EM算法

Bohao.Zou

EM问题描述

上面我们先假设学校所有学生的身高服从正态分布 $N(\mu, \sigma^2)$ 。实际情况并不是这样的，男生和女生分别服从两种不同的正态分布，即男生 $\in N(\mu_1, \sigma_1^2)$ ，女生 $\in N(\mu_2, \sigma_2^2)$ ，(**注意：EM算法和极大似然估计的前提是一样的，都要假设数据总体的分布，如果不知道数据分布，是无法使用EM算法的**)。那么该怎样评估学生的身高分布呢？

简单啊，我们可以随便抽 100 个男生和 100 个女生，将男生和女生分开，对他们单独进行极大似然估计。分别求出男生和女生的分布。

假如某些男生和某些女生好上了，纠缠起来了。咱们也不想那么残忍，硬把他们拉扯开。这时候，你从这 200 个人（的身高）里面随便给我指一个人（的身高），我都无法确定这个人（的身高）是男生（的身高）还是女生（的身高）。用数学的语言就是，抽取得到的每个样本都不知道是从哪个分布来的。那怎么办呢？

EM的意思是“**Expectation Maximization**” ，具体方法为：

- 先设定男生和女生的身高分布参数(初始值)，例如男生的身高分布为 $N(\mu_1 = 172, \sigma_1^2 = 5^2)$ ，女生的身高分布为 $N(\mu_2 = 162, \sigma_2^2 = 5^2)$ ，当然了，刚开始肯定没那么准；
- 然后计算出每个人更可能属于第一个还是第二个正态分布中的（例如，这个人的身高是180，那很明显，他极大可能属于男生），这个是属于Expectation 一步；
- 我们已经大概地按上面的方法将这 200 个人分为男生和女生两部分，我们就可以根据之前说的极大似然估计分别对男生和女生的身高分布参数进行估计（这不变成了**极大**似然估计了吗？**极大即为Maximization**）这步称为 Maximization；
- 然后，当我们更新这两个分布的时候，每一个学生属于女生还是男生的概率又变了，那么我们就再需要调整E步；
- .....如此往复，直到参数基本不再发生变化或满足结束条件为止。

EM公式推导（基础知识）

2.1.1 凸函数

设是定义在实数域上的函数，如果对于任意的实数，都有：

$$f'' \geq 0$$

那么是凸函数。若不是单个实数，而是由实数组成的向量，此时，如果函数的 Hesse 矩阵是半正定的，即

$$H'' \geq 0$$

是凸函数。特别地，如果 $f'' > 0$ 或者 $H'' > 0$ ，称为严格凸函数。

2.1.3 期望

对于离散型随机变量 X 的概率分布为 $p_i = p\{X = x_i\}$ ，数学期望 $E(X)$ 为：

$$E(X) = \sum_i x_i p_i$$

p_i 是权值，满足两个条件 $1 \geq p_i \geq 0$， $\sum_i p_i = 1$ 。

若连续型随机变量X的概率密度函数为 $f(x)$ ，则数学期望 $E(X)$ 为：

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

设 $Y = g(X)$ ，若 X 是离散型随机变量，则：

$$E(Y) = \sum_i g(x_i) p_i$$

若 X 是连续型随机变量，则：

$$E(X) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

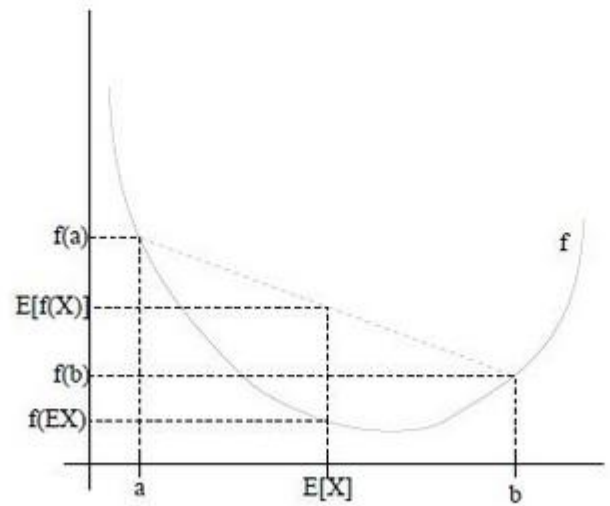
2.1.2 Jensen不等式

如下图，如果函数 f 是凸函数， x 是随机变量，有 0.5 的概率是 a，有 0.5 的概率是 b， x 的期望值就是 a 和 b 的中值了那么：

$$E[f(x)] \geq f(E(x))$$

其中， $E[f(x)] = 0.5f(a) + 0.5f(b)$， $f(E(x)) = f(0.5a + 0.5b)$ ，这里 a 和 b 的权值为 0.5， $f(a)$ 与 a 的权值相等， $f(b)$ 与 b 的权值相等。

特别地，如果函数 f 是严格凸函数，当且仅当： $p(x = E(x)) = 1$ (即随机变量是常量) 时等号成立。



注：若函数 f 是凹函数，Jensen不等式符号相反。

2.2 EM算法的推导

对于 m 个相互独立的样本 $x=(x^{(1)},x^{(2)},\dots x^{(m)})$ ，对应的隐含数据 $z=(z^{(1)},z^{(2)},\dots z^{(m)})$ ，此时 (x,z) 即为完全数据，样本的模型参数为 θ ，则观察数据 $x^{(i)}$ 的概率为 $P(x^{(i)}|\theta)$ ，完全数据 $(x^{(i)},z^{(i)})$ 的似然函数为 $P(x^{(i)},z^{(i)}|\theta)$ 。

假如没有隐含变量 z ，我们仅需要找到合适的 θ 极大化对数似然函数即可：

$$\theta=arg\max_{\theta}L(\theta)=arg\max_{\theta}\sum_{i=1}^{m}logP(x^{(i)}|\theta)$$

增加隐含变量 z 之后，我们的目标变成了找到合适的 θ 和 z 让对数似然函数极大：

$$\theta,z=arg\max_{\theta,z}L(\theta,z)=arg\max_{\theta,z}\sum_{i=1}^{m}log\sum_{z^{(i)}}P(x^{(i)},\ z^{(i)}|\theta)$$

不就是多了一个隐变量 z 吗？那我们自然而然会想到分别对未知的 θ 和 z 分别求偏导，这样做可行吗？

理论上是可行的，然而如果对分别对未知的 θ 和 z 分别求偏导，由于 $logP(x^{(i)}|\theta)$ 是 $P(x^{(i)},\ z^{(i)}|\theta)$ 边缘概率(建议没基础的同学网上搜一下边缘概率的概念)，转化为 $logP(x^{(i)}|\theta)$ 求导后形式会非常复杂（可以想象下 $log(f_1(x)+f_2(x)+\dots$ 复合函数的求导)，所以很难求解得到 θ 和 z 。那么我们想一下可不可以将加号从 log 中提取出来呢？我们对这个式子进行缩放如下：

$$\sum_{i=1}^{m}log\sum_{z^{(i)}}P(x^{(i)},\ z^{(i)}|\theta)=\sum_{i=1}^{m}log\sum_{z^{(i)}}Q_i(z^{(i)})\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})} \quad (1)$$

$$\geq\sum_{i=1}^{m}\sum_{z^{(i)}}Q_i(z^{(i)})log\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})} \quad (2)$$

上面第(1)式引入了一个未知的新的分布 $Q_i(z^{(i)})$ ，满足：

$$\sum_{z}Q_i(z)=1,0\leq Q_i(z)\leq 1$$

第(2)式用到了 Jensen 不等式(对数函数是凹函数)：

$$log(E(y))\geq E(log(y))$$

其中：

$$E(y)=\sum_{i}\lambda_i y_i,\lambda_i\geq 0,\sum_{i}\lambda_i=1$$

$$y_i=\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})}$$

$$\lambda_i=Q_i(z^{(i)})$$

也就是说 $\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})}$ 为第 i 个样本， $Q_i(z^{(i)})$ 为第 i 个样本对应的权重，那么：

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

上式我实际上是我们构建了 $L(\theta,z)$ 的下界，我们发现实际上就是 $log\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})}$ 的加权求和，由于上面讲过权值 $Q_i(z^{(i)})$ 累积和为1，因此上式是 $log\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})}$ 的加权平均，也是我们所说的期望，**这就是Expectation的来历啦。**下一步要做的就是寻找一个合适的 $Q_i(z)$ 最优化这个下界(M步)。

EM算法公式推导

EM算法公式推导

假设 θ 已经给定，那么 $logL(\theta)$ 的值就取决于 $Q_i(z)$ 和 $p(x^{(i)},z^{(i)})$ 了。我们可以通过调整 其中：
这两个概率使下界逼近 $logL(\theta)$ 的真实值，当不等式变成等式时，说明我们调整后的下界能够等价于 $logL(\theta)$ 了。由 Jensen 不等式可知，等式成立的条件是随机变量是常数，则有：

$$\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})}=c$$

其中 c 为常数，对于任意 i ，我们得到：

$$P(x^{(i)},\ z^{(i)}|\theta)=cQ_i(z^{(i)})$$

方程两边同时累加和：

$$\sum_z P(x^{(i)},\ z^{(i)}|\theta)=c\sum_z Q_i(z^{(i)})$$

由于 $\sum_z Q_i(z^{(i)})=1$ 。从上面两式，我们可以得到：

$$\sum_z P(x^{(i)},\ z^{(i)}|\theta)=c$$

$$Q_i(z^{(i)})=\frac{P(x^{(i)},\ z^{(i)}|\theta)}{c}=\frac{P(x^{(i)},\ z^{(i)}|\theta)}{\sum_z P(x^{(i)},\ z^{(i)}|\theta)}=\frac{P(x^{(i)},\ z^{(i)}|\theta)}{P(x^{(i)}|\theta)}=P(z^{(i)}|x^{(i)},\ \theta)$$

边缘概率公式：$P(x^{(i)}|\theta)=\sum_z P(x^{(i)},\ z^{(i)}|\theta)$

条件概率公式：$\frac{P(x^{(i)},\ z^{(i)}|\theta)}{P(x^{(i)}|\theta)}=P(z^{(i)}|x^{(i)},\ \theta)$

从上式可以发现 $Q(z)$ 是已知样本和模型参数下的隐变量分布。

如果 $Q_i(z^{(i)})=P(z^{(i)}|x^{(i)},\ \theta))$ ，则第 (2) 式是我们的包含隐藏数据的对数似然的一个下界。
如果我们能极大化这个下界，则也在尝试极大化我们的对数似然。即我们需要极大化下式：

$$arg\max_\theta \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)})log\frac{P(x^{(i)},\ z^{(i)}|\theta)}{Q_i(z^{(i)})}$$

至此，我们推出了在固定参数 θ后分布 $Q_i(z^{(i)})$ 的选择问题， 从而建立了 $logL(\theta)$ 的下界，这是 E 步，接下来的M 步骤就是固定 $Q_i(z^{(i)})$ 后，调整 θ ，去极大化 $logL(\theta)$ 的下界。

去掉上式中常数的部分 $Q_i(z^{(i)})$ ，则我们需要极大化的对数似然下界为：

$$arg\max_\theta \sum_{i=1}^{m}\sum_{z^{(i)}} Q_i(z^{(i)})logP(x^{(i)},\ z^{(i)}|\theta)$$

EM算法流程

2.3 EM算法流程

现在我们总结下EM算法的流程。

输入：观察数据 $x = (x^{(1)}, x^{(2)}, \dots x^{(m)})$ ，联合分布 $p(x, z|\theta)$ ，条件分布 $p(z|x, \theta)$ ，极大迭代次数 J 。

1) 随机初始化模型参数 θ 的初值 θ^0

2) for j from 1 to J ：

- E步：计算联合分布的条件概率期望：

$$Q_i(z^{(i)}) := P(z^{(i)}|x^{(i)}, \ \theta))$$

- M步：极大化 $L(\theta)$,得到 θ ：

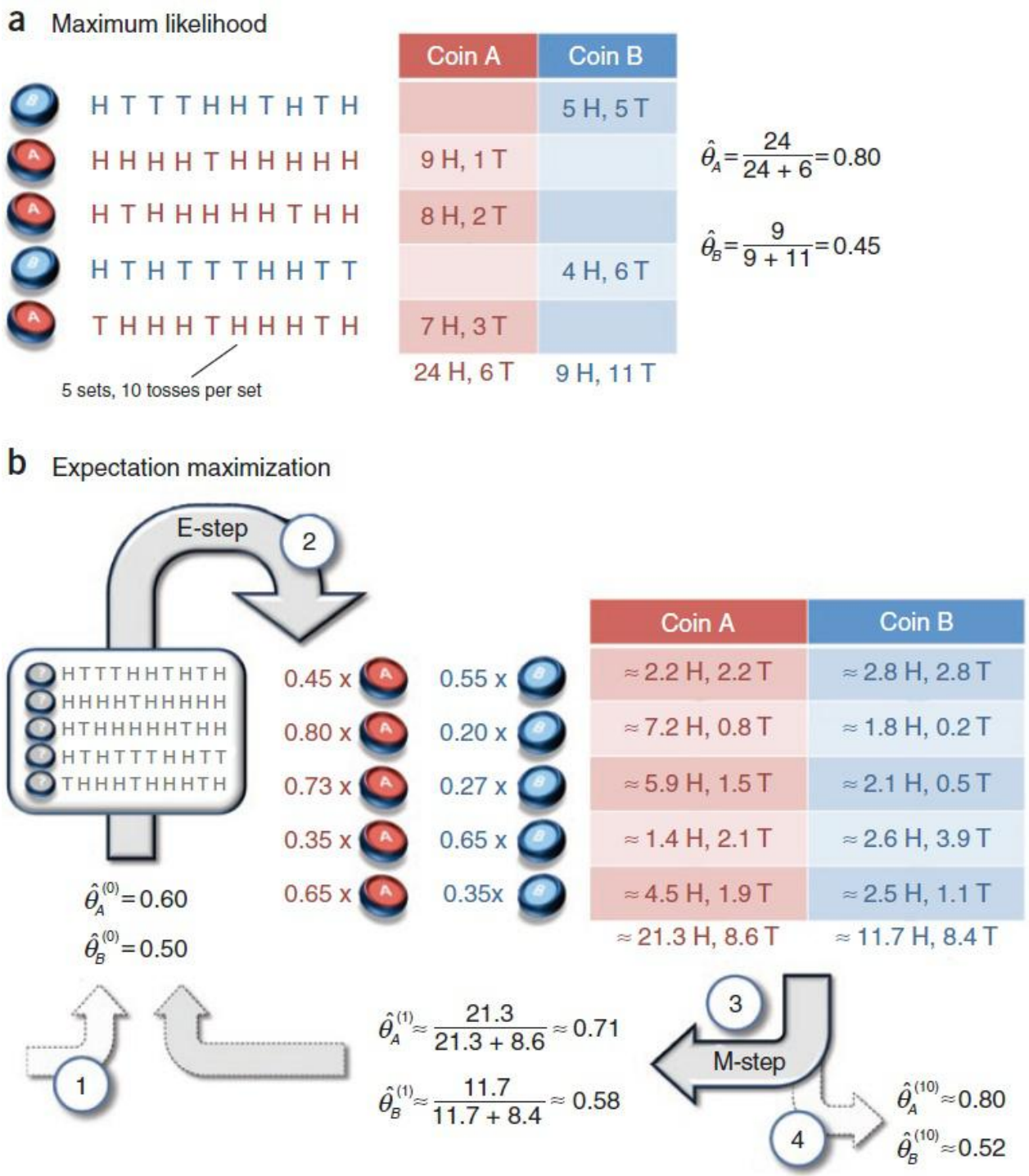
$$\theta := arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) logP(x^{(i)}, \ z^{(i)}|\theta)$$

- 重复E、M步骤直到 θ 收敛

输出：模型参数 θ

EM算法实例

- 假设有两枚硬币A、B，以相同的概率随机选择一个硬币，进行如下的掷硬币实验：共做 5 次实验，每次实验独立的掷十次，结果如图中 a 所示，例如某次实验产生了H、T、T、T、H、H、T、H、T、H (H代表正面朝上)。a 是在知道每次选择的是A还是B的情况下进行，b是在不知道选择的是A还是B的情况下进行，问如何估计两个硬币正面出现的概率？



EM算法实例

CASE b

由于并不知道选择的是硬币 A 还是硬币 B，因此采用EM算法。

E步：初始化 $\hat{\theta}_A^{(0)} = 0.60$ 和 $\hat{\theta}_B^{(0)} = 0.50$ ，计算每个实验中选择的硬币是 A 和 B 的概率，例如第一个实验中选择 A 的概率为：

$$P(z = A|y_1, \theta) = \frac{P(z = A, y_1|\theta)}{P(z = A, y_1|\theta) + P(z = B, y_1|\theta)} = \frac{(0.6)^5 * (0.4)^5}{(0.6)^5 * (0.4)^5 + (0.5)^{10}} = 0.45$$

$$P(z = B|y_1, \theta) = 1 - P(z = A|y_1, \theta) = 0.55$$

计算出每个实验为硬币 A 和硬币 B 的概率，然后进行加权求和。

M步：求出似然函数下界 $Q(\theta, \theta^i)$ ， y_j 代表第 j 次实验正面朝上的个数，μ_j 代表第 j 次实验选择硬币 A 的概率，$1 - \mu_j$ 代表第 j 次实验选择硬币B的概率 。

$$\begin{aligned} Q(\theta, \theta^i) &= \sum_{j=1}^{5} \sum_{z} P(z|y_j, \theta^i) log P(y_j, z|\theta) \\ &= \sum_{j=1}^{5} \mu_j log(\theta_A^{y_j}(1 - \theta_A)^{10-y_j}) + (1 - \mu_j) log(\theta_B^{y_j}(1 - \theta_B)^{10-y_j}) \end{aligned}$$

针对L函数求导来对参数求导，例如对 θ_A 求导：

$$\begin{aligned} \frac{\partial Q}{\partial \theta_A} &= \mu_1(\frac{y_1}{\theta_A} - \frac{10 - y_1}{1 - \theta_A}) + \cdots + \mu_5(\frac{y_5}{\theta_A} - \frac{10 - y_5}{1 - \theta_A}) = \mu_1(\frac{y_1 - 10\theta_A}{\theta_A(1 - \theta_A)}) + \cdots + \mu_5(\frac{y_5 - 10\theta_A}{\theta_A(1 - \theta_A)}) \\ &= \frac{\sum_{j=1}^{5} \mu_j y_j - \sum_{j=1}^{5} 10\mu_j\theta_A}{\theta_A(1 - \theta_A)} \end{aligned}$$

求导等于 0 之后就可得到图中的第一次迭代之后的参数值:

$$\hat{\theta}_A^{(1)} = 0.71$$

$$\hat{\theta}_B^{(1)} = 0.58$$

当然，基于Case a 我们也可以用一种更简单的方法求得：

$$\hat{\theta}_A^{(1)} = \frac{21.3}{21.3 + 8.6} = 0.71$$

$$\hat{\theta}_B^{(1)} = \frac{11.7}{11.7 + 8.4} = 0.58$$

第二轮迭代：基于第一轮EM计算好的 $\hat{\theta}_A^{(1)}, \hat{\theta}_B^{(1)}$ ，进行第二轮 EM，计算每个实验中选择的硬币是 A 和 B 的概率（E步），然后在计算M步，如此继续迭代.....迭代十步之后

$\hat{\theta}_A^{(10)} = 0.8, \hat{\theta}_B^{(10)} = 0.52$