# Discussion 6
## EM Algorithm
## EM in mixture model

田鹏超

tianpch@shanghaitech.edu.cn

# EM Algorithm - Precisely

EM is a general procedure for learning from partly observed data

Given observed variables X, unobserved Z (X={F,A,H,N}, Z={S})

Define $Q(\theta'|\theta) = E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

current

M step new

Iterate until convergence:

- E Step: Use X and current $\theta$ to calculate P(Z|X,$\theta$)
- M Step: Replace current $\theta$ by

$$\theta \leftarrow \arg\max_{\theta'} Q(\theta'|\theta)$$

P(S=1 |

P(S=2 |

$P(S=1|F,A,H,N,\theta)$

$= \frac{P(S=1,F,A,H,N,\theta)}{\sum_t P(S,A,F,H,N,\theta)}$

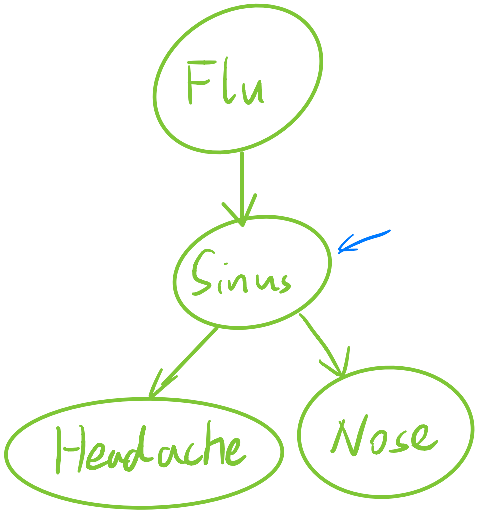
$= \frac{P(S=1|F,A)P(F)P(A)P(H|S=1)P(N|S=1)}{\sum_t \ldots}$

$\theta_{S|F,A}$ $\theta_F$ $\theta_a$ $\theta_{N|S}$ $\theta_{h|s}$

Guaranteed to find local maximum.
Each iteration increases $E_{P(Z|X,\theta)}[\log P(X, Z|\theta')]$

Q =

eg 1.




$X = \{F, N\}$ Observed variables

$Z = \{S, H\}$ Latent variables

$\{F, H, N\}$ 0/1 binary variables

$S \in \{0, 1, 2\}$

There are $K$ training examples in total.

① Derive E step.

② Derive M step.

K

(1) E-step

For each training example k, calculate $\Pr(Z_k|X_k,\theta) = \Pr(S_k, H_k|F_k, N_k, \theta)$.

$$\Pr(S_k = l, H_k = t|F_k, N_k, \theta) = \frac{\Pr(S_k = l, H_k = t, f_k, n_k|\theta)}{\sum_{l=0}^{2}\sum_{t=0}^{1}\Pr(S_k = l, H_k = t, f_k, n_k|\theta)}$$

$$= \frac{\Pr(S_k = l|f_k,\theta)\Pr(H_k = t|S_k = l,\theta)\Pr(n_k|S_k = l)\Pr(f_k|\theta)}{\sum_{l=0}^{2}\sum_{t=0}^{1}\Pr(S_k = l|f_k,\theta)\Pr(H_k = t|S_k = l,\theta)\Pr(n_k|S_k = l)\Pr(f_k|\theta)}$$

$$= \frac{\theta_{s|f}^{l|i}\theta_{h|s}^{t|l}\theta_{n|s}^{i|l}\theta_f}{\sum_{l=0}^{2}\sum_{t=0}^{1}\theta_{s|f}^{l|i}\theta_{h|s}^{t|l}\theta_{n|s}^{i|l}\theta_f}$$

$l \in \{0,1,2\}$ $\quad t \in \{0,1\}$

(2) M-step

Let $l(\theta') = \log\Pr(X, Z|\theta')$ $\quad \prod_{k=1}^{K} Pr(X_k, Z_k|\theta')$

$$Q(\theta'|\theta) = \mathbb{E}_{\Pr(Z|X,\theta)}[\log\Pr(X, Z|\theta')]$$

$$\hat{\theta}' = \operatorname*{argmax}_{\theta'} Q(\theta'|\theta)$$
$$= \operatorname*{argmax}_{\theta'} \mathbb{E}_{\Pr(Z|X,\theta)}[\log\Pr(X, Z|\theta')]$$
$$= \operatorname*{argmax}_{\theta'} \mathbb{E}_{\Pr(Z|X,\theta)}[l(\theta')]$$

$f \in \{0,1\}$

Update $\theta'_f$ :

$$\frac{\partial Q(\theta'|\theta)}{\partial\theta'_f} = \mathbb{E}_{\Pr(Z|X,\theta)}\left[\frac{\sum_{k=1}^{K}\sigma(f_k = 1)}{\theta'_f} - \frac{\sum_{k=1}^{K}\sigma(f_k = 0)}{1-\theta'_f}\right] = 0$$

$\Rightarrow$

$$\theta'_f = \frac{\sum_{k=1}^{K}\sigma(f_k = 1)}{K}$$

Update $\theta'^{l|i}_{s|f}$: $= Pr(S_k = l \mid F_k = i, \theta)$

$$\frac{\partial Q(\theta'|\theta)}{\partial\theta'^{l|i}_{s|f}} = \mathbb{E}_{\Pr(Z|X,\theta)}\left[\frac{\partial l(\theta')}{\partial\theta'^{l|i}_{s|f}}\right]$$

$$= \frac{\partial\sum_z \Pr(Z|X,\theta)l(\theta')}{\partial\,\theta'^{l|i}_{s|f}}$$

$$= \sum_{k=1}^{K}\sum_{l=0}^{2}\sum_{t=0}^{1}\Pr(S_k = l, H_k = t|F_k, N_k, \theta)\left[\frac{\sigma(S_k = l, F_k = i)}{\theta'^{l|i}_{s|f}} - \frac{\sigma(S_k = 2, F_k = i)}{1-\sum_{j=0}^{1}\theta'^{j|i}_{s|f}}\right]$$

$$= \sum_{k=1}^{K}\left[\frac{\sigma(F_k = i)\sum_{t=0}^{1}Pr(S_k = l, H_k = t|F_k, N_k, \theta)}{\theta'^{l|i}_{s|f}} - \frac{\sigma(F_k = i)\sum_{t=0}^{1}Pr(S_k = 2, H_k = t|F_k, N_k, \theta)}{\theta'^{2|i}_{s|f}}\right]$$

$= 0$

$\Rightarrow$

$\theta_{s|f}$ $\quad \left\{\theta^{0|0}_{s|f},\ \theta^{1|0}_{s|f},\ \theta^{0|1}_{s|f},\ \theta^{1|1}_{s|f}\right\}$ $\quad \theta^{2|0}_{s|f} = 1-\theta^{0|0}_{s|f}-\theta^{1|0}_{s|f}$

$$\theta'^{l|i}_{s|f} \propto \sum_{k=1}^{K}\sigma(F_k = i)\sum_{t=0}^{1}Pr(S_k = l, H_k = t|F_k, N_k, \theta)$$

4个参数

$\Rightarrow$

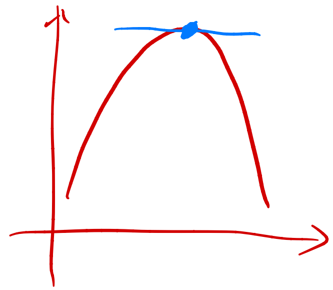
$$\theta'^{l|i}_{s|f} = \frac{\sum_{k=1}^{K}\sigma(F_k = i)\sum_{t=0}^{1}Pr(S_k = l, H_k = t|F_k, N_k, \theta)}{\sum_{l=0}^{2}\sum_{k=1}^{K}\sigma(F_k = i)\sum_{t=0}^{1}Pr(S_k = l, H_k = t|F_k, N_k, \theta)}$$
$$= \frac{\sum_{k=1}^{K}\sigma(F_k = i)\sum_{t=0}^{1}Pr(S_k = l, H_k = t|F_k, N_k, \theta)}{\sum_{k=1}^{K}\sigma(F_k = i)}$$

K

$$\theta'^{\,t|i}_{h|s} = \frac{\sum_{k=1}^{K} P_r(S_k=i,\ H_k=t \mid F_k,\ N_k,\ \theta)}{\sum_{t=0}^{1}\sum_{k=1}^{K} P_r(S_k=i,\ H_k=t \mid F_k,\ N_k,\ \theta)}$$

$$\theta'^{\,1|l}_{n|s} = \frac{\sum_{k=1}^{K} \delta(N_k=1) \sum_{t=0}^{1} P_r(S_k=l,\ H_k=t \mid F_k,\ N_k,\ \theta)}{\sum_{k=1}^{K}\sum_{t=0}^{1} P_r(S_k=l,\ H_k=t \mid F_k,\ N_k,\ \theta)}$$

详细过程略.



$$K = \{0,\ 1,\ 2\}$$

$$\theta_0 = P(X=0),\quad \theta_1 = P(X=1),\quad \overset{\theta_2}{1-\theta_0-\theta_1} = P(X=2)$$

$$L(\theta) = \boxed{\theta_0^{\alpha_0}\cdot\theta_1^{\alpha_1}\cdot\theta_2^{\alpha_2}} \leftarrow \sum_{i=1}^{N} \theta_1^{\delta(y_i=1)}\ \theta_2^{\delta(y_i=2)}\ \theta_0^{\delta(y_i=0)}$$

$$l(\theta) = \log L(\theta) = \alpha_0\log\theta_0 + \alpha_1\log\theta_1 + \alpha_2\log\theta_2$$

$$\theta_2 = 1-\theta_0-\theta_1$$

$$\frac{\partial\, l(\theta)}{\partial\,\theta_0} = \frac{\alpha_0}{\theta_0} - \frac{\alpha_2}{1-\theta_0-\theta_1} = 0$$

$$\Rightarrow\ \theta_i \propto \alpha_i \quad \Rightarrow\ \theta_i = \frac{\alpha_i}{\alpha_0+\alpha_1+\alpha_2}$$

# use EM to solve Gaussian mixture model

probability density function of one-dimensional Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Joint probability density function for N-dimension variable X.

$$f(x) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-u)^T\Sigma^{-1}(X-u)\right), X = (x_1, x_2, \dots, x_n)$$

Gaussian Mixture Model (GMM) with K Gaussian model

$$p(x)=\sum_{k=1}^{K} p(k) p(x \mid k)=\sum_{k=1}^{K} \pi_{k} N\left(x \mid u_{k}, \Sigma_{k}\right)$$

How to use EM Algorithm to solve GMM?

To solve GMM, it's actually to figure out parameters $\theta = (\mu, \Sigma, \pi)$

First, assume the latent variables $Z = (z_1, \ldots, z_K)$ is a binary K-dimensional variable having only a single component equal to 1.
In fact, the latent variable describes the probability of selecting the k-th Gaussian model for each sample.

$$p\left(z_k = 1 | \theta\right) = \pi_k$$

$$p\left(y \mid z_k = 1, \theta\right) = N\left(y \mid \mu_k, \Sigma_k\right)$$

$$p(y) = \sum_z p(z) p(y \mid z) = \sum_{\mathrm{k}=1}^{K} \pi_k N\left(y \mid \mu_k, \Sigma_k\right)$$

For T training examples in total, $Y = (y_1, \ldots, y_T)$. If Z is known the well-informed data should be:

$$(y_t, z_{t,1}, z_{t,2} \ldots z_{t,K}), t = 1, 2 \ldots T$$

However, Z is unkown, we don't know which Gaussian model y is sampled from.

# E-step

$E(Z|X, \theta)$

$$
\begin{aligned}
E\left(z_{t,k} \mid y_t, \mu^i, \Sigma^i, \pi^i\right) &= p\left(z_{t,k} = 1 \mid y_t, \mu^i, \Sigma^i, \Pi^i\right) \\
&= \frac{p\left(z_{t,k} = 1, y_t \mid \mu^i, \Sigma^i, \Pi^i\right)}{p\left(y_t\right)} \\
&= \frac{p\left(z_{t,k} = 1, y_t \mid \mu^i, \Sigma^i, \pi^i\right)}{\sum_{k=1}^{K} p\left(z_{t,k} = 1, y_t \mid \mu^i, \Sigma^i, \pi^i\right)} \\
&= \frac{p\left(y_t \mid Y_{t,k} = 1, \mu^i, \Sigma^i, \pi^i\right) p\left(z_{t,k} = 1 \mid \mu^i, \Sigma^i, \pi^i\right)}{\sum_{k=1}^{K} p\left(yt \mid z_{t,k} = 1, \mu^i, \Sigma^i, \pi^i\right) p\left(z_{t,k} = 1 \mid \mu^i, \Sigma^i, \pi^i\right)} \\
&= \frac{\pi_k^i N\left(y_t; \mu_k^i, \Sigma_k^i\right)}{\sum_{k=1}^{K} \pi_k^i N\left(y_t; \mu_k^i, \Sigma_k^i\right)}
\end{aligned}
$$

$$Q\left(\mu, \Sigma, \pi, \mu^i, \Sigma^i, \pi^i\right) = E_Z\left[\ln \underbrace{p(y, Z \mid \mu, \Sigma, \pi)}_{L} \mid Y, \mu^i, \Sigma^i, \pi^i\right]$$

The likelihood functions is:

$$\begin{aligned}
L(\mu, \Sigma, \pi) &= p(y, Z \mid \mu, \Sigma, \pi) \\
&= \prod_{t=1}^{T} p\left(y_t, z_{t,1}, z_{t,2} \ldots z_{t,K} \mid \mu, \Sigma, \pi\right) \\
&= \prod_{t=1}^{T} \prod_{k=1}^{K} \left(\pi_k N\left(y_t; \mu_k, \Sigma_k\right)\right)^{z_{t,k}} \\
&= \prod_{k=1}^{K} \pi_k^{\sum_{t=1}^{T} z_{t,k}} \prod_{t=1}^{T} \left(N\left(y_t; \mu_k, \Sigma_k\right)\right)^{Y_{t,k}}
\end{aligned}$$

# M-step

$$\mu^{i+1}, \Sigma^{i+1}, \pi^{i+1} = \arg\max Q\left(\mu, \Sigma, \pi, \mu^{i}, \Sigma^{i}, \pi^{i}\right)$$

Set the derivative with respect to $\mu_k$, $\Sigma_k$, $\pi_k$ seperately to 0.

$$\mu_k^{i+1} = \frac{\sum_{t=1}^{T} \frac{\pi_k^i N\left(y_t; \mu_k^i, \Sigma_k^i\right)}{\sum_{k=1}^{K} \pi_k^i N\left(y_t; \mu_k^i, \Sigma_k^i\right)} y_t}{E(\gamma_{t,k} | y_t, \mu^i, \Sigma^i, \pi^i)}, k = 1, 2 \ldots K$$

$$\Sigma_k^{i+1} = \frac{\sum_{t=1}^{T} \frac{\pi_k^i N\left(y_t; \mu_k^i, \Sigma_k^i\right)}{\sum_{k=1}^{K} \pi_k^i N\left(y_t; \mu_k^i, \Sigma_k^i\right)} \left(y_t - \mu_k^i\right)^2}{E(\gamma_{t,k} | y_t, \mu^i, \Sigma^i, \pi^i)}, k = 1, 2 \ldots K$$

$$\pi_k^{i+1} = \frac{E\left(\gamma_{t,k} | y_t, \mu^i, \Sigma^i, \Pi^i\right)}{T}, k = 1, 2 \ldots K$$

EM

while

E-step $\swarrow \theta$ $P_r(Z | X, \theta)$ $\theta' \rightarrow L(\theta')$

M-step $Q$ $\frac{\partial Q}{\partial \theta'} \rightarrow$ update $\theta'$

update $\theta$