EM算法——最大期望算法

——吴泽邦 吴林谦 万仔仁 余淼 陈志明 秦志勇

食堂的大师傅炒了一份菜，要等分成两份给两个人吃

——显然没有必要拿来天平一点一点的精确的去称分量，最简单的办法是先随意的把菜分到两个碗中，然后观察是否一样多，把比较多的那一份取出一点放到另一个碗中，这个过程一直迭代地执行下去，直到大家看不出两个碗所容纳的菜有什么分量上的不同为止

EM算法就是这样，假设我们估计知道A和B两个参数，在开始状态下二者都是未知的，并且知道了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

- **最大期望算法**（Expectation-maximization algorithm，又译期望最大化算法）在统计中被用于寻找，依赖于不可观察的隐性变量的概率模型中，参数的最大似然估计。

- 在统计计算中，最大期望算法是在概率模型中寻找参数最大似然估计或者最大后验估计的算法，其中概率模型依赖于无法观测的隐藏变量。最大期望经常用在机器学习和计算机视觉的数据聚类领域。

期望值（EXPECTED VALUE)

- 在概率和统计学中，一个随机变量的期望值是变量的输出值乘以其机率的总和，换句话说，期望值是该变量输出值的平均数
- 如果X是在概率空间（Ω, P）中的一个随机变量，那么它的期望值E[X]的定义是

$$\mathrm{E}[\mathrm{X}] = \int_{\Omega} X \, dP$$

- 离散: $\mathrm{E}[\mathrm{X}] = \sum_i p_i x_i$
- 连续: $\mathrm{E}[\mathrm{X}] = \int_{-\infty}^{\infty} x f(x) dx$

最大似然估计

某位同学与一位猎人一起外出打猎，一只野兔从前方窜过．只听一声枪响，野兔应声到下，如果要你推测，这一发命中的子弹是谁打的？

——你就会想，只发一枪便打中，由于猎人命中的概率一般大于这位同学命中的概率，看来这一枪是猎人射中的

最大似然思路

最大似然估计

- 假设我们需要调查我们学校的男生和女生的身高分布。你在校园里随便地活捉了100个男生和100个女生。男左女右，首先统计抽样得到的100个男生的身高。假设他们的身高是服从高斯分布的。但是这个分布的均值u和方差∂^2我们不知道，这两个参数就是我们要估计的。记作$\theta=[u, \partial]^T$。
- 数学语言：

  在学校那么多男生（身高）中，我们独立地按照概率密度$p(x|\theta)$抽取100了个（身高），组成样本集X，我们想通过样本集X来估计出未知参数θ。概率密度$p(x|\theta)$我们知道了是高斯分布$N(u,\partial)$的形式，其中的未知参数是$\theta=[u, \partial]^T$。

  抽到这100个人的概率：

  似然函数： $L(\theta) = L(x_1,x_2,...x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$

最大似然估计

- 上例中，在学校那么男生中，我一抽就抽到这100个男生（表示身高），而不是其他人，那是不是表示在整个学校中，这100个人（的身高）出现的概率最大啊。那么这个概率怎么表示？哦，就是上面那个似然函数L(θ)。所以，我们就只需要找到一个参数θ，其对应的似然函数L(θ)最大，也就是说抽到这100个男生（的身高）概率最大。这个叫做θ的最大似然估计量

最大似然估计

- 设总体X是离散型随机变量，其概率函数为 $p(x;\theta)$，其中θ是未知参数．设$X_1,X_2,\ldots X_n$为取自总体X的样本，$X_1,X_2,\ldots X_n$ 的联合概率函数为:

$$\prod_{i=1}^{n} p(x_i|\theta)$$ θ为常量，$X_1,X_2,\ldots X_n$为变量

- 若已知样本取值为$x_1,x_2,\ldots x_n$，则事件

  $\{X_1=x_1, X_2=x_2,\ldots X_n=x_n\}$发生的概率为$\prod_{i=1}^{n} p(x_i|\theta)$

- 显然上面的概率随θ改变而改变，从直观上来讲，既然样本值$x_1,x_2,\ldots x_n$出现，即表示其出现的概率相对较大，而使得$\prod_{i=1}^{n} p(x_i;\theta)$取较大的值，不妨看做$\theta$的函数

- 似然函数：$L(\theta) = L(x_1,x_2,\ldots x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$

最大似然估计

- 如何求L(θ)最大值?
- 考虑到有累乘，不妨取对数，这里是因为lnL函数的单调性和L函数的单调性一致，因此L(θ)的最大值转换为lnL(θ)的最大值

$$\mathbf{H}(\boldsymbol{\theta}) = \boldsymbol{lnL}(\boldsymbol{\theta}) = \boldsymbol{ln}\prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p}(\boldsymbol{x_i}|\boldsymbol{\theta}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{lnp}(\boldsymbol{x_i}|\boldsymbol{\theta})$$

- 求最值,可转换为求解下面的方程

$$\frac{\boldsymbol{dlnL}(\boldsymbol{\theta})}{\boldsymbol{d\theta}} = \mathbf{0}$$

似然方程

EXAMPLE

- 设某工序生产的产品的不合格率为 p，抽 n个产品作检验，发现有T个不合格，试求p的极大似然估计.

分析：设X是抽查一个产品时的不合格个数，则X服从参数的二点分布b(1,p).抽查n个产品，得样本$X_1,X_2,\cdots X_3$，其观察值为$x_1,x_2,\ldots,x_3$，加入样本有T个不合格，表示$x_1,x_2,\ldots,x_3$中有T个取值为1，n-T个取值为0.按照离散分布场合方法，求p的极大似然估计

解：

1. 写出似然函数 $L(p)=\prod_{i=1}^{n}p^{x_i}(1-P)^{1-x_i}$

2. 对$L(p)$取对数，得对数似然函数：

$$l(p)=\sum_{i=1}^{n}[x_i\ln p+(1-x_i)\ln(1-p)]=n\ln(1-p)+\sum_{i=1}^{n}x_i[\ln p-\ln(1-p)]$$

3. 写出似然方程

$$\frac{dl(p)}{dp}=-\frac{n}{1-p}+\sum_{i=1}^{n}x_i(\frac{1}{p}+\frac{1}{1-p})=-\frac{n}{1-p}+\frac{1}{p(1-p)}\sum_{i=1}^{n}x_i=0$$

4. 解似然方程得：$\hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{x}$

5. 验证在$\hat{p}=\bar{x}$时，$\frac{d^2l(p)}{dp^2}<0$，这表明$\hat{p}=\bar{x}$可使似然函数达到最大

小结

极大似然估计，只是一种概率论在统计学的应用，它是参数估计的方法之一。说的是已知某个随机样本满足某种概率分布，但是其中具体的参数不清楚，参数估计就是通过若干次试验，观察其结果，利用结果推出参数的大概值。最大似然估计是建立在这样的思想上：已知某个参数能使这个样本出现的概率最大，我们当然不会再去选择其他小概率的样本，所以干脆就把这个参数作为估计的真实值。

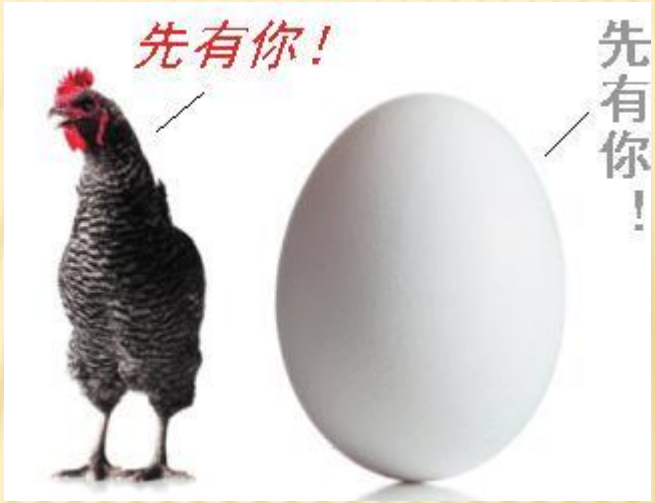
最大期望算法

- 继续回到身高的例子，我抽到这200个人中，某些男生和某些女生一见钟情，已经好上了，怎么着都不愿意分开，这时候，你从这200个人里面随便给我指一个人，我都无法确定这个人是男生还是女生。
- 也就是说你不知道抽取的那200个人里面的每一个人到底是从男生的那个身高分布里面抽取的，还是女生的那个身高分布抽取的。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抽取的。

最大期望算法

- 这个时候，对于每一个样本或者你抽取到的人，就有两个东西需要猜测或者估计的了，一是这个人是男的还是女的？二是男生和女生对应的身高的高斯分布的参数是多少？
- 只有当我们知道了哪些人属于同一个高斯分布的时候，我们才能够对这个分布的参数作出靠谱的预测；反过来，只有当我们对这两个分布的参数作出了准确的估计的时候，才能知道到底哪些人属于第一个分布，那些人属于第二个分布

先有鸡还是先有蛋



亲，还记得ppt开始分菜的厨师么？

为了解决这个你依赖我，我依赖你的循环依赖问题，总得有一方要先打破僵局，说，不管了，我先随便整一个值出来，看你怎么变，然后我再根据你的变化调整我的变化，然后如此迭代着不断互相推导，最终就会收敛到一个解

EM算法的基本思想

EM：EXPECTTATION MAXIMIZATION

- 依然用身高的例子
- Expectation：我们是先随便猜一下男生（身高）的正态分布的参数：如均值和方差是多少。例如男生的均值是1米7，方差是0.1米，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1米8，那很明显，他最大可能属于男生的那个分布）
- Maximization：有了每个人的归属，或者说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最大似然那样，通过这些被大概分为男生的n个人来重新估计第一个分布的参数，女生的那个分布同样方法重新估计
- 这时候，两个分布的概率改变了，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为止

QUESTIONS

- 你老迭代迭代的，你咋知道新的参数的估计就比原来的好啊?
- 为什么这种方法行得通呢?
- 有没有失效的时候呢?
- 什么时候失效呢?
- 用到这个方法需要注意什么问题呢?

EM算法推导

- 假设我们有一个样本集$\{x^{(1)},...,x^{(m)}\}$，包含m个独立的样本。但每个样本i对应的类别$z^{(i)}$是未知的（相当于聚类），也即隐含变量。故我们需要估计概率模型p(x,z)的参数θ，但是由于里面包含隐含变量z，所以很难用最大似然求解，但如果z知道了，那我们就很容易求解了。

EM算法推导

- 这里把每个人（样本）的完整描述看做是三元组$y_i=\{x_i,z_{i1},z_{i2}\}$，
- x_i是第i个样本的观测值
- z_{i1}和z_{i2}表示利用男女哪个高斯分布，隐含变量

z_{ij}在x_i由第j个高斯分布产生时值为1，否则为0 。

例如一个样本的观测值为1.8，来自男生高斯分布，则样本表示为{1.8, 1, 0}。

即若z_{i1}和z_{i2}的值已知，也就是说每个人我已经标记为男生或者女生了

- 对于参数估计，我们本质上还是想获得一个使似然函数最大化的那个参数θ，现在与最大似然不同的只是似然函数式中多了一个未知的变量z

$$\sum_i \log p(x^{(i)};\theta) = \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)};\theta) \quad (1)$$

$$= \sum_i \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \quad (2)$$

$$\geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \quad (3)$$

- 也就是说我们的目标是找到适合的θ和z让L(θ)最大

- （1）式最大化，也就是最大化似然函数，但是可以看到里面有“和的对数”，求导后形式会非常复杂，所以很难求解得到未知参数z和θ。
- （2）式只是分子分母同乘以一个相等的函数，还是有“和的对数”啊，还是求解不了
- （3）式变成了“对数的和”，那这样求导就容易了。我们注意点，还发现等号变成了不等号

为什么能这么变？ → Jensen不等式

- Jensen不等式

  f凸函数： E[f(X)] >= f(E[X])

  f凹函数： E[f(X)] <= f(E[X])

- f(x) = log x, 二次导数为$-1/x^2<0$, 为凹函数

（注意：国内外凹凸函数定义不同，本处采用国际定义）

EM算法流程

- **E步骤**：根据参数初始值或上一次迭代的模型参数记$\theta^{(n)}$,来求一个分布q(z)，使得L(q,θ)最大化

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

- **M步骤**：固定q(z)，求一个θ，记为$\theta^{(n+1)}$,使得L(q,θ)最大

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

PROBLEMS

- 局部最优
- 收敛速度

怎么解决?

THANK YOU