EM algorithm (Expectation Maximization)

EM 算法. 1977.

Maximum likelihood from incomplete data via the EM model

极大似然估计 ML

已知.

model: 样本服从的分布模型

samples: 随机抽取的样本

ML →

求知.

模型参数

样本集 $x_1, \cdots, x_n$ $p(x_i|\theta)$

$$L(\theta) = L(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

$$\hat{\theta} = \arg\max L(\theta)$$

求法:

$$H(\theta) = \log(L(\theta)) = \sum_{i=1}^{n} \log p(x_i|\theta)$$

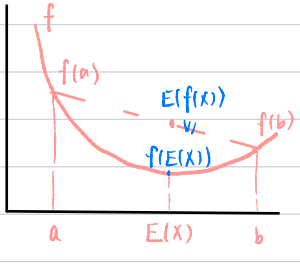
let

$$H(\theta) = 0 \quad \text{求解 } \theta.$$

Jesen 不等式.

$f \mapsto R$, $f'' \geq 0$ (f 是凸函数)

$$E(f(x)) \geq f(E(x))$$ x是常量时, 取等号.



EM algorithm

已知

① 分布模型 (several)
② 随机抽取的样本

EM →

求知.

① 每个样本属于哪一个分布
② 模型参数

样本集 $x_1, \cdots, x_m$

x_i 对应类别 z_i 未知.

EM 又被称为数据添加技术，所添加的数据被称为潜在数据.

观测数据 $(x_1, \cdots, x_m)$

隐含数据 $(z_1, \cdots, z_m)$

完整数据 $Y = (X, Z) = \{(x_1, z_1), \cdots, (x_m, z_m)\}$

推导 (EM).

$$\max l(\theta) = \sum_i \log P(x^{(i)}, \theta) = \sum_i \log \sum_{z^{(i)}} P(x^{(i)}, z^{(i)}; \theta)$$

$$= \sum_i \log \sum_{z^i} Q_i(z^i) \frac{P(x^i, z^i; \theta)}{Q_i(z^i)}$$

↓ Jasen

$$\geq \sum_i \sum_{z^i} \log Q(z^i) \frac{P(x^i, z^i; \theta)}{Q_i(z^i)}$$

思路:

求 $l(\theta)$ 的下界，通过 θ, Q 的更新，使下界逐步提升，通近 $\max l(\theta)$.

用下界通近 $\max l(\theta)$.

E 步骤 (estimation)

固定 θ，调整 Q 使下界上升.

下界最大取等. 即 $\frac{P(x^i, z^i; \theta)}{Q_i(z^i)} = C$ (常数)

$\Rightarrow Q_i(z^i) = P(x^i, z^i; \theta)$ 因为要满足 $\sum_z Q_i(z^i) = 1$

M 步骤 (maximization)

固定 Q，优化 θ. (与 ML 一致).