\boldsymbol{\alpha},\boldsymbol{\theta})$

  et, pour $\mathbf{y} = (y_1,\ldots,y_n)$ une instance des données non observées,

$$P\left(\mathbf{y}\middle|\mathcal{D}_o,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right) = \prod_{i=1}^n P\left(y_i\middle|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right)$$

- Enfin, les valeurs manquantes étant dans un ensemble discret,

$$Q\left(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right) = \sum_y\left[\sum_{i=1}^n \ln p\left(\mathbf{x}_i,y_i\middle|\boldsymbol{\alpha},\boldsymbol{\theta}\right)\right]\prod_{i=1}^n P\left(y_i\middle|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right)$$

le cnam

EM pour modèles de mélange (2)

- Par la définition des données manquantes,

$$p(\mathbf{x}_i, y_i|\boldsymbol{\alpha},\boldsymbol{\theta}) = p(\mathbf{x}_i|\, y_i,\boldsymbol{\alpha},\boldsymbol{\theta})\, P(\, y_i|\boldsymbol{\alpha},\boldsymbol{\theta}) = \alpha_{y_i}\, p_{y_i}\left(\mathbf{x}_i|\,\boldsymbol{\theta}_{y_i}\right)$$

- Aussi, la distribution de la variable manquante est donnée par

$$P\left(y_i|\,\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right) = \frac{\alpha_{y_i}^t\, p_{y_i}\left(\mathbf{x}_i|\,\boldsymbol{\theta}_{y_i}^t\right)}{\sum_{j=1}^m \alpha_j^t\, p_j\left(\mathbf{x}_i|\,\boldsymbol{\theta}_j^t\right)}$$

- Nous pouvons ainsi obtenir l'expression de $Q\left(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right)$, à maximiser par la méthode des multiplicateurs de Lagrange tenant compte de la contrainte $\sum_{j=1}^m \alpha_j = 1$

le cnam

EM pour mélange gaussien

- Considérons des composantes lois normales multidimensionnelles

$$p_j\left(\mathbf{x}|\,\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_j\right|^{1/2}}\, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)}$$

- Données manquantes : $\{Y_i\}_{i=1}^n : y \in \{1,\ldots,m\}$, $y_i = j$ si l'observation x_i a été générée par la composante j du mélange
- Logarithme de la vraisemblance totale ($\boldsymbol{\theta}$ correspond aux $\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j$) :

$$\ln p(\mathcal{D}_o,\mathcal{D}_m|\,\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{i=1}^n \ln p(\mathbf{x}_i, y_i|\,\boldsymbol{\alpha},\boldsymbol{\theta}) = \sum_{i=1}^n \ln \alpha_{y_i}\, p_{y_i}\left(\mathbf{x}_i|\,\boldsymbol{\theta}_{y_i}\right)$$

- Espérance du logarithme de la vraisemblance :

$$Q\left(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t\right) = \sum_y \left[\ \sum_{i=1}^n \ln \alpha_{y_i}\, p_{y_i}\left(\mathbf{x}_i|\,\boldsymbol{\theta}_{y_i}\right)\right] \prod_{i=1}^n \frac{\alpha_{y_i}^t\, p_{y_i}\left(\mathbf{x}_i|\,\boldsymbol{\theta}_{y_i}^t\right)}{\sum_{j=1}^m \alpha_j^t\, p_j\left(\mathbf{x}_i|\,\boldsymbol{\theta}_j^t\right)}$$

le cnam

EM pour mélange gaussien (2)

- Après quelques calculs on obtient

$$Q(\boldsymbol{\alpha},\boldsymbol{\theta};\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t) = \sum_{j=1}^{m}\sum_{i=1}^{n} P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t)\ln\alpha_j + \sum_{j=1}^{m}\sum_{i=1}^{n} P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t)\ln p_j(\mathbf{x}_i|\boldsymbol{\theta}_j)$$

avec $P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t) = \alpha_j^t p_j(\mathbf{x}_i|\boldsymbol{\theta}_j^t) \Big/ \sum_{l=1}^{m}\alpha_l^t p_l(\mathbf{x}_i|\boldsymbol{\theta}_l^t)$

- Les équations de mise à jour résultantes sont

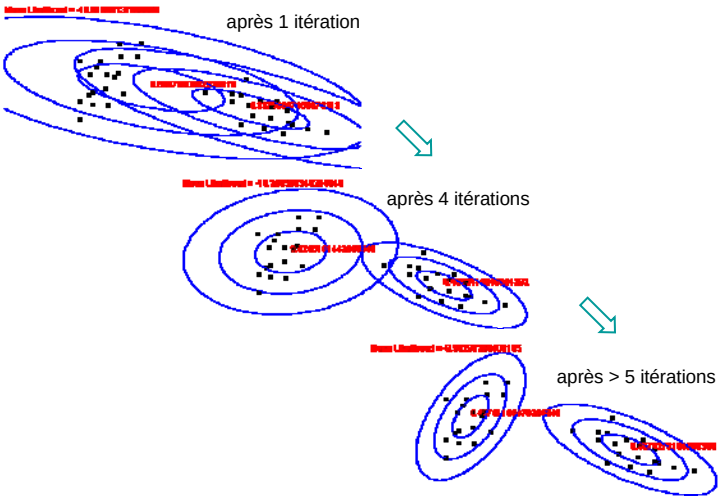
$$\alpha_j^{t+1} = \frac{1}{n}\sum_{i=1}^{n} P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t) \qquad \boldsymbol{\mu}_j^{t+1} = \frac{\sum_{i=1}^{n}\mathbf{x}_i P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t)}{\sum_{i=1}^{n} P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t)}$$

$$\boldsymbol{\Sigma}_j^{t+1} = \frac{\sum_{i=1}^{n} P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t)(\mathbf{x}_i-\boldsymbol{\mu}_j^{t+1})(\mathbf{x}_i-\boldsymbol{\mu}_j^{t+1})^T}{\sum_{i=1}^{n} P(j|\mathbf{x}_i,\boldsymbol{\alpha}^t,\boldsymbol{\theta}^t)}$$

- Exemples : [Akaho]

le cnam

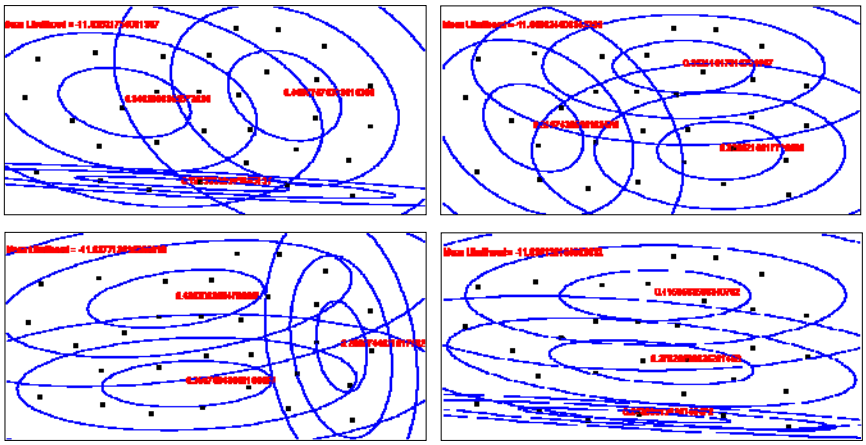
Exemple : 2 gaussiennes 2D



[Akaho]

le cnam

Exemple : données peu structurées



[Akaho]

Résultats très différents obtenus avec 4 initialisations différentes

le cnam

Mélanges et classification automatique

- Si on considère que chaque composante d'un mélange représente un groupe (*cluster*), trouver les paramètres du modèle permet d'identifier les groupes ($\mathbf{x}_i$ sera affecté à la composante $j^* = \arg\max_j P\left(j \middle| \mathbf{x}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}\right)$) et de modéliser chaque groupe
- Approche CML (*Classification Maximum Likelihood*) : on introduit des vecteurs-indicateurs $\mathbf{q}_i$, avec $q_{ij} = 1$ si $\mathbf{x}_i$ est généré par la composante j et $q_{ij} = 0$ sinon (voir aussi [CG92])
- Centres mobiles et EM pour modèles de mélange :
  - Mélange de lois normales multidimensionnelles avec $\alpha_j = 1/n, \forall j$, matrices de variances-covariances identité et $P\left(j \middle| \mathbf{x}_i, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\theta}}\right) = 1$ si et seulement si $\mathbf{x}_i$ est plus proche de $\boldsymbol{\mu}_j$

  ⇒ L'estimation des paramètres du modèle (seuls restent les $\boldsymbol{\mu}_j$) se réduit à l'algorithme de classification par centres mobiles

le cnam

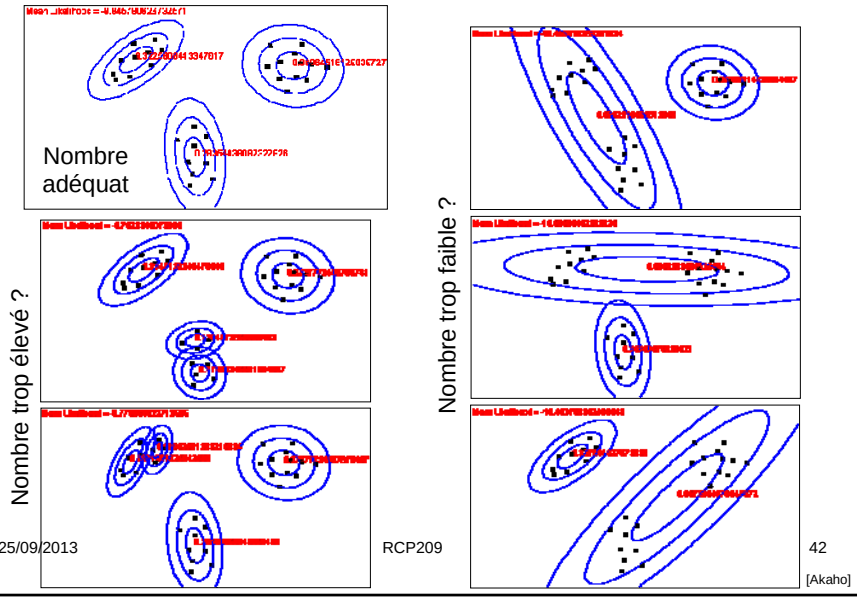
Choix du nombre de composantes

- Difficulté : pour un modèle de mélange, le maximum de la vraisemblance augmente en général avec le nombre de composantes, la vraisemblance ne peut donc pas servir au choix du nombre de composantes (sélection de modèle)
- Types de méthodes (voir la synthèse [OBM05]) :
  - Tests d'hypothèses : inspirés par LRTS (*likelihood ratio test statistic*, ne s'applique pas tel quel) pour *m* composantes vs. *m*+1 composantes
  - Critères d'information : pénaliser proportionnellement au nombre de paramètres (par exemple AIC, BIC)
  - Évaluer le degré de séparation entre composantes (les composantes peu séparées seront fusionnées) : mesure d'entropie (par exemple [CS96]), critères issus de la classification floue
  - Modifications de EM (par exemple [FJ02]) pour estimer le nombre de composantes en même temps que les paramètres

25/09/2013 RCP209 41

le cnam

Exemple : nombre de composantes



25/09/2013 RCP209 42

[Akaho]

le cnam

Paramétrique vs. non paramétrique

- Avantages des méthodes paramétriques :
  - Si les hypothèses sont valides, bonne estimation avec des échantillons de taille comparativement faible
  - Après construction du modèle, faible coût de l'estimation de la densité en un point précis
- Avantages des méthodes non paramétriques :
  - Généralité due à l'absence d'hypothèses sur le nombre et les types de lois
  - Convergence garantie vers la vraie densité… si l'échantillon est suffisant !
  - Paramètre unique… mais difficile à choisir ! De plus, une même valeur peut ne pas convenir à l'ensemble du domaine

le cnam

Références

[Akaho] Akaho S. Applet mélange de gaussiennes. (actuellement sur http://cedric.cnam.fr/~crucianm/appletMixMod.html)

[CG92] G. Celeux, G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14, 315-332, 1992.

[CS96] G. Celeux, G. Soromenho. An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. *Classification Journal*, vol. 13, pp. 195-212, 1996.

[DLR77] A. P. Dempster, N. M. Laird, D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, vol. 39, 1:1-38, 1977.

[FJ02] Figueiredo, M. A. and Jain, A. K. 2002. Unsupervised Learning of Finite Mixture Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 24, 3 (Mar. 2002), 381-396.

[OBM05] A. Oliveira-Brochado, F. V. Martins. Assessing the Number of Components in Mixture Models: a Review. *FEP Working Papers* 194, Universidade do Porto, 2005.