Python
机器学习实战

EM算法和GMM

EM算法

期望最大算法

国际权威学术组织ICDM在2006年12月评选出数据挖掘领域的十大经典算法

无监督学习

思想简单，推导复杂

数据挖掘十大算法：一览表 IDMer

排名	挖掘主题	算法	得票数	发表时间	作者	讲解人
1	分类	C4.5	61	1993	Quinlan, J.R	*Hiroshi Motoda*
2	聚类	K-Means	60	1967	MacQueen, J.B	*Joydeep Ghosh*
3	统计学习	SVM	58	1995	Vapnik, V.N	*Qiang Yang*
4	关联分析	Apriori	52	1994	Rakesh Agrawal	*Christos Faloutsos*
5	统计学习	EM	48	2000	McLachlan, G	*Joydeep Ghosh*
6	链接挖掘	PageRank	46	1998	Brin, S.	*Christos Faloutsos*
7	集装与推进	AdaBoost	45	1997	Freund, Y.	*Zhi-Hua Zhou*
8	分类	kNN	45	1996	Hastie, T	*Vipin Kumar*
9	分类	Naïve Bayes	45	2001	Hand, D.J	*Qiang Yang*
10	分类	CART	34	1984	L.Breiman	*Dan Steinberg*

EM算法

解决什么问题（应用）

高斯混合分布

K-Means聚类

HMM

……

本质

极大似然估计法求解未知参数的最优解

隐变量（多项分布的参数）

分布的参数

用极大极大似然法估计分布的参数

设X~N（μ, σ²），μ, σ²是未知参数，x_1、x_2、x_n是来自X的一组样本，求μ, σ²的最大似然估计

- 概率密度函数

$$f(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- 联合概率密度函数

$$p(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

EM算法

- 似然函数

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

- 对数似然

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

EM算法

- 偏导数为0

$$\begin{cases} \dfrac{\partial \ln L}{\partial \mu} = 0 \\ \dfrac{\partial \ln L}{\partial \sigma^2} = 0 \end{cases}$$

$$即:\begin{cases} \dfrac{1}{\sigma^2}[\sum_{i=1}^{n} x_i - n\mu] = 0 \\ -\dfrac{\text{n}}{2\sigma^2} + \dfrac{1}{(2\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases}$$

- 求解

符合对正态分布的估计

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

EM算法

身高的分布参数

- 设X~N（μ, σ^2)
- μ, σ^2是未知参数
- x_1、x_2、x_{100}是来自100个人的身高
- 性别已知

按照正态分布的公式

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

增加隐变量后身高的分布参数

- 设X~N（μ, σ^2)
- μ, σ^2是未知参数
- x_1、x_2、x_{100}是来自100个人的身高
- 性别未知

- 两个分布、两个分布的参数

EM算法

对于每一个你抽取到的人，有两个东西需要估计或者猜测：

- 是男的是女
- 男生和女生的身高的高斯分布参数

高斯分布说：只要告诉我男生女生的分布参数，我就能判断一个人是男是女

最大似然说了：只要告诉我是男是女，我就能告诉你男生女生的高斯分布参数

鸡生蛋蛋生鸡?

EM算法

仍然用身高的例子

Exception：先随便假设各个正态分布参数（均值和方差）。求出多项分布参数；

Maximization：用E步得到的多项分布参数，重新估计各个正态分布的参数；

这时候，两个分布的概率的就变了，接着继续调整E步、M步，如此反复，直到收敛

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

}

随机变量的函数的期望

定理 设随机变量 Y 是随机变量 X 的函数 $Y=g(X)$，这里 g 是连续函数，那么

(1) 若X是离散型随机变量，且 X 的概率分布为

$$P\{X=x_i\}=p_i,\quad i=1,2,\cdots$$

则 $E(Y)=E[g(X)]=\sum_i g(x_i)p_i$.

(2) 若X是连续型随机变量，且其概率密度为 $f(x)$，

则 $E(Y)=E[g(X)]=\int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x$.

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

EM算法框架

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

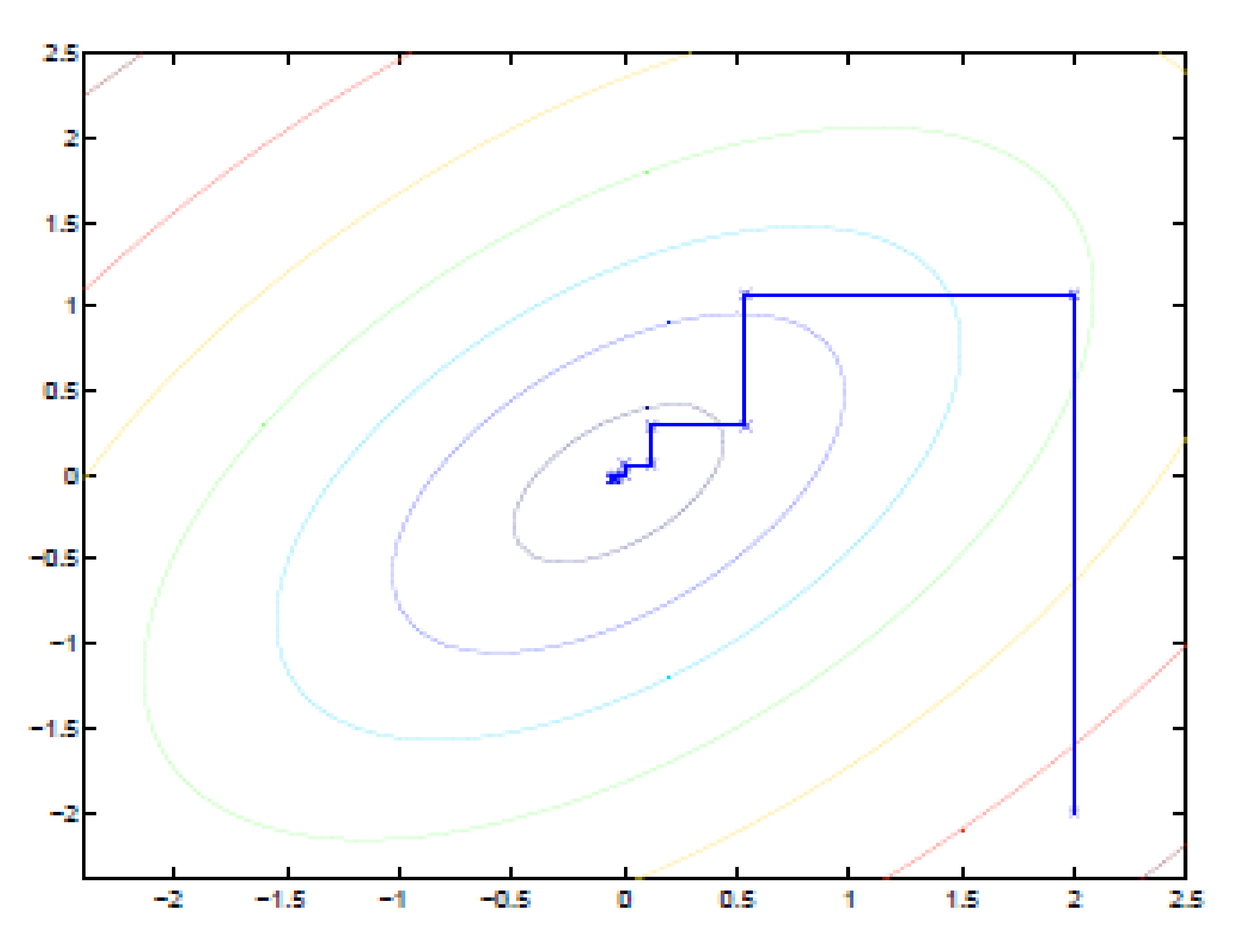
}

EM算法

EM算法和坐标上升法的对比

假设我们想估计知道A和B两个参数，在开始状态下二者都是未知的，但如果知道了A的信息就可以得到B的信息，反过来知道了B也就得到了A。

可以考虑首先赋予A某种初值，以此得到B的估计值，然后从B的当前值出发，重新估计A的取值，这个过程一直持续到收敛为止。



EM算法

EM算法和聚类

Kmeans和EM算法

数学推导

EM算法

极大似然估计

$$
\begin{aligned}
\sum_i \log p(x^{(i)};\theta) &= \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) & (1)\\
&= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} & (2)\\
&\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} & (3)
\end{aligned}
$$

※ (1)全概率公式

※ 由于(1)有和的对数，求导后形式复杂，不能用求偏导并令导数为零的方法

※ (3) Jensen不等式

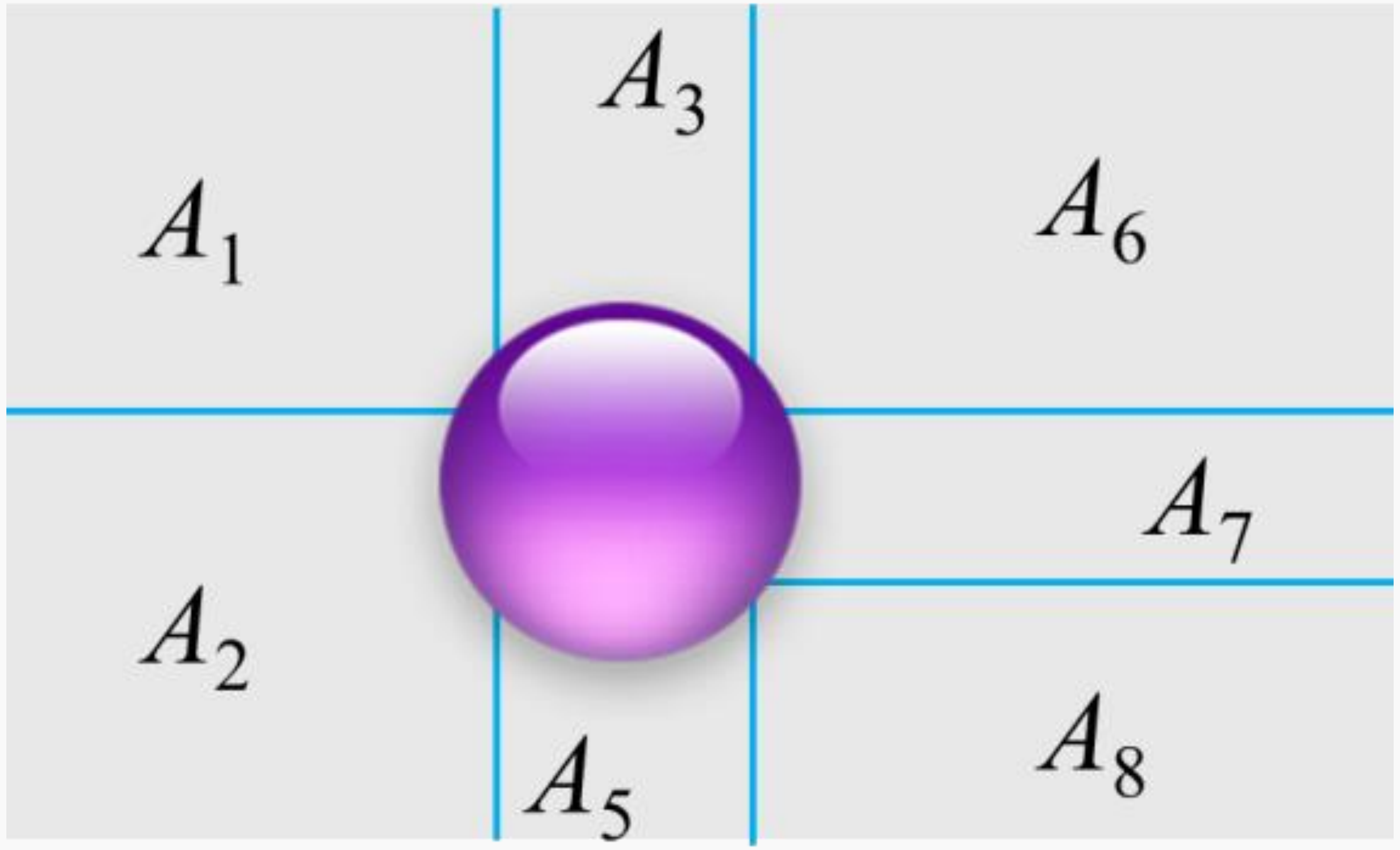
※ (3) 是对数的和，容易求导，但是从等号变成了不等号

EM算法

全概率公式

$$P(B) = P(A_1B + A_2B + \cdots + A_nB)$$
$$= \sum_{i=1}^{n} P(A_iB)$$

$$P(B) = \sum_{i=1}^{n} P(A_i)\, P(B \mid A_i)$$



高斯混合模型可以看作M个高斯密度函数的线性组合

EM算法

Jensen不等式

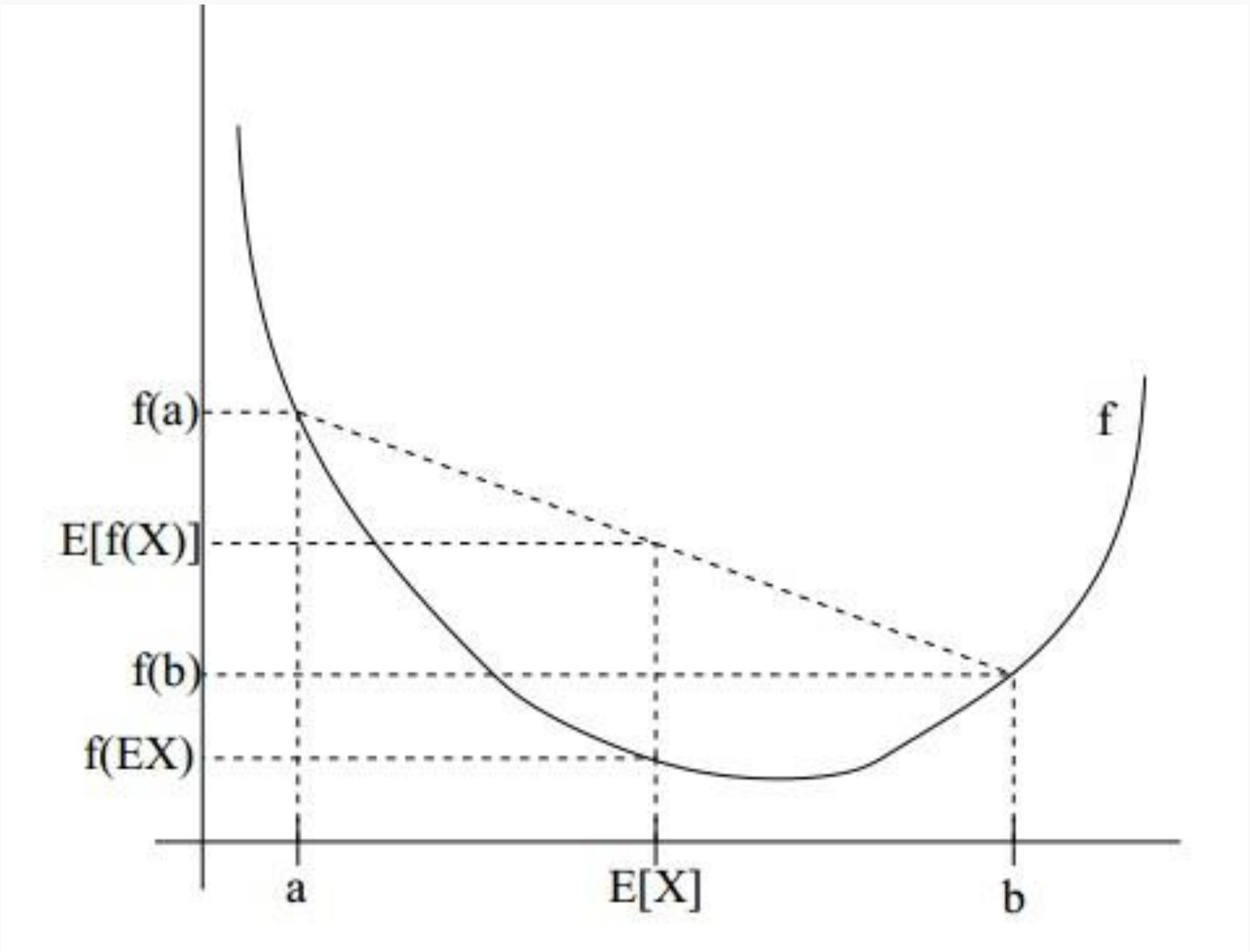
对于一个随机变量X

如果f是凸函数，那么：E[f(X)]>=f(E[X])

如果f是凹函数，那么：E[f(X)]<=f(E[x])

※ 如果f是严格凸函数，当且仅当当自变量X是常数的时候，等式成立

※ log是凹函数（二阶导数小于0）



EM算法

由Jensen不等式的结论得到

$$\frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = c \qquad \sum_z Q_i(z^{(i)}) = 1$$

$$\begin{aligned} Q_i(z^{(i)}) &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{\sum_z p(x^{(i)}, z; \theta)} \\ &= \frac{p(x^{(i)}, z^{(i)}; \theta)}{p(x^{(i)}; \theta)} \\ &= p(z^{(i)} | x^{(i)}; \theta) \end{aligned}$$

※ 为了让Jensen不等式的等号成立，推出z的分布就是z的条件分布

※ 每个样例的两个概率比值都是c

EM算法

于是有了EM算法框架

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

}

※ EM算法推导也只能推导这步，具体再M步的公式推导下去，就要结合模型了

实战

EM算法

EM算法框架

Repeat until convergence {

(E-step) For each i, set

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta).$$

(M-step) Set

$$\theta := \arg\max_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

}

※ EM算法推导也只能推导这步，具体再M步的公式推导下去，就要结合模型了

EM算法

高斯混合分布

$$\sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \phi, \mu, \Sigma)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} Q_i(z^{(i)} = j) \log \frac{p(x^{(i)}|z^{(i)} = j; \mu, \Sigma) p(z^{(i)} = j; \phi)}{Q_i(z^{(i)} = j)}$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2}(x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\right) \cdot \phi_j}{w_j^{(i)}}$$

$$\begin{cases} \mu_k = \dfrac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k) x_i \\ \Sigma_k = \dfrac{1}{N_k} \sum_{i=1}^{N} \gamma(i,k)(x_i - \mu_k)(x_i - \mu_k)^T \\ \pi_k = \dfrac{1}{N} \sum_{i=1}^{N} \gamma(i,k) \\ N_k = N \cdot \pi_k \end{cases}$$