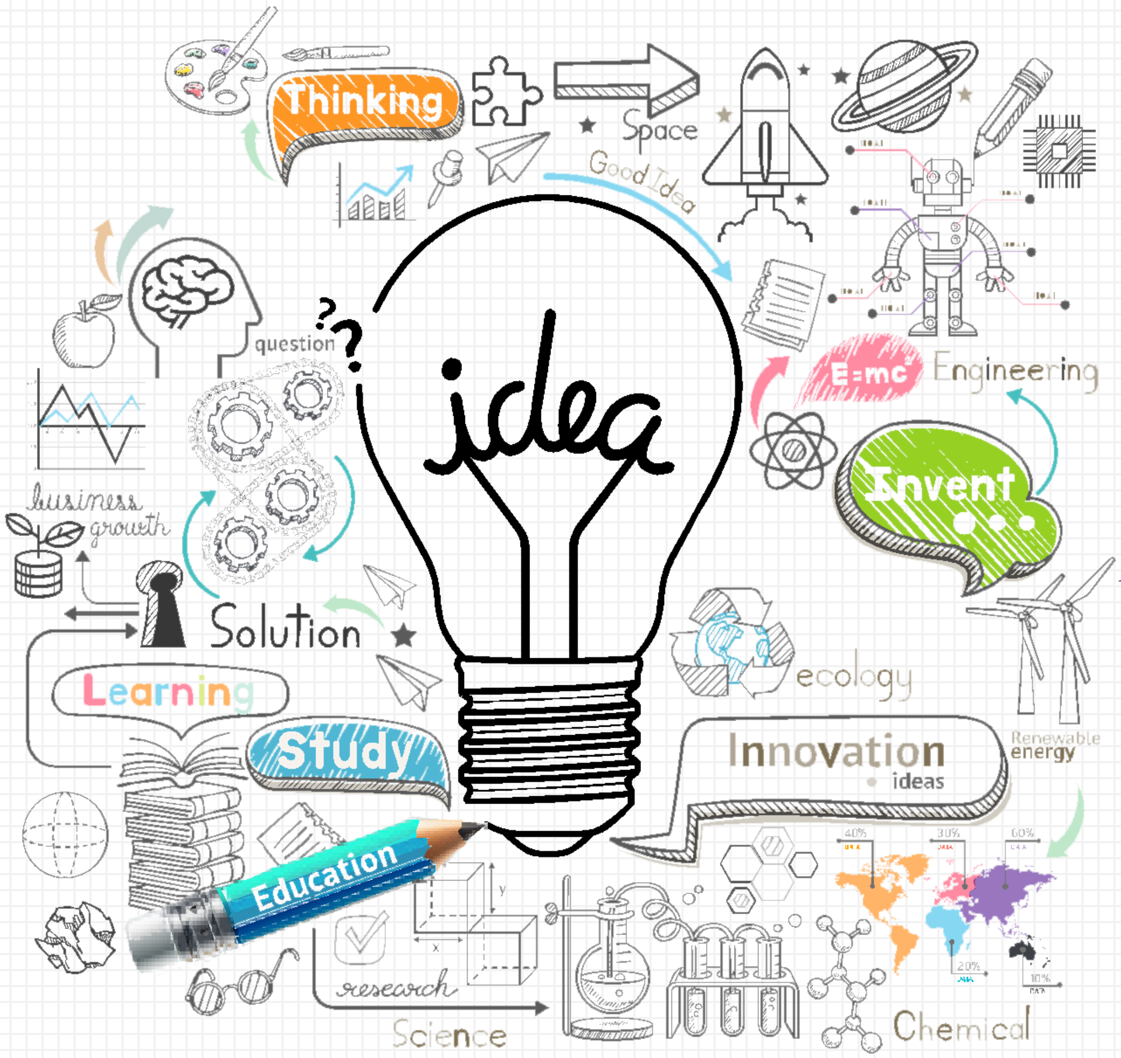
QQ：1366420642

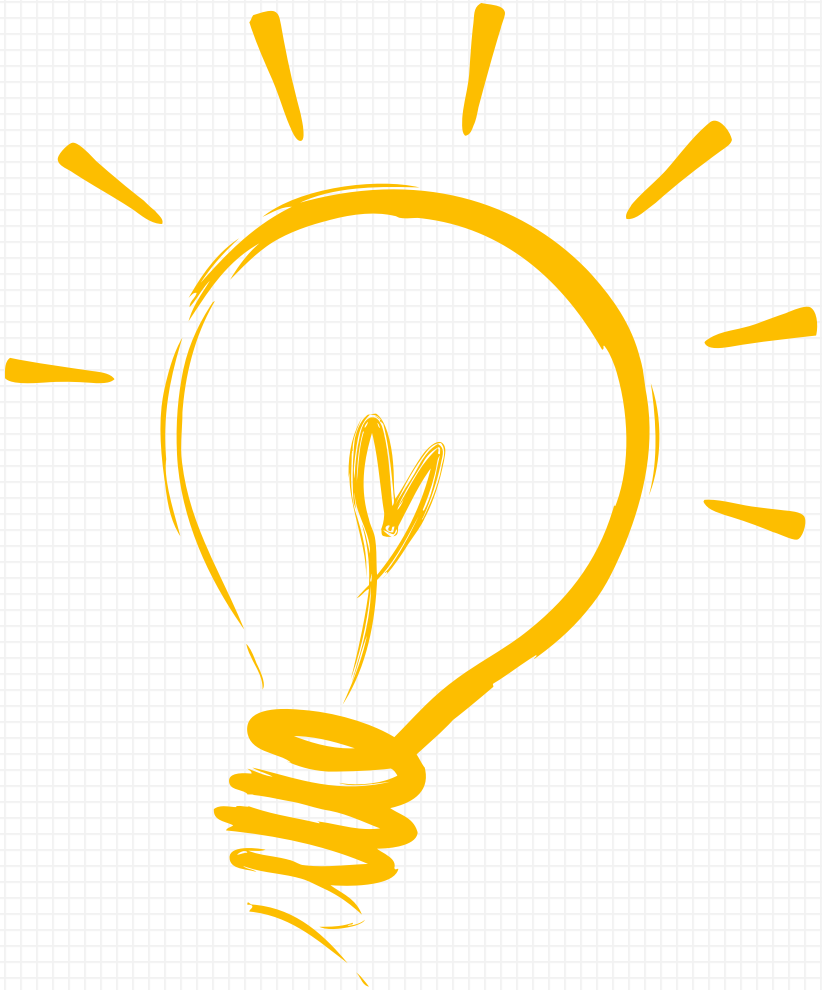


数据挖掘算法

Expection Maximization

QQ：1366420642

CONTENTS



QQ：1366420642

di yi zhang jie

第一章节

算法背景

1.1 基本知识

EM算法思想： 一种**迭代式**的算法，用于**含有隐变量**的概率参数模型的**极大似然估计或极大后验概率估计.**

隐变量：比如聚类问题，样本 x 的特征是可观察到的。其还有一个隐藏属性：所属类别 **z**。（ x , z) 整体是一个完整的观测样本，记为 y。

极大似然估计或极大后验概率估计：比如 x 表示某个样本， 生成于具有参数$\boldsymbol{\theta}$的模型，现在我们有大量样本,想以此样本估计参数$\boldsymbol{\theta}$的情况。

$$p(x|\theta)p(\theta) = p(\theta|x)p(x) \Rightarrow p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \Rightarrow \text{argmax} \quad p(\theta|x)$$

极大似然估计的想法是，哎？我为什么能拿到手里的这个 x ， 而不是其他的什么东西呢 ? 那是因为 θ 产生我手里的 x 是概率最大的，我才能拿到它!

$$\text{argmax} \quad p(\theta|x) \Leftrightarrow \text{argmax} \quad p(x|\theta)$$

1.1 基本知识

$$p(x|\theta)p(\theta) = p(\theta|x)p(x) \Rightarrow p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \Rightarrow \text{argmax}\ \ p(\theta|x)$$

极大后验概率估计，直接从公式入手，因为你观察到样本 x 出现的样本频率(概率)是可以计算的，固定的，那么直接忽略就可以了。

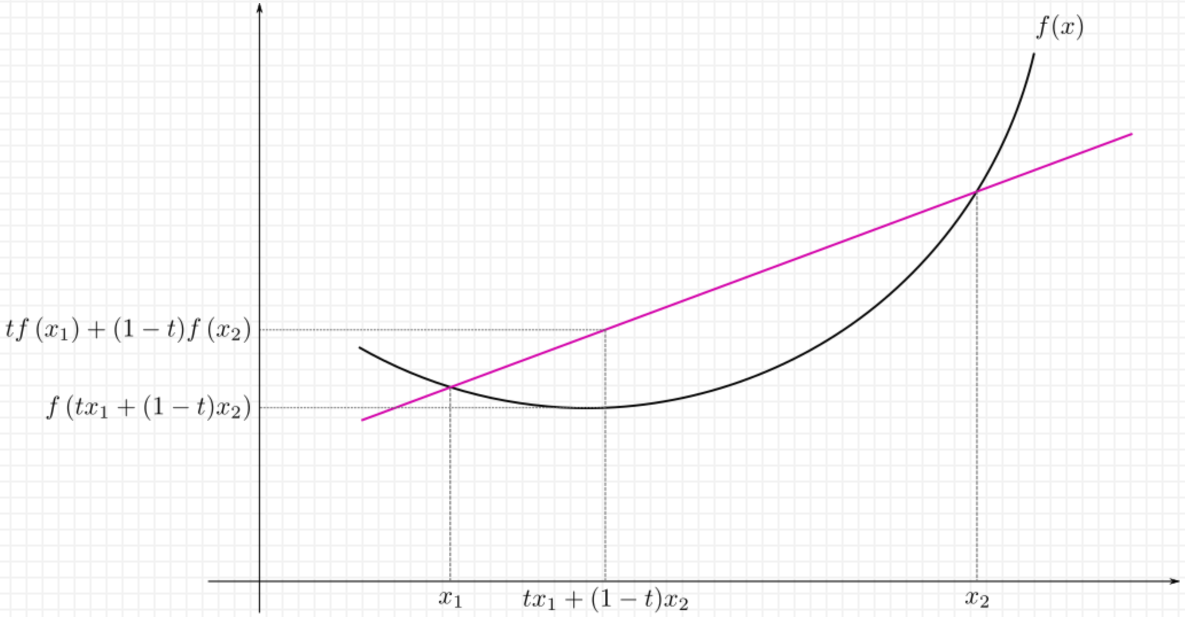
$$\text{argmax}\ \ p(\theta|x) \Leftrightarrow \text{argmax}\ \ p(x|\theta)$$

1.1 基本知识

Jensen不等式:

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

$$f\big(E(X)\big) \leq E(f(X))$$



- https://baike.baidu.com/item/%E7%90%B4%E7%94%9F%E4%B8%8D%E7%AD%89%E5%BC%8F/397409
- https://en.wikipedia.org/wiki/Jensen%27s_inequality
- https://www.probabilitycourse.com/chapter6/6_2_5_jensen's_inequality.php
- **Jensen不等式证明KL散度大于等于0，然后推导出min KL散度 等价于 min 交叉熵。**

di yi zhang jie

第二章节

算法原理

2.1 算法推导

- 独立同分布的可观察样本 X ，隐变量 Z，完整观测样本(X,Z)，当然还有参数θ。那么生成模型为: $p(x,z|\theta)$。
- 根据样本$\{x_i\}$，用极大似然估计求解θ

$$LL(\theta)=\sum_{i=1}^{N}\ln p(x_i|\theta)=\sum_{i=1}^{N}\ln\sum_{z_i}p(x_i,z_i|\theta)$$

我们假设第 i 个样本对应的隐变量 z_i 也具有某种分布：$Q_i(z_i)$

$$\sum_{i=1}^{N}\ln\sum_{z_i}p(x_i,z_i|\theta)=\sum_{i=1}^{N}\ln\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}$$

2.1 算法推导

$$\sum_{i=1}^{N} \ln \sum_{z_i} p(x_i, z_i|\theta) = \sum_{i=1}^{N} \ln \boxed{\sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i|\theta)}{Q_i(z_i)}}$$

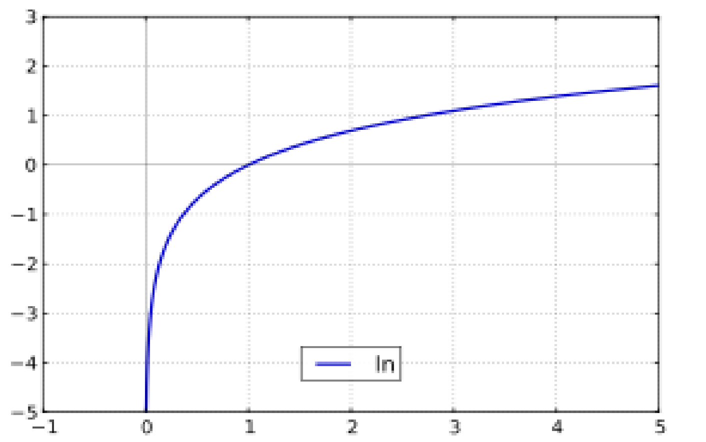
$$E(X) = \sum P(x)x$$

$$E[f(X)] = \boxed{\sum P(x)f(x)}$$

$$\sum_{z_i} Q_i(z_i) \frac{p(x_i, z_i|\theta)}{Q_i(z_i)} \Rightarrow E\left(\frac{p(x_i, z_i|\theta)}{Q_i(z_i)}\right)$$

2.1 算法推导

$$\sum_{i=1}^{N}\ln\sum_{z_i}p(x_i,z_i|\theta)=\sum_{i=1}^{N}\ln\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}=\sum_{i=1}^{N}\ln[E\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)]$$



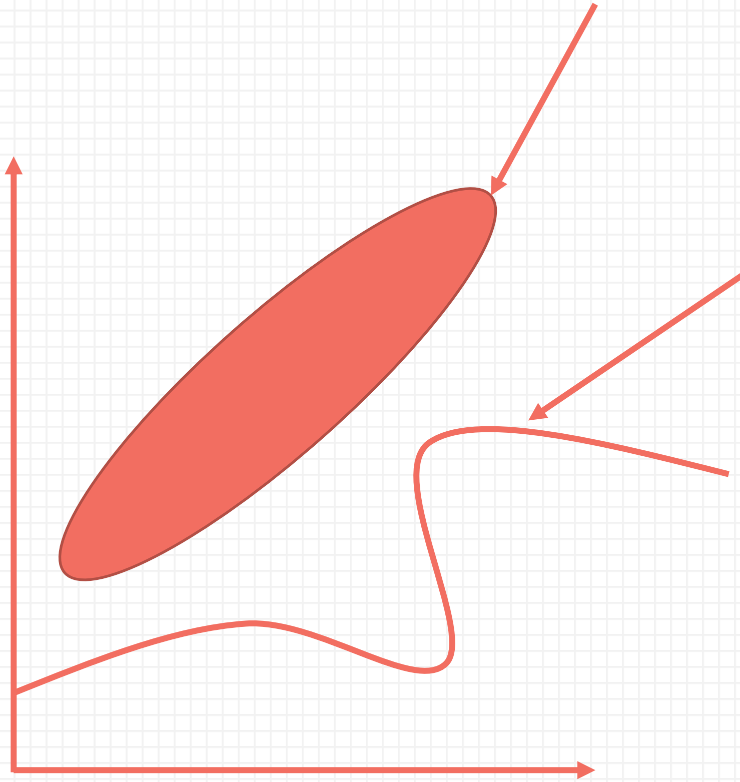
$$\ln\big(E(X)\big)\geq E(\ln(X))$$

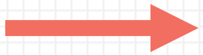
$$\sum_{i=1}^{N}\ln[E\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)]\geq\sum_{i=1}^{N}E\left[\ln\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)\right]=\sum_{i=1}^{N}\sum_{z_i}Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)$$

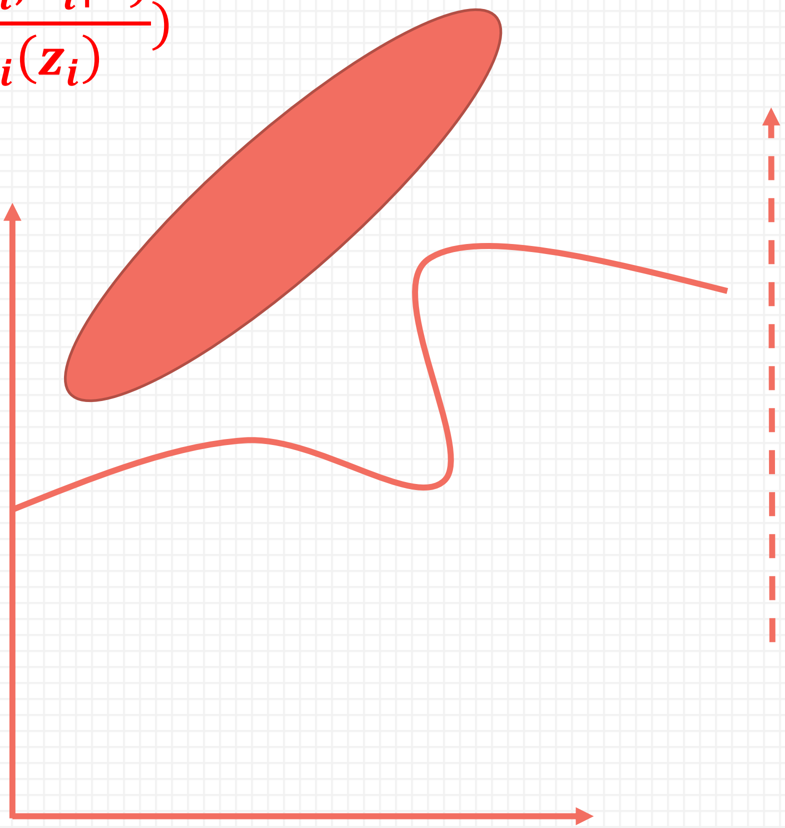
$$\boldsymbol{LL(\theta)=\sum_{i=1}^{N}\ln p(x_i|\theta)=\sum_{i=1}^{N}\ln\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\geq\sum_{i=1}^{N}\sum_{z_i}Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)}$$

2.1 算法推导

$$LL(\theta) \geq \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\ln\left(\frac{p(x_i, z_i|\theta)}{Q_i(z_i)}\right)$$







2.1 算法推导

$$\sum_{i=1}^{N}\ln\sum_{z_i}Q_i(z_i)\frac{p(x_i,z_i|\theta)}{Q_i(z_i)} =?= \sum_{i=1}^{N}\sum_{z_i}Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)$$

$$\frac{p(x_i,z_i|\theta)}{Q_i(z_i)} = C \quad \Rightarrow p(x_i,z_i|\theta) = C\,Q_i(z_i) \ \Rightarrow \boxed{\sum_{z_i}p(x_i,z_i|\theta) = C\sum_{z_i}Q_i(z_i) = C}$$

$$\sum_{z_i}p(x_i,z_i|\theta) = C$$

$$Q_i(z_i) = \frac{p(x_i,z_i|\theta)}{C} = \frac{p(x_i,z_i|\theta)}{\sum_{z_i}p(x_i,z_i|\theta)} = \frac{p(x_i,z_i|\theta)}{p(x_i|\theta)} = p(z_i|x_i;\theta)$$

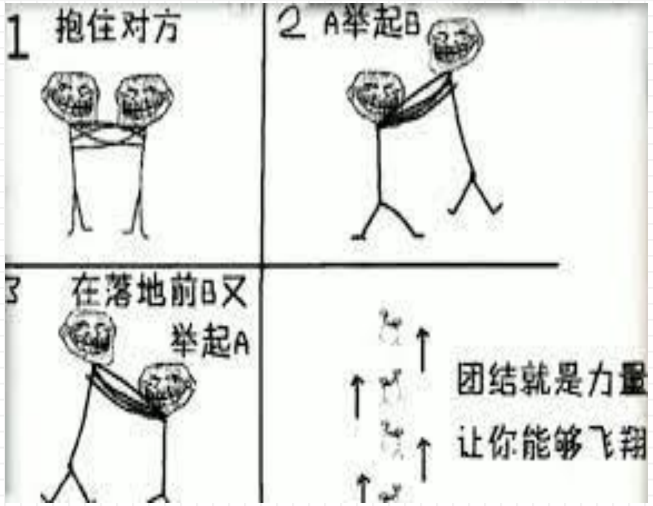
2.3 算法流程

Input :独立同分布的可观察样本 X ，隐变量 Z，完整观测样本(X,Z)，初始化参数θ。

- **隐变量Z，不是隐了吗，怎么还能输入？**答：这里的**隐变量并不是未知变量**，举个例子，预测明天的天气，这是明天的天气就是个隐变量，它是有选项的，比如雷阵雨、晴天、多云等等。只是我们不知道明天天气具体是这些选项的哪一个，需要预测分析，这是隐变量。

E步： $Q_i(z_i)=p(z_i|x_i;\theta)$

M步： $\text{argmax}_\theta \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)$



2.4 收敛性分析

迭代公式有了，但我们的目标是，要求得θ和$Q_i(z_i)$**使$\sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta)}{Q_i(z_i)}\right)$逐渐变大，以促使$LL(\theta)$变大**，能满足吗?

即保证:

$$\sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta^{t+1})}{Q_i(z_i)}\right) \geq \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\ln\left(\frac{p(x_i,z_i|\theta^{t})}{Q_i(z_i)}\right)$$

从而:

$$LL(\theta^{t+1}) \geq LL(\theta^{t})$$

Reference : https://sm1les.com/2019/03/13/expectation-maximization/

di san zhang jie

第三章节

算法分析

3.1 算法分析

- 优点:
  - 降维打击，算法是抽象的，是一种求解某种算法的算法。
  - 简单。

- 缺点:
  - 对初始值敏感：需要初始化参数θ，直接影响收敛效率以及能否得到全局最优解。
  - 非凸分布难以优化，迭代次数多，容易陷入局部最优。

di si zhang jie

第四章节

算法拓展

4.1 算法拓展

高斯混合模型(Guassian Mixed Model) : 一个典型的用EM算法求解的迭代式聚类模型。

Reference : https://www.hrwhisper.me/machine-learning-guassian-mixed-model/

di wu zhang jie

第五章节

案例实操

5.1 案例实操

Pass

QQ：1366420642

THANKS

谢谢观看