4.6 最优分类超平面与支持向量机

只要一个样本集线性可分，就存在无数的解，哪一个解更好呢？最优超平面。直接求出参数的解析解。

定义：一个超平面，如果它能够将训练样本没有错误地分开，并且两类训练样本中离超平面最近的样本与超平面之间的距离是最大的，则把这个超平面称作最优分类超平面(optimal seperating hyperplane)，简称最优超平面(optimal hyperplane)。两类样本中离分类面最近的样本到分类面的距离称作分类间隔(margin)，最优超平面也称作最大间隔超平面，如图 4-9 所示。

支持向量机是什么意思？如何解读？

support vector machine

分类器三要素：

1. 判别函数

$$g(x_i) = w^T x_i + b \gtrless 0 \quad \begin{cases} y_i = +1 \\ y_i = -1 \end{cases}$$

内积 $(w \cdot x)$ is $w^T x$

2. 不同的准则

$$\min_{w,b} \frac{1}{2} \|w\|^2$$

$$\text{s.t.} \quad y_i [(w \cdot x_i) + b] - 1 \geq 0$$

$$i = 1, 2 \ldots N$$

3. 优化

拉格朗日泛涵

准则怎么来的？

$$g(x) = \text{sgn}(g(x))$$

$$g(x_i) = w^T x_i + b$$

对w,b作任何正的尺度调整，都不会改变分类面

$$k g(x_i) = k w^T x_i + k b \quad , \quad k > 0$$

$$g(x_i) = 0 \Rightarrow k g(x_i) = 0$$

为了使这个问题有唯一解，需要把w,b的尺度确定下来

$$\begin{cases} w^T x_i + b > 0, & y_i = +1 \\ w^T x_i + b < 0, & y_i = -1 \end{cases}$$

$$y_i(w^T x_i + b) > 0$$

此式只给出了分类面，但无法刻画两个集合之间最大的间隔距离，我们希望刻画两个集合之间的间隔，而不仅仅是中间分类面

对w和b进行尺度缩放，使得

$$\begin{cases} k w^T x_i + kb \geq \varepsilon & y_i = +1 \\ k w^T x_i + kb \leq -\varepsilon & y_i = -1 \end{cases}$$

引入余量

取 $\varepsilon=1$，（取2也行）

$$y_i(w^T x_i + b) \geq 1, \quad i=1,\dots N$$

$$g(x_i)=1$$
$$g(x_i)=-1$$

叫做两个边界超平面，不是分类面

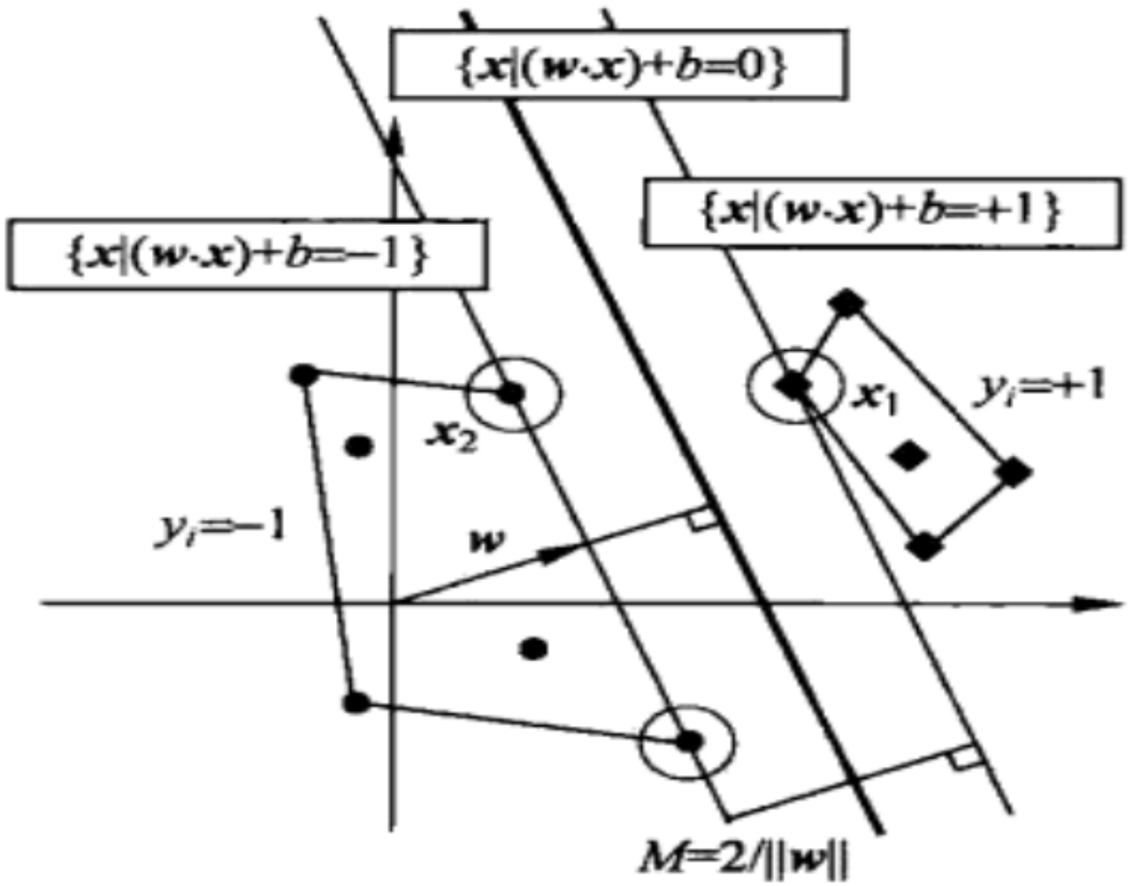

图 4-10 规范化的最优分类面

两个边界超平面的距离是 $\frac{2}{\|w\|}$

高数平行直线距离公式

我们希望离边界超平面越远越好，也就是你的分类面在哪个方向，才能使得边界超平面距离最远，最能将两类分得开。边界超平面平行于分类面。

分类面、边界超平面？

边界超平面距离越大，则是||w||越小

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$s.t.\ y_i(w^T x_i + b) - 1 \geq 0$$

这是一个在不等式约束下的优化问题，可以通过拉格朗日法求解。对每个样本引入一个拉格朗日系数

$$\alpha_i \geqslant 0, \quad i=1,\cdots,N \tag{4-74}$$

每个样本对应一个α_i可看作样本的权重

拉格朗日法求解上述优化问题

$$\min_{w,b} \max_{\alpha} L(w,b,\alpha) = \frac{1}{2}(w \cdot w) - \sum_{i=1}^{N} \alpha_i \{[y_i(w \cdot x_i) + b] - 1\}$$

最优解在鞍点

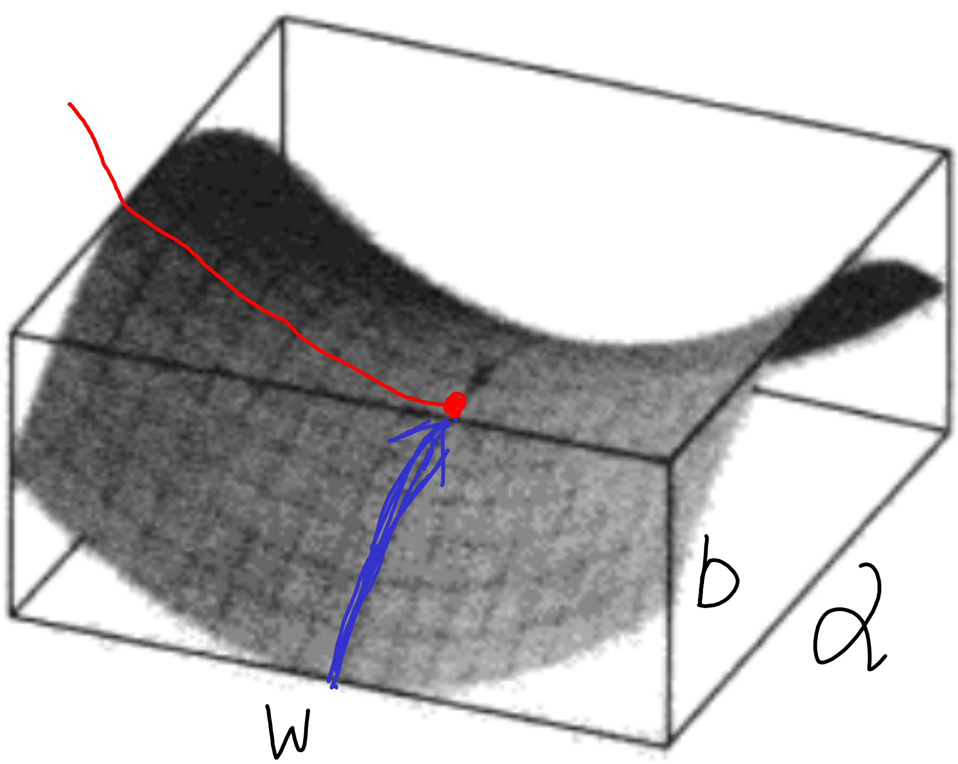

图 4-11 鞍点示意图

$$\frac{\partial L}{\partial w} = 0 \quad , \quad \frac{\partial L}{\partial b} = 0$$

$$\frac{\partial L}{\partial w} = w - \sum a_i y_i x_i = 0$$

$$w^* = \sum a_i^* y_i x_i$$

$$\frac{\partial L}{\partial b} = \sum y_i a_i^* = 0$$

代入准则(4-72)

$$L = \frac{1}{2}(\alpha_1 y_1 x_1 + \alpha_2 y_2 x_2 \ldots)^T(\alpha_1 y_1 x_1 + \alpha_2 y_2 x_2 \ldots) - \sum \alpha_i y_i w^T x_i$$

$$- \sum \alpha_i y_i b + \sum \alpha_i$$

$$\because \sum y_i \alpha_i^* = 0$$

n^2 项

$$= \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$- \sum \alpha_i y_i w^T x_i + \sum \alpha_i$$

$$w^* = \sum \alpha_j^* y_j x_j$$

代入

最优超平面的权重向量为所有样本的加权组合

$$= \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$- \sum \alpha_i y_i \left(\sum \alpha_j y_j x_j\right)^T x_i + \sum \alpha_i$$

$$\boxed{\begin{aligned} &\sum \alpha_i y_i \left(\sum \alpha_j y_j x_j\right)^T x_i \\ &= \sum \alpha_i \alpha_j y_i y_j x_i x_j \end{aligned}}$$

$$= -\frac{1}{2}\sum \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum \alpha_i$$

得到最优超平面的对偶问题

$$\max_{\alpha} \quad Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

$$\text{s. t.} \quad \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$\alpha_i \geqslant 0, \quad i = 1, \cdots, N$$

通过求解对偶问题α_i的解，进而求出原问题，w的解。

$$\frac{\partial Q(\alpha)}{\partial \alpha_i} = 1 - \sum_{j \neq i} \alpha_j y_i y_j x_i^T x_j$$

$$+ \alpha_i y_i y_i x_i^T x_i = 0$$

求出的 α_i^* 代入 $w = \sum \alpha_i^* y_i x_i$

如何得来

n^2项

$$\frac{\partial \sum \alpha_i \alpha_j}{\partial \alpha_i} = \frac{\partial(\alpha_1\alpha_1 + \alpha_1\alpha_2 \cdots)}{\partial \alpha_i}$$

分子与α_i有关的有 α_i^2及 $\alpha_i\alpha_j$

和$\alpha_j\alpha_i$，$i \neq j$

求导之后 $2\alpha_i + 2\sum_{i\neq j} \alpha_j$

求出了w还未求b

KKT

根据最优化理论中的库恩-塔克(Kuhn-Tucker)条件，式(4-75)中的拉格朗日泛函的鞍点处满足

$$\alpha_i\{y_i[(w \cdot x_i)+b]-1\}=0, \quad i=1,2,\cdots,N \tag{4-83}$$

$$\alpha_i\{y_i[(w\cdot x_i)+b]-1\}=0,\quad i=1,2,\cdots,N$$

$$\min_{w,b}\frac{1}{2}\|w\|^2$$

$$\text{s.t.}\quad y_i[(w\cdot x_i)+b]-1\geqslant 0,\quad i=1,2,\cdots,N$$

都要满足，只有当后面这块为0, α_i 才会>0，否则如果后面这块>0,成立则 α_i 就得=0

后面这块=0,则就是图示上的两条边界超平面，此时的x_i都是落在边界超平面的样本

$$w^*=\sum\alpha_i^* y_i x_i$$

只有那些 $\alpha_i>0$ 的样本才参与运算

对于这些支持向量来说，有

$$y_i[(w^*\cdot x_i)+b^*]-1=0 \tag{4-84}$$

因为已经求出了 w^*，所以 b^* 可以用任何一个支持向量根据式(4-84)的方程求得。在实际的数值计算中，人们通常采用所有 α_i 非零的样本用式(4-84)求解 b^* 后再取平均。

对比 4.4 节中的感知器算法，我们也可以把最优超平面等价地看作是在限制权值尺度的条件下求余量的最大化。感兴趣的读者可以自己尝试分析这一关系。