# 第六章支持向量机chap6周志华chap7 bishop

王星

中国人民大学统计学院
E-mail:wangxingwisdom@126.com

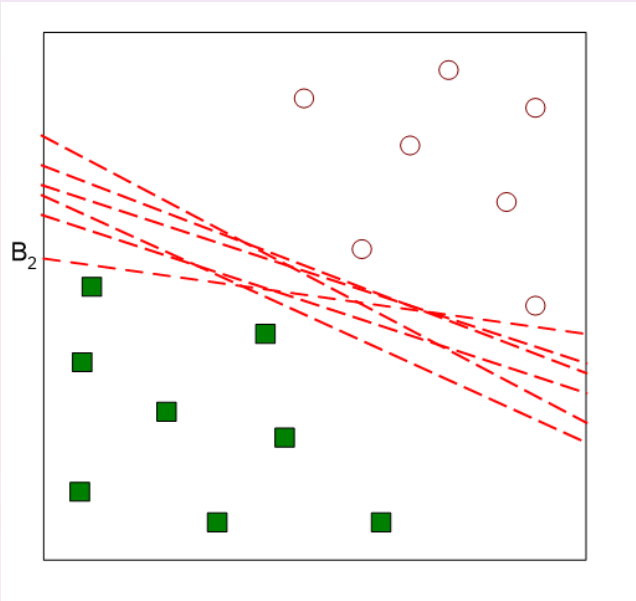
June 4, 2021

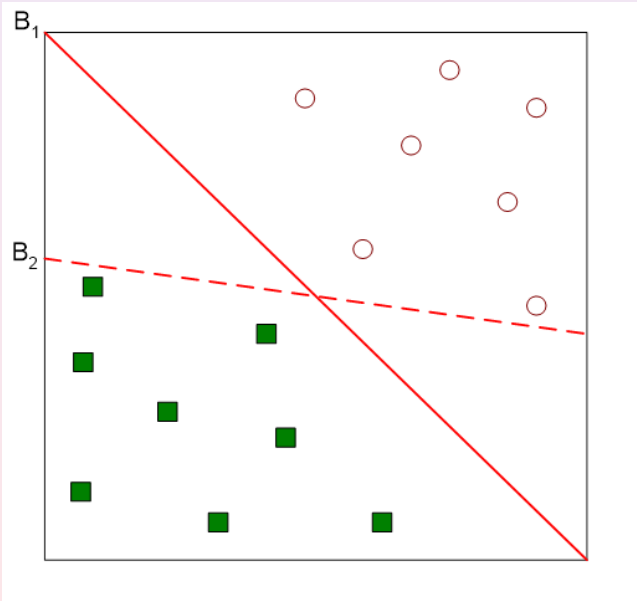
# CONTENT

- SVM
  - 间隔与支持向量机(ZZH6.1)
  - 对偶问题(ZZH6.2)
  - KKT条件(附录B)
  - 核函数(ZZH6.3)(BS 6.1-6.2)
  - SVMs for linearly separable data(BS7.1)
  - 软间隔与正则化(ZZH 6.4)
  - 松弛变量Slack variable (7.1.1)
  - 与逻辑回归的关系Relation to logistic regression (7.1.2)
  - 多类Multiclass SVMs* (7.1.3)
  - SVMs for regression(SVR) (6.5)(7.1.4)
- RVM*
  - RVM for regression (7.2.1)
  - Analysis of sparsity (7.2.2)
  - RVM for classification (7.2.3)

# 间隔与支持向量1/4

给出训练样本$\{(x_1, y_1), \cdots, (x_m, y_m)\}, y_i \in \{-1, +1\}$,分类学习最基本的想法就是基于训练数据集$D$在样本空间中找到一个划分超平面，但发现可以用好多个超平面(hyper plane) 将这些不同的点分开。但是哪个超平面是最合适的呢？从直觉上来看，我们似乎应该将红线作为我们的划分平面。因为它看起来“容忍度”更高，换句话这个超平面的鲁棒性最强



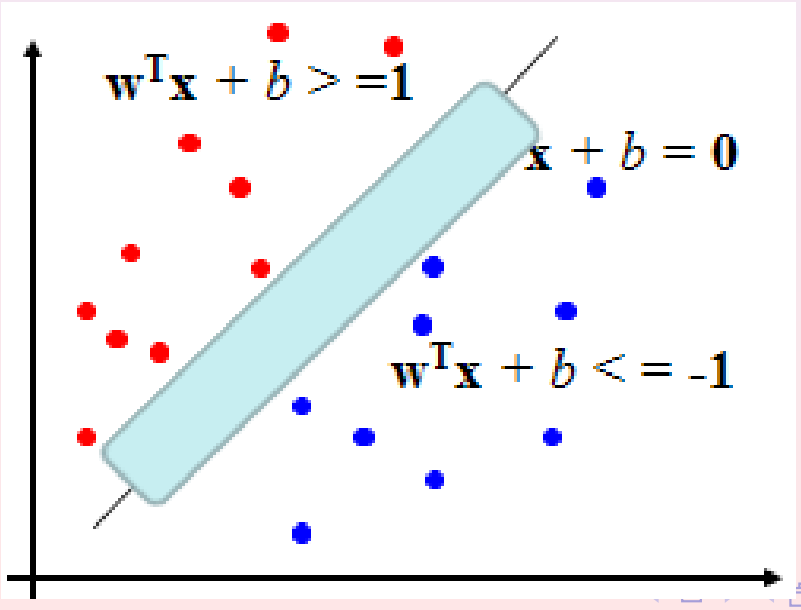



# 分类间隔与支持向量2/4

- 样本空间任意一点$x$到超平面(w,b)的距离可表达为：

$$r = \frac{|w^T x + b|}{\|w\|};$$

- 感知器是1956年Frank Rosenblatt提出的迭代学习算法特点：“在线”“错误驱动”程序
- Binary classification can be viewed as the task of separating classes in feature space: 所有的训练数据都分布在中间管道的一侧



# 分类间隔与支持向量4/4

- 让决策面上的点正好等于1或-1
- 找出一个能划分类别的超平面，使分类间隔(margin)最大

$$\max_{w,b} \frac{2}{\|w\|}, \quad s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, \cdots, m.$$

- 换成另外一种形式（支持向量机的基本型）：

$$\min_{w,b} \frac{1}{2}\|w\|^2, \quad s.t. y_i(w^T x_i + b) \geq 1, i \geq 1, i = 1, 2, \cdots, m.$$

SVM分类问题大致有三种：

- 线性可分问题;
- 近似线性可分问题;
- 线性不可分问题;

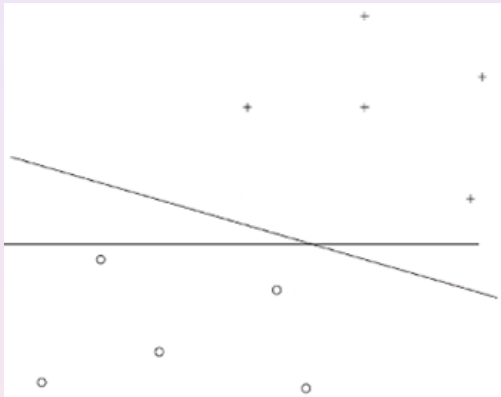


图 1.1.2

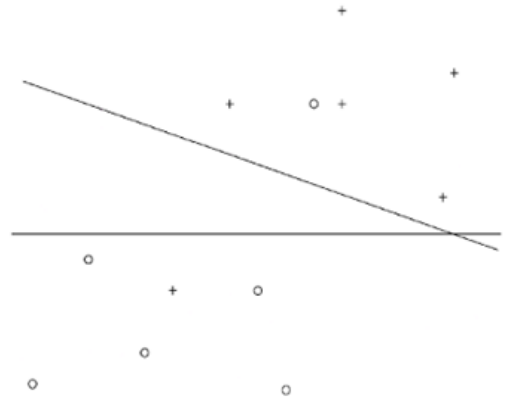


图 1.1.3

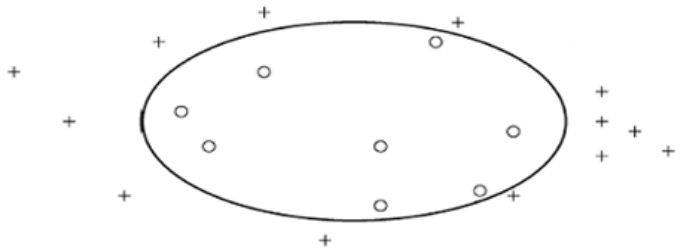


图 1.1.4

# 对偶问题

- 对基本型问题应用拉格朗日乘子法得到其“对偶问题”，求解对偶问题更高效。基本型的拉格朗日函数为

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(w^Tx_i+b))$$

- 令$L$对$w$和$b$偏导为0，可得：

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i;$$

$$0=\sum_{i=1}^{m}\alpha_i y_i$$

# 对偶问题的解

代入$L$得对偶问题为

$$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, i = 1, 2, \cdots, m.$$

上述过程需满足KKT条件
解出$\alpha$后，求出$w$和$b$,得到模型：

$$f(x) = w^T x + b$$

$$= \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$$

# KKT条件(Karash-Kuhn-Tucker)

在求解有约束条件的优化问题时，可以应用拉格朗日乘子法去求取最优值；如果含有不等式约束，可以应用KKT 条件去求取。对于含有不等式约束的优化问题，把所有的不等式约束、等式约束和目标函数全部写为一个式子

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

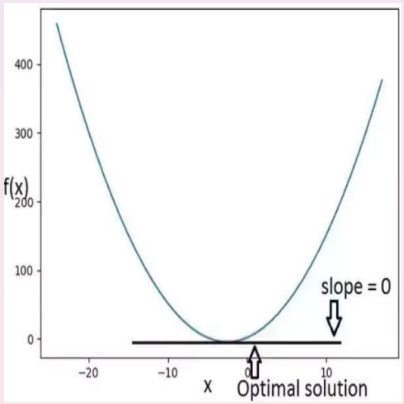
对任意训练样本$(x_i, y_i)$，总有

- $\alpha_i = 0$或$y_i f(x_i) = 1$.若$\alpha_i = 0$,就不会对$f(x)$有影响;
- $\alpha_i > 0$或$y_i f(x_i) \equiv 1$.所对应的样本点位于最大间隔边界上，是一个支持向量。

# 补充：库恩塔克条件的理解

KKT最优化条件是Karush[1939]，以及Kuhn和Tucker[1951]先后独立发表出来的。这组最优化条件在Kuhn 和Tucker发表之后才逐渐受到重视，因此许多情况下只记载成库恩塔克条件（Kuhn-Tucker conditions)

- 库恩塔克条件(Kuhn-Tucker conditions)是非线性规划领域里最重要的理论成果之一，是确定某点为极值点的必要条件。如果所讨论的规划是凸规划，那么库恩-塔克条件也是充分条件。
- 回顾【无约束优化问题的极值】（函数的最大值/最小值）通常发生在斜率为零的点上。




- 为了找到极值，只需要搜索斜率为零的点:如果$x^*$是无约束优化问题的极值，那么：

$$\nabla f(x^*) = 0$$

# 约束的极值优化问题

- 等式约束的极值优化问题

$$\min f(x)$$
$$s.t. g(x) = 0$$

- 如果$x^*$是上述最优化问题的极值，那么

$$\nabla f(x^*) = \lambda \nabla g(x^*), g(x^*) = 0$$

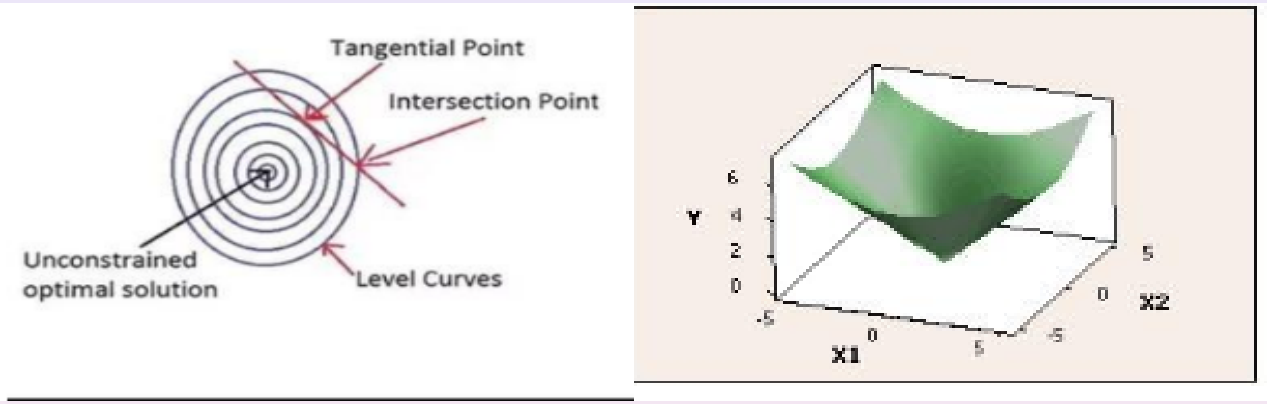
- 不等式约束的极值优化问题

$$\min f(x)$$
$$s.t. g(x) \leq 0$$

- 如果$x^*$是不等式约束优化问题的极值，那么KKT条件
  - 1.原可行性：$g(x^*) \leq 0$;
  - 2.对偶可行性：$\alpha \geq 0$;
  - 3.互补松弛条件：$\alpha g(x^*) = 0$;
  - 4.拉格朗日平稳性：$\nabla f(x^*) = \alpha \nabla g(x^*)$

# 等式约束下的KKT条件的直观理解

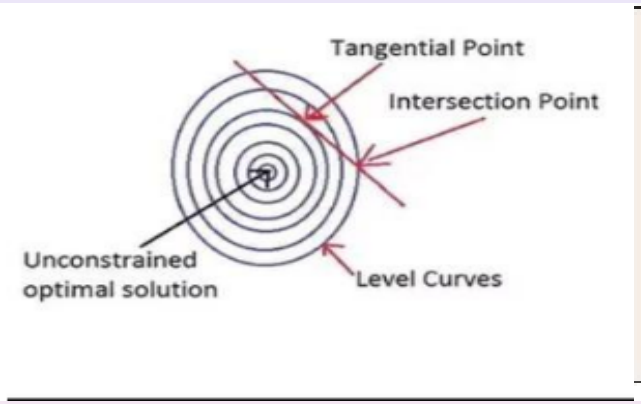
- 拉格朗日平稳性：图中显示了具有等式约束的优化问题的等高线图（它是通过绘制2D格式上的目标函数值的常量切片来表示3D表面的图）

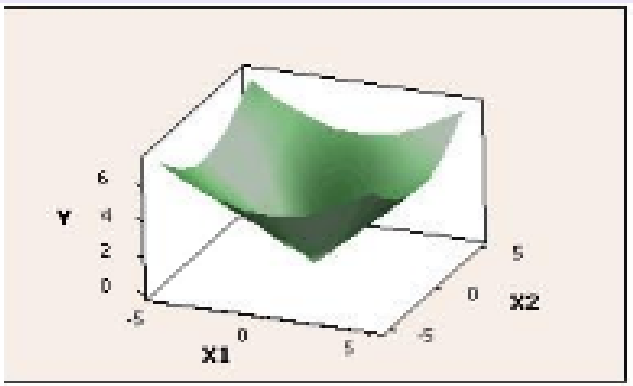



- 等高线图推断出最优问题只有两类可行点：1.切点; 2.交点。
切点是水平曲线（等高线）和约束线彼此相切的点;交点是水平曲线和约束线相交的点。
- 在等高线图中，如果从交叉点（沿约束线）向左移动，则目标函数值会增加。它表明该问题在这方面有改进的余地。同样，如果从交叉点向右移动，目标函数值会减小。但是，对于上图中的切线点，从切线点向右或向左移动只会降低目标函数值。这意味着约束优化问题的极值总是落在切点上。

# 等式约束下的KKT条件的直观理解

- 拉格朗日平稳性：图中显示了具有等式约束的优化问题的等高线图（它是通过绘制2D格式上的目标函数值的常量切片来表示3D表面的图）

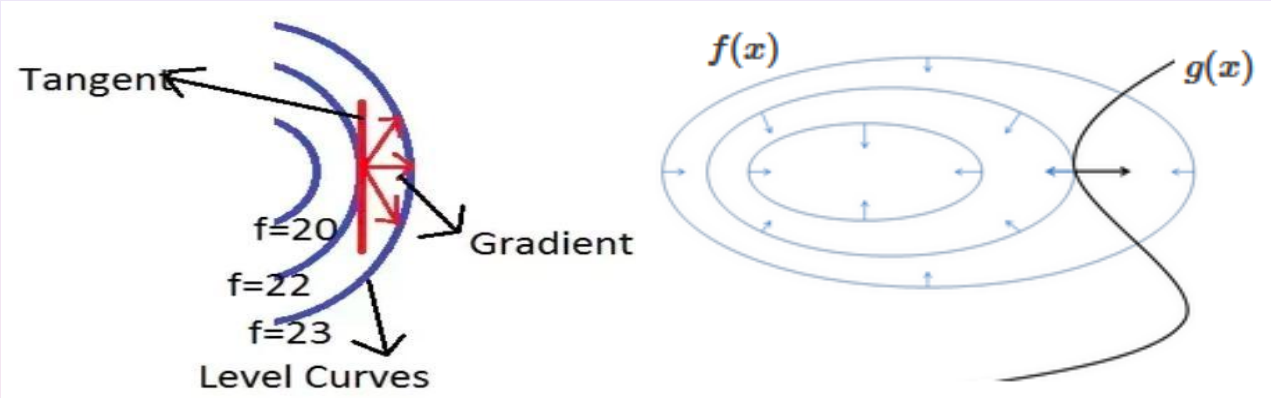



- 等高线图推断出最优问题只有两类可行点：1.切点; 2.交点。
切点是水平曲线（等高线）和约束线彼此相切的点;交点是水平曲线和约束线相交的点。
- 在等高线图中，如果从交叉点（沿约束线）向左移动，则目标函数值会增加。它表明该问题在这方面有改进的余地。同样，如果从交叉点向右移动，目标函数值会减小。但是，对于上图中的切线点，从切线点向右或向左移动只会降低目标函数值。这意味着约束优化问题的极值总是落在切点上。
- 结论1：约束优化问题的极值总是发生在切点上。

# 等式约束下的KKT条件的直观理解

- 函数的梯度指向函数增加最大的方向。在上面的等高线图中，从一个水平曲线移动到另一个水平曲线的最短路径是垂直方向，水平曲线与函数中没有立即变化的方向相切。这意味着任何点处函数的梯度都垂直于该点的函数水平曲线。

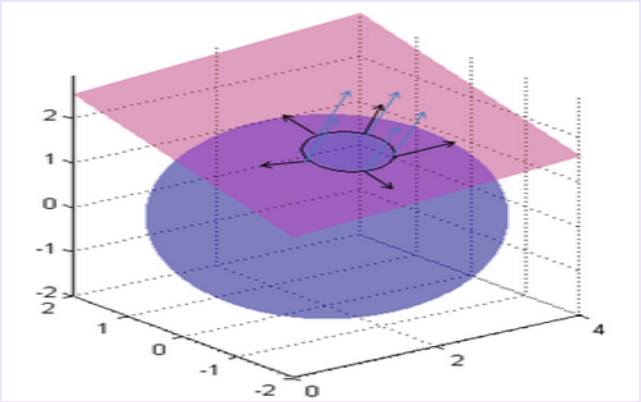



- 结论2：函数的梯度和函数的水平曲线的相切是正交的
- 通过结合结论1和2，可以得出结论，约束的梯度（$\nabla g$）和目标函数的梯度（$\nabla f$）在极值处（切线点）方向是相同或者相反的,表达如下：

$$\nabla f(x^*) = \alpha \nabla g(x^*)$$

其中$\alpha$可以等于0，表示$f(x)$本身的极值点刚好在切点上。除此之外，满足条件的极值点还满足等式方程$g(x) = 0$联立，解出这个方程组，就可以得到问题的解析解了。

# 多等式约束下的KKT条件的直观理解




- 图中的平面和球面分别代表了两个约束$g_1(x)$和$g_2(x)$，那么这个问题的可行域就是它们相交的圆。其中蓝色箭头表示平面的梯度，黑色箭头表示球面的梯度，相交的圆的梯度就是二者的线性组合，极值点的地方目标函数的梯度和约束的梯度的线性组合在一条直线上.对于$m, m \geq 2$ 个等式约束的情形，满足如下式子：

$$\nabla f(x) = \sum_{i=1}^{m} \alpha_i \nabla g_i(x)$$

$$g_1(x) = 0 \quad \cdots$$

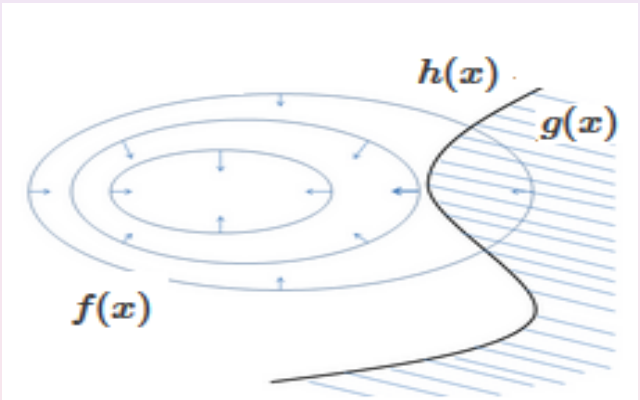
$$g_m(x) = 0$$

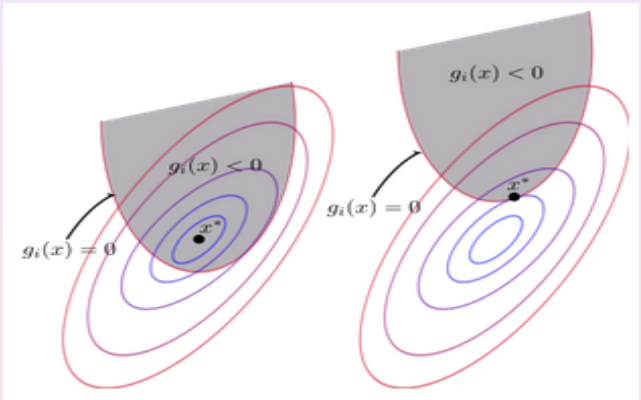
# 不等式约束下的KKT条件的直观理解

如果问题中既有等式约束，又有不等式约束怎么办呢？

$$\min f(x)$$
$$s.t. h(x) = 0$$
$$g(x) \leq 0$$

等式约束下的目标函数如图所示：

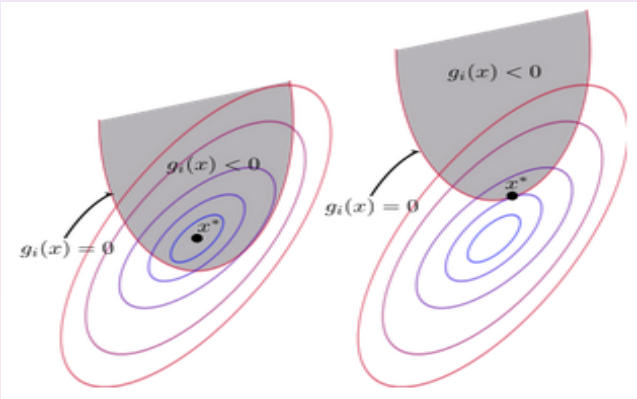



- 阴影部分就是可行域，也就是说可行域从原来的一条线变成了一块区域。那么能取到极值点的地方可能有两种情况：
  1.$f(x)$的极值点本身就在可行域里面(如右图)
  2. 极值点是在$h(x)$和等值线相切的地方

- 对于第一种情况，不等式约束就相当于没有，对$f(x)+\lambda h(x)$用拉格朗日乘子法：

$$\begin{aligned}\bigtriangledown f(x)+\lambda \bigtriangledown h(x)&=0\\ h(x)&=0\\ g(x)&\leq 0\end{aligned}$$




- 对于第二种情况，如果不是相切，那么同样的，对任意一个在可行域中的点，如果在它附近往里走或者往外走，$f(x)$一般都会变大或者变小，所以绝大部分点都不会是极值点,除非这个点刚好在交界处，且和等值线相切；或者这个点在可行域内部，但是本身就是$f(x)$的极值点。

- 对于第二种情况，不等式约束就变成等式约束了，
对$f(x)+\lambda h(x)+\mu g(x)$ 用拉格朗日乘子法：

$$\begin{aligned}\nabla f(x) &= \lambda \nabla h(x) + \mu \nabla g(x)\\ h(x) &= 0\\ g(x) &= 0\\ \mu &\geq 0\end{aligned}$$

为什么不是$\mu \neq 0$ 而是$\mu \geq 0$。后面的约束比前面的更强。
看“不等式约束”那张图，已经知道了问题中的可行域是在$g(x) \leq 0$的一侧，而$g(s)$的梯度是指向大于0 的一侧，也就是不是可行域的一侧。而求的问题是极小值，所以$f(x)$在交点处的梯度是指向可行域的一侧，也就是说两个梯度一定是相反的。所以也就可以确定这里的系数一定是大于0 的。而等式约束由于不知道$h(x)$的梯度方向，所以对它没有约束，那么为什么$\mu$还能等于0 呢，因为极值点可能刚好在$g(s)$上。

# KKT条件

最好把两种情况用同一组方程表示出来。对比一下两个问题，不同的是第一种情况$\mu = 0$ 且$g(x) \leq 0$, 第二种情况中有$\mu \geq 0$ 且$g(x) = 0$ 。综合两种情况，可以写成$\mu g(x) = 0$ 且$\mu \geq 0$ 且$g(x) \leq 0$:

$$\begin{aligned} \bigtriangledown f(x) + \lambda \bigtriangledown h(x) + \mu \bigtriangledown g(x) &= 0 \\ \mu g(x) &= 0 \\ \mu &\geq 0 \\ h(x) &= 0 \\ g(x) &\leq 0 \end{aligned}$$

它的含义是这个优化问题的极值点一定满足这组方程组。（不是极值点也可能会满足，但是不会存在某个极值点不满足的情况）它也是原来的优化问题取得极值的必要条件，解出来了极值点之后还是要代入验证的。

# Lagrangian Duality拉氏对偶问题解法

- The Primal Problem

Primal:

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

Non-slack

slack

可行性区域

The generalized Lagrangian:

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

the $\alpha$'s ($\alpha_i \ge 0$) and $\beta$s are called the Lagarangian multipliers

Lemma:

$$\max_{\alpha, \beta, \alpha_i \ge 0} \mathcal{L}(w, \alpha, \beta) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{o/w} \end{cases}$$

Primal: Problem

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

# Primal Problem, Duality

- Recall the Primal Problem:

$$\max_{\alpha,\beta,\alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta)$$

- The Dual Problem:

$$\min_w \max_{\alpha,\beta,\alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

- **Theorem (weak duality):**

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

- **Theorem (strong duality):**

  Iff there exist a saddle point of $\mathcal{L}(w, \alpha, \beta)$, we have

$$d^* = p^*$$

# KKT

- If there exists some saddle point of $\mathcal{L}$, then the saddle point satisfies the following "Karush-Kuhn-Tucker" (KKT) conditions:

$$\frac{\partial}{\partial w_i}\mathcal{L}(w,\alpha,\beta)=0, \quad i=1,\ldots,n$$

$$\frac{\partial}{\partial \beta_i}\mathcal{L}(w,\alpha,\beta)=0, \quad i=1,\ldots,l$$

$$\alpha_i g_i(w)=0, \quad i=1,\ldots,k$$

$$g_i(w)\le 0, \quad i=1,\ldots,k$$

$$\alpha_i \ge 0, \quad i=1,\ldots,k$$

Karush-Kuhn-tucker
互补条件

- **Theorem**: If $w^*$, $\alpha^*$ and $\beta^*$ satisfy the KKT condition, then it is also a solution to the primal and the dual problems.

# 对偶问题

- 对基本型问题应用拉格朗日乘子法得到其“对偶问题”，求解对偶问题更高效。基本型的拉格朗日函数为

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(w^Tx_i+b))$$

- 令$L$对$w$和$b$偏导为0，可得：

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i;$$

$$0=\sum_{i=1}^{m}\alpha_i y_i$$

# 对偶问题的解

代入$L$得对偶问题为

$$\max \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s.t. \sum_{i=1}^{m} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, i = 1, 2, \cdots, m.$$

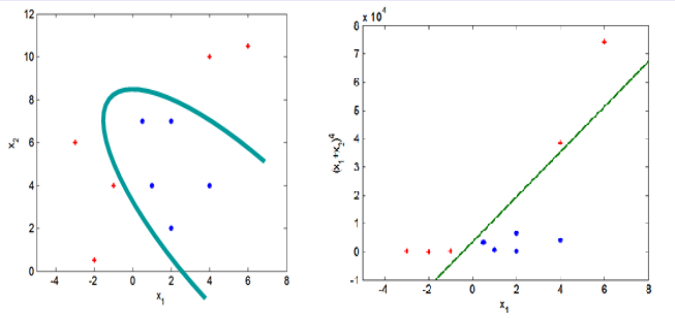
上述过程需满足KKT条件，解出$\alpha$后，求出$w$和$b$,得到模型：

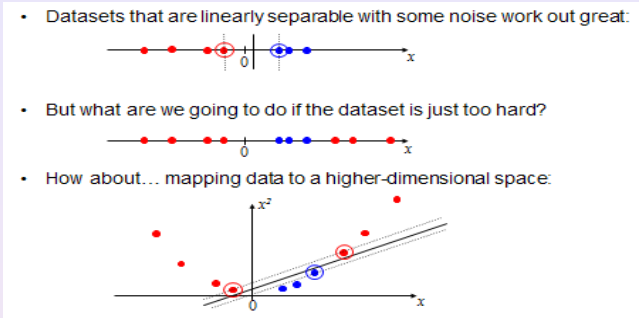
$$f(x) = w^T x + b$$

$$= \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b$$

# 核方法

如果遇见不可线性可分的情况，怎么办？






对待原始数据无法线性可分的问题，一个合适的思路是将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分,于是拉格朗日乘子法转换的对偶问题式子可转换为如下的对偶问题：

$$\min_{\alpha}\left[\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(x_i)^T\phi(x_j)-\sum_{i=1}^{m}\alpha_i\right]$$

$$=\left[\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i,x_j)-\sum_{i=1}^{m}\alpha_i\right]$$

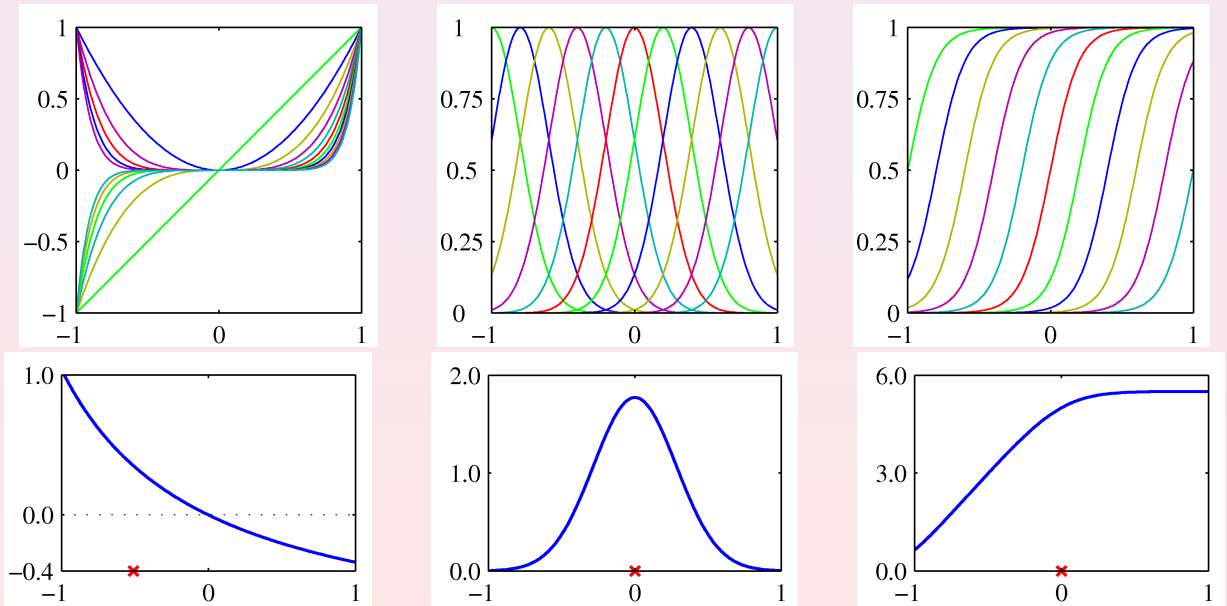
$$s.t.\sum_{i=1}^{m}\alpha_i y_i=0,\quad 0\le\alpha_i\le C, i=1,2,\cdots,m.$$

# Constructing Kernels - First Approach

由于样本$x_i$ 和$x_j$ 映射到特征空间之后的内积因为维数可能很高，所以比较难直接计算。为避开这个障碍，设计了“核函数”（kernel function），这个函数使得$x_i$ 和$x_j$ 在特征空间的内积等于它们在原始样本空间中通过核函数$k(x_i, x_j)$ 计算的结果。

- choose feature space mapping $\phi(x)$

$$k(x, x') = \phi(x)^T \phi(x') = \sum_{i=1}^{M} \phi_i(x)\phi_i(x') \tag{6.10}$$



# Constructing Kernels - Second Approach

- construct kernel function directly and verify its validity
- simple example

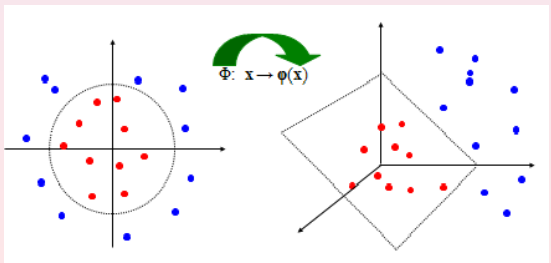
$$k(x,z) = (x^T z)^2 \tag{6.11}$$

in 2-D case corresponds to

$$k(x,z) = \phi(x)^T \phi(z) \tag{6.11}$$

with $\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T$

To test validity without having to construct the function $\phi(x)$ explicitly, one can use the condition:
Function $k(x,x')$ is a valid kernel $\Longleftrightarrow K \geq 0 \qquad \forall\{x_n\}$



# Combining Kernels

Given valid kernels $k_1(x, x')$ and $k_2(x, x')$ the following kernels will also be valid:

$$k(x, x') = ck_1(x, x') \tag{6.13}$$

$$k(x, x') = f(x)k_1(x, x')f(x') \tag{6.14}$$

$$k(x, x') = q(k_1(x, x')) \tag{6.15}$$

$$k(x, x') = \exp(k_1(x, x')) \tag{6.16}$$

$$k(x, x') = k_1(x, x') + k_2(x, x') \tag{6.17}$$

$$k(x, x') = k_1(x, x')k_2(x, x') \tag{6.18}$$

$$k(x, x') = k_3(\phi(x), \phi(x')) \tag{6.19}$$

$$k(x, x') = x^T \mathbf{A} x' \tag{6.20}$$

$$k(x, x') = k_a(x_a, x'_a) + k_b(x_b, x'_b) \tag{6.21}$$

$$k(x, x') = k_a(x_a, x'_a)k_b(x_b, x'_b) \tag{6.22}$$

with corresponding conditions on $c, f, q, \phi, k_3, \mathbf{A}, x_a, x_b, k_a, k_b$

# 常用核函数和半正定

表 6.1 常用核函数

| 名称 | 表达式 | 参数 |
| --- | --- | --- |
| 线性核 | $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$ | |
| 多项式核 | $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$ | $d \geqslant 1$ 为多项式的次数 |
| 高斯核 | $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$ | $\sigma > 0$ 为高斯核的带宽(width) |
| 拉普拉斯核 | $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right)$ | $\sigma > 0$ |
| Sigmoid 核 | $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\beta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \theta)$ | tanh 为双曲正切函数, $\beta > 0, \theta < 0$ |

[Mercer 定理]：任何半正定的函数都可以作为核函数。所谓半正定的函数$k(x_i, x_j)$，是指拥有训练数据集合$(x_1, x_2, ..., x_m)$，定义一个矩阵的元素$a_{ij} = k(x_i, x_j)$，这个矩阵是$m \times m$ 的，如果这个矩阵是半正定的，那么$k(x_i, x_j)$就称为半正定的函数。
这个mercer定理不是核函数必要条件，只是一个充分条件，即还有不满足mercer 定理的函数也可以是核函数

# SVMs for linearly separable data in general 1/7

Consider the **two-class** classification problem using linear models of the form

$$f(\mathbf{x}) = \mathbf{x}^T \phi(\mathbf{x}) + b \qquad (7.1)$$

where $\phi(\mathbf{x})$ denotes a fixed feature-space transformation.$\mathbf{x}_1, \cdots, \mathbf{x_m}$ are the **training data**,with corresponding **target value** $y_1, \cdots, y_m$ , where $y_i \in \{-1, 1\}$.
We assume that the training data set is **linearly separable** in feature space,

$$\mathbf{w}^T x + b = 0;$$

$\mathbf{w} = (w_1, \cdots, w_d)$为法向量，决定了超平面的方向；$b$为位移项，决定了超平面和原点之间的距离，划分超平面由$\mathbf{w}, b$ 唯一决定：so that there must exist $\mathbf{w}, b$ s.t.

$$\begin{aligned} f(\mathbf{x}_i) > 0 \quad y_i = 1 \\ f(\mathbf{x}_i) < 0 \quad y_i = -1 \end{aligned}$$

that is, $f(\mathbf{x}_i) \cdot y_i > 0, \forall i$.

一般而言，

**[Margin]**

The *distance* of a point $\mathbf{x}_i$ to the decision surface is

$$\frac{y_i f(\mathbf{x}_i)}{\|\mathbf{w}\|} = \frac{y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b)}{\|\mathbf{w}\|} \qquad (7.2)$$

The *margin* is the perpendicular distance(垂直距离) to the closet point $\mathbf{x}_i$.

We wish to optimize the parameters $\mathbf{w}$ and $b$ to maximize this distance.

# SVMs for linearly separable data in general 3/7

The **maximum margin** solution is found by solving

$$\arg\max_{\mathbf{w},b}\{\frac{1}{\|\mathbf{w}\|}\min_{i}[y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b)]\} \quad (7.3)$$

Note that,if $\mathbf{w}\to\kappa\mathbf{w}$ and $b\to\kappa b$,the distance is unchanged. For the closet point,let

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_n)+b)=1 \quad (7.4)$$

So we have to solve the optimization problem

$$\arg\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|_2^2 \quad (7.6)$$

$$subject\ \ to\ \ y_i(\mathbf{w}^T\phi(\mathbf{x}_n)+b)\geq 1 \quad (7.5)$$

# SVMs for linearly separable data in general 4/7

Give the Lagrangian function

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|_2^2 - \sum_{i=1}^{m} \alpha_i \{y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) - 1\} \tag{7.7}$$

We can obtain

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \phi(\mathbf{x}_i) \quad 0 = \sum_{i=1}^{m} \alpha_i y_i$$

Eliminate $\mathbf{w}$ and $b$ ,and use the *dual representation* $k(x, x') = \phi(x)^T \phi(x')$,the optimization problem transfers into

$$\tilde{L}(\mathbf{a}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{7.10}$$

subject to

$$\alpha_i \geq 0, \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

# SVMs for linearly separable data in general 5/7

For new data $\mathbf{x}$,we use $\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \phi(\mathbf{x}_i)$ to give

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b \tag{7.13}$$

In Appendix E,we show that the following 3 properties hold:

$$\alpha_i \geq 0 \tag{7.14}$$

$$y_i f(\mathbf{x}_i) - 1 \geq 0 \tag{7.15}$$

$$\alpha_i (y_i f(\mathbf{x}_i) - 1) = 0 \tag{7.16}$$

Thus for every data point,either $\alpha_i = 0$ or $y_i f(\mathbf{x}_i) = 1$.
These satisfy $y_i f(\mathbf{x}_i) = 1$,that is $\alpha_i \neq 0$,we call them **SUPPORT VECTORS**,and they lie on the maximum margin hyperplanes in feature space. 它们对应于特征空间中位于最大边距超平面上的点

To determine $b$,we use

$$y_s\left(\sum_{i\in S}\alpha_i y_i k(\mathbf{x}_i,\mathbf{x}_s)+b\right)=1 \tag{7.17}$$

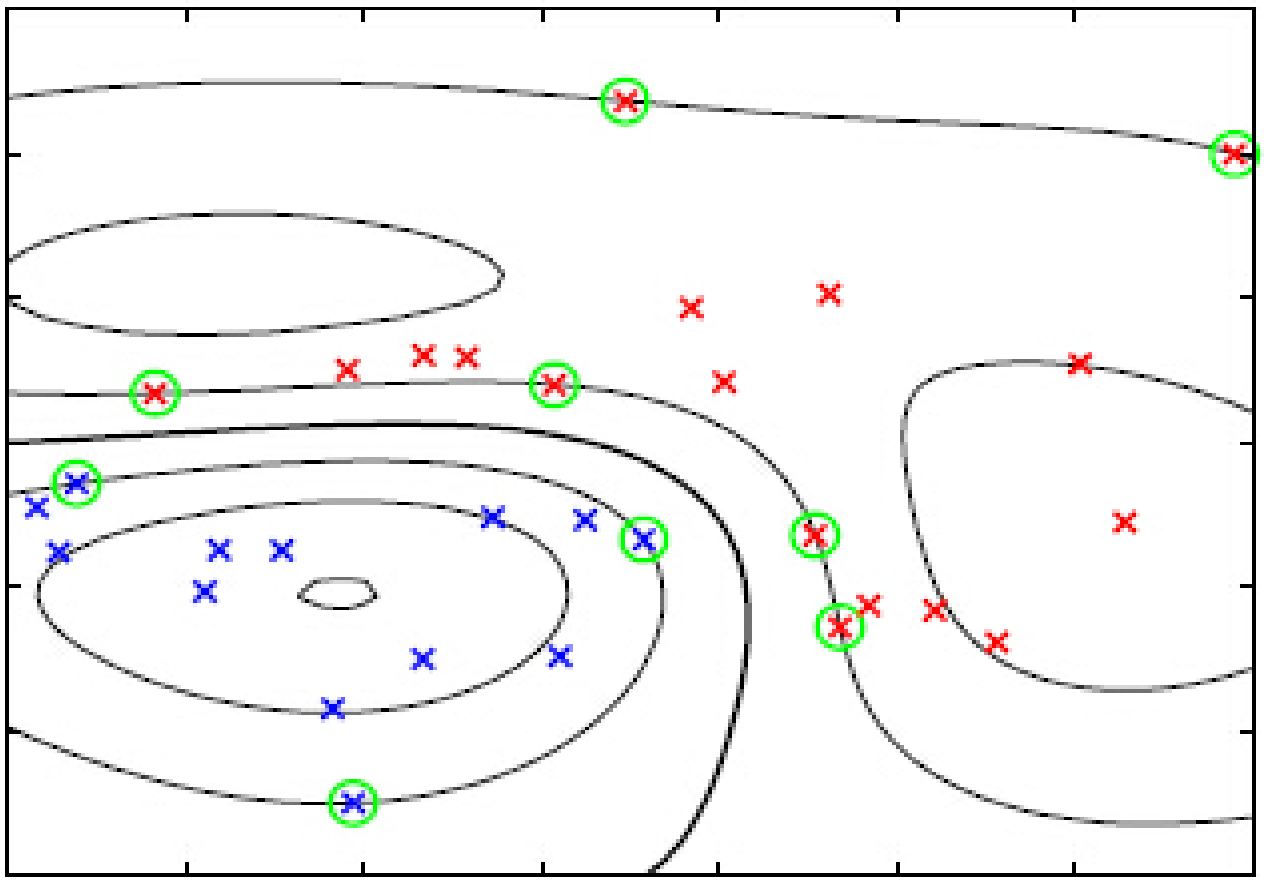
or make use of $y_i^2=1$,then give

$$b=\frac{1}{N_S}\sum_{s\in S}(y_s-\sum_{i\in S}\alpha_i y_i k(\mathbf{x}_i,\mathbf{x}_s))$$

where $S$ denotes the set of SVs.

# SVMs for linearly separable data in general 7/7

二维空间见两个类别模拟数据，使用高斯核函数的支持向量机得到的常数$f(x)$的轮廓线，决策边界、间隔线和支持向量



# SVM方法是对函数复杂性的$K$最优选择的依据

- 使分类间隔最大实际上就是对推广能力的控制，这是SVM的核心思想之一。Vapnik 证明，在$N$ 维空间中，设样本分布在一个半径为$R$ 的超球范围内，则正则超平面$f(x,w,b) = sgn((wx)+b)$ 构成的指示函数集的VC 维满足下面的界

$$h \leq \min[\frac{R^2}{\rho^2}], n+1$$

其中$\rho$是等于$2/w$的分离间隔。因此使$w$最小就是使VC维的上界最小，从而实现SRM准则中对函数复杂性的选择。

- 结构风险最小化原则

Vapnik深入研究后得到如下结论，：经验风险Remp (w)和实际风险R(w) 之间以至少1 −η 的概率满足如下的关系：

$$R(w) \le R_{emp}(w) + \sqrt{\frac{h(\ln(2l/h)+1-\ln(\eta/4))}{l}}$$

VC置信度（VC confidence）

*h* 是VC 维， *l* 是样本数

Remp (w)最小 → 期望风险R(w) 最小

结构风险最小原则：为了使得期望风险R(w) 最小，应设法使得右边的两个同时最小。

# 软间隔与正则化

当超平面无法完全划分开训练样本时，该如何处理？
在现实任务中往往很难确定合适的核函数使得训练样本在特征空间中线性可分，为了缓解该问题，一个合理的办法是允许支持向量机在一些样本上出错。这种策略被称为“软间隔”（soft margin），它允许某些样本不满足不等式约束。
在最大化间隔的同时，不满足约束的样本应尽可能少，于是，优化目标可改写为：

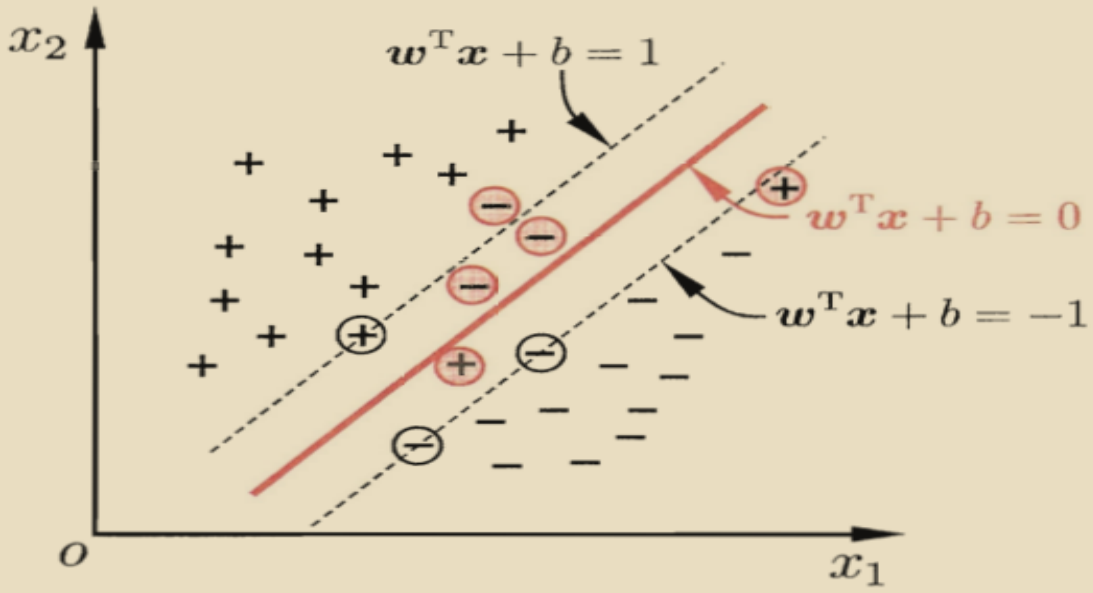



图 6.4 软间隔示意图. 红色圈出了一些不满足约束的样本.

# Non-separable problem:Slack variable松弛变量

We now modify this approach so that data points are allowed to be on the ' wrong ' side of the margin boundary,but with a penalty that increases with the increases with the distance from that boundary. To do this,we introduce **slack variables** $\xi_i \geq 0$.And the exact classification constraints (7.5) are replaced with

$$y_i f(\mathbf{x}_i) \geq 1 - \xi_i, i = 1, \cdots, m \tag{7.20}$$

- $\xi_i = 0$ : *correctly classified* and are either *on the margin* or *on the correct side* of the margin; 被正确分类的点，其位置在边缘上或在边缘的正确一侧
- $0 < \xi_i \leq 1$ : *correctly classified* and lie *inside the margin*; 位于边缘内部被分类正确的
- $\xi_i > 1$ : *misclassified*.错分的

The slack variables give a soft margin and allow for overlapping class distributions.

# Slack variable1/7

Our goal is to maximize the margin while softly penalizing the misclassified points.So we minimize

$$C\sum_{i=1}^{m}\xi_i+\frac{1}{2}\|\mathbf{w}\|_2^2 \tag{7.21}$$

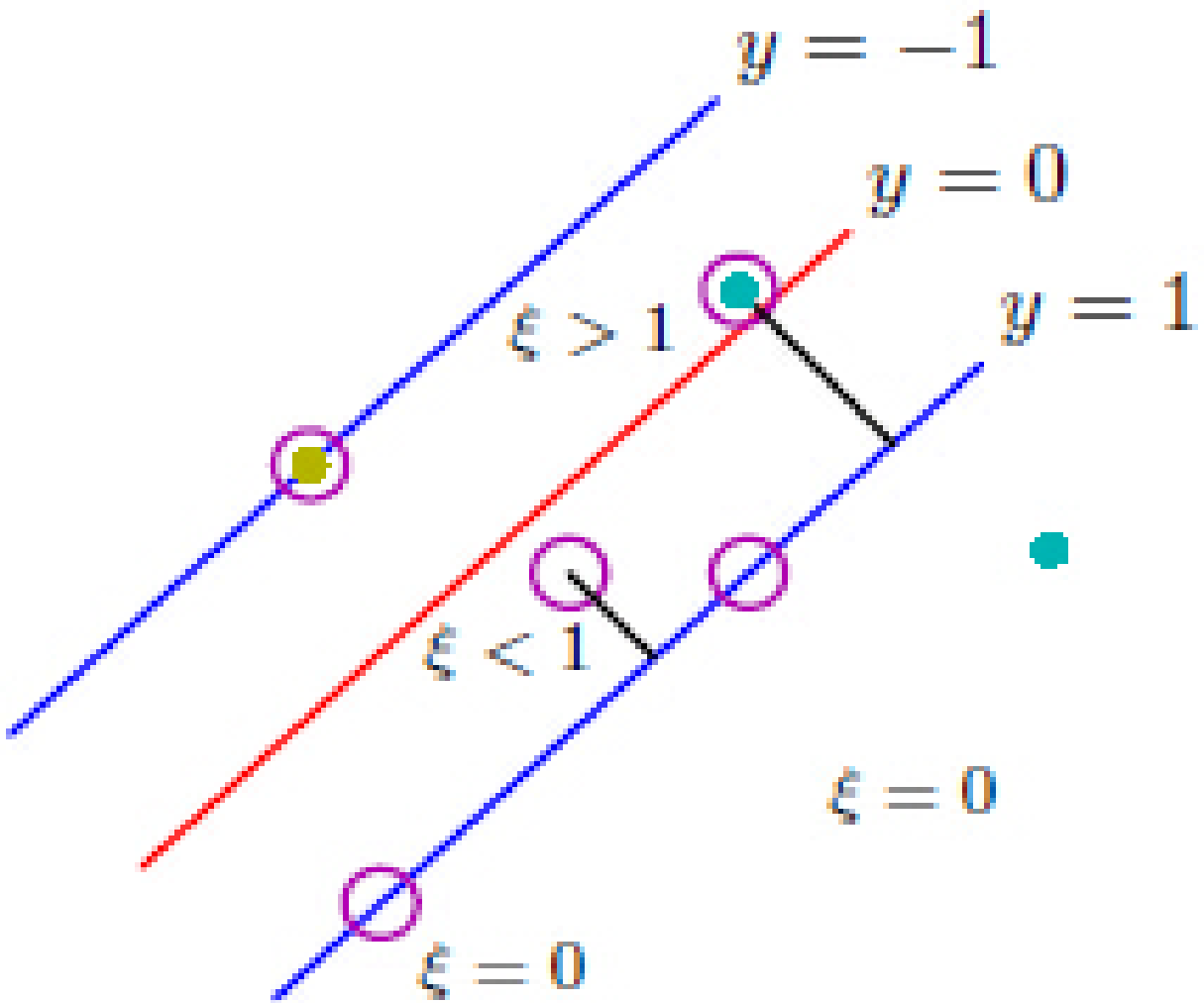
subject to

$$\xi_i \geq 0$$

$$y_i f(\mathbf{x}_i) \geq 1-\xi_i$$

where the parameter $C>0$ controls the trade-off between the slack variable penalty and the margin.If $C\to\infty$ ,we will recover the earlier SVM for separate data.

# Slack variable2/7



Lagrangian function:

$$L(\mathbf{w}, b, \alpha, \mu) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\{y_i f(\mathbf{x}_i) - 1 + \xi_i\} - \sum_{i=1}^{m}\mu_i\xi_i \tag{7.22}$$

And the KKT conditions are given by

$$\begin{array}{cc} \alpha_i \geq 0 & \mu_i \geq 0 \\ y_i f(\mathbf{x}_i) - 1 + \xi_i \geq 0 & \xi_i \geq 0 \\ \alpha_i(y_i f(\mathbf{x}_i) - 1 + \xi_i) = 0 & \mu_i\xi_i = 0 \end{array}$$

and so we can obtain

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \phi(\mathbf{x}_i) \quad \sum_{i=1}^{m}\alpha_i y_i = 0 \quad \alpha_i = C - \mu_i$$

Eliminate $\mathbf{w}$,$b$ and $\xi_n$ ,we obtain the dual Lagrangian in the form

$$\tilde{L}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{7.32}$$

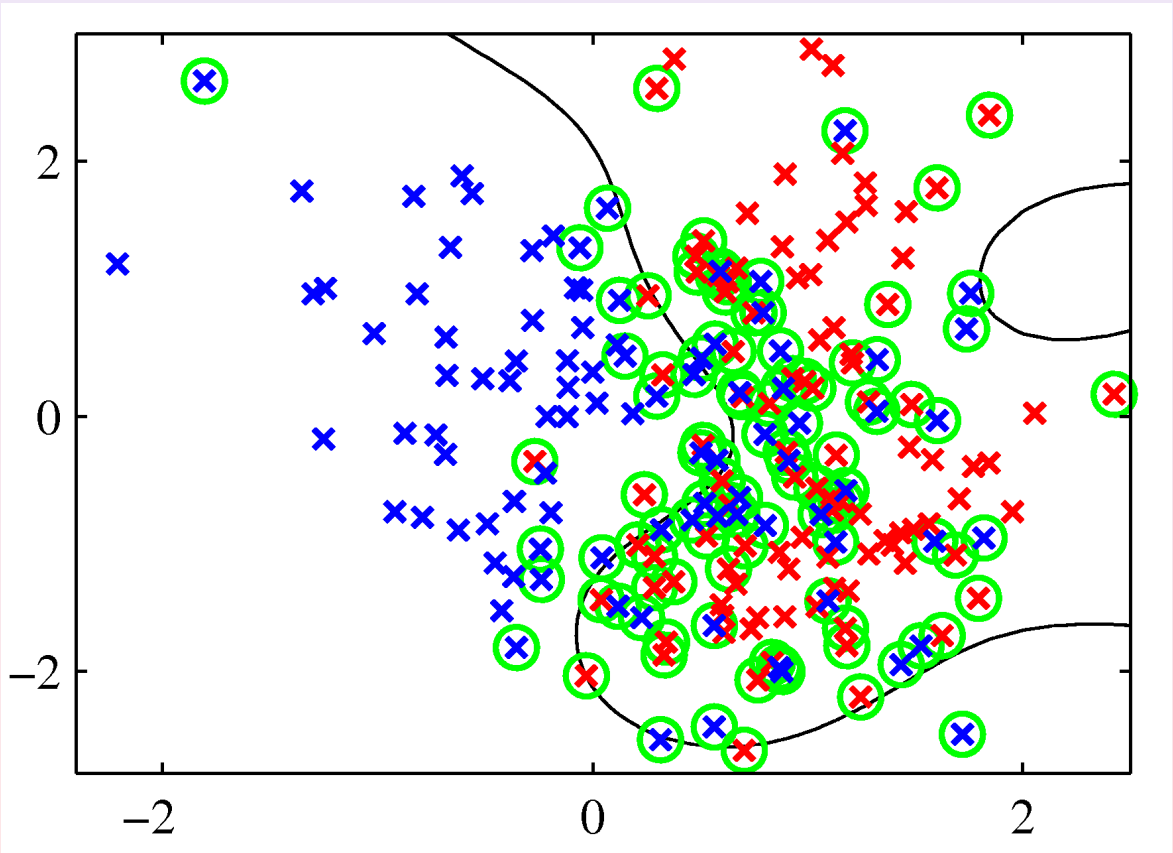
subject to the *box constraint*

$$0 \leq \alpha_i \leq C$$
$$\sum_{i=1}^{m} \alpha_i y_i = 0$$

For new data $\mathbf{x}$,we also give $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i y_i k(\mathbf{x}, \mathbf{x}_i) + b$

# Slack variable5/7

- $\alpha_i \geq 0$ : **support vectors**,and satisfy $y_i f(\mathbf{x}_i) = 1 - \xi_i$
- $0 < \alpha_i < C$ : $\xi_i = 0$ and hence lie *on the margin*
- $\alpha_i = C$ : *inside the margin* ,either $0 < \xi_i \leq 1$ correctly classified or $\xi_i > 1$ misclassified



- $v - SVM$

Minimize

$$\tilde{L}(\alpha) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{i=1}^{m}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{7.38}$$

subject to

$$0 \leq \alpha_i \leq \frac{1}{m} \quad \sum_{i=1}^{m} = 0 \quad \sum_{i=1}^{m}\alpha_i \geq v$$

This approach has the advantage that the parameter $v$ ,which replaces $C$ , can be interpreted as

- an upper bound on the fraction of *margin errors*边缘错误分类的上界
- a lower bound on the fraction of support vectors支持向量数据点比例的下界

The margin errors means points for which $\xi_i > 0$ and hence lie on the wrong side of the margin boundary.

# 支持向量机总结Slack variable:summary

- 核函数将特征空间映射到线性空间Kernel functions correspond to inner products in feature spaces that can have high or even infinite dimensionality.
- 支持向量机也无法控制维度灾难，因为特征的取值也会限制有效性But SVM can non manage to avoid the curse of dimensionality, because there are constraints amongst the feature values that restrict the effective dimensionality of feature space.
- 支持向量机没有可能性的输出，只有分类决策面SVM does not provide probabilistic output but instead makes classification decisions for new input vectors. One approach to address that is to fit a logistic sigmoid to the output of a previously trained SVM. $p(t=1|\mathbf{x}) = \sigma(Ay(\mathbf{x}) + B)$.But this is equivalent to assuming that the output $f(\mathbf{x})$ represents the log-odds of $\mathbf{x}$ belonging to class $y = 1$.So it leads to poor approximation to posterior.