4.6 最优分类超平面与支持向量机

- 主要内容有
  - 线性可分的SVM
  - （非线性可分）SVM

1 线性可分 SVM

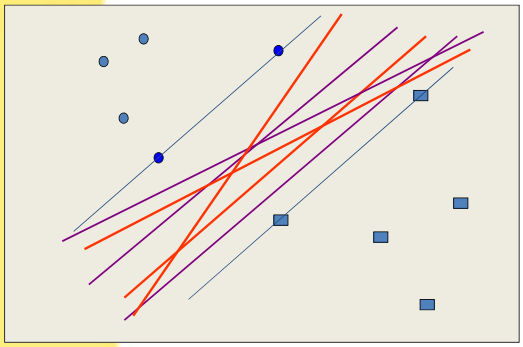
- 所谓线性可分的分类问题是指:
  - 对n维空间的l个数据样本 $(x_1,y_1), (x_2,y_2)\ldots\ldots (x_l,y_l)$,
    - 其中x_i为n维空间的数据样本;
      $y_i\in\{-1,1\}$为数据样本的类别.取值为-1属于I类,为1则为II类.
  - 存在并求解n维空间内超平面的分类面
    - $f(x)=w^Tx-b=0$

使得如下关系成立

$$y_i = g(x_i) = \begin{cases} -1 & f(x_i) < 0, \text{分类为I类} \\ 1 & f(x_i) \geq 0, \text{分类为II类} \end{cases}$$

- 对线性分类问题,模式识别领域已有充分研究,得到了许多有效的算法.
  - 但在实际中,传统的线性分类算法,如模式识别的分类学习方法、感知器:

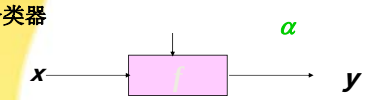
    分类面仅考虑了对现有样本的分类面的存在与求解,
    - 未考虑其泛化能力,因此泛化能力较弱
    - 分类面部不唯一,到底何为最优的分类面?
    - 何为最优的分类,最优指标的意义?

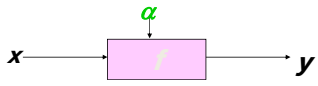


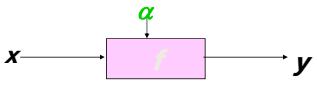
- 对有污染的或近似线性分类的数据样本,可能严格线性分类意义下的分类面不存在.
  - 因此鲁棒性较差.
- 对海量数据,求解分类面的时间代价大.

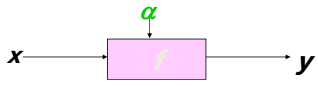
– 因此,发展泛化能力强,鲁棒性佳,求解的代价小的新的分类算法得到重视.

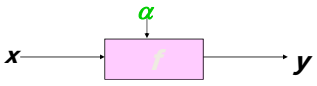
- 最优分类面SVM
- 广义最优分类面 SVM











A. 最优分类面SVM

- SVM对于线性分类的问题描述为:
  - 对n维空间的l个数据样本(x_1,y_1), (x_2,y_2)…… (x_l,y_l),
  - 存在并求解n维空间内满足条件:
    - 经验风险最小(错分最少)
    - 泛化能力最大(空白最大)

  的最优分类面
    - $f(x)=w^Tx-b=0$

线性分类器

f(x,w,b) = sign(w. x - b)

- denotes +1
- denotes -1

如何划分这两类数据？

线性分类器

f(x,w,b) = sign(w. x - b)

- denotes +1
- denotes -1

如何划分这两类数据？

线性分类器

f(x,w,b) = sign(w. x - b)

- denotes +1
- denotes -1

如何划分这两类数据？

线性分类器

f(x,w,b) = sign(w. x - b)

- denotes +1
- denotes -1

如何划分这两类数据？

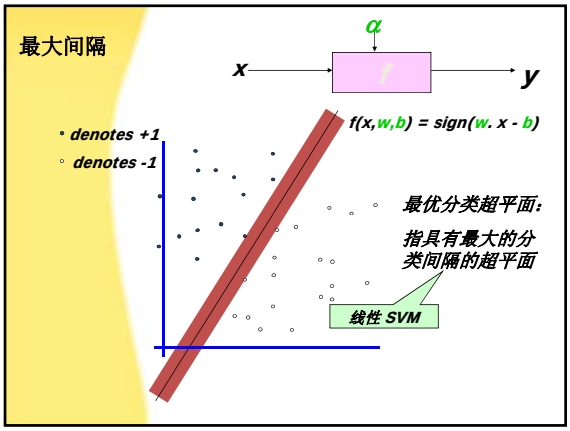
线性分类器

f(x,w,b) = sign(w. x - b)

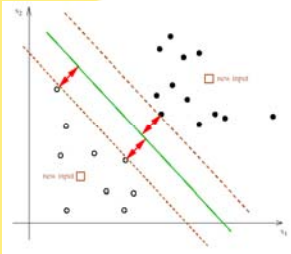
- denotes +1
- denotes -1

如何划分这两类数据？

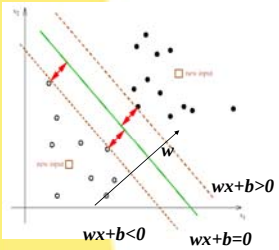
最大间隔



最大间隔



最大间隔原则



性质1：决策函数（**w**, b）与（c**w**, cb）相同

性质2：但是由(**w**, b) 与 (c**w**, cb)确定的决策函数值不同.

间隔的计算（点到超平面的距离）：

$$d = \frac{|wx+b|}{\|w\|}$$

最大间隔原则

根据性质1，我们规定决策超平面在两边最近的点取值分别为+1 和 －1 这样决策超平面就被唯一确定了.

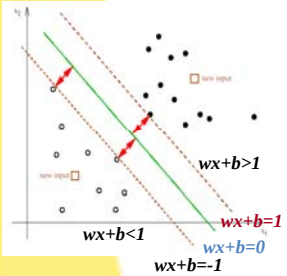
记 x_+ 和 x_- 分别为距超平面最近的正类样本和负类样本，即

$$\mathbf{w}\mathbf{x}_+ + b = +1 \text{ and } \mathbf{w}\mathbf{x}_- + b = -1$$

则间隔为：

$$\frac{1}{2}\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{x}_+ + \frac{b}{\|\mathbf{w}\|} - \frac{\mathbf{w}}{\|\mathbf{w}\|}\mathbf{x}_- - \frac{b}{\|\mathbf{w}\|}\right) = \frac{1}{2\|\mathbf{w}\|}(\mathbf{w}\mathbf{x}_+ + b - \mathbf{w}\mathbf{x}_- - b) = \frac{1}{\|\mathbf{w}\|}$$

$$\left.\begin{array}{ll}\mathbf{w}\mathbf{x}_i + b \geq +1 & \text{for } y_i = +1 \\ \mathbf{w}\mathbf{x}_i + b \leq -1 & \text{for } y_i = -1\end{array}\right\} \Leftrightarrow y_i(\mathbf{w}\mathbf{x}_i + b) - 1 \geq 0 \quad \text{for all } i$$



给定线性可分的训练点集：
(x_i, y_i), i=1,2,…N;

$x_i \in R^n$ $y_i \in \{+1,-1\}$

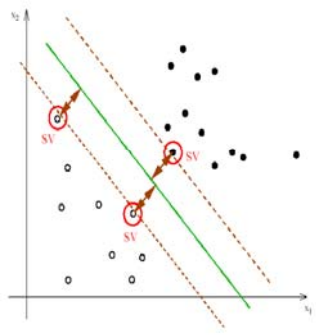
求最优分类超平面就转换为如下数学规划：

Min $\|w\|^2$

S. t.

$$y_i(\mathbf{w}\mathbf{x}_i + b) - 1 \geq 0, \quad i = 1,..N$$

线性支持向量机



在间隔平面上的点成为支持向量. 超平面只与这些向量有关，而与其他正确分类的样本无关

分类问题的数学表示

已知：训练集包含 l 个样本点：

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(X\times Y)^l$$

说明： $x_i\in X=R^n$ 是输入指标向量，或称输入，或称模式，其分量称为特征，或属性，或输入指标；

$y_i\in Y=\{1,-1\}$ 是输出指标，或输出.

问题：对一个新的模式x ，推断它所对应的输出 y =1 还是 -1.

实质：找到一个把R^n上的点分成两部分的规则.

分类学习方法

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。

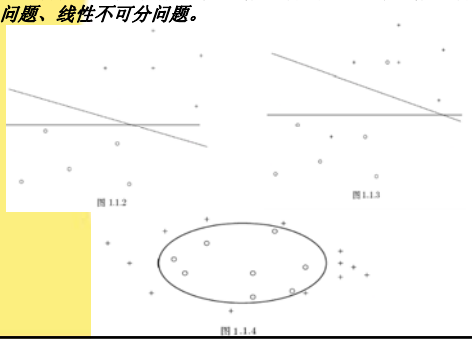


图 1.1.2　图 1.1.3

图 1.1.4

- 最优分类面问题可以表示成约束优化问题
  - Minimize
  - Subject to

$$\Phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w\cdot w)$$

$$y_i((w\cdot x_i)+b)\geq 1, i=1,...,l$$

- 定义Lagrange函数

$$L(w,b,\alpha)=\tfrac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(y_i\cdot((x_i\cdot w)+b)-1)$$

- Lagrange函数

$$L(w,b,\alpha)=\tfrac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i(y_i\cdot((x_i\cdot w)+b)-1)$$

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

$$\sum_{i=1}^{l}a_i y_i=0\qquad w=\sum_{i=1}^{l}\alpha_i y_i x_i$$

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

$$\alpha_i\geq 0,\ i=1,...,l,\ \text{and}\ \sum_{i=1}^{l}\alpha_i y_i=0$$

$$f(x)=\operatorname{sgn}(\sum_{i=1}^{l}y_i\alpha_i\cdot(x\cdot x_i)+b)$$

求解原始问题

为求解原始问题，根据最优化理论，我们转化为对偶问题来求解

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

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

$$\alpha_i\geq 0, i=1\cdots l$$

对偶问题

α_i 为原始问题中与每个约束条件对应的Lagrange乘子。这是一个不等式约束条件下的二次函数寻优问题，存在唯一解 α^*

线性可分问题

根据最优解 $\alpha^*=(a_1^*,\cdots,a_l^*)^T$

计算 $w^*=\sum_{i=1}^{l}y_i\alpha_i^* x_i$，*选择* a^* *的一个正分量* α_j^*，*并据此计算*

$$b^*=y_j-\sum_{i=1}^{l}y_i\alpha_i^*(x_i\cdot x_j)$$

构造分划超平面 $(w^*\cdot x)+b^*=0$，*决策函数* $f(x)=\operatorname{sgn}((w^*\cdot x)+b^*)$

事实上，α^* 的每一个分量 α_i^* 都与一个训练点相对应。而分划超平面仅仅依赖于 α_i^* 不为零的训练点 (x_i,y_i)，而与对应于 α_i^* 为零的那些训练点无关。

称 α_i^* 不为零的这些训练点的输入 x_i 为支持向量(SV)

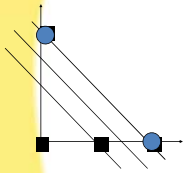
一个简单的例子:

$x_1=(0,0),\ y_1=+1$

$x_2=(1,0),\ y_2=+1$

$x_3=(2,0),\ y_3=-1$

$x_4=(0,2),\ y_4=-1$



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

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

$$\alpha_i \geq 0, i = 1 \cdots l$$

$$max \quad Q(\alpha) = (\alpha_1+\alpha_2+\alpha_3+\alpha_4) - \frac{1}{2}(\alpha_2^2 - 4\alpha_2\alpha_3 + 4\alpha_3^2 + 4\alpha_4^2)$$

$$s.t. \quad \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0$$

```
H=[0 0 0 0;0 1 -2 0;0 -2 4 0;0 0 0 4];
f=-1*[1;1;1;1];
Aeq=[1 1 -1 -1];   beq=0
lb=[0 0 0 0];      ub=[inf inf inf inf];
options=optimset('LargeScale','off');
x0=[1 0 1 0];
alpha=quadprog(H,f,[],[],Aeq,beq,lb,ub,x0,options)
```

$$w^* = \sum_{i=1}^{l} y_i \alpha_i^* x_i$$

$$\alpha = \begin{pmatrix} 0 \\ 4 \\ 3 \\ 1 \end{pmatrix}, \quad w = 4\begin{pmatrix} 1 \\ 0 \end{pmatrix} - 3\begin{pmatrix} 2 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 2 \end{pmatrix} = \begin{pmatrix} -2 \\ -2 \end{pmatrix}$$

$$w^T x_2 + b = 1 \Rightarrow b = 3$$

决策超平面为 $(w^* \cdot x) + b^* = 0$

$$-2x_1 - 2x_2 + 3 = 0$$

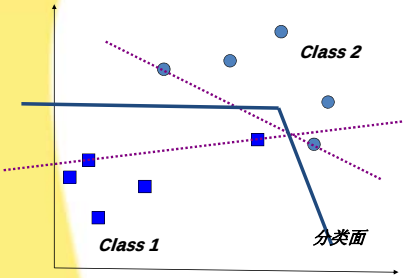
B. 广义最优分类面SVM

- 对于许多实际问题,前面讨论的最优分类面的条件

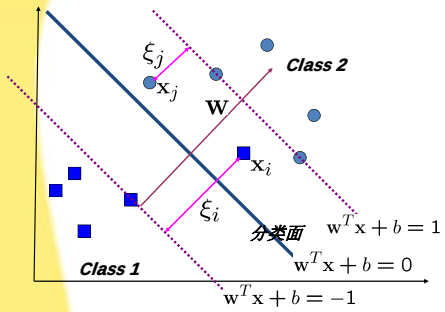
$$y_i g(x_i) = y_i (w^T x_i - b) \geq 1 \qquad i = 1,2,\ldots,n$$

过于严格.

- *对存在数据污染、近似线性分类的情况,可能并不存在一个最优的线性分类面.*
  - ✓ *如对于下图所示的分类问题,不存在严格的线性分类面*



- *此外,对于海量数据,求解最优分类面的时间代价大.*
- *因此,需要发展允许有一定范围内的“错分”,又有较大分界区域的最优分类面.*
  - ✓ *实际上,广义最优分类面是在分类准确性与泛化特性上寻求一个平衡点.*
  - ✓ *上图所示的分类问题可以找到如下图所示的广义最优分类面。*



- “允许有一定的错分”的分类问题通过引入松弛变量ξ_i, 可表示为如下优化问题：

$$\min_{w,b,\xi_i} \sum_{i=1}^{m} \xi_i$$

相应的约束条件为

$$y_i g(x_i) = y_i(w^T x_i - b) \geq 1 - \xi_i \qquad i = 1,2,...,n$$

因此,SVM的广义最优分类问题可表示为如下优化问题:

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \qquad C > 0$$

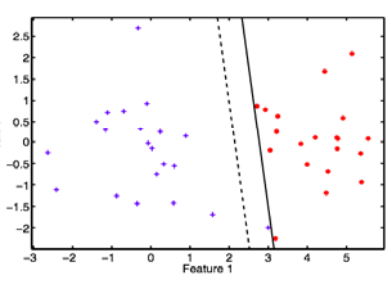
约束条件

$$y_i g(x_i) = y_i(w^T x_i - b) \geq 1 - \xi_i \qquad i = 1,2,...,n$$
$$\xi_i \geq 0 \qquad i = 1,2,...,n$$

其中参数C是一个误差与间隔之间的平衡参数.

- ✓ *C越大,表示分类越严格,允许错分的样本受到的限制越大,错分的样本数少,越过拟合。*
- ➢ *按照上述目标函数进行分类所得到的分类面也称为软间隔最优分类超平面。*

- 控制参数的作用如下图所示



虚线与实线为C=1和10^4所解得的分类面

- 下面讨论基于拉格朗日松弛法, 求解广义最优分类面问题.
  - 首先引入拉格朗日乘子

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i[1 - \xi_i - y_i(w^T x_i - b)] - \sum_{i=1}^{m} \pi_i \xi_i$$

其中α_i与π_i为拉格朗日乘子,满足

$$\alpha_i, \eta_i \geq 0$$

- ✓ *根据拉格朗日松弛法的Kuhn-Tucker条件,有*

$$\nabla_w = w - \sum_i \alpha_i y_i x_i = 0$$
$$\partial L/\partial b = \sum_i \alpha_i y_i = 0$$
$$\partial L/\partial \xi_i = C - \alpha_i - \pi_i = 0$$
$$\alpha_i \geq 0 \quad \alpha_i[1 - \xi_i - y_i(wx_i - b)] = 0$$
$$\pi_i \geq 0 \quad \pi_i \xi_i = 0$$

可得

$$w=\sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0 \quad 0 \le \alpha_i \le C$$

$$\alpha_i[1-\xi_i - y_i(wx_i - b)] = 0$$

$$\pi_i \ge 0 \quad \pi_i \xi_i = 0$$

$$C - \alpha_i - \pi_i = 0$$

- *只要确定α_i,便可解出w,b,ξ_i, π_i*

– 将上述条件代入拉格朗日函数L中, 有

$$L=\frac{1}{2}\|w\|^2+\sum_i \alpha_i - \sum_i \alpha_i y_i x_i^T w - b\sum_i \alpha_i y_i + \sum_i (C-\pi_i-\alpha_i)\xi_i$$

其中w,b,ξ_i, π_i可根据α_i计算出,即优化变量只有α_i.

- *因此,SVM广义最优分类面的优化目标函数可变换为*

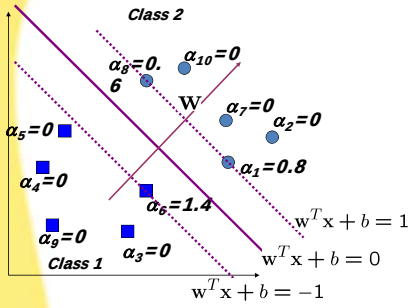
$$L=\sum_i \alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

相应的约束条件为

$$\sum_i \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C$$

- SVM广义最优分类面的上述等效优化问题是一个不等式约束下二次函数寻优的问题, 存在唯一解.
  – 容易证明, 解中将只有一部分(通常是少部分)α_i不为零, 对应的样本就是**支持向量**.
    - 支持向量（α_i不为零）与非支持向量（α_i不零）如下图所示。



- 解上述问题后得到的最优分类函数是

$$g(\mathbf{x}) = \mathrm{sgn}\{\mathbf{w}^T\mathbf{x}+b\} = \mathrm{sgn}\left\{\sum_{i=1}^{n}\alpha_i^* y_i(\mathbf{x}_i^T\mathbf{x})+b^*\right\}$$

式中的求和实际上只对支持向量进行.

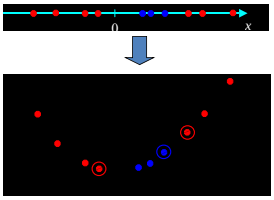
- *b^*是分类阈值,可以用任一个支持向量(满足(1)中的等号)求得,或通过两类中任意一对支持向量取中值求得.*
- *注意到分类函数仅依赖于测试点和支持向量样本点之间的内积。*

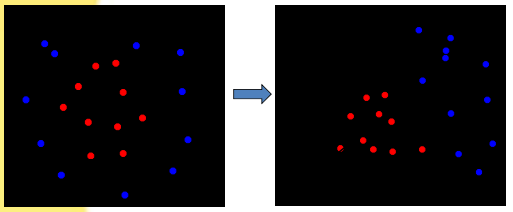
3 支持向量机（SVM）

- 非线性分类问题一直是分类领域的困难问题, 主要的困难在于难于构造非线性的分类判别函数.
  – 实际上, 非线性可分的数据样本有可能在**适当的函数变换**

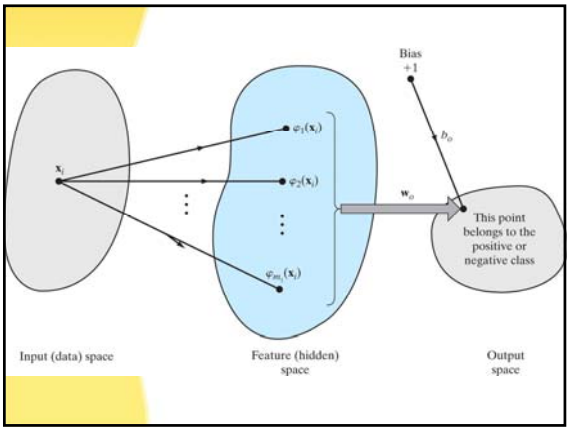
下在高维空间有可能转化为线性可分.

- *如下图所示的2个例子.*





- 因此对非线性问题，可以把样本x映射到某个高维特征空间H，并在H中使用线性分类器.



- 设实函数

$$\phi(x): \mathbf{R}^n \to \mathbf{H}$$

为向量x对应的特征空间**H**中的特征向量.

– 将特征向量$\phi(x)$代替输入向量x，则可以得到相应的分类函数与非线性分类的广义最优分类的目标函数

$$g(\mathbf{x}) = \text{sgn}\left\{\sum_{i=1}^{n} \alpha_i y_i \left[\phi^T(\mathbf{x}_i)\phi(\mathbf{x})\right] + b\right\}$$

$$L = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \left[\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)\right]$$

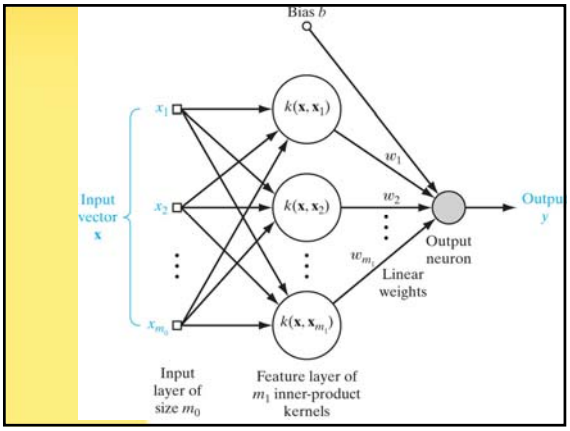
相应的函数优化的约束条件仍然为

$$\sum_i \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C$$

- 由上式可以看出，分类函数与优化函数中只是涉及样本x的特征向量$\phi(x)$之间的内积，而未直接需要求解出特征向量$\phi(x)$.
  - 因此，实际上，在高维的特征空间**H**中只需进行特征向量$\phi(x)$内积运算，而这种内积运算是可以用原空间中的函数实现的，我们甚至不需要知道变换$\phi(x)$的形式.
  - 根据泛函理论，只要一种核函数$K(x_i, x_j) = \phi^T(x_i)\,\phi(x_j)$满足Mercer条件，它就对应某一空间中的内积.
  - 因此，在最优分类中采用适当的内积函数就可以实现某一非线性变换后的线性分类，而计算的复杂度没有增加.
  - 正是这一思路产生了现在非常有效的非线性分类的SVM.

- SVM的这一特点提供了解决算法可能导致的“维数灾难”问题的方法：
  - 在构造判别函数时，不是对输入空间的样本作非线性变换，然后在特征空间中求解；
    - 而是先在输入空间比较向量(例如求点积或是某种距离)，对结果再作非线性变换.
  - 这样，大的工作量将在输入空间而不是在高维特征空间中完成.
  - SVM分类函数形式上类似于一个NN，输出是s中间节点的线性组合，每个中间节点对应一个支持向量，如下图所示.



- 基于核函数$K(x_i, x_j)$，上述非线性SVM的分类函数与广义最优分类的目标函数分别为

$$g(\mathbf{x}) = \operatorname{sgn}\left\{\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right\}$$

$$L = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

  - ***因此,非线性分类的SVM方法最后集中到核函数的选取.***
    - ✓ ***选取适宜核函数是成功进行非线性分类的关键.***

- 下面对偶问题的最优解做一些推导.
  - 定义

$$w(\alpha) = \sum_i \alpha_i y_i \phi(\mathbf{x}_i)$$

$$F_i = w(\alpha)\cdot\phi(\mathbf{x}_i) - y_i = \sum_j \alpha_j y_j K(\mathbf{x}_i, \mathbf{x}_j) - y_i$$

则对偶问题的Lagrange函数可以写成:

$$L = \frac{1}{2}\|w(\alpha)\|^2 + \sum_i \alpha_i - \sum_i \delta_i \alpha_i + \sum_i \xi_i (C - \alpha_i) - b\sum_i \alpha_i y_i$$

  - ***因此KKT条件为***

$$L = \frac{1}{2}\|w\|^2 + \sum_i \alpha_i - \sum_i \alpha_i y_i x_i^T w - b\sum_i \alpha_i y_i + \sum_i (C - \pi_i - \alpha_i)\xi_i$$

$$\frac{\partial L}{\partial \alpha_i} = (F_i - b) y_i - \delta_i + \xi_i = 0$$

$$\delta_i \alpha_i = 0 \quad \text{且} \quad \delta_i \geq 0$$

$$\xi_i (\alpha_i - C) = 0 \qquad \forall\ i$$

- 由此,我们可以推导出如下关系式:
  - 若α_i=0, 则$\delta_i \geq 0$, ξ_i=0 $\Rightarrow$ $(F_i - b_i)\, y_i \geq 0$
  - 若$0 < \alpha_i < C$, 则δ_i=0, ξ_i=0 $\Rightarrow$ $(F_i - b_i)\, y_i = 0$
  - 若$\alpha_i = C$, 则δ_i=0, $\xi_i \geq 0$ $\Rightarrow$ $(F_i - b_i)\, y_i \leq 0$
- 由于KKT条件是最优解应满足的充要条件,所以目前提出的一些算法几乎都是以是否违反KKT条件作为迭代策略的准则.