对偶问题

前言

上一节，我们使用 LDA 对比出了 SVM，并简单介绍了关于 SVM 的基本概念。超平面、支持向量、间隔...现在我们将继续深入这些，以及我们末尾提到的这个公式：

$$\begin{array}{ll}\min_{\mathbf{w},b} & \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{s.t.} & y_i(\mathbf{w}^\top x_i + b) \geq 1, \quad i = 1,2,...,m.\end{array}, \quad (6.6)$$

这个公式与上一节提到的 LDA 损失函数是一样的，它们都是我们模型追求的优化目标，我们只需求解它就可以。但是我们希望求解上式，从而得到最大间隔划分超平面所对应的模型，也就是：

$$f(x) = \mathbf{w}^\top \mathbf{x} + b, \quad (6.7)$$

这里 $\mathbf{w}$ 和 b 是模型参数，并且(6.6)是一个典型的二次规划(convex quadratic programm)问题，怎么处理呢？简单，对(6.6)进行再变形，使用拉格朗日乘子法可得到其“对偶问题”(dual problem)。

不过开始之前，先解决你们肯定会有的两个疑问：

什么是凸二次规划问题？

什么是二次规划问题？对于目标函数，我们的函数首先是凸二次函数，也就是 $\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2$ ；除此以外还要求约束是凸约束，也就是：

- 要么等式约束($\mathbf{A}\mathbf{x} = \mathbf{b}$ ， $\mathbf{A}$ 是矩阵， $\mathbf{b}$ 是向量)，线性等式约束是凸的
- 要么不等式约束($\mathbf{G}\mathbf{x} \leq \mathbf{h}$ ， G 是矩阵， $\mathbf{h}$ 是向量)，线性不等式约束也是凸的

很明显 $\text{s.t.} \quad y_i(\mathbf{w}^\top x_i + b) \geq 1$ 是线性不等式约束，属于凸约束。他有几个很明显的特点：第一，若存在局部最优解那就是全局最优解；第二，有各种高效求解方法。

为什么解决凸二次规划要引入拉格朗日对偶？什么是拉格朗日对偶化？

很简单的问题：

- 第一，方便引入核函数
- 第二，简化运算，减少约束条件（KKT条件）
- 第三，在简化运算的基础上，原始问题和对偶问题的最优解是相等的，还能有核可用，一举两得

至于什么是拉格朗日对偶化？我可能需要先解释一下对偶问题？

简单来说，对偶问题是与原始优化问题相对应的另一个优化问题，通过特定的变换规则从原始问题衍生而来，与原始问题的最优解相同，其常用于求解约束优化问题，也就是上边的 $\mathrm{s.t.}$ 。

而对于拉格朗日对偶化，在不改变原始优化的解的基础上，引入新的乘子和新的约束，使得从原始问题的优化变量维度与特征维度相关，到对偶问题的优化变量维度与样本数量相关（后边看到 α 乘子就明白了，我们开始关注样本总数），当样本数量小于特征维度时，对偶问题计算量更小。

对偶问题

现在我们回到对偶问题上。具体来说，对于上式子(6.6)的每个约束添加我们提到的乘子 α ，则我们得到了下面的变式：

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2+\sum_{i=1}^{m}\alpha_i\big(1-y_i(\mathbf{w}^\top x_i+b)\big), \quad (6.8)$$

1. 求偏导

其中 $\alpha=(\alpha_1;\alpha_2;...;\alpha_m)$ ， α 的个数 m 与样本数量一致，令 $L(\mathbf{w},b,\alpha)$ 对 $\mathbf{w}$ 和 b 的偏导为 0 得：

- **关于 $\mathbf{w}$ 的偏导：**

$$\nabla_{\mathbf{w}}L=\mathbf{w}-\sum_{i=1}^{m}\alpha_i y_i x_i=0 \quad \Rightarrow \quad \mathbf{w}=\sum_{i=1}^{m}\alpha_i y_i x_i, \quad (6.9)$$

这表示 $\mathbf{w}$ 是训练样本的线性组合。

- **关于 (b) 的偏导：**

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{m}\alpha_i y_i=0 \quad \Rightarrow \quad \sum_{i=1}^{m}\alpha_i y_i=0, \quad (6.10)$$

这是一个新约束，确保偏置 b 的合理性。

2. 代入拉格朗日函数

将式(6.9)代入(6.8)，即可将 $L(\mathbf{w},b,\alpha)$ 中的 $\mathbf{w}$ 和 b 消去，再考虑式子(6.10)的约束，得到式子(6.6)的对偶问题

将 $\mathbf{w}=\sum_{i=1}^{m}\alpha_i y_i x_i$ 和 $\sum_{i=1}^{m}\alpha_i y_i=0$ 代回 L，并简化：

$$L=\frac{1}{2}||\mathbf{w}||^2-\sum_{i=1}^{m}\alpha_i y_i(\mathbf{w}\cdot x_i)-b\underbrace{\sum_{i=1}^{m}\alpha_i y_i}_{=0}+\sum_{i=1}^{m}\alpha_i$$

代入 $\mathbf{w}$：

$$||\mathbf{w}||^2=\left(\sum_{i=1}^{m}\alpha_i y_i x_i\right)\cdot\left(\sum_{j=1}^{m}\alpha_j y_j x_j\right)=\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$

$$\sum_{i=1}^{m}\alpha_i y_i(\mathbf{w}\cdot x_i)=\sum_{i=1}^{m}\alpha_i y_i\left(\sum_{j=1}^{m}\alpha_j y_j x_j\cdot x_i\right)=\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$

代入 L：

$$L=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{m}\alpha_i$$

合并项得到：

$$L=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$

加上(6.10)的约束，最终：

$$\begin{aligned}&\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^{\top}x_j\\&\text{s.t.}\quad\sum_{i=1}^{m}\alpha_i y_i=0\\&\qquad\alpha_i\geq 0,\quad i=1,2,\dots,m.\end{aligned}\quad,\quad(6.11)$$

这就是对偶化的原问题，这里需要注意几个关键点：

1. 关注 $x_i^{\top}x_j$，这个以后和核函数有关，但是现在不需要，我提一嘴
2. $\sum_{i=1}^{m}\alpha_i y_i=0$ 是求偏导的得到的约束
3. $i=1,2,\dots,m.$ 来自 $\alpha=(\alpha_1;\alpha_2;\dots;\alpha_m)$

3. 回到原模型上

之后只需根据(6.11)解出 α，求出 $\mathbf{w}$ 和 b 得到模型，将(6.9)代入(6.7)：

$$\begin{aligned}f(x)&=\mathbf{w}^{\top}\mathbf{x}+b\\&=(\sum_{i=1}^{m}\alpha_i y_i x_i^{\top})\mathbf{x}+b\end{aligned}\qquad(6.12)$$

4. 关于 KKT 条件

从对偶问题(6.11)解出的 α_i 是式(6.8)中的拉格朗日乘子，它恰对应着训练样本 (x_i,y_i)，一个 α 对应一对样本数据，也就是我提到的 3。注意式(6.6)有不等式约束，因此上述过程需满足 KKT(Karush-Kuhn-Tucker) 条件，即：

$$\begin{cases} \alpha_i \geq 0;\text{，} ① \\ \\ y_i f(x_i) - 1 \geq 0; ② \\ \\ \alpha_i (y_i f(x_i) - 1) = 0; ③ \end{cases}$$

你可能不理解这几个条件怎么来的，先逐个解释一下三个不等式怎么来的：

1. $\alpha_i \geq 0$ **对偶可行性**：不需要解释，这是(6.11) s.t. 给的条件
2. $y_i f(x_i) - 1 \geq 0$ **原始可行性**：是由(6.3)的 $\begin{cases} \mathbf{w}^\top x_i + b \geq +1，& y_i = +1 \\ \mathbf{w}^\top x_i + b \leq -1，& y_i = -1 \end{cases}$ 变成(6.6)的 $\text{s.t.} \quad y_i(\mathbf{w}^\top x_i + b) \geq 1$ 移项得来
3. $\alpha_i (y_i f(x_i) - 1) = 0$ **互补松弛**：这个可有的说。首先我们需要清楚一件事，α_i 和 $y_i f(x_i) - 1$ 不能同时为 0，也就是有两个情况：
   a. $\alpha_i \geq 0$ 时，$y_i f(x_i) - 1 = 0$：这时候就是我们在 间隔与支持向量 的(6.3) $\begin{cases} \mathbf{w}^\top x_i + b \geq +1，& y_i = +1 \\ \mathbf{w}^\top x_i + b \leq -1，& y_i = -1 \end{cases}$，“使这个方程组满足等号的，也就是我们上边说的支持向量⊕和⊖”。就代表这个点就是支持向量！
   b. 而当 $\alpha_i = 0$ 时，$y_i f(x_i) - 1 \geq 0$：也就是说不满足支持向量的要求，这个点不是支持向量！不会对 $f(x)$ 超平面有任何影响。

现在你明白了么？*KKT 条件在某些方面帮我们识别了哪些是支持向量，哪些不是支持向量！*

5. 求解(2)对偶问题——SMO算法

我们该怎么求解下面这么一坨东西呢？

$$\begin{aligned} \max_\alpha & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j x_i^\top x_j \\ \text{s.t.} & \sum_{i=1}^m \alpha_i y_i = 0 \qquad\qquad , \quad (6.11) \\ & \alpha_i \geq 0， \quad i = 1, 2, \ldots, m. \end{aligned}$$

我们注意到这个式子有 $\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j x_i^\top x_j$，这就是一个二次项，并且约束中既有线性等式约束也有线性不等式约束，诶？这不是二次规划问题么？我们依旧可以使用二次规划的一些算法来解决......但是，该问题的规模正比于训练样本数，这会在实际任务中造成很大的开销，为了避开这个障碍，我们需要利用问题本身的特性，使用一些巧妙的算法，SMO (Sequential Minimal Optimization) 就是其中一个著名的代表。

SMO 算法的基本思路有点像是坐标下降法（但不是）：先固定一些参数，迭代更新其他参数；之后再固定更新后的参数，迭代更新之前固定的参数...反复如此。

再具体一点：先固定 α_i 之外的参数，然后求 α_i 上的极值。由于有约束 $\sum_{i=1}^{m}\alpha_i y_i=0$，若固定 α_i 之外的其他变量，则 α_i 可有其他变量推导出来（也就是说实际上你知道一个变量后，另一个也就知道了），所以，SMO 算法每次选择两个变量 α_i 和 α_j 并固定其他参数，表示其中一个 α，另一个也就是表示出来了（$\alpha_1 y_1+\alpha_2 y_2=\varsigma$）。就这样，不断迭代直至收敛。

注：为什么会选择两个 α 变量呢？想象一下，如果我们一次更新一个 α，会破坏等式约束 $\sum_{i=1}^{m}\alpha_i y_i=0$。所以 SMO 一次选择 **两个变量** (α_i,α_j) 进行更新，这样我们可以在更新时保持等式成立。

下面是 SMO 算法的推导过程：

1）SMO 算法的推导

SMO 算法推导参考：知乎文章和配套视频

先选择 α_1 和 α_2 作为我们的变量，其他变量都固定（怎么选这个值我会在最后说，这里先推导，“*怎么选择 α_1 和 α_2 ？*”）

我们已经知道两个变量 α_1 和 α_2 之间存在约束 $\sum_{i=1}^{n}\alpha_i y_i=0$，那么其实只需要搞定其中一个变量，另一个就可以表示出来，整个式子就变成了一个一元二次方程问题，代入消元即可。

原式变换

下面将(6.11)进行变换：

$$\begin{aligned}\text{原式}&=\max_\alpha\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j\\&=\min_\alpha\Big[\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j-\sum_{i=1}^{m}\alpha_i\Big]\end{aligned}\quad,\qquad (6.11.1)$$

取出 α_1 和 α_2

变换，将 α_1 和 α_2 的组合单独提出，后边的求和 $\sum$ 从 3 开始（我把 min 去了，方便推导，后边再加上也不迟）：

$$
\begin{aligned}
\text{原式} &= \min_{\alpha}[\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^{\top}x_j - \sum_{i=1}^{m}\alpha_i] \\
&= \frac{1}{2}[\alpha_1 y_1 \alpha_1 y_1 x_1^T \cdot x_1 + 2\alpha_1 y_1 \alpha_2 y_2 x_1^T \cdot x_2 \\
&+ 2\sum_{j=3}^{m}\alpha_1 y_1 \alpha_j y_j x_1^T \cdot x_j + \alpha_2 y_2 \alpha_2 y_2 x_2^T \cdot x_2 + 2\sum_{j=3}^{m}\alpha_2 y_2 \alpha_j y_j x_2^T \cdot x_j] \\
&+ 2\sum_{j=3}^{m}\alpha_2 y_2 \alpha_j y_j x_2^T \cdot x_j + \sum_{i=3}^{m}\sum_{j=3}^{m}\alpha_i y_i \alpha_j y_j x_i^T \cdot x_j - [\alpha_1 + \alpha_2 + \sum_{j=3}^{m}\alpha_j]
\end{aligned}
$$

核函数代换

核代换（虽然我们还没说到核，但是不耽误我们代换，先把它当作“令”看待吧）：“令” $x_i^T \cdot x_j = \kappa_{ij}$，方便后续的核函数带入处理非线性问题：

$$
\begin{aligned}
\text{原式} = \frac{1}{2}[&\alpha_1 y_1 \alpha_1 y_1 \kappa_{11} + 2\alpha_1 y_1 \alpha_2 y_2 \kappa_{12} + 2\sum_{j=3}^{m}\alpha_1 y_1 \alpha_j y_j \kappa_{1j} \\
&+ \alpha_2 y_2 \alpha_2 y_2 \kappa_{22} + 2\sum_{j=3}^{m}\alpha_2 y_2 \alpha_j y_j \kappa_{2j} + \sum_{i=3}^{m}\sum_{j=3}^{m}\alpha_i y_i \alpha_j y_j \kappa_{ij}] \\
&- [\alpha_1 + \alpha_2 + \sum_{j=3}^{m}\alpha_j]
\end{aligned}
$$

去常数

进一步简化，其中 $\sum_{i=3}^{m}\sum_{j=3}^{m}\alpha_i y_i \alpha_j y_j \kappa_{ij}$ 和 $\sum_{j=3}^{m}\alpha_j$ 是常数，直接扔掉，不影响我们求导。

$$
\begin{aligned}
\therefore \text{原式} = \frac{1}{2}[&\alpha_1^2 \kappa_{11} + 2\alpha_1 y_1 \alpha_2 y_2 \kappa_{12} + 2\sum_{j=3}^{m}\alpha_1 y_1 \alpha_j y_j \kappa_{1j} \\
&+ \alpha_2^2 \kappa_{22} + 2\sum_{j=3}^{m}\alpha_2 y_2 \alpha_j y_j \kappa_{2j}] - [\alpha_1 + \alpha_2]
\end{aligned}
$$

用 α_2 表示 α_1 （核心）

$$
\because \sum_{i=1}^{m}\alpha_i y_i = 0 \implies \alpha_1 y_1 + \alpha_2 y_2 + \sum_{i=3}^{m}\alpha_i y_i = 0
$$

且 α_i $(i \geq 3)$ 是固定的，所以 $\alpha_1 y_1 + \alpha_2 y_2 - \varsigma = 0 \implies \alpha_1 y_1 + \alpha_2 y_2 = \varsigma$， （6.14） （6.16），这部分对应西瓜书P125，其中：

$$\varsigma=-\sum_{k\neq 1,2}\alpha_k y_k\text{，}\quad (6.15)$$

所以 $\alpha_1=y_1(\varsigma-\alpha_2 y_2)$，且 $y_1^2=1$（无论取 +1 还是 -1）

我们把 $\alpha_1=y_1(\varsigma-\alpha_2 y_2)$ 代入上边推了半天的原式，得：

$$\begin{aligned}\text{原式}=L(\alpha_2)=\frac{1}{2}&[(\varsigma-\alpha_2 y_2)^2\kappa_{11}+2(\varsigma-\alpha_2 y_2)\alpha_2 y_2\kappa_{12}\\&+2\sum_{j=3}^m(\varsigma-\alpha_2 y_2)\alpha_j y_j\kappa_{1j}+\alpha_2^2\kappa_{22}+2\sum_{j=3}^m\alpha_2 y_2\alpha_j y_j\kappa_{2j}]\\&-[y_1(\varsigma-\alpha_2 y_2)+\alpha_2]\end{aligned}$$

求导

还记得原来的要求么？求最小值 $\min_\alpha$，都一元二次还不好求极值么？真是的，直接求导！

$$\begin{aligned}\frac{\partial L}{\partial\alpha_2}=\frac{1}{2}&\Big[-2y_2(\varsigma-\alpha_2 y_2)\kappa_{11}+(2\varsigma y_2\kappa_{12}-4\alpha_2\kappa_{12})\\&-2\sum_{i=3}^m y_2\alpha_i y_i\kappa_{1i}+2\alpha_2\kappa_{22}+2\sum_{j=3}^m y_2\alpha_j y_j\kappa_{2j}\Big]\\&-[-y_1 y_2+1]\end{aligned}$$

展开，再化简，让其等于 0：

$$\begin{aligned}\frac{\partial L}{\partial\alpha_2}=&\,y_1 y_2-1-\varsigma y_2\kappa_{11}+\alpha_2\kappa_{11}\\&+\varsigma y_2\kappa_{12}-2\alpha_2\kappa_{12}+\alpha_2\kappa_{22}\\&-\sum_{i=3}^m y_2\alpha_i y_i\kappa_{1i}+\sum_{j=3}^m y_2\alpha_j y_j\kappa_{2j}=0\end{aligned}$$

移项得到：

$$\begin{aligned}\alpha_2(\kappa_{11}+\kappa_{22}-2\kappa_{12})&=1-y_1 y_2+\varsigma y_2\kappa_{11}-\varsigma y_2\kappa_{12}+\sum_{i=3}^N y_2\alpha_i y_i\kappa_{1i}-\sum_{j=3}^N y_2\alpha_j y_j\kappa_{2j}\\&=y_2\left[y_2-y_1+\varsigma\kappa_{11}-\varsigma\kappa_{12}+\sum_{i=3}^N\alpha_i y_i\kappa_{1i}-\sum_{i=3}^N\alpha_i y_i\kappa_{2i}\right]\ \text{（公式1）}\end{aligned}$$

转换超平面方程(6.12)，代入求导的式子

之后我们观察，还有什么可以换的？诶，还记得我们之前推导过的(6.12)，这玩意好像也可以展开：

$$
\begin{aligned}
f(x) &= \mathbf{w}^\top \mathbf{x} + b \\
&= (\sum_{i=1}^{m} \alpha_i y_i x_i^\top)\mathbf{x} + b \\
&= \alpha_1 y_1 x_1^T \cdot x_1 + \alpha_2 y_2 x_2^T \cdot x_1 + \sum_{i=3}^{m} \alpha_i y_i x_i^T \cdot x_1 + b
\end{aligned} \tag{6.12}
$$

移项得到：

$$
\sum_{i=3}^{m} \alpha_i y_i x_i^T \cdot x_1 = f(x) - (\alpha_1 y_1 x_1^T \cdot x_1 + \alpha_2 y_2 x_2^T \cdot x_1 - b)
$$

核代换 $x_i^T \cdot x_j = \kappa_{ij}$ ，于是得到：

$$
f(x_1) - \alpha_1 y_1 \kappa_{11} - \alpha_2 y_2 \kappa_{12} - b = \sum_{i=3}^{m} \alpha_i y_i \kappa_{1i} \text{ (公式2)}
$$

$$
f(x_2) - \alpha_1 y_1 \kappa_{12} - \alpha_2 y_2 \kappa_{22} - b = \sum_{i=3}^{m} \alpha_i y_i \kappa_{2i} \text{ (公式3)}
$$

我们把 公式2 和 公式3，代入到 公式1 中，得：

$$
\begin{aligned}
\alpha_2(\kappa_{11} + \kappa_{22} - 2\kappa_{12}) = y_2 [&y_2 - y_1 + \varsigma\kappa_{11} - \varsigma\kappa_{12} + f(x_1) \\
&- \alpha_1 y_1 \kappa_{11} - \alpha_2 y_2 \kappa_{12} - b - f(x_2) + \alpha_1 y_1 \kappa_{12} \\
&+ \alpha_2 y_2 \kappa_{22} + b] \text{ (公式4)}
\end{aligned}
$$

那这里有个常量 ς ， ς 等于多少呢？我们说 SMO 算法其实是个迭代的算法，我们会给 α_1 和 α_2 附上初始值为 α_1^{old} 和 α_2^{old} ，从前面的推导可知 $\alpha_1 y_1 + \alpha_2 y_2 = \varsigma$

所以：$\alpha_1^{old} y_1 + \alpha_2^{old} y_2 = \varsigma$ ，将其代入 公式4，得：

$$
\begin{aligned}
\alpha_2(\kappa_{11} + \kappa_{22} - 2\kappa_{12}) = y_2 \big[&y_2 - y_1 + \alpha_1^{old} y_1 \kappa_{11} - \alpha_2^{old} y_2 \kappa_{11} - \alpha_1^{old} y_1 \kappa_{12} - \alpha_2^{old} y_2 \kappa_{12} \\
&+ f(x_1) - \alpha_1^{old} y_1 \kappa_{11} - \alpha_2^{old} y_2 \kappa_{12} - b \\
&- f(x_2) + \alpha_1^{old} y_1 \kappa_{12} + \alpha_2^{old} y_2 \kappa_{22} + b\big] \text{(公式5)}
\end{aligned}
$$

化简，合并同类项，得 公式6：

$$
\begin{aligned}
\alpha_2^{new,unc}(\kappa_{11} + \kappa_{22} - 2\kappa_{12}) = y_2 &\big[(f(x_1) - y_1) - (f(x_2) - y_2)\big] \\
&+ \alpha_2^{old} y_2 (\kappa_{11} + \kappa_{22} - 2\kappa_{12})
\end{aligned}
$$

令：$E_1 = f(x_1) - y_1$，$E_2 = f(x_2) - y_2$，$\eta = \kappa_{11} + \kappa_{22} - 2\kappa_{12}$

最终迭代式

最终我们得到了这个 α 迭代式

$$\alpha_2^{new,unc}=\alpha_2^{old}+\frac{y_2(E_1-E_2)}{\eta}$$

但是还差一步，我们需要由 α 得到 b 。很简单，还记得我们的(6.6)的 $\text{s.t.} \quad y_i(\mathbf{w}^\top x_i+b)\geq 1$ ，我们讨论当且仅当支持向量的时候等号成立，所以有（这部分我们就回到了西瓜书）：

$$y_s\left(\sum_{i\in S}\alpha_i y_i \mathbf{x}_i^T\mathbf{x}_s+b\right)=1, \quad (6.17)$$

其中 $S=i|\alpha_i>0, i=1,2,...,m$ 为所有支持向量的下标集，直接将迭代求得的 α 带入就可以。理论上，可选取任意支持向量并通过求解式(6.17)获得 b ，但现实任务中常采用一种更鲁棒的做法：使用所有支持向量求解的平均值：

$$b=\frac{1}{|S|}\sum_{s\in S}\left(y_s-\sum_{i\in S}\alpha_i y_i\mathbf{x}_i^T\mathbf{x}_s\right), \quad (6.18)$$

2）“*怎么选择 α_1 和 α_2 ？*”

这里西瓜书给了我们答案（来自西瓜书）：直观来看，KKT 条件违背的程度越大，则变量更新后可能导致的目标函数值减幅越大，于是，SMO 先选取违背 KKT 条件程度最大的变量，第二个变量应选择一个使目标函数值减小最快的变量，但由于比较各变量所对应的目标函数值减幅的复杂度过高，因此 SMO 采用了一个启发式：

> 使选取的两变量所对应样本之间的间隔最大，一种直观的解释是，这样的两个变量有很大的差别，与对两个相似的变量进行更新相比，对它们进行更新会带给目标函数值更大的变化

详细的变量选取，可以参考上文给的知乎文章，或者参考李航老师《统计学习方法》P147