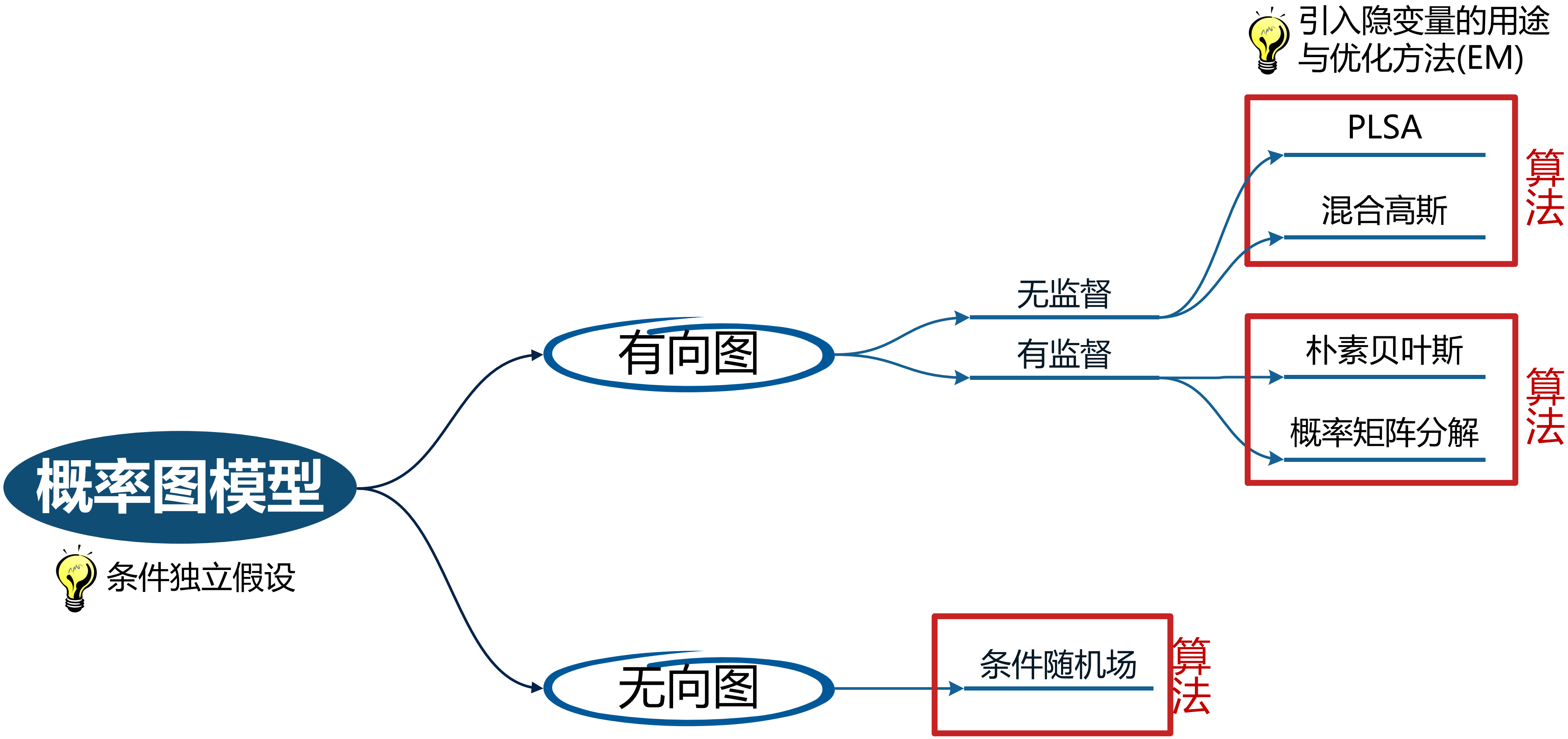
课程专题二：概率图模型方法



EM算法的一般形式

EM算法推导

数据$X=\{(x_i,c_i)\}_{\mathrm{i=1}}^{\mathrm{N}}$，$x_i$出现了$c_i$次

$$L(\theta)=\ln P(X|\theta)=\sum_{i=1}^{N}c_i\ln P(x_i|\theta)$$

$$=\sum_{i=1}^{N}c_i\ln\sum_{z}P(x_i,z|\theta)$$

常数，θ'是上一轮的参数

$$=\sum_{i=1}^{N}c_i\ln\sum_{z}\boxed{P(z|x_i,\theta')}\frac{P(x_i,z|\theta)}{P(z|x_i,\theta')}$$

$$\geq\sum_{i=1}^{N}c_i\sum_{z}P(z|x_i,\theta')\ln\frac{P(x_i,z|\theta)}{P(z|x_i,\theta')}=l(\theta|\theta')$$

等价于优化

$$\sum_{i=1}^{N}c_i\sum_{z}P(z|x_i,\theta')\ln P(x_i,z|\theta)$$

EM算法

Iterate until convergence

- E step: Calculate $P(z|x_i,\theta')$ for each example x_i, Use this to construct

$$\sum_{i=1}^{N} c_i \sum_z P(z|x_i,\theta') \ln P(x_i,z|\theta)$$

- M step: Replace current θ' by

$$\theta' \leftarrow \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} c_i \sum_z P(z|x_i,\theta') \ln P(x_i,z|\theta)$$

解此最优化问题：有可能找不到解析（Closed Form）的最优解；
也可以朝增大的方向优化，这就是Generalized EM算法

EM算法推导：为什么使用 $P(z|x_i,\theta')$

最大似然的下界为

$$l(\theta|\theta') = \sum_{i=1}^{N} c_i \sum_{z} P(z|x_i,\theta') \ln \frac{P(x_i,z|\theta)}{P(z|x_i,\theta')}$$

$$l(\theta'|\theta') = \sum_{i=1}^{N} c_i \sum_{z} P(z|x_i,\theta') \ln \frac{P(x_i,z|\theta')}{P(z|x_i,\theta')}$$

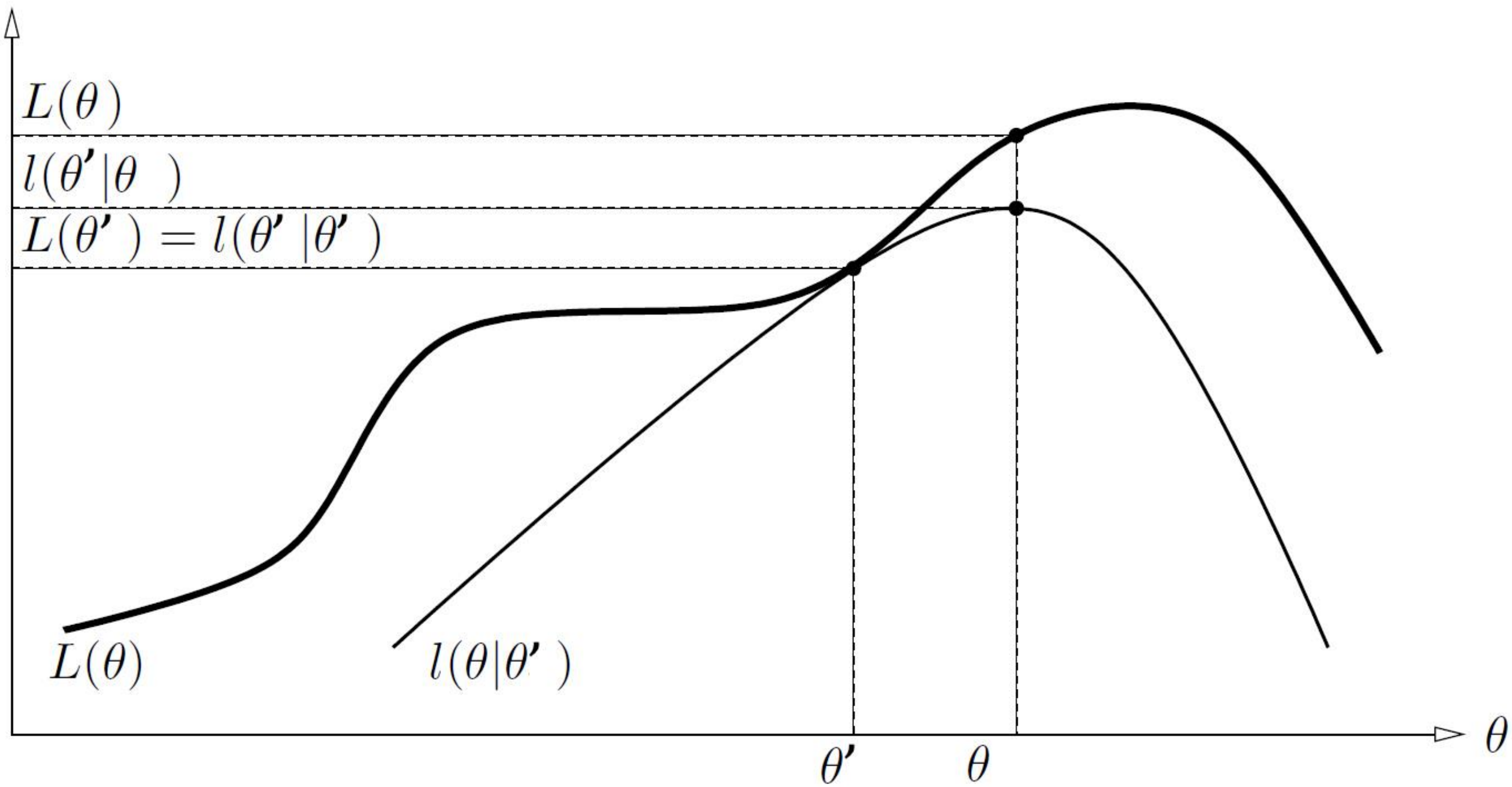
$$= \sum_{i=1}^{N} c_i \sum_{z} P(z|x_i,\theta') \ln P(x_i|\theta')$$

$$= \sum_{i=1}^{N} c_i \ln P(x_i|\theta') \sum_{z} P(z|x_i,\theta')$$

$$= \sum_{i=1}^{N} c_i \ln P(x_i|\theta') = L(\theta')$$

下界在θ'时，等于$L(\theta')$

EM算法推导：为什么使用 $P(z|x_i,\theta')$



构造出一个“紧”的下界

EM应用于PLSA

- 数据$X=\{(x_i,c_i)\}_{\mathrm{i=1}}^{\mathrm{N}}$，$x_i$出现了$c_i$次

$$\sum_{i=1}^{N} c_i \sum_{z} P(z|x_i,\theta')\ln P(x_i,z|\theta)$$

- 应用于PLSA，数据$X=\left\{\left((d_i,w_j),c_{ij}\right)\right\}$

$$\sum_{i,j} c_{ij} \sum_{z} P\left(z\middle|d_i,w_j,\theta'\right)\ln P(d_i,w_j,z|\theta)$$

$$=\sum_{i,j} c_{ij} \sum_{z} P\left(z\middle|d_i,w_j,\theta'\right)\ln P(d_i|\theta)P(z|d_i,\theta)P(w_j|z,\theta)$$

EM应用于PLSA

- 模型的输出结果： $P(z|d_i)$ $P(w_j|z)$
- 引入隐变量的目的
  - **获得可观察变量与隐变量的关系**