C1-6

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

2016.8.28

第七集：最优间隔分类器问题

- 最优间隔分类器(Optimal Margin Classifier)
- 主优化问题与对偶优化问题(Primal/Dual Optimization Problem)
- 支持向量机的对偶问题(SVM Dual)

第七集：最优间隔分类器问题

- 最优间隔分类器
  - 在上一讲中说到支持向量机要解决如下的优化问题
    - 在给定训练集的情况下找到对应的超平面对于数据及进行分类，同时使得最坏样本情况下的距离尽可能大

$$\begin{aligned} \max_{\gamma,w,b} \quad & \gamma \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq \gamma, \quad i = 1, \ldots, m \\ & ||w|| = 1. \end{aligned}$$

在这里使用了几何距离进行目标表述（实际上在||w||=1的前提下函数距离等价于几何距离）。第一个约束条件说明我们要使得所有样本的几何距离的最小值最大化（在这里面假定了数据集是线性可分的）

    - 然而对于一个优化问题而言，我们更加希望它是一个凸优化问题。对于一个凸优化问题来说需要满足：目标函数必须是凸的，不等式约束函数也必须是凸的，等式约束函数必须是仿射的
      - 在这个例子中目标函数和不等式约束函数都是线性函数，为凸函数；但是等式约束是模，模并不是仿射函数（因为不能表达成线性函数和常数的和），因此这个问题并不是凸优化问题，除非修改等式约束

第七集：最优间隔分类器问题

- 最优间隔分类器
  - 原问题中会出现等式约束的原因在于在不等式约束中我们采用的是函数距离的定义，而只有在等式约束||w||=1成立的时候该不等式约束才能成立。与此同时我们的优化目标是几何距离，这也就使得在设计约束时不能直接用函数距离这一条约束，而变成了等价的这两条约束
    - 综上可知，如果我们能将优化目标从几何距离变成等价的函数距离，那么将只剩下不等式约束，不再需要等式约束||w||=1。由此有了下面等价的优化问题

$$\max_{\gamma,w,b} \quad \frac{\hat{\gamma}}{||w||}$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

此时虽然解决掉了等式约束，但是此时的目标函数不是凸函数。因为很明显可以看出w的取值范围为R^n-{0}，这并不是一个凸集。因此原函数不为凸函数

第七集：最优间隔分类器问题

- 最优间隔分类器
  - 由此还得继续改写优化问题的表述形式。由于函数间隔γ采取的是如下的定义：

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$

    - 这也就使得函数间隔γ^是可以直接通过参数w, b进行表达的，因此在一个优化问题中同时出现γ^, w, b作为参数的优化是冗余的。由于我们最后的目标是得到参数w, b，因此我们完全可以固定住γ^优化w, b。而这最简单的方式就是将函数间隔γ^设为1.
      - 因此在目标函数中最大化γ^/||w||的问题变成了最大化1/||w||的问题，从而等价于最小化||w||的问题，因此可将原优化问题转换为如下形式的凸优化问题（通过这个凸优化问题对应的解w,b即可得到对应的最优间隔分类器(Optimal Margin Classifier)，其解可以通过二次函数优化软件得出）：

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

第七集：最优间隔分类器问题

- 主优化问题与对偶优化问题
  - 下面将讨论一般意义上的凸优化问题的求解方法，一般的优化问题均可以转换为如下的形式：

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

这里面的g(ω)、h(ω)只是表示它是一个约束条件而已，i只是为了说明这样的约束条件的序号，k、l只是表示约束条件的数目，约束条件相互之间(比如$g_1(x)$,$g_2(x)$,$h_1(x)$之间)并没有任何关联。这样的优化问题表达形式称为主优化问题(primal optimization problem)

第七集：最优间隔分类器问题

- 主优化问题与对偶优化问题
  - 对于上述的优化问题而言，如果采取逐个击破的办法必定顾头不顾尾，无法保证算法的准确性和效果
  - 因此我们需要将上面的目标函数和约束合成一个函数，之后再对这个函数用合理的手段进行优化，由此有了下面广义Lagrange算子(generalized Lagrangian)的表达形式

$$\mathcal{L}(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

  - 在这个式子中α,β均为Lagrange乘子，考虑下面的等式：

$$\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta)$$

  - 可得：

$$\theta_{\mathcal{P}}(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$$

因为只要有一项不等式或等式约束不满足，只需使得该项的系数α,β趋于无穷，对应的θp(ω)的值就会趋于无穷；如果全部满足，后两项的最大值均为0，对应的θp(ω)的值等于f(ω)

第七集：最优间隔分类器问题

- 主优化问题与对偶优化问题
  - 因此原主优化问题可做如下形式的等价转换

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \dots, k \\ & h_i(w) = 0, \quad i = 1, \dots, l. \end{aligned} \longrightarrow \min_w \theta_{\mathcal{P}}(w) = \min_w \max_{\alpha,\beta:\alpha_i \ge 0} \mathcal{L}(w, \alpha, \beta)$$

    - 对此我们可以将该主问题的最优解(value of the primal problem)记为p*
  - 与此同时我们发现对于这样的问题是难解的，因为其中α,β系数一般很多，求解起来非常复杂。由此我们考虑下面的形式：

$$\max_{\alpha,\beta:\alpha_i \ge 0} \theta_{\mathcal{D}}(\alpha, \beta) = \max_{\alpha,\beta:\alpha_i \ge 0} \min_w \mathcal{L}(w, \alpha, \beta)$$

    - 该问题被称为原优化问题的对偶优化问题(dual optimization problem)，并将其最优解记为d*

第七集：最优间隔分类器问题

- 主优化问题与对偶优化问题
  - 现在来探讨一下对偶优化问题的性质：
    - 性质：对偶优化问题是主优化问题的下界，即：

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \min_{w} \mathcal{L}(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta:\alpha_i\geq 0} \mathcal{L}(w,\alpha,\beta) = p^*$$

      - 证明：

假设ω为原优化问题的可行解，此时

$$\max_{\alpha,\beta:\alpha i\geq 0} \mathcal{L}(\omega,\alpha,\beta) = f(\omega)$$

因此要证

$$\mathrm{d}^* = \max_{\alpha,\beta:\alpha i\geq 0} \min_{\omega} \mathcal{L}(\omega,\alpha,\beta) \leq \min_{\omega} \max_{\alpha,\beta:\alpha i\geq 0} \mathcal{L}\left(\omega,\alpha,\beta\right) = \mathrm{p}^*$$

只需

$$\max_{\alpha,\beta:\alpha i\geq 0} \min_{\omega} \mathcal{L}(\omega,\alpha,\beta) \leq \min_{\omega} f(\omega)$$

由于

$$\mathcal{L}\left(\omega,\alpha,\beta\right) = f\left(\omega\right) + \sum_{i=1}^{k} \alpha i g i\left(\omega\right) + \sum_{i=1}^{l} \beta i h i\left(\omega\right) \leq f\left(\omega\right)$$

因此原命题成立

第七集：最优间隔分类器问题

- 主优化问题与对偶优化问题
  - 由于对偶优化问题是主优化问题的下界，而我们需要求解的是主优化问题的下界
    - 由于对偶优化问题相对好解，因此我们希望通过寻找对偶优化问题的上界来衡量主优化问题的下界；对此最好的结果是对偶优化问题的上界恰好就是主优化问题的下界，这称为强对偶性，即$d^*=p^*$
      - 在强对偶性下求解对偶优化问题等价于求解主优化问题
    - 事实证明，对于凸优化问题强对偶性通常成立（不绝对），而其他情况下一般不成立（同样不绝对）
      - 对于一个凸优化问题而言，只要满足Slater条件，强对偶性必然成立（充分不必要）
        - Slater条件：存在定义域上的一点ω，使得所有约束严格成立，即对于任意的i有：$g_i(\omega)<0$且$h_i(\omega)=0$

第七集：最优间隔分类器问题

- 主优化问题与对偶优化问题
  - 假设ω*为主优化问题的最优解，α*,β*为对偶优化问题的最优解，在强对偶性成立下有如下的性质：
    - 1、p*=d*=L(ω*,α*,β*)，从而求解对偶优化问题等价于求解主优化问题
    - 2、KKT条件(Karush-Kuhn-Tucker (KKT) conditions)

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,l \\
\alpha_i^* g_i(w^*) &= 0, \quad i=1,\ldots,k \\
g_i(w^*) &\leq 0, \quad i=1,\ldots,k \\
\alpha^* &\geq 0, \quad i=1,\ldots,k
\end{aligned}
$$

从而求解主优化问题等价于求解KKT条件，从KKT条件中可以看到原主优化问题的最优解ω*可以通过L(ω,α,β)求导，令其导数为0用α*,β*表达出来；原对偶优化问题的最优解β*也可以通过L(ω,α,β)求导，令其导数为0代入ω*的表达式用α*表达出来。从而通过前两个式子可以成功地把ω*,β*用α*表达出来。最后将α*代入对偶优化问题求解，即可得到α*的正确结果，进而可以得到ω*,β*的结果

而单纯从第三个式子可以看出，两个数的乘积为0，因此对于任意的i，αi*和gi(ω*)中至少一项为0（在一般情况下求解时一般令后者为0（不绝对），从而第四个式子gi(ω*)≤0成为了活动约束(active constraint)，可直接忽略，简化运算）。

第七集：最优间隔分类器问题

- 支持向量机的对偶问题
  - 下面将用上述介绍的优化问题的解法来解决之前讲到的最优间隔分类器的优化问题

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m \end{aligned}$$

    - 首先我们需要对不等式约束进行如下的修改从而将这个问题变为标准形式：

$$g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0.$$

      - 由此即可通过KKT条件对问题进行求解

第七集：最优间隔分类器问题

- 支持向量机的对偶问题
  - 接下来使用KKT条件对原问题进行求解
    - 求解第一步：写出广义Lagrange算子的表达形式

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i \left[y^{(i)}(w^T x^{(i)} + b) - 1 \right]$$

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

      - 由于此处没有等式约束，因此这里只有参数ω和α，并没有β。与此同时，还可以观察到在这个例子中参数ω实际上由超平面斜率ω和截距项b组成
    - 求解第二步：对于参数ω进行微分
      - 在这里参数ω实际上由超平面斜率ω和截距项b组成，因此需要对ω和b分别求偏微分

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

$$\frac{\partial}{\partial b}\mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

This implies that

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}.$$

第七集：最优间隔分类器问题

- 支持向量机的对偶问题
  - 接下来使用KKT条件对原问题进行求解
    - 求解第三步：对于参数β进行微分
      - 由于此处没有β，故省略这步
    - 求解第四步：用参数α替换原式中参数ω和β

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,l \\
\alpha_i^* g_i(w^*) &= 0, \quad i=1,\ldots,k \\
g_i(w^*) &\le 0, \quad i=1,\ldots,k \\
\alpha^* &\ge 0, \quad i=1,\ldots,k
\end{aligned}$$

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^T x^{(j)} - b\sum_{i=1}^{m}\alpha_i y^{(i)}$$

- 根据之前对于参数b求偏导的结果可进行进一步化简 $\sum_{i=1}^{m}\alpha_i y^{(i)} = 0$

$$= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^T x^{(j)}$$

第七集：最优间隔分类器问题

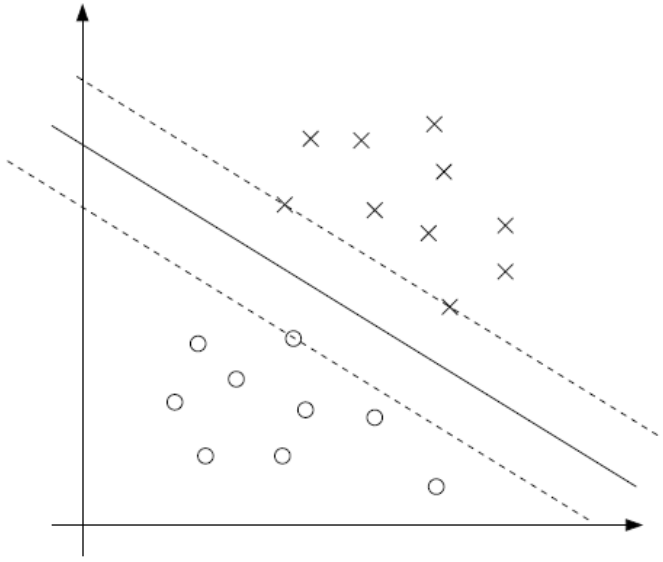
- 支持向量机的对偶问题
  - 接下来使用KKT条件对原问题进行求解
    - 求解第五步：写出原问题等价的对偶优化问题

$$\begin{aligned}
\max_\alpha \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle \\
\text{s.t.} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0,
\end{aligned}$$

尖括号表示内积

$$\begin{aligned}
\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, n \\
\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) &= 0, \quad i = 1, \ldots, l \\
\alpha_i^* g_i(w^*) &= 0, \quad i = 1, \ldots, k \\
g_i(w^*) &\leq 0, \quad i = 1, \ldots, k \\
\alpha^* &\geq 0, \quad i = 1, \ldots, k
\end{aligned}$$

在这里采用一般做法直接令gi(ω)=0从而可以直接满足第三、四个KKT条件，也就是直接令不等式约束为0。对应到这个例子中，不等式约束的物理意义是选择一个能将样本点完全分类的超平面且与样本点的函数距离至少为1（起初假定了样本点是线性可分的）。因此不等式约束为0等价于选择了恰好能将样本点完全分类的超平面（如图中实线所示），这个超平面与最近样本点的函数距离恰好为1（如图中虚线所示，这些边缘点也被称为支持向量(support vector)，因为只有在这些样本点上参数α的值不为0，其余样本点参数α的值为0 ）



第七集：最优间隔分类器问题

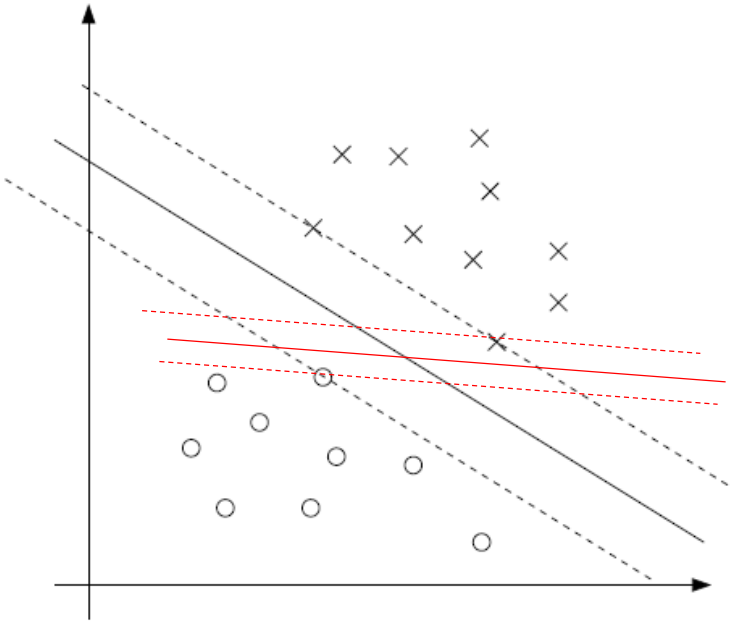
- 支持向量机的对偶问题
  - 从而最终推出了支持向量机等价的对偶优化问题的表达形式
    - 而这个问题是相对好解的，可以采用之后讲到的SMO算法进行求解

$$\max_\alpha \quad W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0,$$

尖括号表示内积



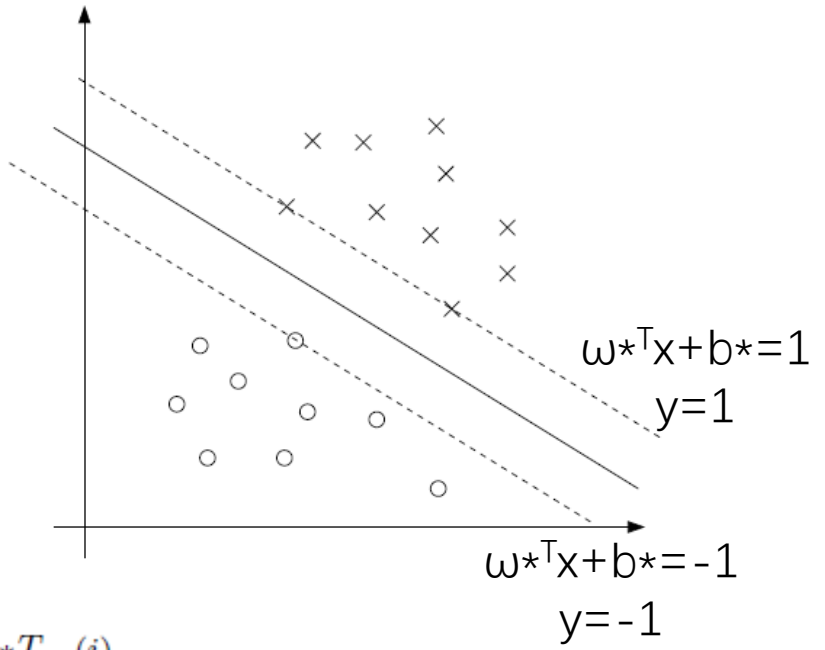
而上面的优化问题的意义在于在找到了如此多的与最近样本点的函数距离恰好为1且能完美分类的所有超平面中，最好的超平面是什么。因为我们知道函数距离可以通过等比例缩放参数ω和b得到，这也就相当于可以对右图进行等比例大小缩放使得函数距离恰好为1（比如说红色实线虽然目前与最近样本点的函数距离小于1，但是如果将图进行等比例放大，函数距离将不断增大，总能等于1）。那么我们自然希望寻找到一条在最小的放缩比例下就可以得到与最近样本点的函数距离为1的直线，这也就是本对偶优化问题的意义所在

第七集：最优间隔分类器问题

- 支持向量机的对偶问题
  - 解决上述的对偶优化问题后即可得到所有的最优α值（记为α*），从而可以得到最优超平面的参数ω*和b*
    - ω*可以直接通过右式得到 $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$
    - b*没有直接的关系式，需要另行考虑

根据之前所述可知图中实线所示的最优超平面与两个类的最近样本点的函数距离均为1。根据定义又知其中一个类的类标y=1，其对应的虚线所示的超平面上样本x满足$\omega^{*T}x+b^{*}=1$，而其他属于该类标的样本x则$\omega^{*T}x+b^{*}>1$，由此可见在类标为1的情况下对应的虚线所示的超平面上样本x可使得$\omega^{*T}x$取最小值 ；而另一个类的类标y=-1，其对应的虚线所示的超平面满足$\omega^{*T}x+b=-1$，而其他属于该类标的样本x则$\omega^{*T}x+b^{*}<-1$，由此可见在类标为-1的情况下对应的虚线所示的超平面上样本x可使得$\omega^{*T}x$取最大值.而实线所示的最优超平面的截距为两虚线所示的超平面截距的均值，因此可得最优超平面的截距b*的表达式为：

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T}x^{(i)} + \min_{i:y^{(i)}=1} w^{*T}x^{(i)}}{2}$$



第七集：最优间隔分类器问题

- 支持向量机的对偶问题
  - 对于一个新的测试数据来说，可以根据下面的方法进行预测：
    - 实际原理非常简单，就是判断该测试数据对应的点在最优超平面的哪一侧，方法是判断下式大于0还是小于0

$$
\begin{aligned}
w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T x + b \\
&= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b.
\end{aligned}
$$

      - 从这个测试方法中可以看到，每一个训练样本xi都会拥有一个系数αi。但根据前面所述可知，除了作为支持向量的样本点系数αi很可能不为0外，其余样本点的系数αi均为0，因此极大地简化了运算
        - 因为要满足KKT条件中的αi*gi(ω*)=0，而除支持向量外其余点的gi(ω*)不为0