# 机器学习算法系列（12）：SVM（5）—对偶

原文地址：支持向量机：Duality

在之前关于 support vector 的推导中，我们提到了 dual ，这里再来补充一点相关的知识。这套理论不仅适用于 SVM 的优化问题，而是对于所有带约束的优化问题都适用的，是优化理论中的一个重要部分。简单来说，对于任意一个带约束的优化都可以写成这样的形式：

$$\begin{aligned} &\min f_0(x) \\ &s.t.\ f_i(x) \le 0, \quad i = 1, \dots, m \\ &\quad h_i(x) = 0, \quad i = 1, \dots, p \end{aligned}$$

形式统一能够简化推导过程中不必要的复杂性。其他的形式都可以归约到这样的标准形式，例如一个 $maxf(x)$ 可以转化为 $min\ -f(x)$等。假如 $f_0, f_1, \dots, f_m$ 全都是凸函数，并且 $h_1, \dots, h_p$ 全都是仿射函数（就是形如 $Ax + b$ 的形式），那么这个问题就叫做凸优化（Convex Optimization）问题。凸优化问题有许多优良的性质，例如它的极值是唯一的。不过，这里我们并没有假定需要处理的优化问题是一个凸优化问题。

虽然约束条件能够帮助我们减小搜索空间，但是如果约束条件本身就是比较复杂的形式的话，其实是一件很让人头痛的问题，为此我们希望把带约束的优化问题转化为无约束的优化问题。为此，我们定义 Lagrangian 如下：

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

它通过一些系数把约束条件和目标函数结合在了一起。当然 Lagrangian 本身并不好玩，现在让我们来让他针对 $\lambda$ 和 $\nu$ 最大化，令：

$$z(x) = \max_{\lambda \succeq 0, \nu} L(x, \lambda, \nu)$$

这里 $\lambda \succeq 0$ 理解为向量 $\lambda$ 的每一个元素都非负即可。这个函数 $z(x)$ 对于满足原始问题约束条件的那些 $x$ 来说，其值等于 $f_0(x)$ ，这很容易验证，因为满足约束条件的$x$ 会使得 $h_i(x) = 0$ ，因此最后一项消掉了，而 $f_i(x) \le 0$ ，并且我们要求了 $\lambda \succeq 0$ ，因此 $\lambda_i f_i(x) \le 0$ ，所以最大值只能在它们都取零的时候得到，这个时候就只剩下 $f_0(x)$ 了。因此，对于满足约束条件的那些 $x$ 来说，$f_0(x) = z(x)$ 。这样一来，原始的带约束的优化问题其实等价于如下的无约束优化问题：

$$\min_x z(x)$$

因为如果原始问题有最优值，那么肯定是在满足约束条件的某个 $x^*$ 取得，而对于所有满足约束条件的 $x$ ，$z(x)$和$f_0(x)$ 都是相等的。至于那些不满足约束条件的 $x$ ，原始问题是无法取到的，否则极值问题无解。很容易验证对于这些不满足约束条件的$x$ 有 $z(x) = \infty$，这也和原始问题是一致的，因为求最小值得到无穷大可以和“无解”看作是相容的。

到这里，我们成功把带约束问题转化为了无约束问题，不过这其实只是一个形式上的重写，并没有什么本质上的改变。我们只是把原来的问题通过 Lagrangian 写作了如下形式：

$$\min_x \max_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

这个问题（或者说原始的带约束的形式）称作 primal problem 。如果你看过之前关于 SVM 的推导，那么肯定就知道了，相对应的还有一个 dual problem ，其形式非常类似，只是把 min 和 max 交换了一下：

$$\max_{\lambda \geq 0, \nu} \min_x L(x, \lambda, \nu)$$

交换之后的 dual problem 和原来的 primal problem 并不相等，直观地，我们可以这样来理解：胖子中最瘦的那个都比瘦骨精中最胖的那个要胖。当然这是很不严格的说法，而且扣字眼的话可以纠缠不休，所以我们还是来看严格数学描述。和刚才的 $z(x)$ 类似，我们也用一个记号来表示内层的这个函数，记：

$$g(\lambda, \nu) = \min_x L(x, \lambda, \nu)$$

并称 $g(\lambda, \nu)$ 为 Lagrange dual function （不要和 L 的 Lagrangian 混淆了）。g 有一个很好的性质就是它是 primal problem 的一个下界。换句话说，如果 primal problem 的最小值记为 $p^*$ ，那么对于所有的 $\lambda \succeq 0$ 和 $\nu$ ，我们有：

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

因为对于极值点（实际上包括所有满足约束条件的点）$x^*$，注意到 $\lambda \succeq 0$ ，我们总是有

$$\sum_{i=1}^{m} \lambda_i f_i(x^*) + \sum_{i=1}^{p} \nu_i h_i(x^*) \leq 0$$

因此

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

于是

$$g(\lambda,\nu)=\min_x L(x,\lambda,\nu)\leq L(x^*,\lambda,\nu)\leq f_0(x^*)=p^*$$

这样一来就确定了 g 的下界性质，于是

$$\max_{\lambda\geq 0,\nu} g(\lambda,\nu)$$

实际上就是最大的下界。这是很自然的，因为得到下界之后，我们自然地就希望得到最好的下界，也就是最大的那一个——因为它离我们要逼近的值最近呀。记 dual problem 的最优值为 $d^*$ 的话，根据上面的推导，我们就得到了如下性质：

$$d^*\leq p^*$$

这个性质叫做 weak duality ，对于所有的优化问题都成立。其中 $p^*-d^*$ 被称作 duality gap 。需要注意的是，无论 primal problem 是什么形式，dual problem 总是一个 convex optimization 的问题——它的极值是唯一的（如果存在的话），并且有现成的软件包可以对凸优化问题进行求解（虽然求解 general 的 convex optimization 实际上是很慢并且只能求解规模较小的问题的）。

这样一来，对于那些难以求解的 primal problem （比如，甚至可以是 NP 问题），我们可以通过找出它的 dual problem ，通过优化这个 dual problem 来得到原始问题的一个下界估计。或者说我们甚至都不用去优化这个 dual problem ，而是（通过某些方法，例如随机）选取一些 $\lambda\succeq 0$ 和 $\nu$ ，带到 $g(\lambda,\nu)$ 中，这样也会得到一些下界（只不过不一定是最大的那个下界而已）。当然要选 λ 和 ν 也并不是总是“随机选”那么容易，根据具体问题，有时候选出来的 λ 和 $\nu$ 带入 $g$ 会得到 $-\infty$，这虽然是一个完全合法的下界，然而却并没有给我们带来任何有用的信息。

故事到这里还没有结束，既然有 weak duality ，显然就会有 strong duality 。所谓 strong duality ，就是

$$d^*=p^*$$

这是一个很好的性质，strong duality 成立的情况下，我们可以通过求解 dual problem 来优化 primal problem ，在 SVM 中我们就是这样做的。当然并不是所有的问题都能满足 strong duality ，在讲 SVM 的时候我们直接假定了 strong duality 的成立，这里我们就来提一下 strong duality

成立的条件。

不过，这个问题如果要讲清楚，估计写一本书都不够，应该也有不少专门做优化方面的人在研究这相关的问题吧，我没有兴趣（当然也没有精力和能力）来做一个完整的介绍，相信大家也没有兴趣来看这样的东西——否则你肯定是专门研究优化方面的问题的了，此时你肯定比我懂得更多，也就不用看我写的介绍啦。

所以，这里我们就简要地介绍一下 Slater 条件和 KKT 条件。Slater 条件是指存在严格满足约束条件的点 $x$ ，这里的“严格”是指 $f_i(x) \leq 0$ 中的“小于或等于号”要严格取到“小于号”，亦即，存在 x 满足

$$
\begin{aligned}
f_i(x) &< 0 \quad i = 1, \dots, m \\
h_i(x) &= 0 \quad i = 1, \dots, p
\end{aligned}
$$

我们有：如果原始问题是 Convex 的并且满足 Slater 条件的话，那么 strong duality 成立。需要注意的是，这里只是指出了 strong duality 成立的一种情况，而并不是唯一情况。例如，对于某些非 convex optimization 的问题，strong duality 也成立。这里我们不妨回顾一下 SVM 的 primal problem ，那是一个 convex optimization 问题（QP 是凸优化问题的一种特殊情况），而 Slater 条件实际上在这里就等价于是存在这样的一个超平面将数据分隔开来，亦即是“数据是可分的”。当数据不可分是，strong duality 不能成立，不过，这个时候我们寻找分隔平面这个问题本身也就是没有意义的了，至于我们如何通过把数据映射到特征空间中来解决不可分的问题，这个当时已经介绍过了，这里就不多说了。

让我们回到 duality 的话题。来看看 strong duality 成立的时候的一些性质。

假设 $x^*$ 和 $(\lambda^*, \nu^*)$ 分别是 primal problem 和 dual problem 的极值点，相应的极值为 $p^*$ 和 $d^*$ ，首先 $p^* = d^*$ ，此时我们可以得到

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

由于两头是相等的，所以这一系列的式子里的不等号全部都可以换成等号。根据第一个不等号我们可以得到 $x^*$ 是 $L(x, \lambda^*, \nu^*)$ 的一个极值点，由此可以知道 $L(x, \lambda^*, \nu^*)$ 在 $x^*$ 处的梯度应该等

于 0 ，亦即：

$$\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) = 0$$

此外，由第二个不等式，又显然$\lambda_i^* f_i(x^*)$ 都是非正的，因此我们可以得到

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \ldots, m$$

这个条件叫做 complementary slackness 。显然，如果 $\lambda_i^* > 0$，那么必定有 $f_i(x^*) = 0$ ；反过来，如果 $f_i(x^*) < 0$ 那么可以得到 $\lambda_i^* = 0$ 。这个条件正是我们在介绍支持向量的文章末尾时用来证明那些非支持向量（对应于 $f_i(x^*) < 0$）所对应的系数 $\alpha_i$ （在本文里对应 $\lambda_i$ ）是为零的.)
再将其他一些显而易见的条件写到一起，就是传说中的 KKT (Karush-Kuhn-Tucker) 条件：

$$\begin{aligned}
f_i(x^*) &\le 0, \quad i = 1, \ldots, m \\
h_i(x^*) &= 0, \quad i = 1, \ldots, p \\
\lambda_i^* &\ge 0, \quad i = 1, \ldots, m \\
\lambda_i^* f_i(x^*) &= 0, \quad i = 1, \ldots, m \\
\nabla f_0(x^*) + \textstyle\sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla h_i(x^*) &= 0
\end{aligned}$$

任何满足 strong duality （不一定要求是通过 Slater 条件得到，也不一定要求是凸优化问题）的问题都满足 KKT 条件，换句话说，这是 strong duality 的一个必要条件。

不过，当原始问题是凸优化问题的时候（当然还要求原函数是可微的，否则 KKT 条件的最后一个式子就没有意义了），KKT 就可以升级为充要条件。换句话说，如果 primal problem 是一个凸优化问题，且存在 $\tilde{x}$ 和 $(\tilde{\lambda}, \tilde{\nu})$ 满足 KKT 条件，那么它们分别是 primal problem 和 dual problem 的极值点并且 strong duality 成立。其证明也比较简单，首先 primal problem 是凸优化问题的话，$g(\lambda, \nu) = min_x L(x, \lambda, \nu)$ 的求解对每一组固定的 $(\lambda, \nu)$ 来说也是一个凸优化问题，由 KKT 条件的最后一个式子，知道 $\tilde{x}$ 是 $min_x L(x, \tilde{\lambda}, \tilde{\nu})$ 的极值点（如果不是凸优化问题，则不一定能推出来），亦即：

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

最后一个式子是根据 KKT 条件的第二和第四个条件得到。由于 g 是 f0 的下界，这样一来，就证明了 duality gap 为零，也就是说，strong duality 成立。 到此为止，做一下总结。我们简要地介绍了 duality 的概念，基本上没有给什么具体的例子。不过由于内容比较多，为了避免文章超长，就挑了一些重点讲了一下。总的来说，一个优化问题，通过求出它的 dual problem ，在只有 weak duality 成立的情况下，我们至少可以得到原始问题的一个下界。而如果 strong duality 成立，则可以直接求解 dual problem 来解决原始问题，就如同经典的 SVM 的求解过程一样。有可能 dual problem 比 primal problem 更容易求解，或者 dual problem 有一些优良的结构（例如 SVM 中通过 dual problem 我们可以将问题表示成数据的内积形式从而使得 kernel trick 的应用成为可能）。此外，还有一些情况会同时求解 dual 和 primal problem ，比如在迭代求解的过程中，通过判断 duality gap 的大小，可以得出一个有效的迭代停止条件。