十月算法班第一讲：概率论与数理统计

管老师

七月在线

June, 2016

主要内容

- 数学基础：微积分选讲
  - 极限
  - 微分与泰勒级数
  - 积分与微积分基本定理
  - 牛顿法
  - 参考资料与作业
- 概率论与数理统计
  - 概率与积分
  - 条件概率与贝叶斯公式
  - 大数定律与中心极限定理
  - 矩估计与极大似然估计
  - 参考资料与作业

极限

通俗语言

函数 f 在 x_0 处的极限为 L

数学记号

$$\lim_{x \to x_0} f(x) = L$$

精确描述:$\epsilon - \delta$ 语言

对于任意的正数 $\epsilon > 0$, 存在正数 δ, 使得任何满足 $|x - x_0| < \delta$ 的 x, 都有

$$|f(x) - L| < \epsilon$$

通俗语言适合于说给对方听，数学记号适合于写给对方看，精确描述比较啰嗦但是非常精确不会造成误解，主要用于证明.

极限: 无穷小阶数

Definition (无穷小阶数)

当 $x \to 0$ 时,

- 如果 $\lim\limits_{x\to 0} f(x) = 0$ 而且 $\lim\limits_{x\to 0} f(x)/x^n = 0$ 那么此时 $f(x)$ 为 n 阶以上无穷小，记为

$$f(x) = o(x^n), x \to 0$$

- 如果 $\lim\limits_{x\to 0} f(x) = 0$ 而且 $\lim\limits_{x\to 0} f(x)/x^n$ 存在且不等于零，那么此时 $f(x)$ 为 n 阶无穷小，记为

$$f(x) = O(x^n), x \to 0$$

为了方便，在不至于引起误解的时候我们回省略掉 $x \to 0$.

所谓无穷小的阶数，就是用我们比较熟悉的多项式类型的无穷小量来衡量其他的无穷小量.

微分学

微分学的核心思想: 逼近.

Definition (函数的导数)

如果一个函数 $f(x)$ 在 x_0 附近有定义，而且存在极限

$$L = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}.$$

那么 $f(x)$ 在 x_0 处可导且导数 $f'(x_0) = L$.

等价定义

无穷小量表述: 线性逼近

如果存在一个实数 L 使得 $f(x)$ 满足,

$$f(x) = f(x_0) + L(x - x_0) + o(x - x_0), x \to x_0.$$

那么 $f(x)$ 在 x_0 处可导且导数 $f'(x_0) = L$.

求导法则

- 链式法则:$\frac{d}{dx}(g \circ f) = \frac{dg}{dx}(f) \cdot \frac{df}{dx}$
- 加法法则:$\frac{d}{dx}(g + f) = \frac{dg}{dx} + \frac{df}{dx}$
- 乘法法则:$\frac{d}{dx}(g \cdot f) = \frac{dg}{dx} \cdot f + g \cdot \frac{df}{dx}$
- 除法法则:$\frac{d}{dx}(\frac{g}{f}) = \frac{\frac{dg}{dx} \cdot f - \frac{df}{dx} \cdot g}{f^2}$
- 反函数求导:$\frac{d}{dx}(f^{-1}) = \frac{1}{\frac{df}{dx}(f^{-1})}$

所有求导法则原则上都可以由链式法则结合二元函数的偏导数来推出来，有兴趣的同学可以思考一下这是为什么

Definition (函数的高阶导数)

如果函数的导数函数仍然可导，那么导数函数的导数是二阶导数，二阶导数函数的导数是三阶导数. 一般地记为

$$f^{(n)}(x) = \frac{d}{dx} f^{(n-1)}(x)$$

或者进一步

$$f^{(n)}(x) = \frac{d^n}{dx^n} f(x)$$

导数是对函数进行线性逼近，高阶导数是对导数函数的进一步逼近，因为没有更好的办法，所以数学家选择继续使用线性逼近.

一元微分学的顶峰：泰勒级数

用多项式逼近的方式描述高阶导数，我们就得到了泰勒级数.

泰勒/迈克劳林级数: 多项式逼近

如果 $f(x)$ 是一个无限次可导的函数，那么在任何一点 x_0 附近我们可以对 $f(x)$ 做多项式逼近:

$$\begin{aligned} f(x_0+\Delta_x) =& f(x_0)+f'(x_0)\Delta_x+\frac{f''(x_0)}{2}\Delta_x^2+\cdots \\ &+\frac{f^{(n)}(x_0)}{n!}\Delta_x^n+o(\Delta_x^n) \end{aligned}$$

在本课中我们不关注对于尾巴上的余项 $o(\Delta_x^n)$ 的大小估计

Definition (单变量函数黎曼积分)

令 $f(x)$ 为开区间 (a,b) 上的一个连续函数，对于任何一个正整数 n 定义,$x_i = a + \frac{i(b-a)}{n}$ 求和式:

$$S_n(f) = \sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i)$$

如果极限 $\lim_{n\to\infty} S_n(f)$ 存在, 那么函数 $f(x)$ 在这个区间上的黎曼积分为

$$\int_a^b f(x)dx = \lim_{n\to\infty} S_n(f)$$

积分学: 理解积分：无穷求和，体积

理解积分

- 代数意义: 无穷求和
- 几何意义: 函数与 X 轴之间的有向面积

此处课堂画图举例说明

积分学: 微积分基本定理: 牛顿-莱布尼茨公式

Theorem (牛顿-莱布尼茨公式)

如果 $f(x)$ 是定义在闭区间 $[a,b]$ 上的可微函数, 那么就有

$$\int_a^b f'(t)dt = f(b) - f(a)$$

不定积分表示为

$$\int f'(t)dt = f(x) + C$$

牛顿-莱布尼茨公式展示了微分与积分的基本关系: 在一定程度上微分与积分互为逆运算.

Example

函数 $\ln(x)$ 的不定积分

令 $f(x) = x\ln(x) - x$, 则 $f'(x) = 1 \cdot \ln(x) + x \cdot \frac{1}{x} - 1 = \ln(x)$.
根据牛顿-莱布尼茨公式我们得到

$$\int \ln(t)dt = \int f'(t)dt = x\ln(x) - x + C$$

牛顿法

很多机器学习或者统计的算法最后都转化成一个优化的问题. 也就是求某一个损失函数的极小值的问题, 在本课范围内我们考虑可微分的函数极小值问题.

优化问题

对于一个无穷可微的函数 $f(x)$，如何寻找他的极小值点.

极值点条件

- 全局极小值: 如果对于任何 $\tilde{x}$, 都有 $f(x_*) \leq f(\tilde{x})$，那么 x_* 就是全局极小值点.
- 局部极小值: 如果存在一个正数 δ 使得，对于任何满足 $|\tilde{x} - x_*| < \delta$ 的 $\tilde{x}$, 都有 $f(x_*) \leq f(\tilde{x})$，那么 x_* 就是局部极小值点.（方圆 δ 内的极小值点）
- 不论是全局极小值还是局部极小值一定满足一阶导数/梯度为零，$f' = 0$ 或者 $\nabla f = 0$.

牛顿法

局部极值算法

我们本节课利用极值点条件，来介绍牛顿法.

- 这种方法只能寻找局部极值
- 这种方法要求必须给出一个初始点 x_0
- 数学原理：牛顿法使用二阶逼近
- 牛顿法对局部凸的函数找到极小值，对局部凹的函数找到极大值，对局部不凸不凹的可能会找到鞍点.
- 牛顿法要求估计二阶导数.

牛顿法

牛顿法：二次逼近

首先在初始点 x_0 处，写出二阶泰勒级数

$$f(x_0+\Delta_x) = f(x_0) + f'(x_0)\Delta_x + \frac{f''(x_0)}{2}\Delta_x^2 + o(\Delta_x^2) \quad (1)$$

$$= g(\Delta_x) + o(\Delta_x^2) \quad (2)$$

我们知道关于 Δ_x 的二次函数 $g(\Delta_x)$ 的极值点为 $-\frac{f'(x_0)}{f''(x_0)}$. 那么本着逼近的精神 $f(x)$ 的极值点估计在 $x_0 - \frac{f'(x_0)}{f''(x_0)}$ 附近，于是定义 $x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$，并重复此步骤得到序列

$$x_n = x_{n-1} - \frac{f'(x_{n-1})}{f''(x_{n-1})}$$

当初始点选的比较好的时候 $\lim\limits_{n\to\infty} x_n$ 收敛于一个局部极值点.

微积分选讲：参考资料

参考资料

- 数学分析教程，常庚哲，史济怀
- 简明微积分，龚升
- 微积分讲义，陈省身

作业

- 作业，数学分析教程，常庚哲，史济怀 (p142:2,3,7,8; p143:3,4,6; p148:2,3,6; p176:8,11; p210:4,5; p211:6)

随机变量与概率：概率密度函数的积分

离散随机变量

假设随机变量 X 的取值域为 $\Omega = \{x_i\}_{i=1}^{\infty}$，那么对于任何一个 x_i，事件 $X = x_i$ 的概率记为 $P(x_i)$.

对于 Ω 的任何一个子集 $S = \{x_{k_i}\}_{i=1}^{\infty}$，事件 $X \in S$ 的概率为

$$P(S) = \sum_{i=1}^{\infty} P(x_i)$$

对于离散随机变量，概率为概率函数的求和.

随机变量与概率：概率密度函数的积分

连续随机变量

假设随机变量 X 的取值域为 $\mathbb{R}$，那么对于几乎所有 $x \in \mathbb{R}$, 事件 $X = x$ 的概率 $P(X = x)$ 都等于 0. 所以我们转而定义概率密度函数 $f : \mathbb{R} \to [0, \infty)$. 对于任何区间 (a, b), 事件 $X \in (a, b)$ 的概率为

$$P((a, b)) = \int_a^b f(x)dx$$

- 对于连续型随机变量，概率为概率密度函数的积分.
- 不论是离散还是连续型随机变量, 概率函数和概率密度函数的定义域即为这个随机变量的值域.
- 作为一个特殊的概率函数，分布函数定义为 $\Phi(x) = P(X < x)$.

我们在此课中只考虑几乎处处连续的概率密度函数，我们不考虑离散，连续混合型的随机变量

随机变量与概率：如何理解概率

事件的概率

- 整个概率空间是一个事件，这个事件一定发生所以全空间的概率为 1
- 事件是随机变量值域的子集 S
- 事件的概率则表示 S 里面概率之和或概率密度之积分.

事件的条件概率

- 条件也是事件，也可表示为随机变量值域的子集:A
- 条件概率里面的事件，又是这个条件的子集:$S \cap A \subset A$
- 事件的条件概率则表示 $S \cap A$ 在 A 里面所占的比例. 故而 $P(S|A) = \frac{P(S \cap A)}{P(A)}$

概率其实就是集合的大小比例，而概率函数或者概率密度函数可以理解为比较大小时候的权重

随机变量与概率：贝叶斯公式

贝叶斯公式

如果 A, B 是两个事件，那么条件概率满足公式

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

利用前面的定义我们知道，事件 A, B 同时发生的概率为 $P(A \cap B)$, 一方面

$$P(A \cap B) = P(B|A)P(A)$$

另一方面对称的有

$$P(A \cap B) = P(A|B)P(B)$$

所以 $P(B|A)P(A) = P(A|B)P(B)$, 两边同时除以 $P(B)$ 就得到了贝叶斯公式.

大数定律和中心极限定理

随机变量的矩

X 是一个随机变量对于任何正整数 n，定义

$$E(X^n) = \int p(x)x^n dx$$

- 当 $n = 1$ 时，$E(X)$ 为随机变量的期望
- 当 $n = 2$ 时，$E(X^2) - E(X)^2$ 为随机变量的方差
- 特征函数，$E(e^{itX}) = \sum_{n=0}^{\infty} \frac{E(X^n)}{n!}(it)^n$.

矩可以描述随机变量的一些特征，期望是 X “中心”位置的一种描述，方差可以描述 X 的分散程度, 特征函数可以全面描述概率分布.

大数定律和中心极限定理

大数定律

X 是随机变量，μ 是 X 的期望，σ 是 X 的方差.$\{X_k\}_{k=1}^{\infty}$ 是服从 X 的独立同分步随机变量，那么 $\bar{X}_n = \frac{\sum_{k=1}^{n} X_k}{n}$ 依概率收敛于 μ. 也就是说对于任何 $\epsilon > 0$ 有

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

大数定律和中心极限定理

中心极限定理

X 是随机变量,$\phi(X)$ 是 X 的特征函数.$\{X_k\}_{k=1}^{\infty}$ 是服从 X 的独立同分步随机变量，那么

$$Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)$$

依分布收敛于正态分布 $N(0,1)$.
也就是说对于任何 $\epsilon > 0$ 有

$$\lim_{n\to\infty} P(Z_n < z) = \Phi(z), \quad \forall z$$

其中 Φ 是标准正态分布的分布函数.

参数估计

参数估计问题

- 已知一个随机变量的分布函数 $X \sim f_\theta(x)$, 其中 $\theta = (\theta_1, \cdots, \theta_k)$ 为未知参数.
- 样本 $X_1, \cdots, X_n$
- 利用样本对参数 θ 做出估计，或者估计 θ 的某个函数 $g(\theta)$
  - 点估计: 用样本的一个函数 $T(X_1, \cdots, X_n)$ 去估计 $g(\theta)$
  - 区间估计: 用一个区间去估计 $g(\theta)$

点估计：矩估计

矩估计的基本原理：大数定律

根据大数定律我们知道, 对于任何随机变量 X, 当样本数 $n \to \infty$ 时, $\frac{1}{n}\sum_{i=1}^{n} X_i$ 收敛于 $E(X)$. 所以

$$a_1(X) \to \alpha_1(X)$$

对于任意的 k 阶矩，令 $Y = X^k$，那么 Y 也是一个随机变量，所以同样满足大数定律，于是

$$a_k(X) = a_1(Y) \to \alpha_1(Y) = \alpha_k(X)$$

而中心矩都可以表示成原点矩的多项式，所以我们同样有

$$m_k(X) \to \mu_k(X)$$

点估计：矩估计

Example (两点分布的参数估计)

X 服从两点分布取值为 $\{-1,1\}$，$P(-1)=1-\theta, P(1)=\theta$. 现在独立重复实验 n 次，得到样本 $X_1,\cdots,X_n$. 请利用矩估计来估计参数 θ.

首先考虑哪一个矩可以用来估计参数 θ. 对于两点分布来说

$$E(X)=(1-\theta)\cdot(-1)+\theta\cdot 1=2\theta-1$$
$$E(X^2)=(1-\theta)\cdot 1+\theta\cdot 1=1$$

我们看到一阶矩 $E(X)$ 与 θ 有简单直接的关系 $\theta=\frac{1+E(X)}{2}$
所以我们使用一阶样本矩估计. 得到一个参数估计量 $\hat{\theta}=\frac{1+\overline{X}}{2}$.

点估计：极大似然估计

极大似然估计

- 给定随机变量的分布与未知参数，利用观测到的样本计算似然函数
- 选择最大化似然函数的参数作为参数估计量.

点估计：极大似然估计

极大似然估计基本原理：最大化似然函数

假设样本 $\{X_1, \cdots, X_n\}$ 服从概率密度函数 $f_\theta(x)$. 其中 $\theta = (\theta_1, \cdots, \theta_k)$ 是未知参数.
当固定 x 的时候，$f_\theta(x)$ 就是 θ 的函数，我们把这个函数称为似然函数, 记为 $L_x(\theta)$ 或 $L(\theta)$.

似然函数不是概率，但是很类似于概率. 当 θ 给定的时候，它是概率密度。当 x 给定，θ 变化的时候，他就类似于在表示，在这个观测量 x 的条件下，参数等于 θ 的可能性 (不是概率). 起个名字叫做似然函数.

点估计：极大似然估计

极大似然估计基本原理：最大化似然函数

假设 $x = (x_1, \cdots, x_n)$ 是样本的观测值. 那么整个样本的似然函数就是

$$L_x(\theta) = \prod_{i=1}^{n} L_{x_i}(\theta)$$

这是一个关于 θ 的函数，选取使得 $L_x(\theta)$ 最大化的 $\hat{(\theta)}$ 作为 θ 的估计量.

最大化似然函数 θ，相当于最大化似然函数的对数 $l_x(\theta) = \ln(L_x(\theta))$. 一般我们求解似然函数或者对数似然函数的驻点方程

$$\frac{dl(\theta)}{d\theta} = 0, (\text{或者} \frac{dL(\theta)}{d\theta} = 0)$$

然后判断整个驻点是否最大点.(求驻点可以用牛顿法，或者梯度法等等)

点估计：极大似然估计

Example (正态分布的参数估计)

X 服从参数为 $\theta = (\mu, \sigma)$ 的正态分布，独立重复实验 n 次得到样本 $X_1, \cdots, X_n$. 请利用极大似然估计来估计参数 θ.

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

$$l(\mu, \sigma) = C - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2} \ln(\sigma^2)$$

所以似然方程为 $\frac{\partial l}{\partial \sigma} = \frac{\partial l}{\partial \mu} = 0$, 也就是

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

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

因此得到极大似然估计量

$$\hat{\mu} = \overline{X}$$

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

点估计：点估计的评判准则

- 相合性 (consistency): 当样本数量趋于无穷时，估计量收敛于参数真实值.
- 无偏性 (bias): 对于有限的样本，估计量所符合的分布之期望等于参数真实值.
- 有效性 (efficiency): 估计值所满足的分布方差越小越好.
- 渐进正态性 (asymptotic normality): 当样本趋于无穷时，去中心化去量纲化的估计量符合标准正态分布.

思考题

请考虑上面例子中对于正态分布方差的极大似然估计是否是无偏估计?

概率与统计：参考资料

参考资料

- 概率统计部分建议参考中科大统计系张卫平老师的课程材料，每一章节比较简明容易阅读：
- http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/,lec(14,15)
- http://staff.ustc.edu.cn/~zwp/teach/Prob-Stat/,lec(4,5,6,7,8)

作业

- 作业：在上述网站中找到今天所讲内容对应的章节并选择阅读，请阅读下面两个在七月问答里面的帖子。
- 我的一个贝叶斯后验估计的例子https://ask.julyedu.com/question/7190
- 我的一个关于 EM 算法的介绍https://ask.julyedu.com/question/7287

谢谢大家!