EM algoritmus

Jan Kracík
jan.kracik@vsb.cz

EM algoritmus

- navrhli Arthur Dempster, Nan Laird a Donald Rubin v roce 1977
- slouží k hledání maximálně věrohodných odhadů z neúplných dat
- častá aplikace: odhadování pravděpodobnostních směsí
- patří mezi nejdůležitější algoritmy ve statistice

Maximálně věrohodný odhad (MLE)

statistický model: $f_\theta(x), \theta \in \Theta$

data: $x_1, x_2, \ldots, x_t$ (i.i.d.)

logaritmická věrohodnostní funkce:

$$l(\theta) = \ln f_\theta(x_1, \ldots, x_t) = \ln \prod_{\tau=1}^{t} f_\theta(x_\tau) = \sum_{\tau=1}^{t} \ln f_\theta(x_\tau)$$

odhad parametru:

$$\hat{\theta}_{MLE} \in \arg\max_{\theta \in \Theta} l(\theta)$$

Příklad 1: MLE

náhodný vektor (X, Y) s hodnotami v $\mathcal{X} \times \mathcal{Y} = \{0, 1\} \times \{0, 1\}$
statistický model:

$$f_\theta(x, y) = \prod_{i,j \in \mathcal{X} \times \mathcal{Y}} \theta_{ij}^{\delta(x,i)\delta(y,j)}$$

$$\theta \in \Theta = \left\{ (\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}) \in \mathbb{R}^4 \,\middle|\, \theta_{ij} > 0, \sum_{i,j} \theta_{ij} = 1 \right\}$$

data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_t, y_t)$$

Příklad 1 (pokračování)

$$
\begin{aligned}
l(\theta) &= \ln \prod_{\tau} f_\theta(x_\tau, y_\tau) = \sum_{i,j} \nu_{ij} \ln \theta_{ij}, \\
\nu_{ij} &= \sum_{\tau} \delta(x_\tau, i)\delta(y_\tau, j)
\end{aligned}
$$

Příklad 1 (pokračování)

$$
\begin{aligned}
l(\theta) &= \ln \prod_{\tau} f_\theta(x_\tau, y_\tau) = \sum_{i,j} \nu_{ij} \ln \theta_{ij}, \\
\nu_{ij} &= \sum_{\tau} \delta(x_\tau, i)\delta(y_\tau, j)
\end{aligned}
$$

minimalizace $l(\theta)$ s využitím Lagrangeových multiplikátorů $\rightarrow$ soustava lin. rovnic:

$$
\begin{aligned}
\nu_{ij} - \lambda\theta_{ij} &= 0 \\
1 - \sum_{i,j} \theta_{ij} &= 0
\end{aligned}
$$

odhad:

$$
\hat{\theta}_{ij} = \frac{\nu_{ij}}{\sum_{u\in\mathcal{X}, v\in\mathcal{Y}} \nu_{uv}} = \frac{\nu_{ij}}{t}
$$

Příklad 2: MLE s chybějícími daty

statistický model stejný jako v příkladu 1

data:

$$(x_1, y_1), \ldots, (x_{t_1}, y_{t_1}),\ x_{t_1+1}, \ldots, x_{t_1+t_2},\ y_{t_1+t_2+1}, \ldots, y_{t_1+t_2+t_3}$$

Příklad 2: MLE s chybějícími daty

statistický model stejný jako v příkladu 1

data:

$$(x_1, y_1), \ldots, (x_{t_1}, y_{t_1}),\ x_{t_1+1}, \ldots, x_{t_1+t_2},\ y_{t_1+t_2+1}, \ldots, y_{t_1+t_2+t_3}$$

marginální hustoty:

$$\begin{aligned} f_\theta(x_\tau) &= \sum_i (\theta_{i0} + \theta_{i1})^{\delta(x_\tau, i)} \\ f_\theta(y_\tau) &= \sum_j (\theta_{0j} + \theta_{1j})^{\delta(y_\tau, j)} \end{aligned}$$

$$l(\theta) = \sum_{i,j} \nu_{ij} \ln \theta_{ij} + \sum_i \xi_i \ln(\theta_{i0} + \theta_{i1}) + \sum_j \eta_j \ln(\theta_{0j} + \theta_{1j})$$

Příklad 2 (pokračování)

$\hat{\theta}_{MLE}$ řešením soustavy nelineárních rovnic

$$\frac{\nu_{ij}}{\theta_{ij}} + \frac{\xi_i}{\theta_{i0} + \theta_{i1}} + \frac{\eta_j}{\theta_{0j} + \theta_{1j}} - \lambda\theta_{ij} = 0$$

$$\sum_{ij} \theta_{ij} - 1 = 0$$

EM algoritmus

statistický model: $f_\theta(x, y), \theta \in \Theta$
data: $x_1, \ldots, x_t$

hledáme maximálně věrohodný odhad

$$\hat{\theta}_{MLE} \in \arg\min_{\theta \in \Theta} \ln f_\theta(x_1, \ldots, x_t)$$

iterační algoritmus

- počáteční aproximace (odhad parametru): $\hat{\theta}_0$, lze volit náhodně
- iterace se skládají ze 2 kroků: E-step, M-step

EM algoritmus: $(k+1)$. iterace

z předchozí iterace: $\hat{\theta}_k$

- E-step: podmíněná stř. hodnota log. věrohodnosti z kompletního modelu vzhledem k $f_{\hat{\theta}_k}(x, y)$

$$q_{k+1}(\theta) := E_{\hat{\theta}_k}[\ln f_\theta((x_1, y_1), \ldots, (x_t, y_t))|x_1, \ldots, x_t]$$

EM algoritmus: $(k+1)$. iterace

z předchozí iterace: $\hat{\theta}_k$

- E-step: podmíněná stř. hodnota log. věrohodnosti z kompletního modelu vzhledem k $f_{\hat{\theta}_k}(x, y)$

$$q_{k+1}(\theta) := E_{\hat{\theta}_k}[\ln f_\theta((x_1, y_1), \ldots, (x_t, y_t))|x_1, \ldots, x_t]$$

- M-step: maximalizace $q_{k+1}(\theta)$ z předchozího kroku

$$\hat{\theta}_{k+1} \in \arg\max_{\theta \in \Theta} q_{k+1}(\theta)$$

Konvergence EM algoritmu

Posloupnost $l(\hat{\theta}_0), l(\hat{\theta}_1), l(\hat{\theta}_2), \ldots$ monotónně roste.

Pro “nalezení” (aproximaci) globálního optima se využívají opakované běhy EM algoritmu s náhodně volenými počátečními aproximacemi $\hat{\theta}_0$.

Zastavovací pravidlo obecně neexistuje. Řídíme se vývojem $l(\hat{\theta}_k)$.

Často doplněn heuristickými postupy, zejména v souvislosti s odhadem směsových modelů.

Konvergence (pro diskrétní n.v.)

$$\begin{aligned} l(\theta) &= \sum_{\tau} \ln f_\theta(x_\tau) = t \sum_{x \in \mathcal{X}} r(x) \ln_\theta f(x) = \\ &= t(H(r(x)) - D(r(x)||f_\theta(x))), \end{aligned}$$

kde

- $r(x)$ ... empirická hustota
- $H(r(x))$ ... entropie $r(x)$
- $D(r(x)||f_\theta(x))$ ... Kullback-Leiblerova divergence hustot $r(x)$ a $f_\theta(x)$

Stačí ukázat, že $D(r(x)||f_{\hat{\theta}_k}(x))$ monotónně klesá.

$q_{k+1}(\theta)$ z E-step lze vyjádřit jako

$$q_{k+1}(\theta) = t \sum_{x,y} f_{\hat{\theta}_k}(y|x) r(x) \ln f_\theta(x,y)$$

pro $\hat{\theta}_{k+1}$ z M-step tedy platí

$$\hat{\theta}_{k+1} \in \arg\min_{\theta \in \Theta} D(f_{\hat{\theta}_k}(y|x) r(x) || f_\theta(x,y))$$

$q_{k+1}(\theta)$ z E-step lze vyjádřit jako

$$q_{k+1}(\theta) = t \sum_{x,y} f_{\hat{\theta}_k}(y|x) r(x) \ln f_\theta(x,y)$$

pro $\hat{\theta}_{k+1}$ z M-step tedy platí

$$\hat{\theta}_{k+1} \in \arg\min_{\theta \in \Theta} D(f_{\hat{\theta}_k}(y|x) r(x) || f_\theta(x,y))$$

pro KL divergenci empirické hustoty a odhadu pak dostáváme

$$D(r(x)||f_{\hat{\theta}_k}(x)) \quad = \quad D(r(x) f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_k}(x,y))$$

$q_{k+1}(\theta)$ z E-step lze vyjádřit jako

$$q_{k+1}(\theta) = t \sum_{x,y} f_{\hat{\theta}_k}(y|x) r(x) \ln f_\theta(x,y)$$

pro $\hat{\theta}_{k+1}$ z M-step tedy platí

$$\hat{\theta}_{k+1} \in \arg\min_{\theta \in \Theta} D(f_{\hat{\theta}_k}(y|x) r(x) || f_\theta(x,y))$$

pro KL divergenci empirické hustoty a odhadu pak dostáváme

$$\begin{aligned} D(r(x)||f_{\hat{\theta}_k}(x)) &= D(r(x) f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_k}(x,y)) \geq \\ &\geq D(r(x) f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_{k+1}}(x,y)) \end{aligned}$$

$q_{k+1}(\theta)$ z E-step lze vyjádřit jako

$$q_{k+1}(\theta) = t \sum_{x,y} f_{\hat{\theta}_k}(y|x) r(x) \ln f_\theta(x,y)$$

pro $\hat{\theta}_{k+1}$ z M-step tedy platí

$$\hat{\theta}_{k+1} \in \arg\min_{\theta \in \Theta} D(f_{\hat{\theta}_k}(y|x) r(x) || f_\theta(x,y))$$

pro KL divergenci empirické hustoty a odhadu pak dostáváme

$$\begin{aligned} D(r(x)||f_{\hat{\theta}_k}(x)) &= D(r(x) f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_k}(x,y)) \geq \\ &\geq D(r(x) f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_{k+1}}(x,y)) \geq \\ &\geq D(r(x)||f_{\hat{\theta}_{k+1}}(x)) \end{aligned}$$

Příklad 2 (pokračování): pomocí EM algoritmu

statistický model:

$$f_\theta(x,y) = \prod_{i,j \in \mathcal{X}\times\mathcal{Y}} \theta_{ij}^{\delta(x,i)\delta(y,j)}$$

data:

$$data = ((x_1,y_1),\ldots,(x_{t_1},y_{t_1}),\ x_{t_1+1},\ldots,x_{t_1+t_2},\ y_{t_1+t_2+1},\ldots,y_{t_1+t_2+t_3})$$

označení:

$$\begin{aligned} T &= t_1+t_2+t_3 \\ \hat{\theta}_k &= (\hat{\theta}_{k;00},\hat{\theta}_{k;01},\hat{\theta}_{k;10},\hat{\theta}_{k;11}) \end{aligned}$$

E-step:

$$
\begin{aligned}
q_{k+1}(\theta) &= E[\ln f_\theta((x_1, y_1), \ldots, (x_T, y_T))|data] = \sum_{i,j} \nu_{ij} \ln \theta_{ij} \\
\nu_{ij} &= \sum_{\tau=1}^{t_1} \delta(x_\tau, i)\delta(y_\tau, j) + \\
&\quad \sum_{\tau=t_1+1}^{t_1+t_2} \delta(x_\tau, i) E_{\hat{\theta}_k}[\delta(y_\tau, j)|x_\tau] + \\
&\quad \sum_{\tau=t_1+t_2+1}^{t_1+t_2+t_3} E_{\hat{\theta}_k}[\delta(x_\tau, i)|y_\tau]\delta(y_\tau, j)
\end{aligned}
$$

$$
\begin{aligned}
E_{\hat{\theta}_k}[\delta(y_\tau, j)|x_\tau] &= \hat{\theta}_{k;x_\tau j}/(\hat{\theta}_{k;x_\tau 0} + \hat{\theta}_{k;x_\tau 1}) \\
E_{\hat{\theta}_k}[\delta(x_\tau, i)|y_\tau] &= \hat{\theta}_{k;iy_\tau}/(\hat{\theta}_{k;0y_\tau} + \hat{\theta}_{k;1y_\tau})
\end{aligned}
$$

M-step:

$$\hat{\theta}_{k+1;ij} = \frac{\nu_{ij}}{\sum_{u \in \mathcal{X}, v \in \mathcal{Y}} \nu_{uv}} = \frac{\nu_{ij}}{T}$$

Poznámka k příkladu 2

Přímé řešení optimalizační úlohy vedlo na soustavu nelineárních rovnic. Nutno řešit numericky . . .

EM algoritmus vede na mnohem jednodušší výpočet podobně jako s kompletními daty.

EM algoritmus chytře využívá přirozenou geometrickou strukturu úlohy!

Pravděpodobnostní směsi

statistický model ve tvaru

$$f_\theta(x) = \sum_{c=1}^{n} \alpha_c m_{\theta_c}(x),$$

kde

- $\alpha_c > 0, \sum_c \alpha_c = 1$ jsou váhy komponent
- $\theta = (\alpha_1, \ldots, \alpha_n, \theta_1, \ldots, \theta_n)$ značí vektor parametrů
- $m_{\theta_c}(x)$ jsou hustoty ze zvolené třídy (např. normální) určené parametrem θ_c, tzv. komponenty

Pravděpodobnostní směsi

využití:

- semiparametrický model pro složitá rozdělení
- modely se skrytou diskrétní veličinou
- klastrování dat

směsový model lze chápat jako marginální distribuci

$$f_\theta(x) = \sum_c f_\theta(x|c) f_\theta(c),$$

kde

$$\begin{aligned} f_\theta(c) &= \alpha_c \\ f_\theta(x|c) &= m_{\theta_c}(x) \end{aligned}$$

odhad směsového modelu lze brát jako odhad z neúplných dat $x_1, \ldots, x_t$ (neznáme indexy komponent $c_1, \ldots, c_t$)

k odhadu parametrů lze použít EM algoritmus

n-rozměrný normální směsový model

$$f(x|\theta) = \sum_{c=1}^{C} \alpha_c m(x|\theta_c)$$

$$m(x|\theta_c) = (2\pi)^{-\frac{n}{2}} |\Sigma_c|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu_c)^T \Sigma_c^{-1}(x-\mu_c)\right)$$

$$\theta_c = (\mu_c, \Sigma_c)$$

$$\theta = (\alpha_1, \ldots, \alpha_C, \theta_1, \ldots, \theta_C)$$

EM algoritmus pro n-rozměrný normální směsový model
inicializace

pro $c = 1, \ldots, C$ náhodně zvol počáteční aproximace parametrů směsi

$$\begin{aligned} \alpha_c^{(0)} &\in (0,1) \\ \mu_c^{(0)} &\in \mathbb{R}^n \\ \Sigma_c^{(0)} &\in \mathbb{R}^{n,n}, \text{ PD} \end{aligned}$$

tak, aby platilo

$$\alpha_1 + \cdots + \alpha_C = 1$$

EM algoritmus pro n-rozměrný normální směsový model
(k+1). iterace, E-step

pro $\tau = 1, \ldots, t$ a $c = 1, \ldots, C$:

$$\begin{aligned} \kappa_{c;\tau}^{(k+1)} &:= f(c_\tau = c | x_\tau, \theta^{(k)}) \\ &= \frac{m(x_\tau | \theta_c^{(k)}) \alpha_c^{(k)}}{\sum_{\tilde{c}} m(x_\tau | \theta_{\tilde{c}}^{(k)}) \alpha_{\tilde{c}}^{(k)}} \end{aligned}$$

EM algoritmus pro n-rozměrný normální směsový model (k+1). iterace, M-step

pro $\tau = 1, \ldots, t$ a $c = 1, \ldots, C$:

$$
\begin{aligned}
\alpha_c^{(k+1)} &= \frac{1}{t} \sum_{\tau=1}^{t} \kappa_{c;\tau}^{(k+1)} \\
\mu_c^{(k+1)} &= \frac{\sum_{\tau=1}^{t} \kappa_{c;\tau}^{(k+1)} x_\tau}{\sum_{\tau=1}^{t} \kappa_{c;\tau}^{(k+1)}} \\
\Sigma_c^{(k+1)} &= \frac{\sum_{\tau=1}^{t} \kappa_{c;\tau}^{(k+1)} \left(x_\tau - \mu_c^{(k+1)} \right)^T \left(x_\tau - \mu_c^{(k+1)} \right)}{\sum_{\tau=1}^{t} \kappa_{c;\tau}^{(k+1)}}
\end{aligned}
$$

Závěrečné poznámky I

- EM má smysl, pokud se sdruženou hustotou $f_\theta(x, y)$ lze pracovat snáze než s $f_\theta(x)$.
- V M-step stačí hledat $\hat{\theta}_{k+1}$ tak, aby

$$D(r(x)f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_{k+1}}(x, y)) \leq D(r(x)f_{\hat{\theta}_k}(y|x)||f_{\hat{\theta}_k}(x, y))$$

  s rovností pouze pro minimum $\rightarrow$ jednodušší implementace.
- formulaci s KL divergencí lze použít jako základ algoritmu pro aproximaci složitých distribucí jednoduššími
- lze kombinovat s numerickými metodami, heuristickými postupy, ...

Závěrečné poznámky II

při praktickém využití je potřeba dořešit

- počet komponent - vhodné nástroje AIC, BIC
- počet iterací - zastavovací pravidlo
- konvergence k lokálnímu maximu - opakované náhodné starty, znáhodněný EM alg.

Kritéria pro volbu modelu

Akaikeho informační kritérium (AIC)
Bayesovské informační kritérium (BIC)

- AIC: $2k - 2\ln(L)$
- BIC: $-2\ln(L) + k\ln(n)$

- L: maximální log. věrohodnost modelu
- k: počet parametrů modelu
- n: počet dat

model s nižší hodnotou AIC (BIC) je vhodnější *s ohledem na počet dat*