Introduction à l'algorithme EM

- Exposition à EM par l'exemple

  Cette partie reprend la présentation faite à EMNLP'2001 par Ted Pedersen lors d'un pannel sur l'algorithme EM Pedersen [2001b]

- Fondements de EM Dempster et al. [1977], Baum [1972] (d'après Jelinek [1998])

- Application à l'estimation de coefficients de pondération dans une mixture de modèles (d'après Berger [2000])

Pedersen [2001a] propose une bonne liste de pointeurs sur EM.

EM par l'exemple

Soit n entités classées dans 4 catégories $Y = (y_1, y_2, y_3, y_4)$

Et soit $\theta = (\frac{1}{2} + \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$, les probabilités associées à chaque catégorie, définies par rapport à un paramètre π que l'on souhaite apprendre.

Alors la probabilité d'observer sur n tirages la classification en 4 classes selon des comptes y_1, y_2, y_3 et y_4 est donnée par la distribution multinomiale:

$$\mathcal{L}(\pi) = p(y_1, y_2, y_3, y_4) = \underbrace{\frac{n!}{y_1!y_2!y_3!y_4!}}_{\alpha} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}$$

Donc: $\log \mathcal{L}(\pi) = -\log \alpha + y_1 \log(\frac{1}{2} + \frac{1}{4}\pi) + (y_2 + y_3) \log(\frac{1}{4}(1-\pi)) + y_4 \log(\frac{1}{4}\pi)$

Rappel sur l'estimation par maximum de vraisemblance

$$\frac{\delta}{\delta\pi}\log\mathcal{L}(\pi) = y_1\frac{1/4}{(2+\pi)/4} + (y_2+y_3)\frac{-1}{1-\pi} + y_4\frac{1}{\pi}$$

Maximum si:

$$\frac{y_1}{2+\pi} - \frac{y_2+y_3}{1-\pi} + \frac{y_4}{\pi} = 0$$

Si on observe pour $n = 197$ les comptes suivants: $(125, 18, 20, 34)$ on trouve que l'équation admet une solution pour $\pi = 0.627$.

La connaissance des y_i est suffisante pour le calcul analytique du maximum de vraisemblance. On parle de statistique suffisante.

Changement des données du problème

En regardant mieux le problème, on s'apperçoit qu'il y a 5 classes pour classer nos n éléments et que y_1 est en fait la composition de deux classes: $x_1 + x_2$. De plus on sait que $p(x_1) = \frac{1}{2}$ et que $p(x_2) = \frac{1}{4}\pi$.

Problème: les expérimentations sont terminées et on ne peut plus faire de mesures des comptes de x_1 et x_2 (pas de chance...)

$\Longrightarrow$ On ne peut pas calculer le maximum de vraisemblance...

On dit qu'on est en présence d'une statistique (ou de données) incomplète(s).

Heureusement (ouf !), EM nous permet de contourner le problème. Son principe: prendre les espérances des comptes manquants afin d'obtenir une statistique suffisante pour calculer le maximum de vraisemblance. Boucler tant qu'on améliore la vraisemblance des données incomplètes.

La recette EM sur notre exemple

Données du nouveau problème:

$X = (x_1, x_2, y_2, y_3, y_4) = (x_1, x_2, 18, 20, 34)$ avec $x_1 + x_2 = y_1 = 125$

et $\theta = (\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$

Le M-STEP (Maximization)

$\log \mathcal{L}(\pi) = -\log \alpha' + x_1 \log(\frac{1}{2}) + (x_2 + y_4) \log(\frac{\pi}{4}) + (y_2 + y_3) \log(\frac{1-\pi}{4})$

D'où: $\frac{\delta}{\delta\pi} \log \mathcal{L}(\pi) = \frac{x_2+y_4}{\pi} - \frac{y_2+y_3}{1-\pi} = 0 \Longrightarrow \pi = \frac{x_2+y_4}{x_2+y_4+y_3+y_2}$

Mais x_2 n'est pas connu: estimons-le $(\hat{x_2})$!

C'est le **Le E-STEP** (Expectation).

Note: $X = (y_1 - \hat{x_2}, \hat{x_2}, y_2, y_3, y_4)$ constitue alors une statistique complète à partir de laquelle on peut calculer le maximum de vraisemblance

La recette EM sur notre exemple

On a 2 catégories, X_1 et X_2 avec pour comptes respectifs (inconnus) x_1 et x_2, sachant qu'il y a eu $y_1 = 125$ tirages indépendants avec:

- p_1, la chance d'avoir un élément classé dans X_1
- $p_2 = (1 - p_1)$ la probabilité d'avoir un élément dans X_2.

Soit $p(x_1)$ la probabilité d'obtenir x_1 occurrences de la classe X_1, si l'on fait l'hypothèse d'une distribution sous-jacente binomiale, on a:

$$p(x_1) = \binom{y_1}{x_1} {p_1}^{x_1} {p_2}^{y_1 - x_1} \text{ et on sait que } E[x_1|y_1] = y_1 p_1$$

Comme $p_1 = \frac{\frac{1}{2}}{\frac{1}{2}+\frac{1}{4}\pi}$, alors $\begin{cases} \hat{x_1} & = 125/\left(2 \times \left(\frac{1}{2} + \frac{1}{4}\pi\right)\right) \\ \hat{x_2} & = 125 - \hat{x_1} \end{cases}$

La recette EM sur notre exemple

INIT	$\pi \longleftarrow 0.5$	$\pi \longleftarrow 0.1$
E-STEP 1	$\hat{x_2} = 25$	$\hat{x_2} = 5.95$
M-STEP 1	$\pi \longleftarrow 0.6082$	$\pi \longleftarrow 0.5125$
E-STEP 2	$\hat{x_2} = 29.15$	$\hat{x_2} = 25.49$
M-STEP 2	$\pi \longleftarrow 0.6243$	$\pi \longleftarrow 0.6102$
E-STEP 3	$\hat{x_2} = 29.74$	$\hat{x_2} = 29.22$
M-STEP 3	$\pi \longleftarrow 0.6264$	$\pi \longleftarrow 0.6245$
E-STEP 4	$\hat{x_2} = 29.82$	$\hat{x_2} = 29.81$
M-STEP 4	$\pi \longleftarrow 0.6267$	$\pi \longleftarrow 0.6267$
E-STEP 5	$\hat{x_2} = 29.82$	$\hat{x_2} = 29.81$
M-STEP 5	$\pi \longleftarrow 0.6268$	$\pi \longleftarrow 0.6267$
E-STEP 6	$\hat{x_2} = 29.82$	$\hat{x_2} = 29.82$
M-STEP 6	$\pi \longleftarrow 0.6268$	$\pi \longleftarrow 0.6268$

EM: la recette

- On observe des réalisations y d'une distribution que l'on essaye de modéliser à l'aide d'un modèle paramétrique de paramètres λ.
- On fait l'hypothèse que y est une statistique incomplète et qu'il existe une variable cachée h dont la connaissance nous donne une statistique complète (y, h).
- **Note:** Cela revient à dire que l'on fait une hypothèse d'une distribution jointe qu'il est plus facile de modéliser que y: $p(y, h|\lambda)$
- L'idée de EM est de faire une estimée de la vraisemblance de la donnée complète $p(y, h)$ à partir de notre estimée courante λ' que nous utilisons pour estimer les nouveaux paramètres λ.
- Formellement:

$$Q(\lambda, \lambda') = E_h \left[\log p(y, h|\lambda)|y, \lambda'\right]$$

EM: la recette

INIT définir des valeurs initiales pour les paramètres λ'

E-STEP (Expectation): exprimer l'espérance (sur h, la variable cachée) de la donnée complète, sachant y (la donnée incomplète) et l'état actuel de nos connaissances sur les paramètres (λ'):

$$Q(\lambda, \lambda') = E_h[\log p_\lambda(y, h)|\lambda', y]$$

M-STEP (Maximization): $\hat{\lambda} = \text{argmax}_\lambda \, Q(\lambda, \lambda')$

LOOP sur E-STEP avec $\lambda' \longleftarrow \hat{\lambda}$ si la convergence n'est pas déjà atteinte

Cette recette, nous amène, lorsqu'elle est applicable à un maximum (souvent local) de la vraisemblance des données incomplètes.

Conditions d'application

- Être capable d'identifier la statistique suffisante et disposer d'un moyen de calculer les espérances des données manquantes
- Pouvoir résoudre le problème de maximisation sur la statistique complète
- Être sensible au fait que le choix des valeurs initiales des données manquantes peut conditionner le résultat de l'apprentissage (dans le cas général, la vraisemblance a de nombreux maxima locaux, et EM n'est garanti de trouver qu'un de ces maxima).

Légitimité de EM

Intéressons-nous au gain du maximum de vraisemblance que l'on obtient en changeant les paramètres λ' par λ $\log p(y|\lambda) - \log p(y|\lambda')$:

$$
\begin{aligned}
&= \overbrace{\sum_h p(h|y,\lambda')}^{1} \log p(y|\lambda) - \overbrace{\sum_h p(h|y,\lambda')}^{1} \log p(y|\lambda') \\
&= \textstyle\sum_h p(h|y,\lambda') \log p(y|\lambda) \frac{p(h,y|\lambda)}{p(h,y|\lambda)} - \sum_h p(h|y,\lambda') \log p(y|\lambda') \frac{p(h,y|\lambda')}{p(h,y|\lambda')} \\
&= \textstyle\sum_h p(h|y,\lambda') \log \frac{p(h,y|\lambda)}{p(h|y,\lambda)} - \sum_h p(h|y,\lambda') \log \frac{p(h,y|\lambda')}{p(h|y,\lambda')} \\
&= \textstyle\sum_h p(h|y,\lambda') \log p(h,y|\lambda) - \sum_h p(h|y,\lambda') \log p(h,y|\lambda') \\
&+ \textstyle\sum_h p(h|y,\lambda') \log p(h|y,\lambda') - \sum_h p(h|y,\lambda') \log p(h|y,\lambda) \\
&\\
&\geq \textstyle\sum_h p(h|y,\lambda') \log p(h,y|\lambda) - \sum_h p(h|y,\lambda') \log p(h,y|\lambda') \\
&= \textstyle\sum_h p(h|y,\lambda') \log \frac{p(h,y|\lambda)}{p(h,y|\lambda')}
\end{aligned}
$$

L'inégalité étant la résultante de l'application de l'inégalité de Jensen (ici $\sum_x p(x) \log \frac{p(x)}{q(x)} \geq 0$, égalité ssi $p = q$).

Légitimité de EM

Résumons:

$$\begin{array}{lrcl} \text{si} & \sum_h p(h|y,\lambda') \log p(h,y|\lambda) & > & \sum_h p(h|y,\lambda') \log p(h,y|\lambda') \\ \text{alors} & \log p(y|\lambda) & > & \log p(y|\lambda') \\ \text{cad} & p(y|\lambda) & > & p(y|\lambda') \end{array}$$

En d'autre termes, si on arrive à trouver λ tel que $\sum_h p(h|y,\lambda') \log p(h,y|\lambda) > \sum_h p(h|y,\lambda') \log p(h,y|\lambda')$, alors le modèle sous le régime λ ne peut que s'améliorer (sur les données d'entraînement !).

Il suffit de maximiser le terme de gauche:

$$\begin{array}{rcl} \Longrightarrow & & \text{argmax}_\lambda \sum_h p(h|y,\lambda') \log p(h,y|\lambda) \\ & = & \text{argmax}_\lambda \sum_h \underbrace{p(h|y,\lambda') \times p(y|\lambda')} \log p(h,y|\lambda) \\ & = & \text{argmax}_\lambda \sum_h p(h,y|\lambda') \log p(h,y|\lambda) \\ & = & \text{argmax}_\lambda Q(\lambda,\lambda') \end{array}$$

Application à l'estimation des coefficients d'un modèle combiné

Le problème: on a N modèles (connus) $p_1(y), p_2(y), \ldots, p_N(y)$ et un corpus d'entraînement $O = y_1, \ldots, y_T$, et:

$$\left\{ p_\lambda(y) = \sum_{i=1}^{N} \lambda_i p_i(y) \text{ avec } \lambda_i \in [0,1] \text{ et } \sum_{i=1}^{N} \lambda_i = 1 \right\}$$

On cherche $\hat{\lambda} = \text{argmax}_\lambda \log p_\lambda(y)$

On peut appliquer EM:

La variable cachée est l'état s ($s \in [1, N]$) dans lequel se trouve le modèle p_λ au moment de la prédiction.

Pensez: dans l'état i, le modèle combiné s'appuie sur le i-ème modèle.

Application à l'estimation d'un modèle combiné

$\hookrightarrow$ Notre fonction auxiliaire (λ' le jeu de paramètres courant) est:

$$Q(\lambda, \lambda') = \sum_y \tilde{p}(y) \sum_{i=1}^{N} p_{\lambda'}(s = i|y) \log p_\lambda(y, s = i)$$

où $\tilde{p}$ est la distribution **empirique**, et:

- λ_i est l'$apriori$ d'être dans l'état i
- $p_\lambda(s = i, y) = \lambda_i \times p_i(y)$ est la probabilité d'être dans l'état i et de générer y
- et $p_{\lambda'}(s = i|y) = \frac{\lambda'_i p_i(y)}{\sum_i \lambda'_i p_i(y)}$ est la probabilité d'être dans l'état i sachant que y est l'observation courante

Application à l'estimation d'un modèle combiné

EM nous dit de maximiser (sur λ) la fonction auxiliaire sous la contrainte que les coefficients somment à 1. On introduit pour cela un multiplicateur de Lagrange α:

$$\begin{array}{lcl}
\frac{\delta}{\delta\lambda_i}\left[Q(\lambda,\lambda') - \alpha\left(\sum_i \lambda_i - 1\right)\right] & = & 0 \\
\sum_y \tilde{p}(y) p_{\lambda'}(s=i|y) \frac{\partial}{\partial\lambda_i}\left[\log \lambda_i p_i(y)\right] - \alpha & = & 0 \\
\sum_y \tilde{p}(y) p_{\lambda'}(s=i|y) \frac{1}{\lambda_i} - \alpha & = & 0 \\
\frac{1}{\lambda_i} \underbrace{\sum_y \tilde{p}(y) p_{\lambda'}(s=i|y)}_{C_i} - \alpha & = & 0
\end{array}$$

où α joue le rôle d'un coefficient de normalisation.

Finalement: $\lambda_i = \frac{C_i}{\sum_i C_i}$

Application à l'estimation d'un modèle combiné

Notre algorithme EM revient donc à:

- Initiliser λ' (n'importe quel choix tel que $\sum_i \lambda'_i = 1$)
- Repéter jusqu'à convergence:
  E-step calcul des comptes C_i selon la formule précédente (c'est une fonction de λ')
  M-step pour tout i, $\lambda_i \leftarrow \frac{C_i}{\sum_i C_i}$

Note: si l'on voit C_i comme le nombre espéré de fois où le modèle i sera utilisé pour générer l'observation (étant donné un jeu de paramètre λ'), alors cet algorithme est assez intuitif.

Retour au modèle Jelinek-Mercer

Rappel: on a N modèles n-grammes (en pratique pour un trigramme, $N = 3$) que l'on combine avec des coefficients λ_i qui dépendent du contexte:

$$p(w_t|\overbrace{w_{t-2}w_{t-1}}^{h_t}) = \sum_{i=1}^{N} \lambda_i(\theta(h_t))p_i(w_t|h_{t,i})$$

où $\theta(h_t)$ est n'importe quelle fonction appliquée à l'historique (ex: $\theta(h) \longrightarrow \log|h|$) et $h_{t,i}$ désigne les $i-1$ derniers mots du contexte h_t.

On souhaite estimer les coefficients λ à maximum de vraisemblance: c'est une instance du modèle précédant à ceci près que les comptes sont conditionnés par les historiques.

Retour au modèle Jelinek-Mercer

Tout serait simple si nous connaissions la contribution $\hat{c}_i(\theta(w''w'))$ de chaque modèle p_i lors d'une prédiction dans le contexte $w''w'$.

On applique la recette EM de la combinaison linéaire de modèles:

E-STEP:

$$\hat{c}_i(\theta(w''w')) = \sum_{t=1:\theta(w_{t-2},w_{t-1})=\theta(w'',w')}^{T} \frac{\lambda_i(\theta(w''w'))p_i(w_t|h_{t,i})}{\sum_{i=1}^{N}\lambda_i(\theta(w''w'))p_i(w_t|h_{t,i})}$$

M-STEP:

$$\lambda_i(\theta(w''w')) \longleftarrow \frac{\hat{c}_i(\theta(w''w'))}{\sum_{i=1}^{N}\hat{c}_i(\theta(w''w'))}$$

Note: En général, les estimées obtenues dépendent grandement des valeurs initiales des λ (optimum local). !!! estimées sur held-out !!!

Un codage possible

Soit S une structure de données associant à $\theta(h)$ à une paire de vecteurs $\lambda = (\lambda_1 \ldots \lambda_N)$ et $c = (c_1 \ldots c_N)$; et soient $\lambda_i[\theta(h)]$ et $c_i[\theta(h)]$ la i-ème valeur de chacun de ces vecteurs. Et Soit M un tableau de N flottants.

Pour toute itération: ($boucle\ sur\ le\ E\text{-}\ et\ le\ M\text{-}step$)
- $\forall h, i : \quad c_i[\theta(h)] \leftarrow 0$ ($init\ des\ c_i\ en\ parcourant\ S$) — E-STEP
- Pour tout mot w_t dans $\mathcal{T}$, $t \in [1, T]$ ($parcours\ du\ corpus$)
  - $somme \leftarrow 0$
  - Pour tout $i \in [1, N]$
    - $M[i] = \lambda_i[\theta(h_t)] \times p_i(w_t|\theta(h_{t,i}))$
    - $somme+ = M[i]$
  - Pour tout $i \in [1, N] : \quad c_i[\theta(h_t)]+ = M[i]/somme$
- $\forall h$ ($parcours\ de\ S$) — M-STEP
  - $Sum \leftarrow 0$
  - $\forall i \in [1, N] : Sum+ = c_i[\theta(h)]$
  - $\forall i \in [1, N] : \lambda_i[\theta(h)] \leftarrow c_i[\theta(h)]/Sum$ ($nouvelle\ valeur$)

Références

L.E. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972.

Adam Berger. Convexity, maximum likelihood and all that. School of Computer Science, Carnegie Mellon University, 2000.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

Frederick Jelinek. *Statistical Methods for Speech Recognition*. The MIT Press, Cambridge, Massachusetts, 1998.

Ted Pedersen. The em algorithm, selected readings. Unpublished notes to accompany the panel discussion on the EM algorithm at EMNLP 2001, 2001a.

Ted Pedersen. A gentle introduction to the em algorithm. Transparents présentés au panel sur l'algorithme EM qui s'est tenu à la conférence Empirical Methods in Natural Language Processing (EMNLP), June 2001b.