Bases de la reconnaissance des formes (SI221)

Classification automatique

M. Sigelle

Année 2012-2013

Dernière mise à jour : 3 janvier 2013

Plan du cours

- **introduction - rappels**
- **estimation des paramètres en données complètes**
- **estimation des paramètres en données incomplètes**

Rappels: chaînes de Markov cachées (HMMs)

◦ **hypothèses : deux processus stochastiques dont l'un est caché**

- M classes (états) : $1..M$
- on n'observe pas directement la séquence d'états

  $q = q_1,\ q_2 \ldots q_t, \ldots q_T$
- mais seulement la séquence de symboles (observations)

  $o = o_1,\ o_2 \ldots o_t, \ldots o_T$ produits par ces états
- durée T de la séquence d'observations : variable

$\rightarrow$ processus sous-jacent : "segmentation" du signal observé

Rappels (suite) : HMMs et GMMs

◦ **indépendance conditionnelle des observations**

$$P(o \ / \ q, \lambda) = P(o_1 \ldots o_T \ / \ q_1 \ldots q_T, \lambda) = \prod_{t=1}^{T} \underbrace{P(o_t \ / \ q_t, \lambda)}_{b_i(o_t)}$$

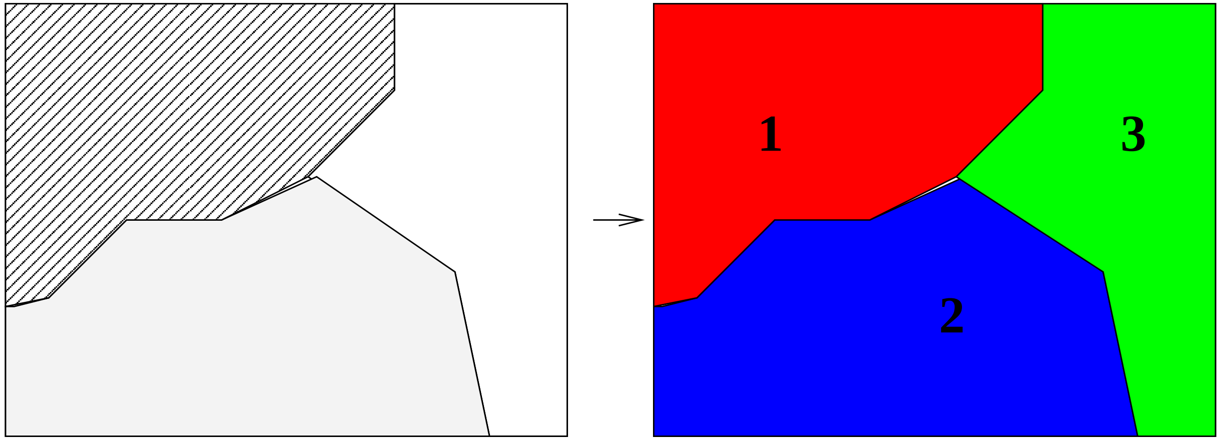
◦ **loi a priori sur q : chaîne de Markov stationnaire (HMMs)**
$\Rightarrow$ **modèle de poids (GMMs)**

$$P(q \ / \ \lambda) = \prod_{t=1}^{T} w(q_t) \qquad \left(\ + \ b_i(o_t) = \mathcal{N}(o_t \ ; \ \mu_i, \sigma_i^2) \ \right)$$

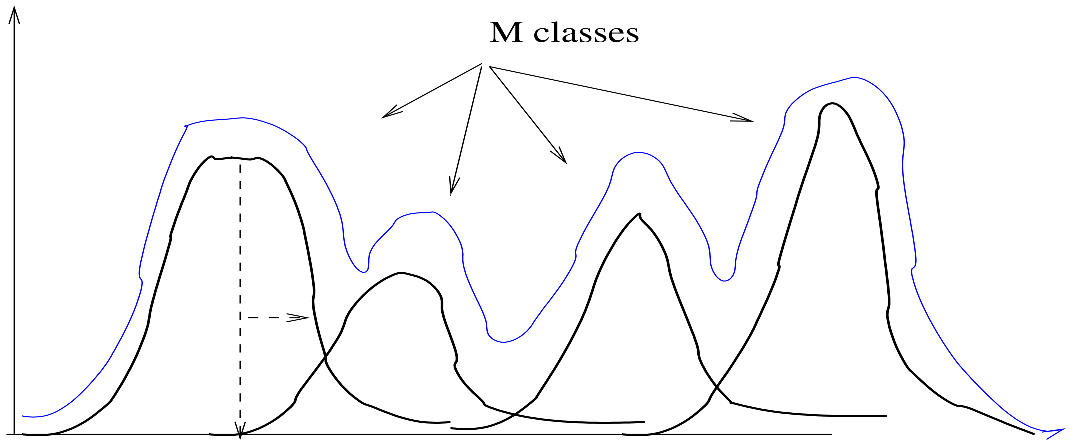
◦ **estimation de mélange/classification bayesienne ponctuelle/ GMM**

Problèmes liés en estimation des paramètres

- **données complètes : on connaît o et la séquence q associée**



- **données incomplètes: on connaît o $\rightarrow$ estimation de mélanges**



$$P(o_t = \xi) = \sum_{i=1}^{M} P(o_t = \xi \ / \ q_t = i) \ P(q_t = i) = \sum_{i=1}^{M} b_i(\xi) \ w_i$$

Estimation des paramètres en données complètes

- **vraisemblance jointe observations - états**

$$\mathcal{L} = P(o,\ q\ /\ \lambda) = P(o\ /\ q, \lambda)\ P(q\ /\ \lambda)$$

$$= \prod_{t=1}^{T} P(o_t\ /\ q_t, \lambda)\ w(q_t)$$

- **si plusieurs échantillons** $(o^{(l)}, q^{(l)})$ **indépendants** $\quad (l = 1..L)$

$$\mathcal{L} = \prod_{l=1}^{L} P(o^{(l)}, q^{(l)}\ /\ \lambda)$$

$$= \prod_{i=1}^{M} w_i^{N_i}\ .\ \prod_{l=1}^{L} \prod_{t=1}^{T^{(l)}} b_{q_t^{(l)}}(o_t^{(l)}) \quad \text{avec}$$

$$N_i = \sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i}$$

Estimation des paramètres en données complètes (suite)

- **maximum de vraisemblance ou (de) log-vraisemblance**

$$
\begin{aligned}
\log \mathcal{L} &= \sum_{i=1}^{M} N_i \log w_i + \sum_{i=1}^{M} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i} \; \log b_i(o_t^{(l)}) \\
&= \sum_{i=1}^{M} N_i \log w_i + \sum_{i=1}^{M} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i} \Big(\frac{(o_t^{(l)} - \mu_i)^2}{2\ \sigma_i^2} - \log \sigma_i \Big)
\end{aligned}
$$

Estimation des paramètres en données complètes (suite)

◦ **probabilités normalisées → estimateurs empiriques**

$$\frac{\partial \log P(o,\ q\ /\ \lambda)}{\partial w_i} = \frac{N_i}{w_i} \quad \Rightarrow$$

$$\widehat{w_i} = \frac{N_i}{S} = \frac{\displaystyle\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i}}{S} \qquad \left(S = \sum_{l=1}^{L} T^{(l)}\right)$$

◦ **lois d'observation gaussiennes → moyenne et variance empiriques**

$$\widehat{\mu_i} = \frac{\displaystyle\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} o_t^{(l)}\ \mathbb{1}_{q_t^{(l)}=i}}{\displaystyle N_i = \sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i}} \qquad \widehat{(\sigma_i)}^2 = \frac{\displaystyle\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \left(o_t^{(l)} - \widehat{\mu_i}\right)^2\ \mathbb{1}_{q_t^{(l)}=i}}{\displaystyle N_i = \sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \mathbb{1}_{q_t^{(l)}=i}}$$

→ lois multi-variées

Estimation des paramètres en données incomplètes

- **on ne connaît que les séquences d'observations (pas les états)**
- **notations : $\tilde{\mathbf{E}}[U] = \mathbf{E}[U \ / \ o, \lambda]$ espérance a posteriori**
- **paramètres (a priori) de poids**

$$\widehat{w_i} = \frac{\tilde{\mathbf{E}}[N_i]}{S} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} \tilde{\mathbf{E}}[\mathbb{1}_{q_t^{(l)}=i}]}{S} = \frac{\sum_{l=1}^{L} \sum_{t=1}^{T^{(l)}} P(q_t^{(l)} = i \ / \ o^{(l)}, \widehat{\lambda})}{S}$$

avec (Bayes + GMM)

$$P(q_t^{(l)} = i \ / \ o^{(l)}, \ \widehat{\lambda}) = P(q_t^{(l)} = i \ / \ o_t^{(l)}, \ \widehat{\lambda})$$

$$= \frac{P(o_t^{(l)} \ / \ q_t^{(l)} = i) \ \widehat{w_i}}{P(o_t^{(l)})} = \frac{P(o_t^{(l)} \ / \ q_t^{(l)} = i) \ \widehat{w_i}}{\sum_{j=1}^{M} P(o_t^{(l)} \ / \ q_t^{(l)} = j) \ \widehat{w_j}}$$

Estimation exacte des paramètres : d'où cela vient-il ?

◦ **on a vu que**

$$\frac{\partial \log P(o,\ q\ /\ \lambda)}{\partial w_i} = \frac{1}{w_i}\ N_i(q)$$

$$\begin{aligned}
\text{or } P(o\ /\ \lambda) &= \sum_q P(o,\ q\ /\ \lambda) \\
\frac{\partial P(o\ /\ \lambda)}{\partial a_{ij}} &= \frac{1}{w_i}\ \sum_q N_i(q)\ P(o,\ q\ /\ \lambda) \\
\frac{\partial \log P(o\ /\ \lambda)}{\partial w_i} &= \frac{1}{w_i}\ \sum_q N_i(q)\ \frac{P(o,\ q\ /\ \lambda)}{P(o\ /\ \lambda)} \\
&= \frac{1}{w_i}\ \tilde{\mathbf{E}}[N_i]
\end{aligned}$$

(normalisation des probabilités de transition $\sum_{i=1}^{M} w_i = 1 \quad \forall i$)

Estimation des paramètres en données incomplètes (suite)

- **paramètres des lois d'observations gaussiennes**

$$\widehat{\mu_i} = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ \tilde{\mathbf{E}}[\mathbb{1}_{q_t^{(l)}=i}]}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbf{E}}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})}$$

$$\widehat{(\sigma_i)^2} = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \left(o_t^{(l)}-\widehat{\mu_i}\right)^2\ \tilde{\mathbf{E}}[\mathbb{1}_{q_t^{(l)}=i}]}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbf{E}}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \left(o_t^{(l)}-\widehat{\mu_i}\right)^2\ P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)}=i\ /\ o^{(l)},\widehat{\lambda})}$$

- **$\Rightarrow$ équations auto-cohérentes $\rightarrow$ non solubles exactement !**

Algorithme EM

- **principe**

$$\lambda^{(\mathbf{n+1})} = \arg\max_{\lambda} \tilde{\mathbf{E}}^{(\mathbf{n})}[\,\log P(q,\ o\ /\ \lambda)\,]$$

- **$\Rightarrow$ la vraisemblance des observations croît au cours des itérations**

- **paramètres (a priori) de poids**

$$w_i{}^{(\mathbf{n+1})} = \frac{\tilde{\mathbf{E}}^{(\mathbf{n})}[N_i]}{S} = \frac{1}{S}\ \sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)} = i\ /\ o^{(l)}, \lambda^{(\mathbf{n})}) \quad \leftarrow \text{ appartenance !}$$

$$= \frac{1}{S}\ \sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \frac{b_i(o_t^{(l)})^{(\mathbf{n})}\ \ w_i{}^{(\mathbf{n})}}{\sum_{j=1}^{M} b_j(o_t^{(l)})^{(\mathbf{n})}\ \ w_j{}^{(\mathbf{n})}}$$

Algorithme EM (suite)

- **paramètres des lois gaussiennes**

$$\mu_i{}^{(\mathbf{n+1})} = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ \tilde{\mathbf{E}}^{(\mathbf{n})}[\mathbb{1}_{q_t^{(l)}=i}]}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} \tilde{\mathbf{E}}^{(\mathbf{n})}[\mathbb{1}_{q_t^{(l)}=i}]} = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} o_t^{(l)}\ P(q_t^{(l)} = i \ / \ o^{(l)}, \lambda^{(\mathbf{n})})}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)} = i \ / \ o^{(l)}, \lambda^{(\mathbf{n})})}$$

$$(\sigma_i)^{2\,(\mathbf{n+1})} = \quad '' \quad = \frac{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} (o_t^{(l)} - \mu_i{}^{(\mathbf{n})})^2\ P(q_t^{(l)} = i \ / \ o^{(l)}, \lambda^{(\mathbf{n})})}{\displaystyle\sum_{l=1}^{L}\sum_{t=1}^{T^{(l)}} P(q_t^{(l)} = i \ / \ o^{(l)}, \lambda^{(\mathbf{n})})}$$

- **exercice : remise à jour des matrices de covariance Σ_i**

Algorithme EM (suite) : propriétés

◦ **converge vers un maximum local de la vraisemblance**

◦ $\rightarrow$ **initialisation :** $[\ w_i, \mu_i, \sigma_i\ ]^{(\mathbf{0})}$

${w_i}^{(\mathbf{0})} = 1/M$

${\mu_i}^{(\mathbf{0})}$ équi-réparties dans le domaine admissible d'observations.

variances ${(\sigma_i)^2}^{(\mathbf{0})}$ en conséquence.

EM (suite) : transparents de Jean-Marie Nicolas

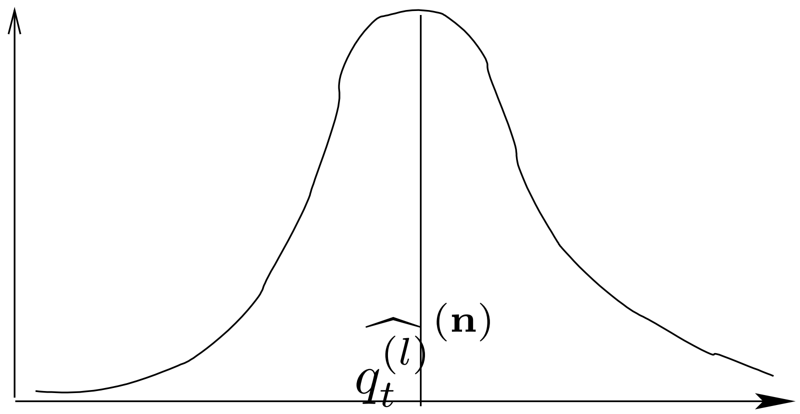
- **pp. 10-11 - 21-24**

2 gaussiennes $\mu_i = (50, 75)$

Même variance ($\sigma = 8$), w_i variable

w_1	m_1	m_2	# 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

Approximation de Viterbi - K-means



$$P(q_t^{(l)} = i \ / \ o^{(l)}, \widehat{\lambda}^{(\mathbf{n})}) \approx \mathbb{1}_{q_t^{(l)} = \widehat{q_t^{(l)}}^{(\mathbf{n})}}$$

$$\text{avec } \widehat{q_t^{(l)}}^{(\mathbf{n})} = \text{ estimateur MAP de } P(q_t^{(l)} \ / \ o_t^{(l)}, \widehat{\lambda}^{(\mathbf{n})})$$

- **estimation des paramètres pour la donnée complète** $(o^{(l)}, \widehat{q^{(l)}}^{(\mathbf{n})})$ **!**
- $\rightarrow$ **schéma itératif EM :**

$\lambda^{(\mathbf{n})} \rightarrow \quad \widehat{q^{(l)}}^{(\mathbf{n})}$ segmentation optimale au sens du MAP : Estimation !

$\widehat{q^{(l)}}^{(\mathbf{n})} \rightarrow \quad \lambda^{(\mathbf{n+1})}$ estimation pour la donnée complète obtenue : Maximization !

Application : K-means

- **hypothèse : mélange de gaussiennes avec :**

  - les poids des classes w_i sont tous égaux

  - les écarts-types des classes σ_i sont tous égaux

- **le schéma itératif EM s'écrit alors :**

$\mu^{(\mathbf{n})} \rightarrow {q^{(l)}}^{(\mathbf{n})}$ segmentation MAP : moyenne la plus proche Estimation !

${q^{(l)}}^{(\mathbf{n})} \rightarrow \mu^{(\mathbf{n+1})}$ estimation : moyennes empiriques des classes Maximization !

- **exercice : remise à jour des poids w_i et des écarts-types σ_i**

Conclusion

◦ **estimation des paramètres en données complètes**

$\rightarrow$ simple

◦ **estimation des paramètres en données incomplètes**

- classification ponctuelle : GMMs $\rightarrow$ facile
- dépendance temporelle (1D) entre états : HMMs $\rightarrow$ assez difficile
- dépendance spatiale (2D) entre états : MRFs $\rightarrow$ très difficile

◦ $\Rightarrow$ **on obtient du même coup une classification associée**

Bibliographie

◦ Livres

Apprentissage artificiel. Concepts et algorithmes A. Cornuéjols et L. Miclet. Eyrolles, 2002

◦ Articles

An Introduction to Hidden Markov Models. L.R. Rabiner and B.H. Juang, IEEE ASSP Magazine, 4-15, Jan. 1986.

A tutorial on Hidden Markov Models and selected applications in Speech Recognition. http://www.ai.mit.edu/courses/6.867-f02/papers/rabiner.pdf. L. Rabiner, Proceedings of the IEEE, 77(2):257-285, Feb. 1989.