EM algoritmus

používá se pro odhad nepozorovaných veličin.
Jde o iterativní algoritmus opakující dva kroky:

- **Estimate**, který odhadne hodnoty nepozorovaných dat, a
- **Maximize**, který maximalizuje věrohodnost vzhledem k datům přes uvažované modely.

Proč zahrnovat do modelu neznámé veličiny

Protože se to hodí.

- Známe model, některé veličiny nemůžeme pozorovat.
- Neznámá nepozorovaná veličina zaviní, že vše souvisí se vším.
- Často se používají směsi gausovských rozložení: na klastrování, na popis funkce při zpracování obrazu, atd.

Estimate

- Mám model (z předchozího kroku, na počátku volíme parametry např. náhodně či rovnoměrnou distribuci).
- Pro každý řádek dat:
  - vložím do modelu evidenci na veličinách, které jsem pozorovala,
  - podívám se na pravděpodobnost veličin, které pozorované nebyly,
  - řádek dat rozdrobím na spoustu dílků, každý s jinými hodnotami nepozorovaných veličin, váha dílku odpovídá pravděpodobnosti situace, součet vah drobků je 1.

Maximize

- Pro některé modely to umíme odminule:
- gausovská distribuce
- bayesovská síť

Směs gausovských distribucí

- 2 distribuce mají parametry: $\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$, na začátku μ náhodně, $\pi = 0.5$, $\sigma =$ výběrový rozptyl
- Estimate – krok:

$$\gamma_i = \frac{\pi \phi_{\theta_2}(y_i)}{(1-\pi)\phi_{\theta_1}(y_i) + \pi \phi_{\theta_2}(y_i)}$$

- Maximize – krok: odhadnout střední hodnoty a rozptyly,

$$\mu_1 = \frac{\sum_{i=1}^{N}(1-\gamma_i)y_i}{\sum_{i=1}^{N}(1-\gamma_i)}$$

$$\sigma_2^2 = \frac{\sum_{i=1}^{N}\gamma_i(y_i-\mu_2)^2}{\sum_{i=1}^{N}\gamma_i}$$

$$\pi = \frac{\sum_{i=1}^{N}\gamma_i}{N}$$

- a iterujeme EM do konvergence.

Směs gausovských distribucí

- Estimate: vložím evidenci, zapíši si distribuci na komponentách C,

$$p_{ij} = P(C = i|x_j) = \alpha \cdot P(x_j|C = i) \cdot P(C = i)$$

  Definujeme součty přes všechny příklady j pro jednotlivé komponenty:

$$p_{ij} = \sum_{j=1}^{N} p_{ij}$$

- Maximize: pro daná data spočteme maximálně věrohodný odhad:

Pro jednorozměrné:

$$\sigma_i^2 \leftarrow \sum_j \frac{p_{ij}}{p_i}(x_j - \mu_i)^2$$

$$\begin{aligned} \mu_i &\leftarrow \sum_j \frac{p_{ij}}{p_i} x_j \\ \Sigma_i &\leftarrow \sum_j \frac{p_{ij}}{p_i}(x_j - \mu_i)(x_j - \mu_i)^T \\ P(C = i) &\leftarrow \frac{p_i}{\sum_{l=1}^{k} p_l} \end{aligned}$$

EM algoritmus

- Lze dokázat, že v každém kroku zvýší věrohodnost modelu.
- Nakonec (možná) najde model s větší věrohodností, než má model původní. Data jsou generovaná náhodně a nemusí úplně přesně vystihovat původní model.
- Za jistých předpokladů se dá dokázat, že EM konverguje k maximu, obecně jako každá gradientní metoda může zůstat v lokálním maximu.
- Narozdíl od většiny gradientních metod nemáme parametr velikost kroku.
- Spíš je problém, že ke konci konverguje pomalu, než že by zůstal v lokálním maximu.

EM algoritmus pro bayesovské sítě

- Základní princip je stejný – Estimate a Maximize.
- Příklad: Dva pytle bonbónů někdo smíchal dohromady. Každý bonbón má nějaký obal *Wrapper* a příchuť *Flavor* a buď v něm jsou dírky *Holes*, nebo ne. V každém pytli byl jiný poměr příchutí, jiný poměr děravých bonbónů k neděravým atd.

Příklad se dá popsat jako naivní bayesovský model.

Příklad

Snědli jsme 1000 bonbónů a zapsali, co jsme pozorovali:

	W=red		W=green	
	H=1	H=0	H=1	H=0
F=cherry	273	93	104	90
F=lime	79	100	94	167

Počáteční parametry modelu zvolíme:

$$\theta^{(0)} = 0.6,\ \theta_{F1}^{(0)} = \theta_{W1}^{(0)} = \theta_{H1}^{(0)} = 0.6,\ \theta_{F2}^{(0)} = \theta_{W2}^{(0)} = \theta_{H2}^{(0)} = 0.4$$

- Odhad θ: kdyby byla pozorovaná, spočteme podíl bonbónů z prvního balíčku ke všem bonbónům.
- Protože jí nepozorujeme, **sčítáme očekávané počty**

$$\theta^{(1)} = \frac{1}{N}\sum_{j=1}^{N} \frac{P(flavor_j|Bag=1)P(wrapper_j|Bag=1)P(holes_j|Bag=1)P(Bag=1)}{\sum_{i=1}^{2} P(flavor_j|Bag=i)P(wrapper_j|Bag=i)P(holes_j|Bag=i)P(Bag=i}$$

(normalizační konstanta dole také záleží na hodnotách parametrů).

Pro bonbón $red, cherry, holes$ dostaneme:

$$\frac{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)}}{\theta_{F1}^{(0)}\theta_{W1}^{(0)}\theta_{H1}^{(0)}\theta^{(0)} + \theta_{F2}^{(0)}\theta_{W2}^{(0)}\theta_{H2}^{(0)}\theta^{(0)}} \approx 0.835055$$

takových bonbónů máme 273, tedy je jejich příspěvek $\frac{273}{N} \cdot 0.835055$.

Podobně spočteme příspěvky dalších sedmi políček a dostaneme:

$$\theta^{(1)} = 0.6124$$

- Odhad θ_{F1} by v plně pozorovaném případě byl ...

- My musíme počítat podíl očekávaných počtů $Bag = 1 \& F = cherry$ a $Bag = 1$, tj.

$$\theta_{F1}^{(1)} = \frac{\sum_{j;Flavor_j=cherry} P(Bag = 1|Flavor_j = cherry, wrapper_j, holes_j)}{\sum_j P(Bag = 1|cherry_j, wrapper_j, holes_j)}$$

- Podobně dostaneme:

$$\theta^{(1)} = 0.6124,\ \theta_{F1}^{(1)} = 0.6684, \theta_{W1}^{(1)} = 0.6483, \theta_{H1}^{(1)} = 0.6558,$$

$$\theta_{F2}^{(1)} = 0.3887, \theta_{W2}^{(1)} = 0.3817, \theta_{H2}^{(1)} = 0.3827$$

Pozn: V Bayesovské síti lze učit parametry tak, že postupně vložíme jeden příklad za druhým a sčítáme pravděpodobnosti pro jednotlivé konfigurace dítěte plus jeho rodičů. Tím dostaneme očekávané četnosti (resp. po vydělení počtem příkladů), z očekávaných četností spočteme parametry podílem odpovídajících četností, tj.

$$\theta_{ijk} \leftarrow \frac{\text{četnost } (X_i = x_{ij} \& pa(X_i) = pa_{ik})}{\text{četnost } (pa(X_i) = pa_{ik})}$$

Obecný EM algoritmus

Máme–li počáteční odhady parametrů $\bar{\theta}^{(0)}$, skryté proměnné Z a pozorovaná *data*, pak můžeme jeden krok EM algoritmu zapsat přiřazením:

$$\bar{\theta}^{(i+1)} \leftarrow argmax_{\bar{\theta}^{(i)}} \sum_{z \in Z} P(Z = z | data, \bar{\theta}^{(i)}) \cdot L(data, Z = z | \bar{\theta}^{(i)})$$