机器学习白板推导系列笔记 (6.5~6.8)

Dexing Huang
dxhuang@bupt.edu.cn
Beijing University of Posts and Telecommunications

日期：2021 年 11 月 7 日

1 约束优化问题

1.1 弱对偶性证明

首先给出原问题

$$\begin{cases} \min\limits_{x\in\mathbb{R}^p} f(x) \\ s.t. \quad m_i(x) \leq 0 \quad i=1,2,\cdots,M \\ \qquad\quad n_j(x)=0 \quad j=1,2,\cdots,N \end{cases} \tag{1}$$

构造 Lagrange 方程

$$\mathcal{L}(x,\lambda,\eta)=f(x)+\sum_{i=1}^{M}\lambda_i m_i(x)+\sum_{j=1}^{N}\eta_j n_j(x) \tag{2}$$

注. $\lambda_i \geq 0, \eta_j$ 无限制

将原问题转化为无约束问题

$$\begin{cases} \min\limits_{x\in\mathbb{R}^p}\max\limits_{\lambda,\eta}\mathcal{L}(x,\lambda,\eta) \\ s.t. \quad \lambda_i \geq 0 \end{cases} \tag{3}$$

简单说明一下为什么无约束问题和原问题等价，在硬间隔 SVM 中其实已经说过．其实对于问题 (3)，已经蕴含了 $m_i(x) \leq 0$ 的约束条件．如果 x 违反了该约束条件，$m_i(x) > 0$，那么 $\max_{\lambda,\eta}\mathcal{L} \to \infty$，如果满足约束那么 $\max_{\lambda,\eta}\mathcal{L} < \infty$，因此

$$\min_x\max_{\lambda,\eta}\mathcal{L}=\min_x\{\max_{\lambda,\eta}\mathcal{L},+\infty\}=\min_x\max_{\lambda,\eta}\mathcal{L}$$

所以也称问题 (3) 为原问题的等价表示.

下面讨论原问题的对偶问题

$$\begin{cases} \max\limits_{\lambda,\eta}\min\limits_{x\in\mathbb{R}^p}\mathcal{L}(x,\lambda,\eta) \\ s.t. \quad \lambda_i \geq 0 \end{cases} \tag{4}$$

由弱对偶性可知

$$\min\max\mathcal{L} \geq \max\min\mathcal{L} \tag{5}$$

上式的证明在 SVM 一节中举了个鸡头凤尾的例子来定性描述，下面进行数学的证明，对于

$\min_x \mathcal{L}(x,\lambda,\eta)$ 有

$$\underbrace{\min_x \mathcal{L}(x,\lambda,\eta)}_{A(\lambda,\eta)} \leq \mathcal{L}(x,\lambda,\eta) \leq \underbrace{\max_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta)}_{B(x)} \tag{6}$$

因此

$$\max_{\lambda,\eta} A(\lambda,\eta) \leq \min_x B(x) \tag{7}$$

那么

$$\max_{\lambda,\eta} \min_x \mathcal{L}(x,\lambda,\eta) \leq \min_x \max_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta) \tag{8}$$

1.2 对偶关系的几何解释

本节从几何的角度来解释对偶关系. 为了演示方便, 将问题简化为二维，原问题

$$\begin{cases} \min\limits_{x\in\mathbb{R}^p} f(x) \\ s.t. \quad m_1(x) \leq 0 \end{cases} \tag{9}$$

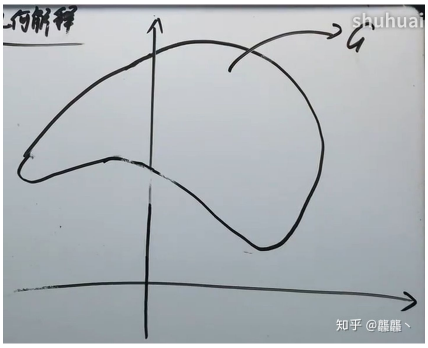
令 D 为定义域, $D = dom\, f \cap dom\, m$, 使用 Lagrange 乘子法

$$\mathcal{L}(x,\lambda) = f(x) + \lambda(x) \tag{10}$$

令 $p^\star = \min f(x)$, $d^\star = \max_\lambda \min_x \mathcal{L}(x,\lambda)$, 令 $G = \{(m_1(x), f(x))|x \in D\} = \{(u,t)|x \in D\}$, 将 $p^\star, d^\star$ 用集合表示

$$\begin{aligned} p^\star &= \min f(x) = \min t = \inf\{t|(u,t) \in G, u \leq 0\} \\ d^\star &= \max_\lambda \min_x \mathcal{L}(x,\lambda) = \max_\lambda \min_x (f(x) + \lambda(x)) = \max_\lambda \inf\{t + \lambda u|(u,t) \in G\} \end{aligned}$$

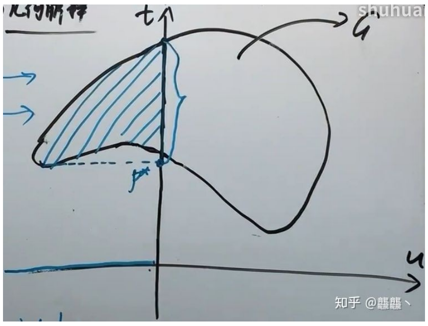
令 $g(\lambda) = \inf\{t + \lambda u|(u,t) \in G\}$, 因此 $d^\star = \max_\lambda g(\lambda)$, 接下来将 $p^\star$ 和 $q_\star$ 表示在图形上, 并假设集合 G 为非凸集(因为非凸集是一般情况). 凸集 (Convex set) 是一个点集合，其中每两点之间的直线点都落在该点集合中[1], 非凸集则不满足该条件.



先表示 $p^\star$, $u < 0$ 因此在左半轴

显然 $p^\star$ 就是左半轴 t 的最小值.

[1]维基百科

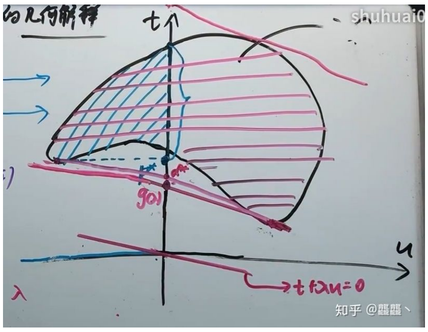


下面表示 $d^\star$

$$g(\lambda) = \inf\{t+\lambda u|(u,t)\in G\}$$

$$d^\star = \max_\lambda g(\lambda)$$

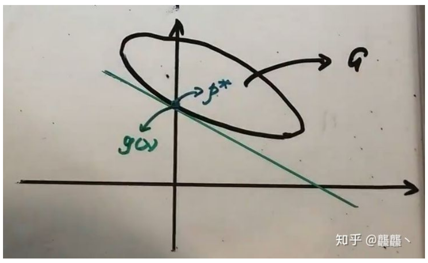
先看 $g(\lambda)$, $t+\lambda u$ 是斜率为 $-\lambda$ 的直线, 截距可以取不同值, 记 $\Delta = t+\lambda u$, 将此直线向上平移, 当与 G 相切时设 $\Delta_1 = t+\lambda u$, 继续上移产生不同的 Δ_i 直到与上边界相切, 不同的截距组成的集合 $\{\Delta_1, \Delta_2, \cdots\} = \{t+\lambda u|(u,t)\in G\}$, 那么 $g(\lambda) = \inf\{t+\lambda u|(u,t)\in G\} = \Delta_1$, 调整斜率 λ, 那么必然可以找到一个 $\lambda^\star$ 使得直线 $t+\lambda u$ 与 G 上两个凸点相切, 如下所示



这样就找到了 $d^\star$, 从图中很容易可以看出 $d^\star \le p^\star$, 又因为 $p^\star = \max f(x)$ 与 $\min_x \max_\lambda \mathcal{L}(x,\lambda)$ 等价, 因此

$$\min_x \max_\lambda \mathcal{L}(x,\lambda) \ge \max_\lambda \min_x \mathcal{L}(x,\lambda) \tag{11}$$

弱对偶关系成立, 当 G 为凸集时, 强对偶关系成立, 从图中也可以很容易得出



另外提一下满足强对偶关系的条件是

$$\text{凸优化问题} + \text{slater 条件} \Rightarrow \text{强对偶关系(充分非必要)}$$

SVM 是一个二次规划问题, 其天生满足强对偶关系.

1.3 slater 条件

slater 条件的描述如下

$$\exists \hat{x} \in relint\, D \quad s.t. \forall i = 1, 2, \cdots, M \quad m_i(\hat{x}) < 0$$

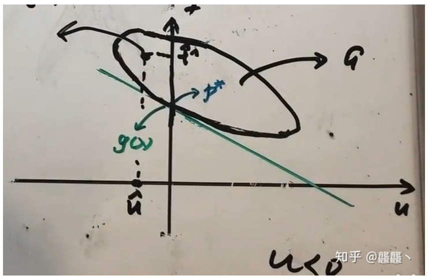
注. relint 指的是定义域内, 不含边界

- 对于大多数凸优化问题, slater 条件成立
- 需要找到一个点, 然后验证每一个 $m_i(x)$ 都小于 0, 这很难, 因此有一个放松的 slater 条件, 如果有 k 个 $m_i(x)$ 是仿射函数, 那么只需要验证其他 $M-k$ 个即可.

又引出了一个仿射函数的概念, 仿射变换 (Affine transformation), 又称仿射映射, 是指在几何中, 对一个向量空间进行一次线性变换并接上一个平移, 变换为另一个向量空间[2], 常见的线性函数都是仿射函数.

对于 SVM 问题, 是凸二次优化问题, 且约束条件是仿射函数, 因此符合 slater 条件.

从几何的角度上来看, slater 条件要求需要至少存在一个点其横坐标在 u(即存在x使得$m_1(x) < 0$) 的负半轴上, 使得切线不会垂直于 u 轴, 即斜率不能为无穷大.



1.4 KKT 条件

原问题表示如下

$$\begin{cases} \min\limits_{x \in \mathbb{R}^p} f(x) \\ s.t. \quad m_i(x) \leq 0 \quad i = 1, 2, \cdots, M \\ \qquad\quad n_j(x) = 0 \quad j = 1, 2, \cdots, N \end{cases}$$

其中约束条件 m_i, n_j 均可微, 令原问题为 $p^\star$, 最优解为 $x^\star$, 使用 Lagrange 乘子法

$$\mathcal{L}(x, \lambda, \eta) = f(x) + \sum_{i=1}^{M} \lambda_i m_i(x) + \sum_{j=1}^{N} \eta_j n_j(x)$$

令 $g(\lambda, \eta) = \min_x \mathcal{L}(x, \lambda, \eta)$, 对偶问题为

$$\begin{cases} \max\limits_{\lambda, \eta} g(\lambda, \eta) \\ s.t. \quad \lambda_{\geq}0 \end{cases}$$

令对偶问题为 $d^\star$, 其最优解为 $\lambda^\star, \eta^\star$. 当目标优化问题为凸优化 (convex) 并满足 slater 条件, 则可以推出其满足强对偶关系, 强对偶与 KKT 条件互为充要条件, 并通过 KKT 条件可以求解强

[2]维基百科

对偶问题

$$convet + slater \Rightarrow strong\ duality \Leftrightarrow KKT$$

KKT 条件分为三类: 可行条件, 互补松弛条件, 梯度为 0

$$\begin{cases} \text{可行条件} \quad m_i(x) \leq 0, n_j(x) = 0, \lambda^\star \geq 0 \\ \text{互补松弛条件} \quad \lambda_i m_i = 0 \\ \text{梯度为0} \quad \frac{\partial \mathcal{L}}{\partial x}|_{x=x^\star} = 0 \end{cases} \tag{12}$$

- 可行条件

  满足约束条件，并且需满足隐含条件 $\lambda \geq 0$

- 互补松弛条件

$$\begin{aligned} d^\star &= \max_{\lambda,\eta} g(\lambda, \eta) \\ &= g(\lambda^\star, \eta^\star) \\ &= \min_x \mathcal{L}(x, \lambda^\star, \eta^\star) \\ &\leq \mathcal{L}(x^\star, \lambda^\star, \eta^\star) \\ &= f(x^\star) + \sum_{i=1}^{M} \lambda_i^\star m_i(x^\star) + \sum_{j=1}^{N} \eta_j^\star n_j(x^\star) \\ &= f(x^\star) + \sum_{i=1}^{M} \lambda_i^\star m_i(x^\star) \\ &\leq f(x^\star) = p^\star \end{aligned} \tag{13}$$

  在强对偶关系中 $d^\star = p^\star$, 因此上述不等号全部改为等号, $f(x^\star) + \sum_{i=1}^{M} \lambda_i^\star m_i(x^\star) \leq f(x^\star)$ 若为等号, 那么

$$\sum_{i=1}^{M} \lambda_i^\star m_i(x^\star) = 0 \tag{14}$$

  又因为 $\lambda_i^\star m_i(x^\star) \leq 0$, 那么$\lambda_i^\star m_i(x^\star) = 0$, 因此将此条件加入 KKT 条件, 也称为互补松弛条件

- 梯度为 0

  $\min_x \mathcal{L}(x, \lambda^\star, \eta^\star) \leq \mathcal{L}(x^\star, \lambda^\star, \eta^\star)$, 如果取等号的话意味着 $\mathcal{L}$ 在 $x^\star$ 处梯度为 0, 即

$$\frac{\partial \mathcal{L}}{\partial x}|_{x=x^\star} = 0 \tag{15}$$

以上说明了 KKT 条件为什么与强对偶关系互为充要条件.