Systèmes Intelligents : Raisonnement et Reconnaissance

James L. Crowley

Deuxième Année ENSIMAG Deuxième Sémestre 2008/2009

Séance 10 4 mai 2009

Mélange de Gaussiennes et L'Algorithme EM (Expectation-Maximization)

Sources Bibliographiques :

"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Mélange de Gaussiens.

Si les événements sont issus d'une composition de "N" phénomènes, la densité $p(\vec{X})$ prendra la forme d'une composition de lois Normales.

Dans un tel cas, on peut approximer par une somme pondérée de densités Normales.

$$p(x) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x; \mu_n, \sigma_n)$$

Un tel somme est connu par la terme "mélange de Gaussiens"
pour chaque Gaussien il faut estimer trois paramètres :

$$\vec{\nu}_n = (\alpha_n, \mu_n, \sigma_n)$$

En total il y a $3 \cdot N$ paramètres à estimer.

$$\vec{\nu} = (\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2, \ldots, \alpha_N, \mu_N, \sigma_N)$$

Si toutes les μ et σ etait fixe (et égale), on pouvait calculer les α_n directement.
Mais parce qu'ils sont libres il faut les estimer par un processus iterative.
Un tel processus est composé de deux étapes.

Ces étapes fournissent une estimation des variables cachées.

Pour un mélange de Gaussiens, les variables cachées sont les "sources" pour des événements.

On suppose chaque événement est issu d'un des N sources.
Nous allons construire une table de probabilités. h(m, n)

h(m, n) = Pr{l'événement E_m est issu de la source N}

les probabilités, h(m, n), nous donnera les facteurs de Mélange, α_n, ainsi que μ_n, σ_n.
L'algorithme d'estimation s'appelle "Expectation-Maximization" ou "EM".

Ebauche de l'Algorithme EM

Soit une ensemble ("training set") de M observations $\mathbf{T} = \{X_m\}$.
Fait une premier estimation des paramètres $\vec{\nu}^{(0)}$ et puis on alterne "E" et "M".

E: Faire une estimation des valeurs manquantes, h(m, n), pour les événements.

$h(m, n)^{(i)} = p(h_m | X_1, X_2, ..., X_M, \vec{\nu}^{(i)})$ pour chaque terme "n".

$$h(m, n)^{(i)} = \frac{\alpha_n^{(i)} \mathcal{N}(X_m; \mu_n^{(i)}, \sigma_n^{(i)})}{\sum_{j=1}^{N} \alpha_j^{(i)} \mathcal{N}(X_m; \mu_j^{(i)}, \sigma_j^{(i)})}$$

M: Recalculer $\vec{\nu}^{(i+1)}$ avec $p(h_m | X_1, X_2, ..., X_M, \vec{\nu}^{(i)})$

$$S_n^{(i+1)} = \sum_{m=1}^{M} h(m,n)^{(i)}$$

$$\alpha_n^{(i+1)} = \frac{1}{M} S_n^{(i+1)} = \frac{1}{M} \sum_{m=1}^{M} h(m,n)^{(i)}$$

$$\mu_n^{(i+1)} = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} X_m$$

$$\sigma^2{}_n^{(i+1)} = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} (X_m - \mu_n^{(i+1)})^2$$

Pour la dérivation l'algorithme EM, il faut introduire le concept de "likelihood" (vraisemblance) .

Maximisation de la Vraisemblance

Nous commençons par le cas ou D = 1. (une seule caractéristique)
Soit une ensemble de M exemples de caractéristiques $\{X_m\}$.
Les exemples sont supposés d'être les échantillons independents .

On souhaite estimer une forme paramétrique pour p(x).
Supposons que $p(x) = \mathcal{N}(x; \mu,\sigma)$.

Pour une loi normale avec D = 1 les paramètres sont

$$\vec{\nu} = (\mu,\sigma)$$

On souhait trouve une estimation $\hat{\nu} = (\hat{\mu},\hat{\sigma})$ qui minimise la probabilité d'erreur.
Pour ceci on définit le Likelihood $L(\vec{\nu} \mid X_1, X_2, ..., X_M)$ de $\hat{\nu}$ étant donnée $\{X_m\}$
Si les X_m sont indépendants,

$$p(X_1, X_2 \mid \vec{\nu}) = p(X_1 \mid \vec{\nu}) \cdot p(X_2 \mid \vec{\nu})$$

En générales pour M événements :

$$p(\{X_m\} \mid \vec{\nu}) = p(X_1, X_2, ..., X_M \mid \vec{\nu}) = \prod_{m=1}^{M} p(X_m \mid \vec{\nu})$$

Nous définissons le "Likelihood" (vraisemblance) de $\vec{\nu}$ étant $\{X_m\}$ comme

$$L(\vec{\nu} \mid \{X_m\}) = L(\vec{\nu} \mid X_1, X_2, ..., X_M) = p(X_1, X_2, ..., X_M \mid \vec{\nu})$$
$$= \prod_{m=1}^{M} p(X_m \mid \vec{\nu})$$

Notre objectif est d'estimer les paramètres $\hat{\nu}$ pour maximiser $L(\vec{\nu} \mid S)$.

$$\hat{\nu} = \arg\max_{\nu} \{ L(\vec{\nu} \mid \{X_m\})\} = \arg\max_{\nu} \{ \prod_{m=1}^{M} p(X_m \mid \vec{\nu})\}$$

Pour simplifier l'analyse, on travaille avec la log : $l(\vec{\nu}) = \text{Log}\{L(\vec{\nu} \mid S)\}$.

$$l(\vec{\nu}) = \text{Log}\{L(\vec{\nu} \mid \{X_m\})\} = \text{Log}\{ p(\{X_m\} \mid \vec{\nu})\} = \sum_{m=1}^{M} \text{Log}\{p(X_m \mid \vec{\nu})\}$$

Si $p(X_m \mid \vec{\nu})$ est une simple loi normale, il suffit de trouver

$$\hat{\nu} = \arg\max_{\nu}\{ \prod_{m=1}^{M} p(X_m \mid \vec{\nu})\}$$

Ceci nous donne les formes habituelles des moments $\hat{\nu} = (\mu,\sigma)$

Maximum de vraisemblance pour le cas univariate

Soit un modèle normale $\mathcal{N}(\mu,\sigma)$. Pour estimer μ,σ :

$$l(\vec{\nu}) = \text{Log}\{p(X_m \mid \vec{\nu})\} = -\frac{1}{2} \text{Log}\{2\pi\sigma^2\} - \frac{1}{2\sigma^2}(X_m - \mu)^2$$

$$\frac{\partial l(\vec{\nu})}{\partial \mu} = \frac{1}{\sigma^2}(X_m - \mu)$$

$$\frac{\partial l(\vec{\nu})}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4}$$

$$\nabla_{\mu,\sigma}\, l(\vec{\nu}) = \begin{pmatrix} \frac{1}{\sigma^2}(X_m - \mu) \\ -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4} \end{pmatrix}$$

La maximum se trouve ou le dérivé est nul.

$$\frac{\partial l(\vec{\nu})}{\partial \mu} = \sum_{m=1}^{M} \frac{1}{\sigma^2}(X_m - \hat{\mu}) = 0.$$

Avec un peu d'algèbre on a :

$$\sum_{m=1}^{M} \frac{1}{\sigma^2}(X_m - \hat{\mu}) = 0.$$

$$\frac{1}{\sigma^2}\sum_{m=1}^{M} X_m = \frac{1}{\sigma^2}\sum_{m=1}^{M} \hat{\mu}$$

$$\sum_{m=1}^{M} X_m = M\,\hat{\mu}$$

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

et de la même façon pour

$$\frac{\partial l(\vec{\nu})}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4} = 0$$

$$\sum_{m=1}^{M} -\frac{1}{2\sigma^2} + \frac{(X_m - \mu)^2}{2\sigma^4} = 0$$

$$\sum_{m=1}^{M} \frac{1}{2\sigma^2} = \sum_{m=1}^{M} \frac{(X_m - \mu)^2}{2\sigma^4}$$

$$\frac{1}{2\sigma^2}\sum_{m=1}^{M} 1 = \frac{1}{2\sigma^4}\sum_{m=1}^{M} (X_m - \mu)^2$$

$$M = \frac{1}{\sigma^2}\sum_{m=1}^{M} (X_m - \mu)^2 \qquad \hat{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M} (X_m - \hat{\mu})^2$$

Le cas multi-variate.

Pour $\vec{x}$ composé de D caractéristiques, avec M exemples d'une classe $\mathbf{T} = \{\vec{X}_m\}$
Issues un modèle normale $\mathcal{N}(\vec{\mu}, \mathbf{C})$.

Le problème est d'estimer les

$$\hat{\nu} = \max_{\nu}\{L(\vec{\nu} \mid \mathbf{T})\} = \max_{\nu}\{\prod_{m=1}^{M} p(\vec{X}_m \mid \vec{\nu})\}$$

Le maximum de $\vec{\nu}$ est trouvé quand le gradient est null.

$$\nabla_{\nu} l(\vec{\nu}) = \sum_{m=1}^{M} \nabla_{\nu} \text{Log}\{p(\vec{X}_m \mid \vec{\nu}) = 0$$

où le gradient est $\nabla_{\nu} = \begin{pmatrix} \frac{\partial}{\partial \nu_1} \\ \dots \\ \frac{\partial}{\partial \nu_{2D}} \end{pmatrix}$

Le gradient nous permet de calculer une solution analytique.
Les estimations de $\hat{\mu}, \hat{C}$ sont obtenus par

$$\text{Log}\{p(\vec{X}_n \mid \vec{\nu})\} = -\frac{1}{2} \text{Log}\{2\pi^n \det(\mathbf{C})\} - \frac{1}{2}(\vec{X}_n - \vec{\mu})^T \mathbf{C}^{-1}(\vec{X}_n - \vec{\mu})$$

$$\sum_{m=1}^{M} \nabla_{\mu,C} \text{Log}\{p(\vec{X}_m \mid \vec{\nu})\} = 0$$

Donne la formule classique :

$$\hat{\mu} = \frac{1}{M}\sum_{m=1}^{M} \vec{X}_m \quad \text{et} \quad \hat{C} = \frac{1}{M}\sum_{m=1}^{M}(X_m - \hat{\mu})(X_m - \hat{\mu})^T$$

L'Algorithme EM (Expectation-Maximization)

L'algorithme EM s'applique à l'estimation de données cachés.
Il est utilisé notamment pour l'estimation des Modèles de Markov Cachées (HMM's) et pour l'estimation des Mélanges de Gaussiens.

Pour un mélange de Gaussiens, les variables cachées sont les sources des événements.
On suppose que chaque événement est produit par un des N sources.
Pour chaque événement E_m On définit la variable "cachée" h_m

$h_m = N$ si l'événement E_m (avec caractéristique X_m) est issu de la source N.
Nous cherchons à estimer

$$p(x) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x; \mu_n, \sigma_n) \quad \text{Tels que} \quad \sum_{n=1}^{N} \alpha_n = 1.$$

Il faut estimer N vecteurs : $\nu_n = (\alpha_n, \mu_n, \sigma_n)$
En total il y a $3 \cdot N$ paramètres à estimer.

$$\vec{\nu} = (\alpha_1, \mu_1, \sigma_1, \alpha_2, \mu_2, \sigma_2, \ldots, \alpha_N, \mu_N, \sigma_N)$$

$$l(\vec{\nu} \mid \{X_m\}) = \text{Log}\,(L(\vec{\nu} \mid X_1, X_2, \ldots, X_M)) = \text{Log}\ \{ \prod_{m=1}^{M} p(X_m|\vec{\nu}\,)$$

$$= \sum_{m=1}^{M} \text{Log}\ \{p(X_m|\vec{\nu}\,)\}$$

$$= \sum_{m=1}^{M} \text{Log}\ \Big(\sum_{n=1}^{N} \alpha_n\, p(X_m \mid \nu_n) \Big)$$

Il faut $\nu_S = \max_{\nu} \{ l(\vec{\nu} \mid \{X_m\})\}$

Il n'y pas de solution analytique, parce qu'il n'y est pas une dérivé pour la logarithme de la somme.

Soit $\mathbf{T} = \{X_m\}$ $h(m,n)$ le donnée Complèt.
On vas maximiser une mesure de qualité $Q(\vec{\nu}^{(i)})$ définit par une esperence conditionnel

$$Q(\vec{\nu}^{(i)}) = E\{ l(\vec{\nu}^{(i)}) \mid \mathbf{T}\} = \text{Log}\ \{ \prod_{m=1}^{M} p(X_m|\vec{\nu}\,) \}$$

L'etape E

Pour chaque événement E_m avec caractéristique X_m,
On suppose qu'il manque l'information: h_m

h_m = n, la source de l'événement E_m.

On ne connaît pas h_m, mais on peut estimer les probabilités $P(h_m = n)$.
par une table h(m,n).

Pour chaque X_m, et son source caché h_m

$$p(h_m, X_m \mid \vec{\nu}) = p(h_m \mid X_m, \vec{\nu})\, p(X_m \mid \vec{\nu})$$

donc

$$p(h_m \mid X_m, \vec{\nu}) = \frac{p(h_m, X_m \mid \vec{\nu})}{p(X_m \mid \vec{\nu})}$$

où

$$p(h_m=n, X_m \mid \vec{\nu}) = \alpha_n \mathcal{N}(X_m; \mu_n, \sigma_n)$$

$$p(X_m \mid \vec{\nu}) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x_m; \mu_n, \sigma_n)$$

Donc

$$p(h_m=n \mid X_1, X_2, ..., X_M, \vec{\nu}_j) = \frac{\alpha_n \mathcal{N}(X_m; \mu_n, \sigma_n)}{\sum_{j=1}^{N} \alpha_j \mathcal{N}(X_m; \mu_j, \sigma_j)}$$

Donc pour chaque itération (i) le premier étape E est :

$$h(m, n)^{(i)} = \frac{\alpha_n^{(i)} \mathcal{N}(X_m; \mu_n^{(i)}, \sigma_n^{(i)})}{\sum_{j=1}^{N} \alpha_j^{(i)} \mathcal{N}(X_m; \mu_j^{(i)}, \sigma_j^{(i)})}$$

L'etape M

Pour le deuxième étape, M, nous allons maximiser le "likelihood" $\text{Log}\{p(X_m \mid \vec{\nu})\}$
On défini la "Expected Complete Data Log Likelihood" pour **T**. = $\{X_m\}$ U h(m,n)

$$\Delta Q = Q(\vec{\nu}^{(i)}) - Q(\vec{\nu}^{(i-1)})$$

Pour chaque cycle dans l'itération nous allons chercher :

$$\vec{\nu}^{(i+1)} = \underset{\nu}{\text{argmax}}\{Q(\vec{\nu} \mid \vec{\nu}^{(i)})\}$$

Ce maximum est donné par :

$$S_n^{(i+1)} = \sum_{m=1}^{M} p(h_m=n \mid X_m, \vec{\nu}^{(i)})) = \sum_{m=1}^{M} h(m, n)$$

$$\alpha_n^{(i+1)} = \frac{1}{M} \sum_{m=1}^{M} P(h_m=n \mid X_m, \vec{\nu}^{(i)}) = \frac{1}{M} S_n^{(i+1)}$$

$$\mu_n^{(i+1)} = \frac{1}{S_n^{(i)}} \sum_{m=1}^{M} p(h_m=n \mid X_m, \vec{\nu}^{(i)})\, X_m = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,.n)\, X_m$$

$$\sigma^2{}_n^{(i+1)} = \frac{1}{S_n^{(i)}} \sum_{m=1}^{M} p(h_m=n \mid X_m, \vec{\nu}^{(i)})\ (X_m - \mu_n^{(i+1)})^2.$$

$$= \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)\ (X_m - \mu_n^{(i+1)})^2$$

Avec notre table de probabilités h(m, n)

$$h(m, n) = P(h_m=n \mid X_m, \vec{\nu}^{(i)})$$

E (Expectation) :

$$h(m, n)^{(i)} := \frac{\alpha_n^{(i)} \mathcal{N}(X_m; \mu_n^{(i)}, \sigma_n^{(i)})}{\sum_{j=1}^{N} \alpha_j^{(i)} \mathcal{N}(X_m; \mu_j^{(i)}, \sigma_j^{(i)})}$$

M: (Maximisation)

$$S_n^{(i+1)} := \sum_{m=1}^{M} h(m, n)^{(i)}$$

$$\alpha_n^{(i+1)} := \frac{1}{M} S_n^{(i+1)}$$

$$\mu_n^{(i+1)} := \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,.n)^{(i)} X_m$$

$$\sigma^2{}_n^{(i+1)} := = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} (X_m - \mu_n^{(i+1)})^2$$

Dans le cas Multivariate (D> 1) la covariance C est composée de σ_{jk}^2:

$$\sigma^2{}_{jkn}^{(i+1)} := = \frac{1}{S_n^{(i+1)}} \sum_{m=1}^{M} h(m,n)^{(i)} (X_{jm} - \mu_{jn}^{(i+1)})(X_{km} - \mu_{kn}^{(i+1)})$$