深度学习班第三期第 1 课

七月在线 管博士

Jun, 2017

主要内容

- 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
- 概率统计选讲
  - 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

主要内容

- 线性代数选讲
  - 线性映射与矩阵
  - 矩阵变换与特征值
  - 奇异值分解
  - 应用举例：PCA
  - 参考资料
- 凸优化选讲
  - 优化与凸优化
  - 凸集与凸函数
  - 对偶问题与 KKT 条件
  - 应用举例：SVM 的最简单形式
  - 参考资料

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 常庚哲和史济怀老师把泰勒级数称为是一元微分学的顶峰，我也同意这个观点

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 $\lim_{n\to\infty} x_n$ 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

- 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

线性空间与基

实系数线性空间是一个由向量组成的集合, 向量之间可以做加减法, 向量与实数之间可以做乘法, 而且这些加，减，乘运算要求满足常见的交换律和结合律. 我们也可以类似地定义其他系数的线性空间.

Example (线性空间)

有原点的平面。

- 如果平面有一个原点 O, 那么平面上任何一个点 P, 都对应着一个向量 $\overrightarrow{OP}$。
- 这些向量以及他们的运算结构放在一起，就组成一个向量空间。

- 原点 O 在空间中引入了线性结构。(向量之间的加法，以及向量与实数的乘法)

线性空间与基

基是线性空间里的一组向量，使得任何一个向量都可以唯一的表示成这组基的线性组合.

Example (坐标空间)

有原点的平面，加上一组基 $\{\vec{X}, \vec{Y}\}$。

- 任何一个向量 $\overrightarrow{OP}$, 都可以唯一表达成 $\overrightarrow{OP} = a\vec{X} + b\vec{Y}$ 的形式。
- (a, b) 就是 P 点的坐标。

- 基给出了定量描述线性结构的方法——坐标系。

线性映射与矩阵

Definition (线性映射)

V 和 W 是两个实线性空间，$T : V \to W$ 如果满足如下条件就是一个线性映射。

$$
\begin{aligned}
&(i) \quad T(v_1 + v_2) = T(v_1) + T(v_2), && \forall v_1, v_2 \in V \\
&(ii) \quad T(\lambda v) = \lambda T(v), && \forall \lambda \in \mathbb{R}, v \in V
\end{aligned}
$$

- 线性映射的本质就是保持线性结构的映射
- 到自身的线性映射 $T : V \to V$ 叫做线性变换

线性映射与矩阵

线性变换的矩阵描述

V, W 分别为 n, m 维的线性空间，
$\alpha = \{\alpha_1, ..., \alpha_n\}, \beta = \{\beta_1, ..., \beta_m\}$ 分别为 V, W 的一组基。
$T : V \to W$ 是一个线性映射。于是 T, α, β 唯一决定一个矩阵 $A_{\alpha,\beta}(T) = [A_{ij}]_{m \times n}$，使得

$$T(\alpha_j) = \sum_{i=1}^{m} A_{ij} * \beta_i, \forall j \in 1, ..., n \tag{3}$$

(3) 等价于

$$T(\alpha_1, ..., \alpha_n) = (\beta_1, ..., \beta_m) \cdot A_{\alpha,\beta}(T) \tag{4}$$

简记为

$$T(\alpha) = \beta \cdot A_{\alpha,\beta}(T) \tag{5}$$

线性映射与矩阵

如果我们选取 V, W 的另外一组基，$\tilde{\alpha} = \alpha \cdot P$, $\tilde{\beta} = \beta \cdot Q$. 那么存在矩阵 $A_{\tilde{\alpha},\tilde{\beta}}(T)$ 使得,

$$T(\tilde{\alpha}) = \tilde{\beta} \cdot A_{\tilde{\alpha},\tilde{\beta}}(T)$$

两边分别代入 $\tilde{\alpha}$ 与 $\tilde{\beta}$ 得到,

$$T(\alpha) \cdot P = T(\alpha \cdot P) = \beta \cdot Q \cdot A_{\tilde{\alpha},\tilde{\beta}}(T)$$

与(5)比较我们得到矩阵变换公式:

$$Q \cdot A_{\tilde{\alpha},\tilde{\beta}}(T) \cdot P^{-1} = A_{\alpha,\beta}(T) \tag{6}$$

小结 (线性映射与矩阵)

- 矩阵是线性映射在特定基下的一种定量描述

$$T(\alpha) = \beta \cdot A_{\alpha,\beta}(T)$$

- 基变换下的矩阵变换公式的推导方法

$$Q \cdot A_{\tilde{\alpha},\tilde{\beta}}(T) \cdot P^{-1} = A_{\alpha,\beta}(T)$$

相似变换 (把矩阵看做线性映射)

如果 $T : V \to V$ 是一个线性变换, 那么对于 V 的两组基 α 与 $\tilde{\alpha} = \alpha \cdot P$，线性变换 T 的矩阵分别为

$$A_\alpha(T) \text{ and } A_{\tilde{\alpha}}(T) = P^{-1} \cdot A_\alpha(T) \cdot P$$

方阵的相似变换

- 如果两个方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^{-1}AP$. 那么这两个方阵就互为相似矩阵
- 相似矩阵的几何意义是同一个线性变换在不同的基下的表达形式
- 当研究对象是线性变换的时候，我们只关心矩阵在相似变换下不变的几何性质。

相似变换 (把矩阵看做线性映射)

相似变换下不变的性质

- 行列式 (det)

$$\begin{aligned}\det(P^{-1}AP) &= \det(P^{-1})\det(A)\det(P)\\ &= \det(P^{-1})\det(P)\det(A)\\ &= \det(A)\end{aligned}$$

- 迹 (trace),$\mathrm{tr}(AB) = \mathrm{tr}(BA)$

$$\mathrm{tr}(P^{-1}AP) = \mathrm{tr}(APP^{-1}) = \mathrm{tr}(A \cdot I) = \mathrm{tr}(A)$$

- 秩 (rank)

相似不变量

相似变换下不变的性质

- 特征值：特征方程 $\det(A-\lambda I)=0$ 的根。
  如果 $\det(A-\lambda I)=0$，那么 $\det(P^{-1}(A-\lambda I)P)=0$, 于是 $\det(P^{-1}AP-\lambda I)=0$

- 特征值是最重要的相似不变量，利用这个相似不变量可以方便的得出上面所有的不变量。

相合变换 (二次型)

假设 V 是一个实系数线性空间，那么线性空间上的度量指的是空间中向量的内积关系 $G(v_1, v_2)$. 如果 $\alpha\{\alpha_1, \cdots, \alpha_k\}$ 是空间 V 的一组基，那么这个内积一般可以用一个对称矩阵 $H_\alpha = [h_{ij}]_{n\times n}$ 来表示.

$$h_{ij} = G(\alpha_i, \alpha_j)$$

这时候对于任意两个向量 v_1, v_2, 如果 $v_1 = \alpha \cdot x_1, v_2 = \alpha \cdot x_2$，那么

$$G(v_1, v_2) = x_1^T H_\alpha x_2$$

相合变换 (二次型)

方阵的相合变换

- 如果两个对称方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^T AP$. 那么这两个方阵就互为相合矩阵
- 相似矩阵的几何意义是同一个内积结构在不同基下的表示形式

相合变换 (二次型)

方阵的正交相似变换

正交相似变换同时满足相似与相合变换的条件，也就是说它同时保持了矩阵的相似与相合不变量。

- 如果两个对称方阵 A 和 $\tilde{A}$ 满足，$\tilde{A} = P^T AP$. 而且 P 是正交矩阵:$P^T = P^{-1}$. 那么这 A 与 $\tilde{A}$ 就互为正交相似.

相合变换 (二次型)

方阵的正交相似标准型

任何一个对称矩阵 A 都可以正交相似于一个对角矩阵 D.

总存在一个正交矩阵 P 使得，$A = P^T D P$.

相合变换 (二次型)

长方矩阵的奇异值分解 (SVD)

对于任何一个矩阵 $B_{m\times n}$, 存在正交矩阵 $P_{m\times m}$,$Q_{n\times n}$. 使得

$$B = PDQ$$

其中 $D_{m\times n}$ 是一个只有对角元素不为零的矩阵.

应用举例：PCA

主成分分析 (PCA)

PCA 的主要目的是降维，也可以起到分类的作用

- 当数据维度很大的时候，如果相信大部分变量之间存在线性关系，那么我们就希望降低维数，用较少的变量来抓住大部分的信息.
- 一般来讲做 PCA 之前要做 normalization 使得变量中心为 0，而且方差为 1.

比较广泛应用于图像识别，文档处理，推荐系统

应用举例：PCA

主成分分析 (PCA)

- 首先计算变量之间的协方差矩阵 Σ(利用样本)
- 找到 Σ 的正交相似标准型

正交相似标准性的求解由计算机完成，我们主要关心他的几何意义

应用举例：PCA

推荐系统

如果一个旅游网站里面有 10000000 个注册用户，以及 40000 个注册酒店. 网站有用户通过本网站点击酒店页面的记录信息.$A = [A_{ij}]_{10000000 \times 40000}, A_{ij}$ 表示第 i 个用户点击 j 酒店的次数.

- 如何评价酒店之间的相似度?
- 给定一个酒店，请找出与它最相似的其他几个酒店?
- 如果要给酒店分类，有什么办法?

线性代数：参考资料

参考资料

- 陶哲轩的讲义
  `http://www.math.ucla.edu/~tao/resource/general/115a.3.02f/`
- 入门教材：线性代数及其应用, 莱 (Lay D.C.) (作者), 刘深泉 (译者)
- 艰深教材：线性代数，李炯生、查建国

作业

- 入门教材：线性代数及其应用, 莱 (Lay D.C.) (作者), 刘深泉 (译者)
  p69:8,13,15,16; p79:7,14
  p185:18,30; p186:9,14,15
  p242:6,17,18; p292:5,6
  p293:19,20,22,25,26; p99:25,26,27,35
  p413:9,11; p423:17,18,21
  p431:1

优化与凸优化简介

优化问题

优化问题的一般形式

$$\begin{aligned} \text{最小化：} & \quad f_0(x) \\ \text{条件：} & \quad f_i(x) \leq b_i, i = 1, \cdots, m. \end{aligned}$$

其中 $f_0(x)$ 为目标函数，条件里的不等式是限制条件.

凸优化问题

一个优化问题如果满足如下条件则为凸优化问题

- 凸优化问题的条件,$f_0, f_1, \cdots, f_m$ 都是凸函数.
- 凸优化问题的特点, 局部最优等价于全局最优.

优化与凸优化简介

举例：

极大似然估计

如果 $L(\mu, \sigma)$ 是一个极大似然估计问题中的似然函数，其中 μ, σ 分别是期望和方差，那么极大似然估计的问题转化为

$$\text{最小化：} \quad -L(\mu, \sigma)$$
$$\text{条件：} \quad \sigma \geq 0$$

最小二乘

如果 $A_{n\times k}$ 是一个矩阵,$b \in \mathbb{R}^n$ 是一个向量, 对于 $x \in \mathbb{R}^k$

$$\text{最小化：} \quad f_0(x) = |Ax - b|^2$$

优化与凸优化简介

凸优化的应用

- 凸优化问题逼近非凸优化问题，寻找非凸问题的初始点
- 利用对偶问题的凸性给原问题提供下界估计
- 凸优化问题可以给非凸问题带来一些启发

凸集合与凸函数

凸集合定义

如果一个集合 Ω 中任何两个点之间的线段上任何一个点还属于 Ω, 那么 Ω 就是一个凸集合.i.e.

$$\lambda x_1 + (1-\lambda)x_2 \in \Omega, \forall x_1, x_2 \in \Omega, \lambda \in (0,1)$$

凸函数定义

如果一个函数 f 定义域 Ω 是凸集，而且对于任何两点. 以及两点之间线段上任意一个点都有

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$

$\forall x_1, x_2 \in \Omega, \lambda \in (0,1)$

凸集合与凸函数

函数的上境图

假设 f 是一个定义在 Ω 上的函数，区域 $\{(x,y):y\geq f(x),\forall x\in\Omega\}$ 就是 f 的上境图.

上境图就是函数图像上方的部分区域.

凸集合与凸函数

一个函数是凸函数当且仅当 f 的上境图是凸集合.

凸集合与凸函数有很多相对应的性质可以由这个结论来进行链接。

凸集合与凸函数

凸组合

对于任何 n 个点 $\{x_i\}_{i=1}^n$, 以及权重系数 $\{w_i\}_{i=1}^n$. 若权重系数非负 $w_i \geq 0$ 而且 $\sum_{i=1}^n w_i = 1$, 则线性组合

$$S = \sum_{i=1}^n w_i x_i$$

为一个凸组合.

凸组合的物理意义可以理解成 n 个重量为 w_i 的点的整体重心.

凸集合与凸函数的对应性质 (凸组合)

凸集合性质

假设 Ω 是一个凸集合，那么 Ω 任意 n 个点的凸组合仍包含于 Ω.

凸函数性质：琴生 (Jensen) 不等式

如果 $f : \Omega \to \mathbb{R}$ 是一个凸函数，则对于任何 $\{x_i \in \Omega\}_{i=1}^n$，以及凸组合 $\sum_{i=1}^{n} w_i x_i$ 都有

$$\sum_{i=1}^{n} w_i f(x_i) \geq f(\sum_{i=1}^{n} w_i x_i)$$

凸集合与凸函数的对应性质 (集合相交)

凸集合性质

任意多个凸集合的交集仍是凸集合.

凸函数性质

- 任意多个凸函数的逐点上确界仍是凸函数.
- 固定一个凸函数的若干个变量，所得的函数仍然是凸函数
- 凸函数的 sublevel set 都是凸集合.

凸集分离定理

凸集分离定理

若 C, D 分别为 $\mathbb{R}^n$ 中的两个不交的非空凸集合,i.e.$C \cap D = \emptyset$, 则一定存在向量 $a \in \mathbb{R}^n$ 以及实数 $b \in \mathbb{R}$ 使得任何 $x_C \in C, x_D \in D$ 有 $a^T x_C \leq b$ 以及 $a^T x_D \geq b$.

定理中不等式的几何意义在于 C, D 分别位于超平面 $a^T x = b$ 的两边.

对偶问题：拉格朗日对偶函数

考虑 $\mathbb{R}^n$ 上的优化问题:

优化问题

$$
\begin{aligned}
&\text{最小化:} f_0(x)\\
&\text{不等条件:} f_i(x) \leq 0, i = 1, \cdots, m\\
&\text{等式条件:} h_i(x) = 0, i = 1, \cdots, p\\
&\text{定义域:} \mathcal{D} = \bigcap_{i=0}^{m} \mathsf{dom} f_i \cap \bigcap_{i=1}^{p} \mathsf{dom} h_i.
\end{aligned}
$$

请注意定义域 $\mathcal{D}$ 指的是使得所有函数 f_i, h_i 有定义的区域。而可行域指的是定义域中满足不等条件与等式条件的那些点. 本课中把这个优化问题称为原问题，优化点称为 x^*, 最优化值为 p^*.

对偶问题：拉格朗日对偶函数

根据原函数与限制条件我们定义拉格朗日量 $L(x,\lambda,\nu):\mathbb{R}^{n+m+p}\to\mathbb{R}$

拉格朗日量

$$L(x,\lambda,\nu)=f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x),$$

根据拉格朗日函数我们定义拉格朗日对偶函数 $g(\lambda,\nu):\mathbb{R}^{m+p}\to\mathbb{R}$

拉格朗日对偶函数

$$\begin{aligned} g(\lambda,\nu) &= \inf_{x\in\mathcal{D}} L(x,\lambda,\nu) \\ &= \inf_{x\in\mathcal{D}}\left(f_0(x)+\sum_{i=1}^{m}\lambda_i f_i(x)+\sum_{i=1}^{p}\nu_i h_i(x)\right) \end{aligned}$$

对偶问题：对偶问题

根据对偶函数，定义对偶问题的一般形式

对偶问题

$$\text{最大化:} g(\lambda, \nu)$$
$$\text{不等条件:} \lambda_i \geq 0, i = 1, \cdots, m$$

我们把对偶问题的最大值点称为 (λ^*, ν^*)，相应的最大值称为 d^*，这里面的对偶函数 g 定义域为 $\mathsf{dom} g = \{(\lambda, \nu) : g(\lambda, \nu) > -\infty\}$. 在 g 的定义域中满足 $\lambda_i \geq 0$ 的那些 (λ, ν) 全体，叫做对偶可行域. 也就是对偶问题的可行域.

对偶性

根据对偶函数的性质我们已经知道在对偶可行域中，$g(\lambda, \nu)$ 总是不大于 p^*. 所以就有

弱对偶性

$$d* \leq p*$$

若对偶性总是对的. 相对而言的强对偶性是指一部分优化问题来说，有更好的结论.

强对偶性

$$d* = p*$$

强对偶性不总成立.

强对偶性条件

第一个强对偶性的条件, 几乎所有的凸优化问题都满足强对偶性.

Slater 条件

对于一个凸优化问题

$$\begin{aligned} &\text{最小化:} f_0(x) \\ &\text{不等条件:} f_i(x) \leq 0, i = 1, \cdots, m \\ &\text{等式条件:} h_i(x) = 0, i = 1, \cdots, p \end{aligned}$$

如果存在一个可行域中的点 x 使得 $f_i(x) < 0, i = 1, \cdots, m$, 那么这个凸优化问题就满足强对偶条件.

凸优化问题求解 (KKT)

我们来看一下如果强对偶性满足的话，这些最优化点应该满足何种条件. 这一部分中我们假定所有的函数都是可微函数.
如果 $x^*, (\lambda^*, \nu^*)$ 分别是原问题与对偶问题的最优解，那么首先这些点应该满足可行域条件

- $f_i(x^*) \leq 0$
- $h_i(x^*) = 0$
- $\lambda_i^* \geq 0$

凸优化问题求解 (KKT)

其次我们已经知道

$$
\begin{aligned}
d^* &= g(\lambda^*, \nu^*) \\
&\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \\
&= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \\
&\leq f_0(x^*) \\
&= p^*
\end{aligned}
$$

于是 $d^* = p^*$ 意味着上述不等式全都是等式.

凸优化问题求解 (KKT)

KKT 条件

- $f_i(x^*) \leq 0,\ i = 1, \cdots, m$
- $h_i(x^*) = 0, i = 1, \cdots, p$
- $\lambda_i^* \geq 0,\ i = 1, \cdots, m$
- $\lambda_i^* f_i(x^*) = 0,\ i = 1, \cdots, m$
- $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$

KKT 条件使用

- 对于凸优化问题,KKT 条件是 $x^*, (\lambda^*, \nu^*)$ 分别作为原问题和对偶问题的最优解的充分必要条件.
- 对于非凸优化问题，KKT 条件仅仅是必要而非充分.

应用举例：支持向量机最简单形式

支持向量机的最简单形式

空间 $\mathbb{R}^n$ 中有可分的两个点集 C,D. 我们希望找到一个最合适的超平面 $a^T x = b$ 对他们进行区分. 也就是说对于 $p \in C, q \in D$, 有

$$a^T p > b, \quad a^T q < b$$

也就是说存在一个正数 t 使得

$$a^T p - b \geq t, \quad a^T q - b \leq -t$$

这个正数 t 一定程度上描述了两个集合被分开的程度.

为了体现“最合适”，一个比较好的想法就是希望当 a 是单位向量的时候，t 越大越好. 于是的到了一个优化问题

应用举例：支持向量机最简单形式

经过若干推导与转化就得到如下凸优化问题

SVM 优化问题

$$
\begin{aligned}
\text{最小化:} \quad & -t \\
\text{不等条件 1:} \quad & -a^T p_i + b \leq -t \\
\text{不等条件 2:} \quad & a^T q_i - b \leq -t \\
\text{不等条件 3:} \quad & -t \leq 0 \\
\text{不等条件 4:} \quad & |a|^2 \leq 1
\end{aligned}
$$

支持向量机 (SVM)

考虑 C 与 D 的凸包 $\overline{C}$ 和 $\overline{D}$

SVM 凸集合思路

若 $p \in \overline{C}, q \in \overline{D}$, 满足 $d(p,q) = d(\overline{C}, \overline{D})$ 则，向量 $\vec{pq}$ 即为所求支撑向量。

支持向量机 (SVM)

定义:

$$P = [p_1, \cdots, p_{n_c}]$$
$$Q = [q_1, \cdots, q_{n_D}]$$

于是:$X^T = [P^T, Q^T]$, 而对于 $\lambda^T P \in \overline{C}$ 与 $\mu^T Q \in \overline{D}$, 这两点之间的距离为

$$(\lambda^T P - \mu^T Q)^T (\lambda^T P - \mu^T Q) = [\lambda^T, \mu^T] K_0 [\lambda^T, \mu^T]^T$$

支持向量机 (SVM)

SVM 凸优化问题 (凸集合思路)

$$
\begin{aligned}
\text{最小化:} \quad & [\lambda^T, \mu^T] K_0 [\lambda^T, \mu^T]^T \\
\text{不等条件:} \quad & \lambda \geq 0 \\
\text{不等条件:} \quad & \mu \geq 0 \\
\text{等式条件:} \quad & \lambda^T E = 1 \\
\text{等式条件:} \quad & \mu^T E = 1
\end{aligned}
$$

我们看到 K_0 集中包含了 x_i 的全部信息. 不同的分类问题对应着不同的 K_0.

线性代数：参考资料

参考资料

- 非常清晰完整的经典教材：凸优化，Stephen Boyd, Lieven Vandenberghe

作业

- 教材：凸优化 (英文版页码与题号) p60:2.32.10,2.12,2.16,2.23,2.26; p113:3.1,3.2,3.12,3.21,3.36,3.39; p189:4.2,4.8,4.10,4.21 p273:5.1,5.4,5.11,5.225.24

谢谢大家!