Support Vector Machine 01 Hard Margin Modeling and Solution

Chen Gong

13 November 2019

众所周知，Support Vector Machine (SVM) 有三宝，间隔，对偶，核技巧。所以，SVM 可以大致被分为三类：hard-margin SVM；soft-margin SVM；kernel SVM。

1 SVM 基本思想

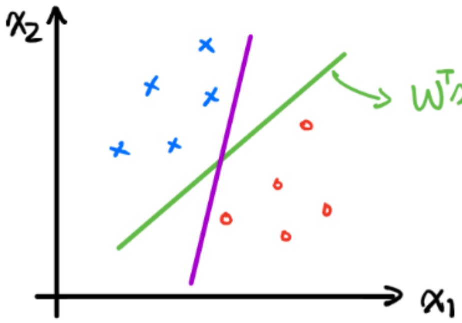

图 1: 二分类问题模型图

支持向量机模型可以被简要的描述为：$f(w) = w^T x + b$。很显然这是一个判别模型。实际上，我们想一想就知道，这样的直线其实有很多的。但是紫色的那条虽然可以做到分类的效果，但是效果也太差了，没有什么鲁棒性，泛化能力并不行。显然，绿色的那条直线要更好一些。那么，SVM 的基本思想可以被简要的概述为，找到一条最好的直线，离样本点距离足够的大。

2 SVM 模型建立

数据集可以描述为 $D = \{(x_i, y_i)\}_{i=1}^N$，其中 $x_i \in \mathbb{R}^p$，$y_i \in \{1, -1\}$。

首先我们希望，把这些点的间隔分得越大越好，并且根据符号函数给不同的值相应的类别标号。那么，我们可以写做：

$$
\begin{aligned}
&\max_{w,b}\ margin\ (w,b) \\
&s.t. \begin{cases} w^T x_i + b > 0 & y_i = +1 \\ w^T x_i + b < 0 & y_i = -1 \end{cases}
\end{aligned} \tag{1}
$$

由于 y_i 和 w_i^Tx+b 是同号的，那么很显然有 $y_i(w_i^Tx+b)>0$，所以，模型被我们改写为：

$$\begin{aligned}&\max_{w,b}\ margin\ (w,b)\\&s.t.\quad y_i(w_i^Tx+b)>0\quad (i=1,2,\cdots,N)\end{aligned}\tag{2}$$

平面上一点到某一直线的矩阵的计算方法比较简单。对于平面上一条直线 $y=w^Tx+b$，点 (x_i,y_i) 到直线的距离，可以被记做：

$$distance=\frac{1}{||w||}|w^Tx+b|\tag{3}$$

我们的希望是离超平面最近的点分得越开越好。离超平面最近的点就是 $\min distance(w,b,x_i)$，这个是针对点 $x_i(i=1,2,\cdots,n)$。然后就是分得越开越好,那么我们可以描述为 $\max\min distance(w,b,x_i)$，这个是针对 w,b 进行优化的。那么我们可以把模型进一步改写为：

$$\begin{aligned}&\max_{w,b}\min_{x_i}\frac{1}{||w||}|w^Tx_i+b|\\&s.t.\quad y_i(w_i^Tx+b)>0\quad (i=1,2,\cdots,N)\end{aligned}\tag{4}$$

对于约束条件 $y_i(w_i^Tx+b)>0\quad (i=1,2,\cdots,N)$，很显然可以得到 $\exists\gamma>0$ 使得 $s.t.\ \min y_i(w_i^Tx+b)=\gamma$。这里很显然我们可以使用一个小技巧来做一些的调整，来使我们方便计算，我们可以把约束条件转换为 $s.t.\ \min\frac{y_i(w_i^Tx+b)}{z}=\frac{\gamma}{z}$。我们很显然可以看到，$w$ 和 b 之间是可以自由放缩的，那么就放缩到令 $\frac{\gamma}{z}=1$，那么就有 $\min y_i(w_i^Tx+b)=1$。于是，模型可以化简为：

$$\begin{aligned}&\max_{w,b}\frac{1}{||w||}\\&s.t.\quad \min_{x_i}y_i(w_i^Tx+b)=1\Longrightarrow y_i(w_i^Tx+b)\geq 1\quad (i=1,2,\cdots,N)\end{aligned}\tag{5}$$

将该优化问题进行等价变换：

$$\begin{aligned}&\min_{w,b}\ \frac{1}{2}w^Tw\\&s.t.\quad y_i(w_i^Tx+b)\geq 1\quad (i=1,2,\cdots,N)\end{aligned}\tag{6}$$

很显然，这是一个凸优化 (Convex Optimization) 问题，目标函数是二次函数，一共有 N 个约束。那么这是一个二次规划问题 (Quadratic Programming)，通常也被描述为 QP 问题。

3 模型求解

在支持向量机的模型求解中，一个非常重要的概念就是将原问题 (Prime Problem) 转换为对偶问题 (Dual Problem)。我们将模型进一步改写为：

$$\begin{aligned}&\max_{w,b}\ \frac{1}{2}w^Tw\\&s.t.\quad 1-y_i(w_i^Tx+b)\leq 0\quad (i=1,2,\cdots,N)\end{aligned}\tag{7}$$

求解带约束的极值，显然需要采用拉格朗日乘子法，我们定义拉格朗日函数为：

$$\mathcal{L}(w,b,\lambda)=\frac{1}{2}w^Tw+\sum_{i=1}^N\lambda_i(1-y_i(w^Tx_i+b))\tag{8}$$

在拉格朗日数乘法里，λ 一定是大于零的数。所以模型为：

$$\begin{aligned} & \min_{w,b} \ \max_{\lambda} \ \mathcal{L}(w,b,\lambda) \\ & s.t. \quad \lambda_i \geq 0 \quad (i=1,2,\cdots,N) \end{aligned} \tag{9}$$

很显然，在这里，我们就将一个带约束的问题转换成了一个无约束的问题。

然而我们需要考虑一个问题，那就是 $\mathcal{L}(w,b,\lambda)$ 是否一定和公式 (7) 等价呢？这需要探究验证一下。

$$\begin{aligned} & if \ 1-y_i(w^Tx_i+b) \geq 0, \ \max_{\lambda} \mathcal{L}(\lambda,w,b) = +\infty \\ & if \ 1-y_i(w^Tx_i+b) \leq 0, \ \max_{\lambda} \mathcal{L}(\lambda,w,b) = 0 \end{aligned} \tag{10}$$

很显然在 $\min_{w,b} \ \max_{\lambda} \ \mathcal{L}(w,b,\lambda)$ 的计算中可以表示为：

$$\min_{w,b} \ \max_{\lambda} \ \mathcal{L}(w,b,\lambda) = \min_{w,b}\{+\infty, \frac{1}{2}w^Tw\} = \frac{1}{2}w^Tw \tag{11}$$

所以在上述的描述中，我们可以得到，实际上 $\min_{w,b} \ \max_{\lambda} \ \mathcal{L}(w,b,\lambda)$ 中隐藏了一个 $1-y_i(w^Tx_i+b) \leq 0$ 的隐藏条件。所以两种写法实际上是等价的。为了方便计算，下面我们需要使用对偶的方法，也就是将模型作如下的转换：

$$\left\{\begin{array}{l} \min_{w,b} \ \max_{\lambda} \ \mathcal{L}(w,b,\lambda) \\ s.t. \quad \lambda_i \geq 0 \end{array}\right. \quad \overset{dual}{\Longrightarrow} \left\{\begin{array}{l} \max_{\lambda} \ \min_{w,b} \ \mathcal{L}(w,b,\lambda) \\ s.t. \quad \lambda_i \geq 0 \end{array}\right. \tag{12}$$

这里我们需要介绍两种对偶关系，所谓：

弱对偶关系就是：$\min\max\mathcal{L} \geq \max\min\mathcal{L}$。

强对偶关系就是：$\min\max\mathcal{L} = \max\min\mathcal{L}$。

大家或许对这个关系会有点懵逼，其实仔细用直觉来想想还是很好接受的，具体的证明过程这里就不再做过多的阐述了。中国有句古话叫：“宁做鸡头不做凤尾”，但是凤就是凤始终要比鸡好。先取 max 就是凤的意思，然后取 min 就是凤尾。同理先取 min 就是鸡的意思，然后取 max 就是鸡头的意思。凤尾肯定比鸡头要好，当然这是直观的理解。而对于强对偶关系，需要我们满足 KKT 条件，这个后面会详细的说。

3.1 估计参数的值

我们的目标是 $\min_{w,b} \ \mathcal{L}(w,b,\lambda)$，那么

$$\frac{\partial \mathcal{L}}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{N}\lambda_i[1-y_i(w^Tx_i+b)] = 0 \tag{13}$$

$$-\sum_{i=1}^{N}\lambda_i y_i = 0 \tag{14}$$

代入到 $\mathcal{L}(w,b,\lambda)$ 中可得，

$$\mathcal{L}(w,b,\lambda) = \frac{1}{2}w^Tw + \sum_{i=1}^{N}\lambda_i(1-y_i(w^Tx_i+b)) \tag{15}$$

$$=\frac{1}{2}w^Tw+\sum_{i=1}^N\lambda_i-\sum_{i=1}^N\lambda_iy_iw^Tx_i-\sum_{i=1}^N\lambda_iy_ib \tag{16}$$

$$=\frac{1}{2}w^Tw+\sum_{i=1}^N\lambda_i-\sum_{i=1}^N\lambda_iy_iw^Tx_i \tag{17}$$

下一步，则是对 w 求偏导，

$$\frac{\partial\mathcal{L}}{\partial w}=\frac{\partial}{\partial w}[\frac{1}{2}w^Tw+\sum_{i=1}^N\lambda_i-\sum_{i=1}^N\lambda_iy_iw^Tx_i]=w-\sum_{i=1}^N\lambda_iy_ix_i=0 \tag{18}$$

解得：

$$w=\sum_{i=1}^N\lambda_iy_ix_i \tag{19}$$

将 w 的值代入到 $\mathcal{L}(w,b,\lambda)$ 中可以得到：

$$\begin{aligned}\mathcal{L}(w,b,\lambda)=&\frac{1}{2}(\sum_{i=1}^N\lambda_iy_ix_i)^T(\sum_{i=1}^N\lambda_iy_ix_i)+\sum_{i=1}^N\lambda_i-\sum_{i=1}^N\lambda_iy_i(\sum_{i=1}^N\lambda_iy_ix_i)^Tx_i\\=&\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\lambda_i\lambda_jy_iy_j(x_i^Tx_j)-\sum_{i=1}^N\sum_{j=1}^N\lambda_i\lambda_jy_iy_j(x_j^Tx_i)+\sum_{i=1}^N\lambda_i\\=&-\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\lambda_i\lambda_jy_iy_j(x_i^Tx_j)+\sum_{i=1}^N\lambda_i\end{aligned} \tag{20}$$

所以，模型被我们改写为：

$$\begin{cases}\max_\lambda -\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N\lambda_i\lambda_jy_iy_j(x_i^Tx_j)+\sum_{i=1}^N\lambda_i\\ s.t. \quad \lambda_i\geq 0,\ \sum_{i=1}^N\lambda_iy_i=0\end{cases} \tag{21}$$

4 KKT 条件

这个 KKT 条件或许会让很多人都感觉一脸懵逼，作者自己也理解了很久才勉强把它看懂的，如果有什么不到位的地方，欢迎发邮件到 gongchen2020@ia.ac.cn 与作者取得联系。深刻理解 KKT 条件需要掌握一些凸优化的知识，支持向量机是一个典型的凸二次优化问题。KKT 条件可以帮助我们理解支持向量机的精髓，什么是支持向量？支持向量机只需要用少量的数据，有很强的鲁棒性，并且可以取得很好的效果。

KKT 条件可以描述为：

$$\begin{cases}\frac{\partial\mathcal{L}}{\partial w}=0,\quad \frac{\partial\mathcal{L}}{\partial b}=0\\ \lambda_i(1-y_i(w^Tx_i+b))=0\\ \lambda_i\geq 0\\ 1-y_i(w^Tx_i+b)\leq 0\end{cases} \tag{22}$$

其中 $\lambda_i(1-y_i(w^Tx_i+b))=0$ 是互补松弛条件 (Complementary Relaxation Condition)。满足 KKT 条件是原问题的对偶 (dual) 问题有强对偶关系的充分必要条件。下面我们用一张图来进行理解 KKT 条件的作用：

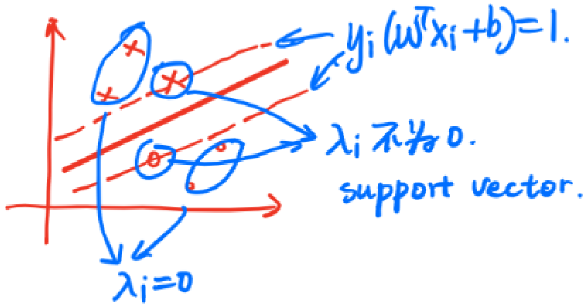

图 2: 支持向量的 KKT 条件

首先，需要明确，离分界面最近的数据点满足这个条件，$y_i(w^Tx_i+b)=1$ 至于为什么？前面的公式 (4) 有详细的分析。那么离分界面最近的数据点就被我们称为支持向量了。在支持向量上 $1-y_i(w^Tx_i+b)=0$，那么 λ_i 可以不为 0。而在其他向量上一定会有 $1-y_i(w^Tx_i+b)<0$ 为了满足 $\lambda_i(1-y_i(w^Tx_i+b))=0$，必然有 $\lambda_i=0$，那么我们就可以理解为这个数据点失去了作用。所以，KKT 条件使得，支持向量机中只有支持向量在模型的优化中有作用，这实在是太棒了。

为了确定这个超平面，我们已经得到了

$$w^* = \sum_{i=1}^{N}\lambda_i y_i x_i \tag{23}$$

但是，现在怎么求 b^* 是一个很尴尬的问题，因为我们在求 $\frac{\partial\mathcal{L}}{\partial b}$ 的时候，并没有看到和 b 相关的等式。但是我们知道只有支持向量会在模型求解中起作用，那么有支持向量 (x_k,y_k) 使得 $1-y_k(w^Tx_k+b)=0$。所以：

$$y_k(w^Tx_k+b)=1 \tag{24}$$

$$y_k^2(w^Tx_k+b)=y_k \tag{25}$$

$$b^* = y_k - w^Tx_k = y_k - \sum_{i=1}^{N}\lambda_i y_i x_i^T x_k \tag{26}$$

那么做到这里，我们的 hard-margin SVM 就已经做完了。模型为 $f(x)=sign(w^{*T}x+b^*)$，超平面为 $w^{*T}+b^*$。其中 $w^*=\sum_{i=1}^N\lambda_iy_ix_i$，$b^*=y_k-\sum_{i=1}^N\lambda_iy_ix_i^Tx_k$。

5 小结

本节主要探究了 Hard-margin SVM 的建模和求解。最终解得对于一个 $\{(x_i,y_i)_{i=1}^N\}$ 的分类问题，使用支持向量机来求解，我们可以得到，分类模型为：

$$f(x)=sign(w^{*T}x+b^*) \qquad \begin{cases} w^*=\sum_{i=1}^N\lambda_iy_ix_i \\ b^*=y_k-\sum_{i=1}^N\lambda_iy_ix_i^Tx_k \end{cases} \tag{27}$$

KKT 条件是原问题的对偶 (dual) 问题有强对偶关系的充分必要条件。它成功的使支持向量机模

型的求解只和支持向量有关，这也是支持向量机的强大之处，运算比较简单，而且具有较强的鲁棒性。

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0, \quad \frac{\partial \mathcal{L}}{\partial b} = 0, \quad \frac{\partial \mathcal{L}}{\partial \lambda} = 0 \\ \lambda_i(1 - y_i(w^T x_i + b)) = 0 \\ \lambda_i \geq 0 \\ 1 - y_i(w^T x_i + b) \leq 0 \end{cases} \tag{28}$$

Support Vector Machine 02 Soft Margin

Chen Gong

15 November 2019

在上一小节中，我们介绍了 Hard-Margin SVM 的建模和求解过程。这个想法很好，但是实际使用过程中会遇到很多的问题。因为，并不一定数据集就可以被很好的分开，而且实际数据没有那么简单，其间有很多的噪声。而 Soft Margin 的基础思想就是允许那么一点点的错误。这样在实际运用中往往可以得到较好的效果。下面我们将进行 Soft Margin SVM 的详细演变过程。

1 Soft Margin SVM

最简单的思路就是在优化函数里面引入一个 loss function。也就是：

$$\min \frac{1}{2}w^Tw + loss\ function \tag{1}$$

那么，我们如何来定义这个 loss function 呢？大致可以分这两种引入的模式：

1. loss = 错误点的个数 = $\sum_{i=1}^N I\{y_i(w^Tx_i+b)<1\}$，这个方法非常容易想到，但是我们马上就发现了一个问题，那就是这个函数不连续的，无法进行优化。这种方法非常容易想到。

2. loss：距离。现在我们做如下定义：

1) 如果 $y_i(w^Tx_i+b)\geq 1$，$loss = 0$。

2) 如果 $y_i(w^Tx_i+b)<1$，$loss = 1-y_i(w^Tx_i+b)$。

那么，我们就可以将 loss function 定义为：

$$loss = \max\{0, 1-y_i(w^Tx_i+b)\} \tag{2}$$

进一步，我们令 $y_i(w^Tx_i+b)=z$，那么：

$$loss_{max} = \max\{0, 1-z\} \tag{3}$$

我们将 loss function 的图像画出来就如下图所示：

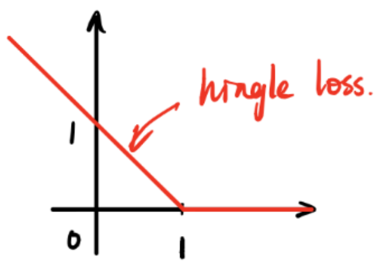

图 1: loss function 的展示图

这个 loss function 已经是连续的了，而且看起来是不是很像书的开着的样子。所以，它有一个非常形象的名字也就是“合页函数” (Hinge loss)。那么到这里，我们的 Soft Margin SVM 可以被定义为：

$$\begin{cases} \min \ \frac{1}{2}w^Tw + C\sum_{i=1}^{N} \max\{0, 1 - y_i(w^Tx_i + b)\} \\ s.t. \quad y_i(w^Tx_i + b) \geq 1 \end{cases} \tag{4}$$

但是，这样写显然不是我们想要的形式，我们需要得到更简便一些的写法。我们引入 $\xi_i = 1 - y_i(w^Tx_i + b)$，$\xi_i \geq 0$。我们仔细的想一想 $\max\{0, 1 - y_i(w^Tx_i + b)\}$ 和 ξ_i 之间的关系。有了 $\xi_i \geq 0$，我们可以得到其实 $\xi_i \geq 0$ 和 $\max\{0, 1 - y_i(w^Tx_i + b)\}$ 实际上是等价的。那么这个优化模型我们可以写成:

$$\begin{cases} \min \ \frac{1}{2}w^Tw + C\sum_{i=1}^{N} \xi_i \\ s.t. \quad y_i(w^Tx_i + b) \geq 1 - \xi_i, \ \xi_i \geq 0 \end{cases} \tag{5}$$

在图像上表示即为:

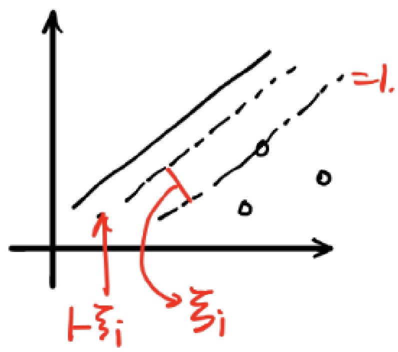

图 2: Soft Margin SVM 模型展示图

在以前的基础上我们增加了一个缓冲区，由于这个缓冲区的存在我们可以允许有点点的误差。而支持向量的区间被放宽到了 $1 - \xi_i$。

Support Vector Machine 03 Weak Duality Proof

Chen Gong

16 November 2019

在前面我们已经展示的 Hard Margin 和 Soft Margin SVM 的建模和求解。前面提到的 SVM 有三宝，间隔，对偶，核技巧。前面我们已经分析了间隔，大家对于其中用到的对偶，虽然我们用比较直觉性的方法进行了解释，但是估计大家还是有点懵逼。这节我们希望给到通用性的证明，这里实际上就是用到了约束优化问题。

1 弱对偶性证明

首先，我们需要证明约束优化问题的原问题和无约束问题之间的等价性。

1.1 约束优化问题与无约束问题的等价性

对于一个约束优化问题，我们可以写成：

$$\left\{\begin{array}{l} \min_{x\in\mathcal{R}^p} f(x) \\ s.t. \quad m_i(x) \leq 0,\ i=1,2,\cdots,N \\ \qquad\quad n_i(x) = 0,\ i=1,2,\cdots,N \end{array}\right. \tag{1}$$

我们用拉格朗日函数来进行表示：

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

我们可以等价的表示为：

$$\left\{\begin{array}{l} \min_x \max_{\lambda,\eta}\ \mathcal{L}(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0,\ i=1,2,\cdots,N \end{array}\right. \tag{3}$$

这就是将一个约束优化问题的原问题转换为无约束问题。那么这两种表达形式一定是等价的吗？我们可以来分析一下：

如果，x 违反了约束条件 $m_i(x) \leq 0$，那么有，$m_i(x) > 0$。且 $\lambda_i > 0$，那么很显然 $max_\lambda\ \mathcal{L} = +\infty$。

如果，x 符合约束条件 $m_i(x) \leq 0$，那么很显然 $max_\lambda\ \mathcal{L} \neq +\infty$。

那么：

$$\min_x \max_{\lambda,\eta}\ \mathcal{L}(x,\lambda,\eta) = \min_x \{\max\ \mathcal{L}, +\infty\} = \min_x \{\max\ \mathcal{L}\} \tag{4}$$

其实大家可以很明显的感觉到，这个等式自动的帮助我们过滤到了一半 $m_i(x) \geq 0$ 的情况，这实际上就是一个隐含的约束条件，帮我们去掉了一部分不够好的解。

1.2 证明弱对偶性

原问题我们可以写为：

$$\begin{cases} \min_x \max_{\lambda,\eta}\ \mathcal{L}(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0,\ i=1,2,\cdots,N \end{cases} \tag{5}$$

而原问题的对偶问题则为：

$$\begin{cases} \min_{\lambda,\eta} \max_x\ \mathcal{L}(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0,\ i=1,2,\cdots,N \end{cases} \tag{6}$$

原问题是一个关于 x 的函数，而对偶问题是一个关于 λ,η 的最小化问题，而弱对偶性则可以描述为：对偶问题的解 $\leq$ 原问题的解。为了简化表达，后面对偶问题的解我们用 d 来表示，而原问题的解我们用 p 来表示。那么我们用公式化的语言表达也就是：

$$\min_{\lambda,\eta}\max_x\ \mathcal{L}(x,\lambda,\eta) = d \leq \min_x \max_{\lambda,\eta}\ \mathcal{L}(x,\lambda,\eta) = p \tag{7}$$

在前面我们使用感性的方法证明了 $\max\min\ \mathcal{L} \leq \min\max\ \mathcal{L}$，下面我们给出严谨的证明：

很显然可以得到：

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

那么，$\min_x\ \mathcal{L}(x,\lambda,\eta)$ 可表示为一个与 x 无关的函数 $A(\lambda,\eta)$，同理 $\max_{\lambda,\eta}\ \mathcal{L}(x,\lambda,\eta)$ 可表示为一个与 λ,η 无关的函数 $B(x)$。显然，我们可以得到一个恒等式：

$$A(\lambda,\eta) \leq B(x) \tag{9}$$

那么接下来就有：

$$\begin{aligned} A(\lambda,\eta) &\leq \min\ B(x) \\ \max\ A(\lambda,\eta) &\leq \min\ B(x) \\ \min_{\lambda,\eta}\max_x\ \mathcal{L}(x,\lambda,\eta) &\leq \min_x\max_{\lambda,\eta}\ \mathcal{L}(x,\lambda,\eta) \end{aligned} \tag{10}$$

弱对偶性，证毕!!

Support Vector Machine 04 Weak Duality Geometric Interpretation

Chen Gong

17 November 2019

上一小节中我们讨论了有关弱对偶性的证明，这一节我们从几何的角度来解释一下有关对偶问题。为了方便描述，我们将对偶问题进行简化为如下形式：

$$\begin{cases} \min_{x\in\mathcal{R}^p} \ \ f(x) \\ s.t. \quad m_i \leq 0 \end{cases} \tag{1}$$

$\mathbb{D}$: 定义域，$D = dom\ f \cap dom\ m_i$，这是一种常见的定义域的表示方法。其中，$x \in \mathbb{D}$。我们将模型表达为拉格朗日函数的形式，

$$\mathcal{L}(x,\lambda) = f(x) + \lambda m_1(x), \qquad \lambda \leq 0 \tag{2}$$

我们将原问题的最优解记为：$p^* = \min \ f(x)$。

我们将对偶问题的最优解记为：$d^* = \max_\lambda \min_x \ \mathcal{L}(x,\lambda)$。

1 模型表述

上述表述中，表达了模型的基本问题，下面我们进一步抽象模型。首先，我们需要描述一个集合：

$$G = \{(m_1(x), f(x))|x \in \mathbb{D}\} \tag{3}$$

为了简化运算，我们需要简化符号，令 $m_1(x) = \mu$，$f(x) = t$。那么，

$$G = \{(\mu, t)|x \in \mathbb{D}\} \tag{4}$$

我们需要想想如何集合话来表示，首先 $p^* = \min \ f(x) = \min \ t$，其中，$\{t|(\mu,t) \in G\}$。那么，我们用 inf 来表示下确界的意思，就有：

$$p^* = inf\,\{t|(\mu,t) \in G, \mu \leq 0\} \tag{5}$$

那么对偶问题，我们可以写成，

$$d^* = \max_\lambda \min_x \ \mathcal{L}(x,\lambda) = \max_\lambda \min_x (t + \lambda\mu) \tag{6}$$

又因为 $(t + \lambda\mu)$ 只和 λ 有关，那么可以被记做 $g(\lambda)$。而且，$g(\lambda)$ 可以被写作，$g(\lambda) = inf(t + \lambda\mu)|(\mu,t) \in G$。在对偶条件中不需要那个 $\mu \leq 0$，因为已经包含在原等式的隐藏条件里了。但是，在原问题中，我们一定不能忘记这个条件。

2 模型表达

2.1 p^* 的几何表示

下一步的主要问题就是，我们需要如何来表达 p^* 和 $d^*(g(\lambda))$。首先我们来看 p^*，其实它的表达还算比较简单。我们来看这个图像的表达式：

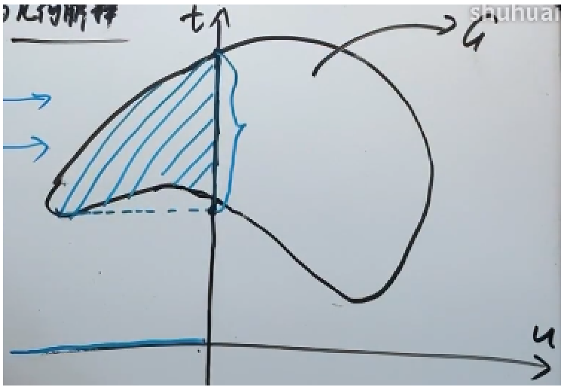

图 1: p^* 的几何表示

我们假设 G 就是 (μ, t) 的定义域的几何表示区间。$p^* = inf\,\{t|(\mu, t) \in G, \mu \leq 0\}$，由于 $\mu \leq 0$，所以我们只看左边一半。那么 t 的值就是坐标纵轴上的一截部分。最小值非常的好确定，就是平行于 μ 轴，最下方的切点。

2.2 $d^*(g(\lambda))$ 的几何表示

这个等式的几何表示就会有点困难了，我们需要分解成两步，第一步确定 $g(\lambda)$ 的几何表达；第二步，确定 d^* 的几何表达。

2.2.1 $g(\lambda)$ 的几何表达

由于 $t + \lambda\mu$ 是一个关于 x 的变量，在这其中 λ 起到的是一个斜率的作用，这个斜率是一直保持不变的。而得到的 $t + \lambda\mu$ 的结果我们记为 $\triangle$。$\triangle$ 也就是 $t + \lambda\mu$ 和 t 轴的交点。那么，也就是一根固定斜率的直线在 t 的方向上进行移动。

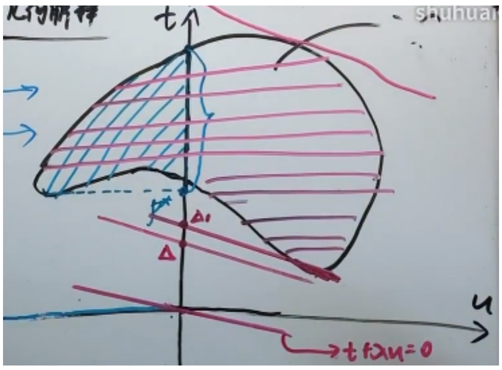

图 2: $g(\lambda)$ 的几何表达

我们可以假设 $t+\lambda\mu$ 与 t 轴的交点是一个集合，这个集合就是 $\{\triangle_1, \triangle_2, \cdots, \triangle_N\}$。

2.2.2 d^* 的几何表达

下一步，我们需要求的是 $d^* = \max_\lambda g(\lambda)$。现在相当于是固定了一个点，然后围着这个点在转。这个点是哪个店呢？就是 $(0, t)$。大家仔细想一想比对一下上图就知道是不是转到与集合 G 相切的时候得到的这个解是最优解，但是这个解一定会比 p^* 得到的解会更小。为什么？用屁股想都知道，一个是横着切，一个是斜着切，哪个会更小？不言而喻了吧。通过这个我们也可以得到，

$$d^* \leq p^* \tag{7}$$

3 小结

上面我们从几何的角度来重新解释了这个问题，其实仔细的想一想也不算很难。但是，强对偶性的证明这个东西有点难，实际上学习机器学习中的 SVM，学到这就差不多够了。如果是强对偶性，我们还需要满足两个条件，也就是 1. 是一个凸优化问题；2. slate 条件。就可以得到 $d^* = p^*$。下一节会进一步解释，但是这只是一个充分必要条件，满足其他的条件也可能是强对偶关系。而 SVM 是一个二次规划问题，那么它一定是一个强对偶问题。

Support Vector Machine 05 Slate & KKT Condition

Chen Gong

18 November 2019

首先，我们整理一下前面得到的有关约束优化的模型。我们可以描述为：

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

其中，

$$D = \left\{ dom \ f \bigcap_{i=1}^{M} dom \ m_i \bigcap_{j=1}^{N} dom \ n_j \right\} \tag{2}$$

我们将模型进行简化可得：

$$\begin{cases} \min \ \ f(x) \\ s.t. \quad m_i(x) \end{cases} \Longrightarrow G = \{(m, f) | x \in D\} = \{(\mu, t) | x \in D\} \tag{3}$$

那么，我们的优化目标为：

$$p^* = inf\{t | (\mu, t) \in G, \mu \leq 0\} \tag{4}$$

$$g(\lambda) = inf(t + \lambda\mu | (\mu, t) \in G) \tag{5}$$

通常来说，凸优化问题，不一定是强对偶问题。往往都是凸优化问题需要加上一些限定条件才可以构成强对偶问题。比如说 slate condition，但是这些条件往往都是充分非必要的。这样的条件有很多种，slate condition 只是其中一种，类似的还有 KKT condition。

1 Slate Condition

下面简述一下 Slate Condition，详细的证明过程就不做过多的描述。也就是 $\exists \hat{x} \in relint \ D, \ s.t. \ \forall i = 1, 2, \cdots, m, \ m_i(\hat{x}) \leq 0$。而 relint 的意思就是，relative interior，相对内部的意思。

而对于绝大部分的凸优化问题，通常 Slate 条件是成立的。而放松的 Slate 条件为：假设 M 中有 k 个仿射函数，$M - k$ 个仿射。而 SVM 是一个典型的凸二次规划问题，也就是目标函数 f 是凸函数，m_i 是仿射函数，n_j 为仿射。那么在几何上是什么意思呢？也就是限制至少有一个点在坐标系的左边，限制直线不是垂直的，这里需要结合 Support Vector Machine 04 中的几何解释来看。

2 KKT Condition

在上文中我们知道了 Convex 和 Slate Condition 可以得到强对偶关系，也就是 $d^* = p^*$。但是这只是一个充分非必要条件。同样的在满足 KKT Condition 的情况下，我们也可以得出是一个强对偶问题，并且这是一个充分必要的条件。

我们在来回顾一下模型的原问题：

$$\begin{cases} \min \ \ f(x) \\ s.t. \quad m_i(x) \leq 0, \ i = 1, 2, \cdots, M \\ \qquad\quad n_j(x) = 0, \ j = 1, 2, \cdots, N \end{cases} \tag{6}$$

而拉格朗日形式的表达为：

$$\mathcal{L}(x, \lambda) = f(x) + \sum_i \lambda_i m_i(x) + \sum_j \eta_j n_j(x) \tag{7}$$

对于对偶问题，我们可以描述对应的 $g(\lambda, \eta) = \min_x \mathcal{L}(x, \eta, \lambda)$；$d* \longleftarrow \lambda^*, \eta^*$。所以对偶问题 (Dual Prob) 也就是：

$$\begin{cases} \max_{\lambda,\eta} \ \ g(\lambda, \eta) \\ s.t. \quad \lambda_i \geq 0, \ i = 1, 2, \cdots, M \end{cases} \tag{8}$$

下面进行 KKT 条件的推导：

首先一定需要满足的是，在可行域以内。所以，一定会有：$m_i(x^*) \leq 0, n_i(x^*) = 0, \lambda^* \geq 0$。并且还需要满足：

$$\begin{aligned} d^* =& \max_{\lambda,\eta} g(\lambda, \eta) = g(\lambda^*, \eta^*) \\ =& \min_x \mathcal{L}(x, \lambda^*, \eta^*) \\ \leq& \mathcal{L}(x, \lambda^*, \eta^*), \quad \forall x \in D \\ =& \mathcal{L}(x^*, \lambda^*, \eta^*) \\ =& f(x^*) + \sum_i \lambda_i m_i(x^*) + \sum_j \eta_j n_j(x^*) \\ =& f(x^*) + \sum_i \lambda_i m_i(x^*) \end{aligned} \tag{9}$$

上式中的 $f(x^*)$ 也就是 p^*，用因为 $\lambda_i m_i(x^*) \leq 0$ 是必然存在的。所以，$d^* \leq f(x^*)$。这就是弱对偶关系，如果是强对偶关系，就需要我们需要在两个小于或等于号那取等才行。

第一，对于 $\forall i = 0, 1, 2, \cdots, M$，都有 $\sum_i \lambda_i m_i = 0$。

第二，$\min \mathcal{L}(x, \lambda^*, \eta^*), \quad \forall x \in D = \mathcal{L}(x^*, \lambda^*, \eta^*)$。也就是：

$$\frac{\partial \mathcal{L}(x, \lambda^*, \eta^*)}{\partial x} \mid_{x=x^*} = 0 \tag{10}$$

所以，KKT 条件就已经完成了，我们总结一下，KKT 条件分成 3 个部分。

1. 可行条件：也就是需要满足定义域的条件，$m_i(x^*) \leq 0, n_i(x^*) = 0, \lambda^* \geq 0$。
2. 互补松弛条件：$\lambda_i m_i = 0$。
3. 梯度为零：$\frac{\partial \mathcal{L}(x, \lambda^*, \eta^*)}{\partial x} \mid_{x=x^*} = 0$。

我们可以对比之前学习的 SVM 的 KKT 条件。