机器学习里的数学第 7 课凸优化进阶

管老师

七月在线

June, 2016

主要内容

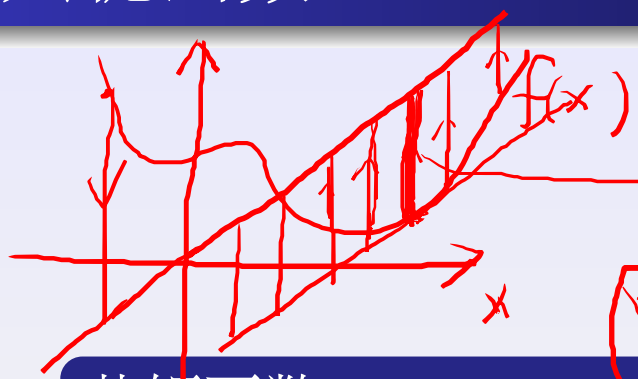
- 共轭函数
  - 共轭函数
- 对偶问题
  - 拉格朗日对偶函数
  - 拉格朗日对偶问题
  - 共轭函数与拉格朗日对偶函数
- 对偶性
  - 弱对偶性与强对偶性
  - 强对偶性成立的几种情况
  - 凸优化问题求解 (KKT)
- 应用举例
  - 支持向量机 (SVM) 的最简单形式
- 总结寄语 (数学部分)

记号

- 本节课常用数学记号

V, W	向量空间
v, w	向量
$\mathbb{R}^n$,$\mathbb{R}^m$	实坐标空间
α,β	V 和 W 的基
$T : V \to W$	向量空间 V 到 W 的线性映射
$A_{\alpha,\beta}(T)$	线性映射 T 在 α 和 β 这两组基下的矩阵
$G(v_1, v_2)$	内积空间 V 上的内积
H_α	G 在基 α 下的矩阵形式

共轭函数



$f^*(y)$. $y=1$

$f^*(y) = y \cdot x - f(x) = \nabla f(x)^T x - f(x)$

$\nabla f(x) = y$

$= \nabla f(x^*)^T x^* - f(x^*)$

共轭函数

如果 $f : \mathbb{R}^n \to \mathbb{R}$ 是一个函数, 那么 f 的共轭函数

$$f^*(y) = \sup_{x \in \mathbf{dom} f} (y^T x - f(x))$$

其中 $f^*(y)$ 的定义域是使得等式右边有上界的那些 y.

$f^{**}(z) = \sup_{y \in \mathrm{dom} f^*} (z^T y - f^*(y))$

$\leq f(z)$

$z^T y - f^*(y) \leq f(z)$

$f^*(y) \geq \sup (y^T x - f(x))$

$f(x) + f^*(y) \geq y^T x$

共轭函数

共轭函数的基本性质

- 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,$f^{**} \leq f$
- 如果 f 是一个凸函数，那么 $f^{**} = f$

共轭函数

$f^*(y) = \sup_x \{y^T x - f(x)\}$ $\forall x, y$ $f(y) \geq y^T x - f(x)$

共轭函数的进一步性质

- $f(x) + f^*(y) \geq x^T y$
- 如果 f 是凸函数而且可微，那么
  $f^*(y) = x^{*T} \nabla f(x^*) - f(x^*)$，其中 x^* 满足 $\nabla f(x^*) = y$.
- 如果 $g(x) = f(Ax + b)$, 则 $g^*(y) = f^*(A^{-T} y) - b^T A^{-T} y$.
- 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

$f(x) = x\ln x$

$f^*(y) = \sup_x(yx - f(x)) = \sup_x(yx - x\ln x)$

$\frac{d}{dx}(yx - x\ln x) = y - \ln x - \frac{x}{x} = y - 1 - \ln x$

$\ln x^x = y - 1, \quad x^x = e^{y-1}$

$f^*(y) = ye^{y-1} - e^{y-1}\ln(e^{y-1})$

$= ye^{y-1} - e^{y-1}(y-1)$

$= e^{y-1}$

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

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

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

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

拉格朗日量在数学与物理中有极为广泛的应用，感兴趣的同学可以在“信息几何学”，“理论力学”中找到它其他的应用.

对偶问题：拉格朗日对偶函数

对偶函数有如下重要性质

对偶函数为原问题提供下界

如果限制 $\lambda_i \geq 0, \forall i = 1, \cdots, m$, 则

$$g(\lambda, \nu) \leq p^*$$

证明

对任意一个 $x \in \mathcal{D}$, 如果 x 在可行域中，那么

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

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.

对偶问题：拉格朗日对偶函数

对偶问题举例

原问题，线性规划

$$\begin{aligned} &\text{最小化:} c^T x \\ &\quad \text{条件:} Ax = b, \text{而且} x_i \geq 0, \forall i = 1, \cdots, n \end{aligned}$$

对偶问题：拉格朗日对偶函数

$f_0 = c^T x$,

$f_i(x) = -x_i \le 0$

$h_i(x) = a_i^T x - b_i = 0$

$A = [a_1 \cdots a_p]$

$A^T \cdot x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_p^T x \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}$

$L(x, \lambda, \nu)$

$= c^T x + \lambda^T(-x) + \nu^T(A^T x - b)$

$g(\lambda, \nu) = \inf\left(c^T x - \lambda^T x + \nu^T A^T x - \nu^T b\right)$

$= \inf\left(\left[c^T - \lambda^T + \nu^T A^T\right] x - \nu^T b\right)$

$= \begin{cases} -\infty \\ -\nu^T b, \quad c^T - \lambda^T + \nu^T A^T = 0 \end{cases}$

对偶问题：拉格朗日对偶函数

对偶问题举例

线性规划的对偶问题

最小化:$b^T\nu$

条件:$A^T\nu - \lambda + c = 0$, 而且 $\lambda_i \geq 0, \forall i = 1, \cdots, n$.

对偶问题：共轭函数与拉格朗日对偶函数

当优化问题的限制条件是线性条件时，可以利用共轭函数的一些性质方便的得到对偶问题

线性约束优化问题的对偶问题

$$\text{最小化:} f_0(x)$$
$$\text{不等条件:} AX \leq b$$
$$\text{等式条件:} Cx = d$$

这里面向量的比较 $u < v$ 指的是 u 里边每一个分量都小于 v 里面对应的分量.

对偶问题：共轭函数与拉格朗日对偶函数

$$\inf(y^T x + f_0(x)) = -\sup(-y^T x - f_0(x)) = -f^*(-y) \qquad y = (A^T\lambda + C^T\nu)$$

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

而这个函数的定义域就是

$$\mathrm{dom} g = \{(\lambda,\nu) : -A^T\lambda - C^T\nu \in \mathrm{dom} f_0^*\}.$$

对偶问题：共轭函数与拉格朗日对偶函数

最小化向量范数

$|x| = |x|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$

$|x|_1 = |x_1| + \cdots + |x_n|$

对偶问题举例

最小化:$|x|$

条件:$Ax = b$

$-b^T \cdot \nu$

$|A^T \nu| \leq 1$

若 $f(x) = |x|$ 那么共轭函数是 $f^*(y) = 0, \forall |y| \leq 1$, 否则 $f^*(y) = +\infty$. 所以对偶问题就是:

② $|y| > 1$, $x_t = t \cdot y$, $t > 0$

$y^T x - |x| = t|y|^2 - t|y|$

$= t|y|(|y| - 1)$

最大化:$-b^T \nu$

条件:$|A^T \nu| \leq 1$

$f^*(y) = \sup(y^T x - |x|)$

① $|y| \leq 1$, $|y^T x| \leq |y| \cdot |x| \leq |x| \Rightarrow y^T x - |x| \leq 0$, 当 $x = \vec{0}$ 时 $y^T x - |x| = 0$

So: $f^*(y) = 0$

对偶问题：共轭函数与拉格朗日对偶函数

最大熵问题

对偶问题举例

$$\text{最小化:} f_0(x) = \sum_{i=1}^{n} x_i \ln x_i$$

$$\text{不等条件:} Ax \leq b$$

$$\text{等式条件:} \mathbf{1}^T x = 1$$

因为 $(f_1(x) + f_2(y))^* = f_1^*(w) + f_2^*(z)$, 我们有

$$f_0^*(y) = \sum_{i=1}^{n} e^{y_i - 1}$$

所以对偶函数为 $g(\lambda, \nu) = -b^T\lambda - \nu - e^{-\nu-1} \sum_{i=1}^{n} e^{-a_i^T \lambda}$

对偶问题：共轭函数与拉格朗日对偶函数

最大熵问题的对偶问题就是

最大熵问题的对偶问题

$$\text{最大化: } -b^T\lambda - \nu - e^{-\nu-1}\sum_{i=1}^{n} e^{-a_i^T\lambda}$$

不等条件:$\lambda \geq 0$

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.

强对偶性条件

根据 Slater 条件，我们前述的几个例子都是满足强对偶性的.

满足强对偶性的例子

- 线性规划
- 最小二乘
- 最大熵问题

这种情况下我们如果发现对偶问题比原问题更容易解决，那么就可以使用对偶问题来解出 $d^* = p^*$

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)

所以我们有

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

, 以及

$$g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*)$$

而因为

$$g(\lambda^*, \nu^*) = \inf(L(x^*, \lambda^*, \nu^*))$$

所以 x^* 是拉格朗日函数在 x 方向的驻点，所以有

$$\nabla_x L(x^*, \lambda^*, \nu^*) = 0$$

. 综上所述我们就的到了 KKT 条件.

凸优化问题求解 (KKT)

KKT 条件

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

其中第四个条件是由 $\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$ 以及第一个和第三个条件共同得到的.

凸优化问题求解 (KKT)

所以我们有

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

, 以及

$$g(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*)$$

而因为

$$g(\lambda^*, \nu^*) = \inf(L(x^*, \lambda^*, \nu^*))$$

所以 x^* 是拉格朗日函数在 x 方向的驻点，所以有

$$\nabla_x L(x^*, \lambda^*, \nu^*) = 0$$

. 综上所述我们就的到了 KKT 条件.

凸优化问题求解 (KKT)

KKT 条件叙述

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

其中第四个条件是由 $\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$ 以及第一个和第三个条件共同得到的.

凸优化问题求解 (KKT)

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.

凸优化问题求解 (KKT)

使用 KKT 条件解决优化问题例子

$$\text{最小化:}(1/2)x^T P x + q^T x + r$$

$$\text{等式条件:}Ax = b$$

KKT 条件是

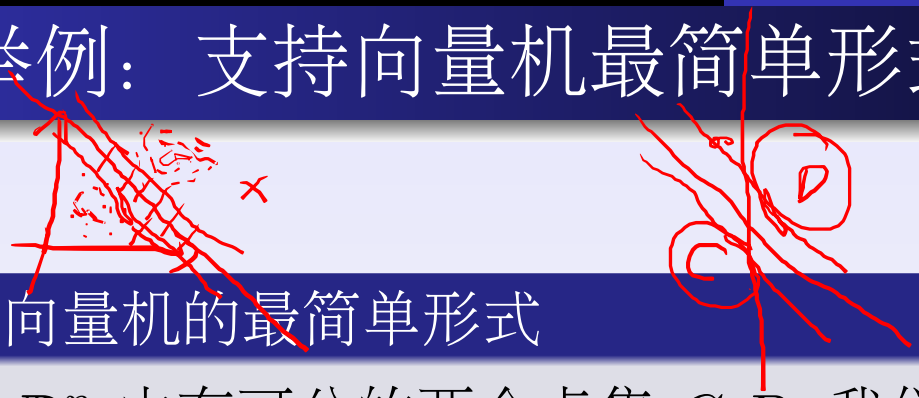
$$Ax^* = b$$

以及

$$Px^* + q + A^T v^* = 0$$

我们求解这个线性方程即可得到索要的结果.

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

SVM 优化问题

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

但是这里面 $|a|^2 = 1$ 并非凸条件，所以我们把它换成 $|a|^2 \le 1$. 就得到一个跟原问题等价的凸优化问题.

应用举例：支持向量机最简单形式

SVM 凸优化问题

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

具体求解，因为这是一个凸优化问题，所以可以选择使用 KKT 条件来解方程求解.

总结寄语 (数学部分)

总结

- 本次课程数学部分涵盖微积分，线性代数，概率统计，凸优化的内容
- 复习目标：
  - 分别对每一部分进行总结，得到清晰知识结构脉络
  - 最好对每一部分都能记住一些有代表性的例子，比如矩阵里边的旋转，拉伸变换
- 使用目标：在工作和学习中遇到相关知识的时候
  - 可以回忆起最基本的想法（比如微分的本质是逼近，矩阵变换本质是换基）
  - 对于技术细节能很快找到参考资料

总结寄语 (数学部分)

小故事

R.Thom 是法国人，35 岁得的 Fields 奖。在一次采访当中，作为数学家的 Thom 同两位古人类学家讨论问题。谈到远古的人们为什么要保存火种时

- 一个人类学家说，因为保存火种可以取暖御寒;
- 另外一个人类学家说，因为保存火种可以烧出鲜美的肉食。
- 而 Thom 说，因为夜幕来临之际，火光摇曳妩媚，灿烂多姿，是最美最美的。

除了上面所有目标，我还希望本次课程中讲述的数学各个分支之间产生的一些相互作用能或多或少给大家一点数学独特美学的享受. 比如微积分与线性代数之间，矩阵与几何之间，凸集合与凸函数之间那些意料之外又在情理之中的关系. 毕竟数学的终极目的还是美.

$$f(u,v)=f_1(u)+f_2(v)$$

$$f^*(w,z)=\sup_{u,v}\{w^Tu+z^Tv-f(u)-f(v)\}$$

$$=\sup_u\{w^Tu+f_1(u)\}+\sup_v\{z^Tv-f_2(v)\}$$

谢谢大家!

$$=f_1^*(w)+f_2^*(z)$$