机器学习算法班第 8 期第 2 课：概率论与凸优化

邓明格

七月在线

mingge_deng@brown.edu

May 6, 2017

主要内容

1. 概率论
   - 概率空间与随机变量
   - 贝叶斯公式
   - 随机变量的特征函数
   - 两大基本定理
   - 参数估计
2. 凸优化
   - 凸优化问题
   - 凸集合与凸函数
   - 保凸运算
   - 共轭函数
   - 带边界优化的对偶问题
   - KKT 条件

概率空间与随机变量

- 古典概率 or 统计定义：频率
- 现代概率 or 公理化定义：测度论 (科尔莫戈罗夫)

概率空间

概率空间三元组 $(\Omega, \mathcal{F}, P)$:

- Ω: 样本空间（最小不可分的独立互斥事件集合）
- $\mathcal{F}$: 事件（Ω 的子集构成）
- P: 测度（事件的概率）

随机变量

随机变量是一个从概率空间到实数域 $\mathcal{R}$ 的可测函数：$X(\omega) : \Omega \to \mathcal{R}$。例如：随机变量 $X > 0$ 对应的集合 $\{\omega \in \Omega : X(\omega) > 0\}$ 是一个可测事件。

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

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

后验分布，先验分布，似然函数

X: 观测数据，θ: 模型参数

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{\int_{\theta'} P(X|\theta')P(\theta')} \tag{2}$$

后验分布 $\propto$ 先验分布 $\times$ 似然函数

随机变量的矩

随机变量 X，对于任何正整数 n，其 k 阶矩定义为

$$E(X^k) = \int p(x)x^k dx \tag{3}$$

$k = 1$，$E(X)$ 为期望；$k = 2$，$E(X^2) - E(X)^2$ 为方差

特征函数

随机变量 X, 其特征函数定义为 $\phi_X(t) = E(e^{itX}) = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!}(it)^k$.

- 随机变量的傅立叶变换
- 包含了所有随机变量矩的信息
- 特征函数可以全面描述概率分布

特征函数的性质

- $\forall X$, $\phi_X(t)$ 存在，且是一致连续函数。
- $\phi(0) = 1$ 且 $|\phi(t)| \leq 1, \forall t$
- $\phi_{\bar{X}}(t) = \phi_{-X}(t)$
- X 中心对称，$\phi_X(t)$ 为实函数。
- 如果 X 的 k 阶矩存在，那么 $\phi_X(t)$ 至少 k 阶可微，且

$$E(X^k) = (-i)^n \phi^{(k)}(0) \tag{4}$$

- X, Y 为独立随机变量，那么 $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$。
- 如果 $\phi_X(t) = \phi_Y(t)$, 那么 X, Y 服从同一个分布。

重要分布的特征函数

- 独点分布 $P(a)=1$， $\phi(t)=e^{iat}$
- 两点分布 $P(-1)=P(1)=\frac{1}{2}$， $\phi(t)=\cos(t)$
- 正态分布 $f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$， $\phi(t)=\exp\left(i\mu t-\frac{\sigma^2t^2}{2}\right)$
- 泊松分布 $P(n)=e^{-\lambda}\frac{\lambda^n}{n!}$， $\phi(t)=e^{-\lambda(1-e^{it})}$

大数定理

X 是随机变量，μ 是 X 的期望，σ 是 X 的方差. $\{X_k\}_{k=1}^{\infty}$ 为 i.i.d. 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

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0 \tag{5}$$

大数定理弱证明（依分布收敛到独点分布 $P(\mu)=1$）

- X 有一阶矩，特征函数 $\phi_X(t)$ 存在一阶泰勒展开:

$$\phi_X(t) = 1 + i\mu t + o(t) \tag{6}$$

- 考虑 $\bar{X}_n$ 的特征函数:

$$\phi_{\bar{X}_n}(t) = E(\exp(it\bar{X}_n)) = \prod_{i=1}^{n} E(\exp(itX/n)) = (1 + i\mu t/n + o(t/n))^n \tag{7}$$

$$\lim_{n\to\infty} \phi_{\bar{X}_n}(t) = \lim_{n\to\infty} (1 + i\mu t/n + o(t/n))^n = e^{i\mu t} \tag{8}$$

中心极限定理

X 是随机变量,μ 是 X 的期望，σ 是 X 的方差. $\{X_k\}_{k=1}^{\infty}$ 为 i.i.d. 随机变量,$\bar{X}_n = \frac{\sum_{k=1}^{n} X_k}{n}$, 那么

$$Z_n = \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \tag{9}$$

依分布收敛于标准正态分布 $N(0,1)$.

中心极限定理证明（作业）

- X 有一, 二阶矩，特征函数 $\phi_X(t)$ 存在二阶泰勒展开。
- 证明 Z_n 的特征函数收敛到标准正态分布的特征函数 $e^{-\frac{1}{2}t^2}$。

参数估计问题

参数估计问题

已知一个随机变量 X 的分布函数 $f_\theta(X)$, 其中 $\theta = (\theta_1, ..., \theta_k)$ 为未知参数，利用已有样本 $X_1, ..., X_n$ 对参数 θ 或者 θ 的函数 $g(\theta)$ 作出估计。

1. 点估计：用样本的一个函数 $T(X_1, ..., X_n)$ 估计 $g(\theta)$

2. 区间估计：用一个置信区间去估计 $g(\theta)$

矩估计

矩估计

基本原理：大数定理，对于任何 i.i.d 变量 X, 当 $n \to \infty$, $\frac{1}{n}\sum_i^n X_i$ 收敛于 $E(X)$, 同理其 k 阶矩也满足大数定理，$\frac{1}{n}\sum_i^n X_i^k$ 收敛于 $E(X^k)$, 可以构造 k 组方程求救。

两点分布的矩估计

X 服从两点分布取值为 $\{-1,1\}$，$P(-1) = 1-\theta$, $P(1) = \theta$, 用样本 $X_1, ...X_n$ 估计参数 θ.

能用低阶矩不用高阶矩！！

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.

似然函数

假设 $X = (X_1, ..., X_n$ 样本的观测值. 那么整个样本的似然函数就是

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i) \tag{10}$$

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

作业：正态分布的参数极大似然估计 $\hat{\mu} = \bar{X}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i = \bar{X})^2$。

点估计的评判

- 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.

作业：为什么统计上估计方差为 $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$。

凸优化问题

优化问题的一般形式

优化问题的一般形式
最小化： $f_0(x)$
不等条件： $f_i(x) \leq 0,\ i = 1...m$
等式条件： $h_i(x) = 0,\ i = 1...p$

- $f_0(x)$ 为目标函数， $f_i(x)$ 和 $h_i(x)$ 是限制条件。
- 优化问题的定义域为目标函数定义域与限制条件定义域的交集。
- 优化问题的可行域为满足所有限制条件的定义域。
- 优化点称为 x^*, 最优化值为 p^*。

凸优化问题

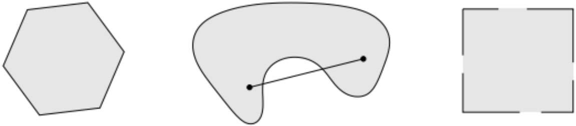
- 优化问题中的目标函数及限制条件均为凸函数。
- 局部最优等价于全局最优。
- 凸优化问题求解工具 (CVX 等)。

凸集合

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.

$$\lambda x_1 + (1-\lambda)x_2 \in \Omega, \forall x_1, x_2 \in \Omega, \lambda \in (0,1) \tag{11}$$



凸函数

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

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2), \forall x_1, x_2 \in \Omega, \lambda \in (0,1) \quad (12)$$

常见凸函数

- 仿射函数 $ax + b$ on $\mathcal{R}^n$
- 指数函数 $e^{\alpha x}, \forall \alpha \in \mathcal{R}$
- 幂函数 x^α on $\mathcal{R}^+, \forall \alpha \geq 1, \quad or \quad \alpha < 0$
- 负熵 $x \log x$ on $\mathcal{R}^+$
- 负对数函数 $-\log x$ on $\mathcal{R}^+$

Hessian 矩阵半正定!!!

凸集合与图函数的关系

函数的上境图

假设 f 是一个定义在 Ω 上的函数，区域 $f(x, y) : y \geq f(x), \forall x \in \Omega$ 就是 f 的上境图，上境图就是函数图像上方的部分区。

凸集合与图函数的关系

- 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。

凸组合与凸闭包

凸组合

对于任何 n 个点 $\{x_i\}_{i=1}^n$ 以及权重系数 $\{\omega_i\}_{i=1}^n$. 若权重系数非负 $\omega_i \geq 0$ 且 $\sum_{i=1}^n \omega_i = 1$, 则线性组合 $S = \sum_{i=1}^n \omega_i x_i$ 为一个凸组合. 凸组合的物理意义可以理解成 n 个重量为 ω_i 的点的整体重心。

集合的凸包

n 个点 $\{x_i\}_{i=1}^n$ 的全部凸组合就构成 $\{x_i\}_{i=1}^n$ 的凸包.

函数的凸闭包

如果 C 是函数 f 的上境图，$\bar{C}$ 是 C 的凸包, 那么以 $\bar{C}$ 为上境图的函数称为 f 的凸闭包。

- 若 g 是 f 的凸闭包，那么 $g \leq f$
- 若 g 是 f 的凸闭包, 那么 $\inf g = \inf f$

Jensen 不等式

凸集合性质

假设 Ω 是一个凸集合，那么 Ω 任何子集的凸包仍包含于 Ω.

Jensen 不等式

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

$$\sum_{i=1}^{n} \omega_i f(x_i) \geq f\left(\sum_{i=1}^{n} \omega_i x_i\right) \tag{13}$$

凸集合保凸运算

- 任意多个凸集合的交集仍是凸集合。
- 凸集合的线性映射仍是凸集合

凸函数保凸运算

- 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。
- 凸函数的非负线性组合仍是凸函数,$f_1 \ldots f_n$ 是凸函数，而且 $\omega_i \geq 0$，则 $\sum_{i=1}^{n} \omega_i f_i$ 也是凸函数。
- 若 $f : \mathcal{R}^n \to \mathcal{R}$ 是凸函数，$A \in \mathcal{R}^{n \times m}, b \in \mathcal{R}^n$，那么复合函数 $g(x) = f(Ax + b)$ 还是凸函数。
- perspective: $g(x, t) = tf(x/t)$, g 是凸函数如果 f 为凸函数。

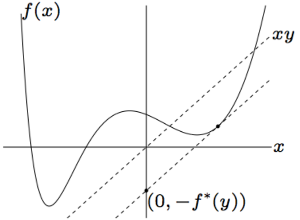
共轭函数

共轭函数

若 $f: \mathcal{R}^n \rightarrow \mathcal{R}$ 是实值函数，那么 f 的共轭函数为

$$f^*(y) = \sup_{x \in dom f} \left(y^T x - f(x) \right) \tag{14}$$

其中 $f^*(y)$ 的定义域是使得等式右边有界的那些 y。



共轭函数性质

- 共轭函数 f^* 是一个凸函数。
- 如果 g 是 f 的凸闭包，那么 $g^* = f^*$
- 对一般的函数 f, $f^{**} \le f$
- 如果 f 是一个凸函数，那么 $f^{**} = f$
- 如果 $g(x) = f(Ax + b)$, 则 $g^*(y) = f^*(A^{-1}y) - b^T A^{-1} y$
- 如果 $f(u, v) = f_1(u) + f_2(v)$, 那么 $f^*(w, z) = f_1^*(w) + f_2^*(z)$

共轭函数实例

- $f(x) = ax - b$,

$$f^*(y) = \sup_{x \in \mathcal{R}} \{yx - ax + b\} = \sup_{x \in \mathcal{R}} \{(y-a)x + b\} = \begin{cases} b, & y = a \\ \infty, & y \neq a \end{cases}$$
$$f^{**}(x) = \sup_{y \in \mathcal{R}} \{xy - f^*(y)\} = xa - f^*(a) = ax - b = f(x) \tag{15}$$

- $f(x) = x\log x, \forall x \in \mathcal{R}^+, f(0) = 0$, 则 $f^*(y) = \sup_{x \in \mathcal{R}^+}\{yx - x\log x\}$, 在 $x = e^{y-1}$ 取极大，那么 $f^*(y) = ye^{y-1} - e^{y-1}(y-1) = e^{y-1}$

- 作业：$f(x) = x^2$, 则 $f^*(y) = \frac{y^2}{2}$
- 作业：$f(x) = |x|$, 则 $f^*(y) = 0 \quad if \quad |y| \leq 1, \quad \infty \quad if \quad |y| > 1$

拉格朗日对偶函数

优化问题的拉格朗日量

考虑优化问题: 最小化 $f_0(x)$
不等条件： $f_i(x) \leq 0,\ i = 1...m$
等式条件： $h_i(x) = 0,\ i = 1...p$
其拉格朗日量定义为

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \tag{16}$$

拉格朗日对偶函数

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

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) = \inf_{x \in \mathcal{D}} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \tag{17}$$

为什么研究拉格朗日对偶函数

- 对偶函数为原问题提供下界!
- 无论原函数如何，拉格朗日对偶函数总为凹函数!

对偶函数性质

如果限制 $\lambda_i \geq 0, \forall i = 1...m$, 则 $g(\lambda, \nu) \leq p^*$.
证明:
$\forall x \in \mathcal{D}$, 如果 x 在可行域中，那么

$$\begin{aligned} g(\lambda, \nu) &\leq f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \\ &=\leq f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \\ &\leq f_0(x) \leq p^* \end{aligned} \tag{18}$$

对偶问题

对偶问题的一般形式

最大化：$g(\lambda, \nu)$
不等条件：$\lambda_i \geq 0, \forall i = 1...m$

- 对偶问题的最大值点称为 (λ^*, ν^*)，相应的最大值称为 d^*，$d^* \leq p^*$。
- 对偶问题的定义域为 $domg = \{(\lambda, \nu) : g(\lambda, \nu) > -\infty\}$。
- 对偶问题的可行域为满足 $\lambda_i \geq 0$ 的 (λ, ν) 全体。
- 弱对偶性：$d^* \leq p^*$
- 强对偶性：$d^* = p^*$

强对偶性条件

- 几乎所有的凸优化问题都满足强对偶性。
- 可能存在两个问题都无可行解的状况，这时强对偶性不成立。所以有必要建立“使其有解”的约束条件，其中之一便是 Slater 条件

Slater 条件

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

原问题与对偶问题间的关系

- 原问题和对偶问题都是可行的，弱对偶性成立，强对偶性不一定成立。
- 原问题和对偶问题都不可行，此时弱对偶性依旧成立，但强对偶性不成立。
- 当原问题是下无界时，即 $p^* = -\infty$，由弱对偶性，$d^* = -\infty$，即对偶问题不可行。
- 反之，当对偶问题上无界时，$d^* = +\infty$，必有 $p^* = +\infty$，即原问题不可行。

对偶问题 \ 原问题	可行	下无界 ($p^* = -\infty$)	不可行($p^* = +\infty$)
可行	√	×	×
上无界 ($d^* = +\infty$)	×	×	√
不可行 ($d^* = -\infty$)	×	√	√

线性规划对偶问题

线性规划原问题

最小化：$c^T x$
等式条件：$Ax = b$
不等条件: $x_i \leq 0, \forall i = 1...n$

线性规划对偶问题

最小化：$b^T \nu$
等式条件：$A^T \nu - \lambda + c = 0$
不等条件：$\nu_i \geq 0, \forall i = 1...n$

线性约束优化对偶问题

线性约束优化原问题

最小化： $f_0(x)$
等式条件： $Cx = d$
不等条件: $Ax \leq b$

线性约束优化对偶问题

$$
\begin{aligned}
g(\lambda, \nu) &= \inf_x \left(f_0(x) + \lambda^T (Ax - b) + \nu^T (Cx - d) \right) \\
&= -b^T \lambda - d^T \nu + \inf_x \left(f_0(x) + (A^T \lambda + C^T \nu)^T x \right) \\
&= -b^T \lambda - d^T \nu - f_0^*(-A^T \lambda - C^T \nu)
\end{aligned}
\tag{19}
$$

最小化向量 L1 范数

最小化向量 L1 范数

最小化：$|x|$
等式条件：$Ax = b$

对偶问题 (作业)

最小化：$b^T\nu$
不等条件：$|A^T\nu| \leq 1$

强对偶性条件:KKT 条件

强对偶性条件:KKT 条件

- 原问题可行域条件 $f_i(x^*) \leq 0$
- 原问题可行域条件 $h_i(x^*) = 0$
- 对偶问题可行域条件 $\lambda_i^* \geq 0$
- $\sum_i^m \lambda_i^* f_i(x^*) = 0 \Rightarrow \lambda_i^* f_i(x^*)\ 0$
- $g(\lambda^*, \nu^*) = \inf(L(x^* \lambda^*, \nu^*)) = L(x^* \lambda^*, \nu^*) \Rightarrow \nabla_x L(x^* \lambda^*, \nu^*) = 0$

KKT 条件使用

- 对于凸优化问题,KKT 条件是 x^*, (λ^*, ν^*) 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。

KKT 求解凸优化问题

优化问题

最小化： $\frac{1}{2}x^T Px + q^T x + r$
等式条件： $Ax = b$

KKT 条件

$Ax^* = b$
$Px^* + q + A^T \nu^* = 0$
求解这个线性方程即可得到结果.

谢谢大家！！