

Méthodes en classification automatique

Modèle de mélange,
l'algorithme EM

Yves Lechevallier
INRIA-Rocquencourt
78153 Le Chesnay Cedex
E_mail : Yves.Lechevallier@inria.fr

Master ISI-10

Classe « homogène »

Classe P_k	Critère
<p>Approche géométrique</p> <div style="border: 1px solid black; padding: 2px; width: fit-content; margin: 5px auto;">d distance</div>	<div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 5px auto;"> $w(P_k) = \sum_{e_i \in P_k} \sum_{e_j \in P_k} d^2(\mathbf{z}_i, \mathbf{z}_j)$ </div>
<p>Modèle probabiliste</p> <div style="border: 1px solid black; padding: 2px; width: fit-content; margin: 5px auto;"> $p(\mathbf{z}/\theta) = \sum_{j=1}^K p(\mathbf{z}/\theta_j) \cdot \pi_j$ </div>	<div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 5px auto;"> $L(P_k / \theta_k) = \prod_{e_i \in P_k} p(\mathbf{z}_i / \theta_k)$ </div>
<p>Prototype</p> <div style="border: 1px solid black; padding: 2px; width: fit-content; margin: 5px auto;">L_k prototype</div>	<div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 5px auto;"> $w(P_k, L_k) = \sum_{e_i \in P_k} D(\mathbf{z}_i, L_k)$ </div>

Master ISI-10

Modèles de mélange

On suppose que la structure probabiliste est connue à l'exception des valeurs des paramètres du modèle de mélange:

- Le nombre K de classes est connu.
- Les probabilités a priori π_j des classes sont connues.
- La forme des lois de probabilités $p(\mathbf{z}/\theta_j)$ conditionnellement aux K classes est connue.
- Les valeurs des paramètres $\theta = (\theta_1, \dots, \theta_K)$ des densités de probabilités sont inconnues.

La loi de probabilité sur D est donnée par $\forall \mathbf{z} \in D \quad p(\mathbf{z}/\theta) = \sum_{j=1}^K p(\mathbf{z}/\theta_j) \cdot \pi_j$

Cette loi de probabilité est la *loi de probabilité d'un mélange de K composants*.

Master ISI-10

Exemple

Les variances et les probabilités a priori sont égales

La taille moyenne des femmes est égale à 1,67
La taille moyenne des hommes est égale à 1,76
 $\mu_1 = 1,67$ et $\mu_2 = 1,76$

$L_k(x)$

Master ISI-10

Loi de probabilité est identifiable

L'objectif est d'utiliser l'échantillon App pour l'estimation des paramètres de cette loi de probabilité

La population théorique Ω est connue. Dans ce cas nous pouvons déterminer la valeur $p(\mathbf{x}/\theta)$ pour tout $\mathbf{x} \in D$

la loi de probabilité $p(\cdot/\theta)$ est **identifiable** si
si $\theta \neq \theta' \Rightarrow \exists \mathbf{z} \in D$ tel que $p(\mathbf{z}/\theta) \neq p(\mathbf{z}/\theta')$



Estimation

Un **estimateur** est une variable aléatoire, fonction d'un N -échantillon d'une distribution de paramètre θ .

$$R(\theta, \hat{\theta}) = \int C(\theta, \hat{\theta}(\mathbf{x})) f(\mathbf{x}/\theta) d\mathbf{x}$$

Où $C(\theta, \hat{\theta}(\mathbf{x}))$ est le coût de choisir $\hat{\theta}(\mathbf{x})$ alors que la vraie valeur est θ .

Si c'est C est un coût quadratique alors

$$R(\theta, \hat{\theta}) = E_{\theta} \left[(\theta - \hat{\theta})^2 \right] = \text{Var}[\hat{\theta}] + (E_{\theta}[\hat{\theta}] - \theta)^2$$



Estimateur du maximum de vraisemblance (1/3)

Supposons que nous avons un **échantillon** $E = (\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_N)$ dont la loi de probabilité p est égale à:

$$p(\mathbf{z}/\theta) = \sum_{j=1}^K p(\mathbf{z}/\theta_j) \cdot \pi_j$$

La **vraisemblance** de cet échantillon est égale à:

$$L(E/\theta) = \prod_{i=1}^N p(\mathbf{z}_i/\theta)$$

L'estimateur $\hat{\theta}$ du **maximum de vraisemblance** est

$$\hat{\theta} = \arg \max_{\theta} L(App/\theta)$$



Estimateur du maximum de vraisemblance (2/3)

Si la **vraisemblance** est **dérivable** en fonction de θ et que les paramètres des composants sont **fonctionnellement indépendants** alors le **gradient du logarithme de la vraisemblance** est égal à:

$$\frac{\partial}{\partial \theta_j} \log(L(App/\theta)) = \sum_{i=1}^N \frac{1}{p(\mathbf{z}_i/\theta)} \cdot \frac{\partial}{\partial \theta_j} p(\mathbf{z}_i/\theta) \cdot \pi_j$$

$P(j/\mathbf{z}_i, \theta)$ est la **probabilité a posteriori** (i.e. la probabilité qu'une observation \mathbf{z} appartienne à la classe j sachant θ fixé)

alors ce **gradient** s'écrit sous une forme plus intéressante:

$$\frac{\partial}{\partial \theta_j} \log(L(App/\theta)) = \sum_{i=1}^N P(j/\mathbf{z}_i, \theta) \cdot \frac{\partial}{\partial \theta_j} \log(p(\mathbf{z}_i/\theta))$$



Estimateur du maximum de vraisemblance (3/3)

Alors l'estimateur $\hat{\theta}$ du **maximum de vraisemblance** doit vérifier les conditions:

$$(I) \sum_{i=1}^N P(j/\mathbf{z}_i, \hat{\theta}) \cdot \frac{\partial}{\partial \theta_j} \log(p(\mathbf{z}_i/\hat{\theta}_j)) = 0 \text{ pour } j = 1, \dots, K$$

On peut généraliser ces résultats en supposant que les probabilités a priori π_j des classes sont inconnues. Dans ce cas l'estimateur $\hat{\pi}_j$ de la **probabilité a priori** π_j et l'estimateur $\hat{P}(j/\mathbf{z}, \theta)$ de la **probabilité a posteriori** de la classe j sont égaux à:

$$(II) \hat{\pi}_j = \frac{1}{N} \sum_{i=1}^N \hat{P}(j/\mathbf{z}_i, \hat{\theta}_j) \text{ et } \hat{P}(j/\mathbf{z}, \hat{\theta}_j) = \frac{p(\mathbf{z}_i/\hat{\theta}_j) \cdot \hat{\pi}_j}{\sum_{i=1}^K p(\mathbf{z}_i/\hat{\theta}_j) \cdot \hat{\pi}_i}$$

Une résolution efficace de ces équations de vraisemblance peut être obtenue par les algorithmes EM (Dempster et al, 1977).



Principe de EM

$App = (\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_N)$ un N -échantillon

Chaque $\mathbf{z}_i = (z_{ib}, z_{im})$ est composé d'une partie complète z_{ib} et d'une partie manquante z_{im} .

θ_i est fixé et on prend comme critère espérance mathématique sur les valeurs manquantes du log de la vraisemblance

$$Q(\theta, \theta') = \int_{D_m} \log[\Pr[z/\theta]] \Pr[z_m/z_b; \theta'] dz_m$$

Initialisation : fixer $\theta_0, \epsilon, t=0$

Étape E (estimation) : calculer $Q(\theta, \theta_0)$

Étape M (maximisation) : $\theta_{t+1} = \arg \max_{\theta} Q(\theta, \theta_t)$

Arrêt si $Q(\theta, \theta_t) - Q(\theta, \theta_{t+1}) < \epsilon$



Algorithme EM en classification

Posons le problème de l'estimation de $\Theta = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$ sous une forme traitable par le principe d'information manquante.

L'échantillon complet s'écrit $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ avec $\mathbf{x}_i = (\mathbf{z}_i, \mathbf{y}_i)$, \mathbf{y}_i est un vecteur qui indique de quelle composante du mélange est issu \mathbf{x}_i . $y_i = (y_{ik}, k = 1, \dots, K)$, $y_{ik} \in \{0, 1\}$ et $\sum_{k=1}^K y_{ik} = 1$

$y_{ik} = 1$ signifie que e_i provient de la composante k .

Log-vraisemblance complétée

$$\sum_{i=1}^N \sum_{k=1}^K y_{ik} \log(\pi_k \cdot f_k(z_i/\theta_k))$$



Algorithme EM

Initialisation $\Theta = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$ et K fixés. $\theta_1, \dots, \theta_K$ dépendent des hypothèses sur les lois

Étape d'Estimation : calcul des probabilités a posteriori et a priori

Pour $i=1, \dots, N$ et $k=1, \dots, K$ on calcule

$$c_{ik} = \Pr[y = k / z_i; \Theta] = \pi_k f_k(z_i/\theta_k) / f(z_i)$$

$$\pi_k = \sum_{i=1}^N \Pr[y = k / z_i; \Theta] / N$$

Étape de Maximisation : calcul de Θ qui maximise

$$\Theta \leftarrow \arg \max_{\Gamma} Q(\Gamma, \Theta) = \sum_{i=1}^N \sum_{k=1}^K c_{ik} \cdot \log[\pi_k \cdot f_k(z_i/\Gamma)]$$



Distributions Gaussiennes

$$c_{ik} = \Pr[y = k / z_i; \Theta] = \pi_k f_k(z_i / \theta_k) / f(z_i)$$

avec $\theta_k = (\mu_k, \Sigma_k)$

moyenne

$$\mu_k = \frac{\sum_{i=1}^N c_{ik} \cdot z_i}{\sum_{i=1}^N c_{ik}}$$

Matrice de variance-covariance

$$\Sigma_k = \frac{\sum_{i=1}^N c_{ik} \cdot (z_i - \mu_k)(z_i - \mu_k)}{\sum_{i=1}^N c_{ik}}$$

Master ISI-10



Stochastique EM

Cet algorithme (Celeux, Diebolt, 1986) introduit une étape supplémentaire entre E et M:

Étape S (stochastique)

Cette étape suit le principe de l'affectation stochastique

Pour chaque z_i , on tire au hasard $y_i = (y_{ik}, k=1, \dots, K)$ avec $y_{ik} = 1$ si e_i est affecté à la classe k , en fonction d'une loi multinominale de paramètres (c_{i1}, \dots, c_{iK}) qui sont les probabilités a posteriori.

La matrice y définit une partition « dure » sur E

Master ISI-10



Classification et EM

Cet algorithme (Celeux, Govaert, 1992) introduit une étape supplémentaire entre E et M:

Étape C (classification)

Cette étape suit le principe du MAP (maximum a posteriori)

Pour chaque z_i est affecté à la composante k du mélange de plus forte probabilité a posteriori, $k = \arg \max \{c_{i1}, \dots, c_{iK}\}$.

La matrice y définit aussi une partition « dure » sur E

Critère optimisé

$$W(P, \Theta) = \sum_{i=1}^N \sum_{k=1}^K \Pr[y = k / z_i; \Theta] \cdot \log[\pi_k f_k(z_i / \Theta)] = \sum_{i=1}^N \sum_{k=1}^K c_{ik} \log[f_k(z_i / \Theta)] + \sum_{k=1}^K n_k \log \pi_k$$

Vraisemblance classifiante

Master ISI-10



Distributions Gaussiennes

$$c_{ik} = \Pr[y = k / z_i; \Theta] = \pi_k f_k(z_i / \theta_k) / f(z_i)$$

avec $\theta_k = (\mu_k, \Sigma_k)$

Critère optimisé par CEM

$$W(P, \Theta) = \sum_{i=1}^N \sum_{k=1}^K c_{ik} \left(-\frac{p}{2} \log(2\pi |\Sigma_k|) - \frac{1}{2} (z_i - \mu_k) \Sigma_k^{-1} (z_i - \mu_k) \right) + \sum_{k=1}^K n_k \log \pi_k$$

Si Σ est égal à I alors CEM est exactement l'algorithme des centres mobiles

Logiciel : MIXMOD

Master ISI-10



K-means dans le cas des mélanges

C'est une **méthode alternative** de la méthode du maximum de vraisemblance.

Cette approche impose que les **probabilités a posteriori** doivent être toutes égales à 0 ou à 1.

Dans le cas de classes très bien séparées les probabilités a posteriori sont toutes proches de 0 ou de 1, et la solution $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ du maximum de vraisemblance peut être alors approchée par une solution $\theta^* = (\theta_1^*, \dots, \theta_k^*)$ dont chaque composant θ_j^* est estimé uniquement à partir des observations de la classe j indépendamment des autres observations.



algorithme

(a) Étape 1

A partir d'une fonction d'affectation ϕ on calcule les estimateurs $\hat{\theta}_j$ du maximum de vraisemblance des classes Q_j qui vérifient:

$$\prod_{e_i \in Q_j} p(\mathbf{z}_i / \hat{\theta}_j) = \max_{\theta} \prod_{e_i \in Q_j} p(\mathbf{z}_i / \theta)$$

(b) Étape 2

A partir des estimateurs $\hat{\theta}_j$ on construit une nouvelle fonction d'affectation ϕ comme ceci:

$$\text{pour } e_i \in E \quad \phi(\mathbf{z}_i) = j \text{ si } p(\mathbf{z}_i / \hat{\theta}_j) = \max_{\ell=1, \dots, K} p(\mathbf{z}_i / \hat{\theta}_\ell)$$

(c) Si la nouvelle fonction d'affectation est différente de la précédente alors on va en (a).



SOFM et STPEM

SOFM : Self-Organizing Feature Maps

STPEM : Stochastic Topology Preserving EM

Mélange de densités $f(\mathbf{z} / \theta) = \sum_{j=1}^K f(\mathbf{z} / \theta_j) \pi_j$

La vraisemblance d'un échantillon est $L(E / \theta) = \prod_{i=1}^N f(\mathbf{z}_i / \theta)$

Le critère optimisé est égal à $W(E, T) = -\sum_{i=1}^T \sum_{j=1, \dots, K} K(\phi_i^*, \phi_j^*) \log f(\mathbf{z}_i, \theta_j)$

d'où la fonction de mise à jour suivante

$$\phi_j(t+1) = \phi_j(t) + \alpha_i K(i, c) \frac{\partial}{\partial \theta} \log(f(\mathbf{z}^{(i)}, \theta_i^{(i)}))_{\theta = \theta_j^{(i)}}$$



SOFM et STPEM

si on suppose que les densités suivent des lois normales alors nous retrouvons la formule de mise à jour de Kohonen

$$\phi_j(t+1) = \phi_j(t) + \alpha_i K(i, c) [z_i' - \phi_j(t)]$$

Dans ce cas la fonction de voisinage utilisée est

$$K(i, c) = \begin{cases} 1 & \text{si } d_{ic} \leq \sigma(t) \\ 0 & \text{sinon} \end{cases}$$

Ambroise C. et Govaert G. Constrained Clustering and Kohonen Self-Organizing Maps. *Journal of Classification* 13, 1996, 299-313.

