机器学习
Machine Learning

第6讲：核函数与SVM

1.核函数

由特征向量构造特征映射函数，即

$$\boldsymbol{x} \Rightarrow \phi(\boldsymbol{x}) = \left[\phi_0(\boldsymbol{x}), \phi_1(\boldsymbol{x}), \cdots, \phi_{M-1}(\boldsymbol{x})\right]^T$$

则核函数的一种定义是

$$k(\boldsymbol{x}, \boldsymbol{x}') = \phi^{\mathrm{T}}(\boldsymbol{x})\phi(\boldsymbol{x}')$$

核函数满足对称性

$$k(\boldsymbol{x}, \boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x})$$

一个简单的核函数例子 $\phi(\boldsymbol{x}) = \boldsymbol{x}$ 则

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}' = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle$$

核函数正定性

设有样本集 $\{\boldsymbol{x}_{\mathrm{n}}\}_{n=1}^{N}$

可定义 $K_{nm} = k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \phi^{\mathrm{T}}(\boldsymbol{x}_n)\phi(\boldsymbol{x}_m)$

则有矩阵 $\mathbf{K} = [K_{nm}]_{N\times N} = [k(\boldsymbol{x}_n, \boldsymbol{x}_m)]_{N\times N}$

$= \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}$

这里 $\boldsymbol{\Phi} = [\phi(\boldsymbol{x}_1), \phi(\boldsymbol{x}_2), \cdots, \phi(\boldsymbol{x}_{\mathrm{N}})]^{\mathrm{T}}$

$\mathbf{K}$ 称为Gram矩阵（核矩阵），是对称的半正定矩阵。

例：可直接构造核函数，分解为特征映射函数的标量积形式

两维的例子：$\mathbf{x} = (x_1, x_2)$

构造：

$$\begin{aligned} k(\mathbf{x}, \mathbf{z}) &= (\mathbf{x}^{\mathrm{T}}\mathbf{z})^2 = (x_1 z_1 + x_2 z_2)^2 \\ &= x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= (x_1^2, \sqrt{2}x_1 x_2, x_2^2)(z_1^2, \sqrt{2}z_1 z_2, z_2^2)^{\mathrm{T}} \\ &= \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{z}). \end{aligned}$$

相当于 $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1 x_2, x_2^2)^{\mathrm{T}}$

则 $k(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^{\mathrm{T}}\mathbf{z})^2$ 是合格的核函数

直接构造并判断有效核函数

设有函数 $k(\mathbf{x}, \mathbf{x}')$ 如下定理说明其是否是核函数

Mercer定理（默瑟定理）

函数 $k(\mathbf{x}, \mathbf{x}')$ 是核函数的充分和必要条件是

对于任意样本集 $\{\mathbf{x}_{\mathrm{n}}\}_{n=1}^{N}$， $N < \infty$ 如下核矩阵

$$\mathbf{K} = [K_{nm}]_{N\times N} = [k(\mathbf{x}_n, \mathbf{x}_m)]_{N\times N}$$

是对称的半正定矩阵。

利用简单核构造新核函数

$k_1(\mathbf{x},\mathbf{x}')$ and $k_2(\mathbf{x},\mathbf{x}')$ 是有效核，如下构造核

$$
\begin{aligned}
k(\mathbf{x},\mathbf{x}') &= ck_1(\mathbf{x},\mathbf{x}') \\
k(\mathbf{x},\mathbf{x}') &= f(\mathbf{x})k_1(\mathbf{x},\mathbf{x}')f(\mathbf{x}') \\
k(\mathbf{x},\mathbf{x}') &= q\left(k_1(\mathbf{x},\mathbf{x}')\right) \\
k(\mathbf{x},\mathbf{x}') &= \exp\left(k_1(\mathbf{x},\mathbf{x}')\right) \\
k(\mathbf{x},\mathbf{x}') &= k_1(\mathbf{x},\mathbf{x}') + k_2(\mathbf{x},\mathbf{x}')
\end{aligned}
$$

$c > 0$

$f(\cdot)$ is any function.

$q(\cdot)$ is a polynomial with nonnegative coefficients

利用简单核构造新核函数

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}' \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b) \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b)
\end{aligned}
$$

$\phi(\mathbf{x})$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$

$\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and k_a and k_b are valid kernel

$\mathbf{A}$ is a symmetric positive semidefinite matrix.

简单核和构造核

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^{\mathrm{T}}\mathbf{x}'$$

$$k(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'\right)^2$$ 相当于$\phi(\mathbf{x})$是纯二阶函数

$$k(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'+c\right)^2 \quad c > 0$$

相当于$\phi(\mathbf{x})$是不高于二阶的函数（见下页）

$$k(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'\right)^M \quad M > 2$$

$$k(\mathbf{x}, \mathbf{x}') = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'+c\right)^M$$

一个核函数例子

$$
\begin{aligned}
& k(\mathbf{x}, \mathbf{z}) \\
= \quad & \left(1+\mathbf{x}^{\mathrm{T}} \mathbf{z}\right)^2=\left(1+x_1 z_1+x_2 z_2\right)^2 \\
= \quad & 1+2 x_1 z_1+2 x_2 z_2+x_1^2 z_1^2+2 x_1 z_1 x_2 z_2+x_2^2 z_2^2 \\
= \quad & \left(1, \sqrt{2} x_1, \sqrt{2} x_2, x_1^2, \sqrt{2} x_1 x_2, x_2^2\right)\left(1, \sqrt{2} z_1, \sqrt{2} z_2, z_1^2, \sqrt{2} z_1 z_2, z_2^2\right)^{\mathrm{T}} \\
= \quad & \phi(\mathbf{x})^{\mathrm{T}} \phi(\mathbf{z}) .
\end{aligned}
$$

构造核

高斯核 $$k(\mathbf{x},\mathbf{x}') = \exp\left(-\|\mathbf{x}-\mathbf{x}'\|^2/2\sigma^2\right)$$

由于

$$\|\mathbf{x}-\mathbf{x}'\|^2 = \mathbf{x}^{\mathrm{T}}\mathbf{x} + (\mathbf{x}')^{\mathrm{T}}\mathbf{x}' - 2\mathbf{x}^{\mathrm{T}}\mathbf{x}'$$

故高斯核写为

$$k(\mathbf{x},\mathbf{x}') = \exp\left(-\mathbf{x}^{\mathrm{T}}\mathbf{x}/2\sigma^2\right)\exp\left(\mathbf{x}^{\mathrm{T}}\mathbf{x}'/\sigma^2\right)\exp\left(-(\mathbf{x}')^{\mathrm{T}}\mathbf{x}'/2\sigma^2\right)$$

故高斯核确实是一个有效核

2* 线性基函数回归的对偶表示和核表示

正则化约束函数为

$$J(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\right\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

对**w**求导为**0**，得

$$\mathbf{w} = -\frac{1}{\lambda}\sum_{n=1}^{N}\left\{\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\right\}\phi(\mathbf{x}_n) =$$

$$= \sum_{n=1}^{N} a_n \phi(\mathbf{x}_n) = \mathbf{\Phi}^{\mathrm{T}}\mathbf{a}$$

这里

$$\mathbf{a} = (a_1, \ldots, a_N)^{\mathrm{T}}$$

线性回归的对偶表示和核表示（续）

上页中的

$$a_n = -\frac{1}{\lambda}\left\{\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) - t_n\right\}$$

将 $\mathbf{w} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{a}$ 带入 $J(\mathbf{w})$

得到对偶目标函数

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{a}$$

定义 $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ 和 $K_{nm} = \phi(\mathbf{x}_n)^{\mathrm{T}}\phi(\mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m)$

线性回归的对偶表示和核表示（续2）

对偶目标函数写成核形式

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a}$$

对a求导为0，得

$$\mathbf{a} = \left(\mathbf{K} + \lambda \mathbf{I}_N\right)^{-1}\mathbf{t}.$$

回归问题解的核表示为

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\phi(\mathbf{x})$$

$$= \sum_{n=1}^{N} a_n k\left(\boldsymbol{x}, \boldsymbol{x}_{\boldsymbol{n}}\right)$$

3.预备：不等式约束的拉格朗日方法 Lagrange Multipliers

$$\max_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\begin{aligned} s.t. \quad & g_j(\boldsymbol{x}) = 0 \qquad j = 1, 2, \cdots, J \\ & h_k(\boldsymbol{x}) \ge 0 \qquad k = 1, 2, \cdots, K \end{aligned}$$

令：

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{\boldsymbol{j=1}}^{\boldsymbol{J}} \lambda_{\boldsymbol{j}} g_j(\boldsymbol{x}) + \sum_{\boldsymbol{k=1}}^{\mathrm{K}} \mu_{\boldsymbol{k}} h_k(\boldsymbol{x})$$

不等式约束的拉格朗日方法（续）

最优解需解如下方程

$$\frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \mathbf{x}} = 0 \text{ 和，} \quad \frac{\partial L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})}{\partial \boldsymbol{\lambda}} = 0$$

并满足如下KKT（Karush-Kuhn-Tucker）条件

$$h_k(\boldsymbol{x}) \geq 0$$

$$\mu_{\mathbf{k}} \geq 0$$

$$\mu_{\mathbf{k}} h_k(\boldsymbol{x}) = 0$$

不等式约束的拉格朗日方法（续）

对函数的最小化，约束条件不变

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$s.t.\quad g_j(\boldsymbol{x}) = 0 \qquad j = 1, 2, \cdots, J$$

$$h_k(\boldsymbol{x}) \geq 0 \qquad k = 1, 2, \cdots, K$$

则拉格朗日函数为

$$L(\boldsymbol{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\boldsymbol{x}) + \sum_{j=1}^{J} \lambda_j g_j(\boldsymbol{x}) - \sum_{k=1}^{K} \mu_k h_k(\boldsymbol{x})$$

则求解方程和KKT条件不变。对 μ_k 参数求L函数最大。

可证明：以上最优问题的求解可分解为如下原解

$$\min_{\mathbf{x}}\left\{\max_{\mu,\lambda,\mu_{\mathbf{k}}\geq 0}\left[L\left(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}\right)\right]\right\}$$

对偶最优解为

$$\max_{\mu,\lambda,\mu_{\mathbf{k}}\geq 0}\left\{\min_{\mathbf{x}}\left[L\left(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\mu}\right)\right]\right\}$$

若有 $f\left(\mathbf{x}\right),h_k\left(\mathbf{x}\right)$ 是凸函数，$g_j(x)$ 是放射函数，且

$h_k\left(\mathbf{x}\right)>0$ 对一些$\boldsymbol{x}$

则：对偶解与原解同解

4. 支持向量机
Support Vector Machine：SVM

针对两类分类问题。线性分类函数

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b$$

分类输出

$$t = sign(y(\boldsymbol{x})) = sign(\boldsymbol{w}^T \phi(\boldsymbol{x}) + b)$$

训练集 $\{\boldsymbol{x}_\mathrm{n}, t_n\}_{n=1}^N \qquad t_n \in \{-1, 1\}$

线性分类：$\phi(\boldsymbol{x}) = \boldsymbol{x} = [x_1, x_2, \cdots, x_\mathrm{D}]^\mathrm{T}$

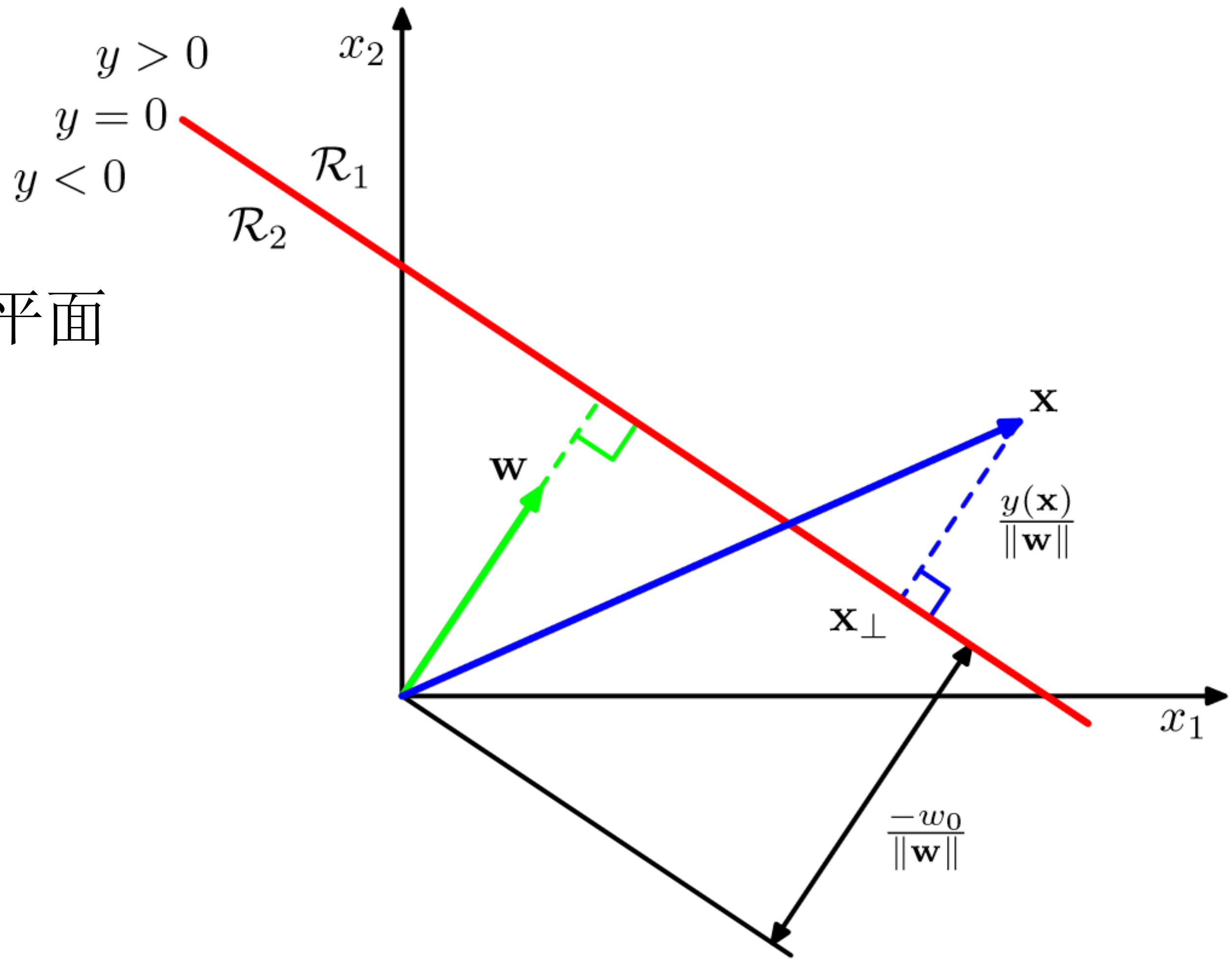
非线性（特征空间线性）分类：

$$\phi(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \cdots, \phi_\mathrm{M}(\boldsymbol{x})]^\mathrm{T}$$

预备知识：点到决策超平面距离

任意点 $\boldsymbol{x}$ 到决策超平面（红线）的距离

$$r = \frac{\left|y(\boldsymbol{x})\right|}{\|\boldsymbol{w}\|} = \frac{\left|\boldsymbol{w}^T\phi(\boldsymbol{x}) + b\right|}{\|\boldsymbol{w}\|}$$



请练习证明此式！

4.1 可分情况的SVM （线性可分或特征空间可分）

学习函数 $y(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + b$

分类边界 $y(\boldsymbol{x}) = 0$

训练集 $\{\boldsymbol{x}_\mathrm{n}, t_n\}_{n=1}^N$ $t_n \in \{-1, 1\}$

可分性指：存在一个分类边界，满足如下

对正样 $y(\boldsymbol{x}_{\boldsymbol{n}}) > 0$ $t_n = 1$

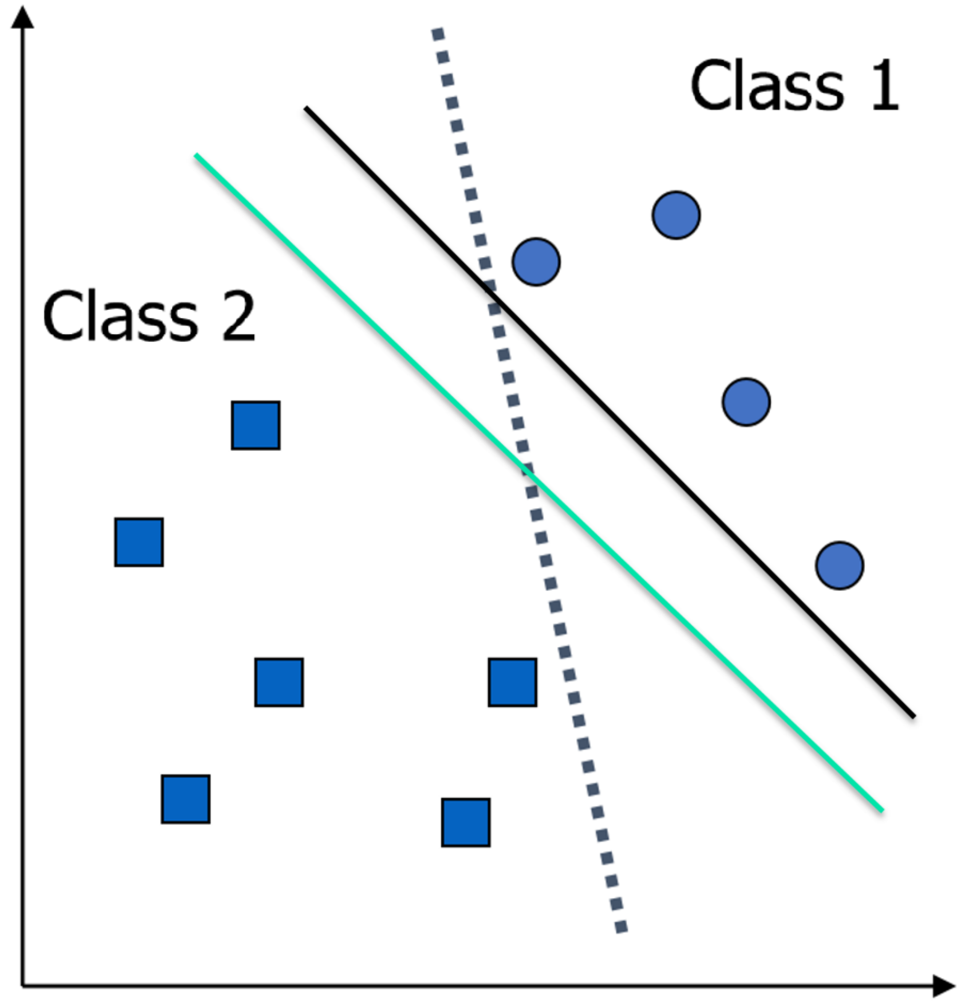
对负样 $y(\boldsymbol{x}_{\boldsymbol{n}}) < 0$ $t_n = -1$

对所有样本 $t_n y(\boldsymbol{x}_{\boldsymbol{n}}) > 0$

可分情况下，分界线不是唯一的，怎样的分界线具有最好泛化性

右边的三条分界线均可正确分类训练集，哪个具有最好泛化性?

感知机可能训练得到任何可能的分界线，受控于初始和样本次序



可分情况的SVM（续1）

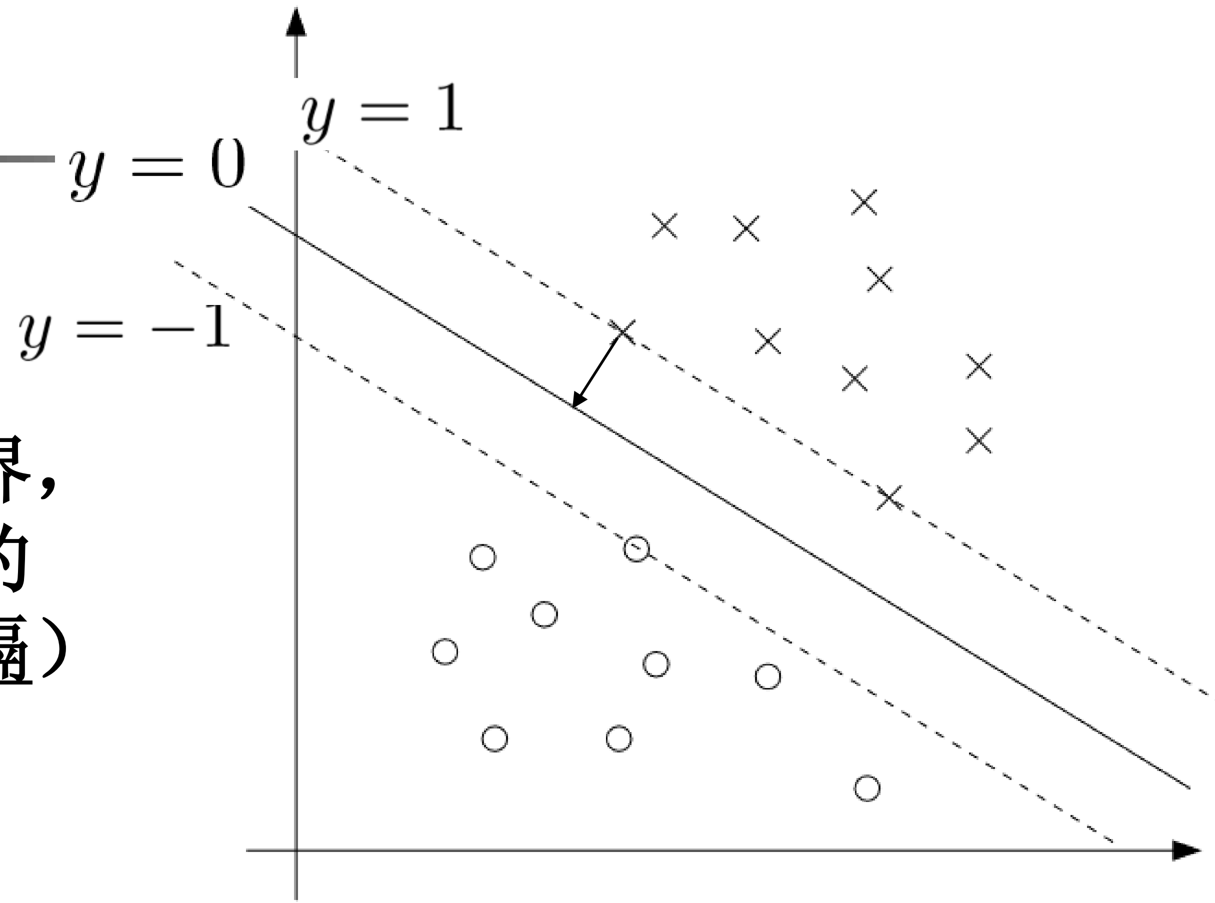
1.对于可分情况，存在一个分类边界，可正确分类所有训练样本，这种分类器不是唯一的。

2.一个更可信的分类器，使得样本点与边界距离尽可能远。

3.对于任意样本，其与分类边界的距离为

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}$$

可分情况的SVM（续2）



求一个分类边界，使样本离边界的最近距离（间隔）最大化，即

（#1）

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_{n}\left[t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\right]\right\}$$

可分情况的SVM（续3）

（#1）的优化问题难以求解。改变表示。

定义函数距离

$$|y(\mathbf{x}_\mathrm{n})| = t_n\left(\mathbf{w}^T\phi(\mathbf{x}_n) + b\right)$$

设 $\mathbf{w} \to c\mathbf{w}$， $c > 0$ 则函数距离改变c倍，距离不变。

强加最近距离点的函数距离为1，即

$$\min_n\left\{t_n\left(\mathbf{w}^T\phi(\mathbf{x}_n) + b\right)\right\} = 1$$

（#2）

可分情况的SVM（续4）

则所有样本满足约束条件

$$t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\geqslant 1,\qquad\qquad n=1,\ldots,N.$$

（#3）

（#1）的优化问题变为

$$\max_{\boldsymbol{w},b}\left\{\frac{1}{\|\boldsymbol{w}\|}\right\}$$

（#4）

可分情况的SVM（续5）

组合并重写（#3）（#4）为约束优化问题。

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

S.T. $$t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right) \geqslant 1$$

$$n = 1, \ldots, N$$

可分情况的SVM（续6）

为解约束最优，构造拉格朗日函数如下

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left\{t_n(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n) + b) - 1\right\}$$

（#5）

$a_n \geqslant 0$ 拉格朗日乘数

$$\mathbf{a} = (a_1, \ldots, a_N)^{\mathrm{T}}$$

可分情况的SVM（续7）

令： $$\frac{\partial L(\mathbf{w}, b, \boldsymbol{a})}{\partial \mathbf{w}} = 0\,, \quad \frac{\partial L(\mathbf{w}, b, \boldsymbol{a})}{\partial b} = 0$$

得：

$$\mathbf{w} = \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n)$$

$$0 = \sum_{n=1}^{N} a_n t_n.$$

（#6）

可分情况的SVM（续8）

将（#6）带入（#5）得到对偶表示，最大化

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

（#7）

约束条件

$$a_n \geqslant 0, \qquad n = 1, \ldots, N,$$

$$\sum_{n=1}^{N} a_n t_n = 0.$$

其中 $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{x}')$ 为核函数

可分情况的SVM（续9）

以上的优化问题，称为二次规划问题。
可用二次规划标准算法实现，运算复杂度$\mathrm{O}\left(M^3\right)$

参数 $\{a_n\}$ 求得后，将（#6）中$\boldsymbol{w}$ 带入，得

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b.$$

分类输出

$$t(\boldsymbol{x}) = sign(y(\boldsymbol{x}))$$

可分情况的SVM（续10）

SVM优化结果分析

解满足的KKT条件为

$$
\begin{aligned}
a_n &\geqslant 0 \\
t_n y(\mathbf{x}_n) - 1 &\geqslant 0 \\
a_n \{t_n y(\mathbf{x}_n) - 1\} &= 0.
\end{aligned}
$$

由KKT条件3

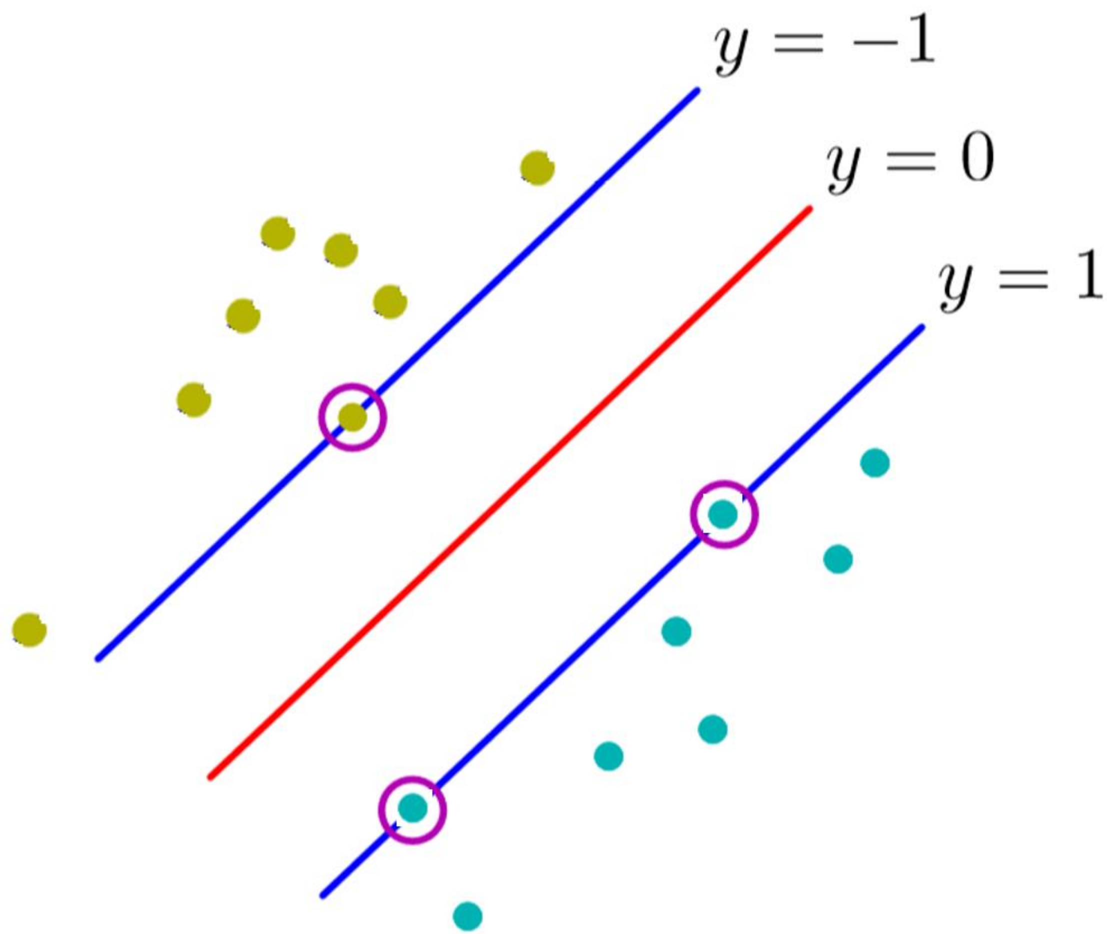
或 $a_n = 0$ 相应样本对结果无贡献

或 $t_n y(\mathbf{x}_n) = 1$ $a_n > 0$

离分类边界最近距离点，**<u>支持向量</u>**

可分情况的SVM（续11）

仅由支持向量
表示输出分类表达式
支持向量数量较少
这是一种稀疏表示。



可分情况的SVM（续12）

设支持向量序号集合为 $\mathcal{S}$

对一个支持向量 $\mathbf{x}_n$ 有 $t_n y(\mathbf{x}_n) = 1$，故

$$t_n \left(\sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1$$

两边同时乘 tn

得 $$b = \frac{1}{N_{\mathcal{S}}} \sum_{n \in \mathcal{S}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

至此所有参数均求得。

可分情况的SVM（续13）

例子：取高斯核函数，则特征空间可分

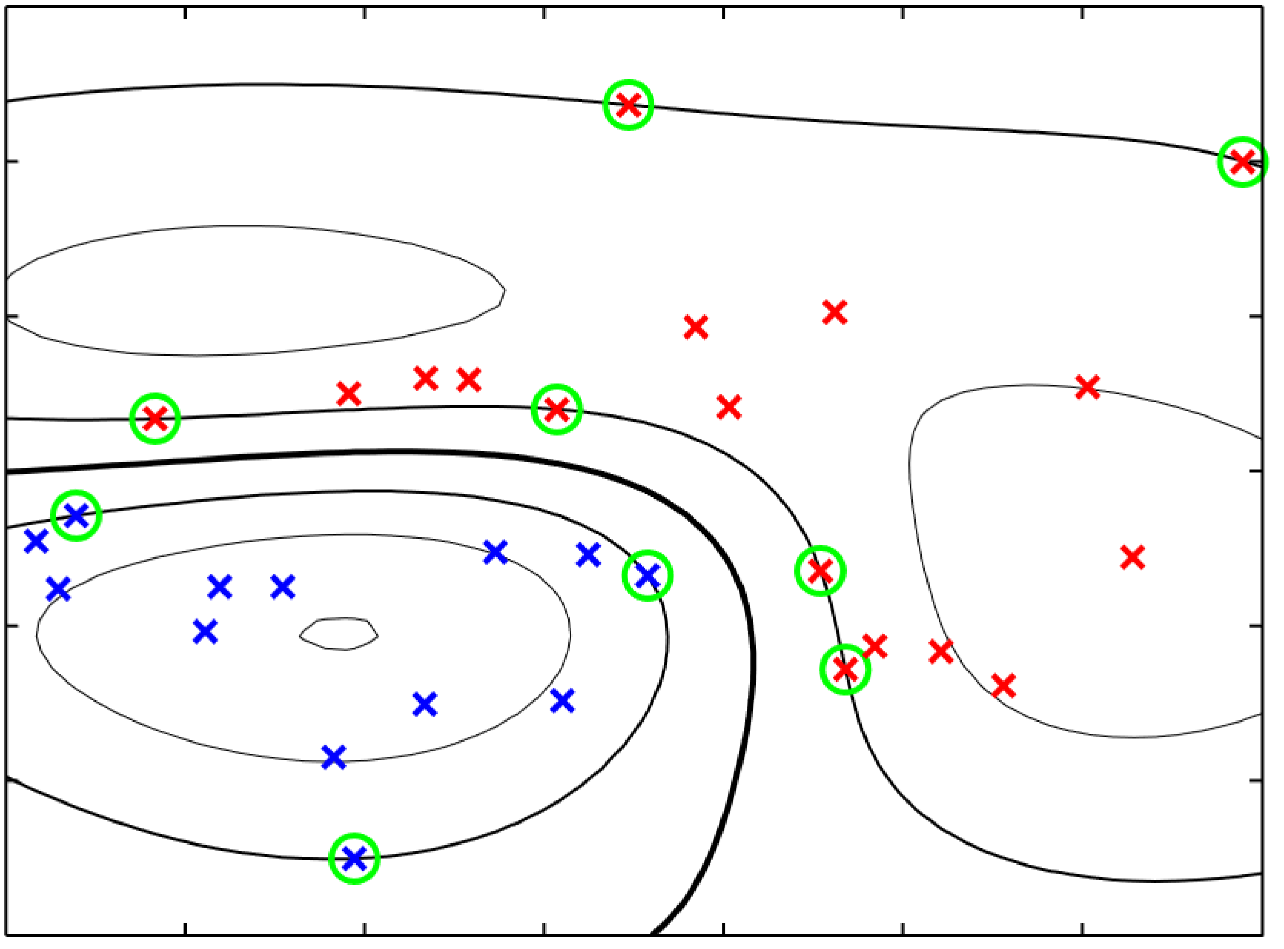
图中标出：

分界线

最小距离线（Margin边界）

$y(\mathbf{x})$等值线

支持向量



4.2 不可分情况的SVM

对于不可分的一般情况，改进SVM方法引入松弛变量（Slack Variables）

$$\xi_n \geqslant 0 \quad n = 1, \ldots, N \qquad (\#1)$$

对于处于最小距离线（margin）上或正确一侧

$$\xi_n = 0$$

对于其他，则令

$$\xi_n = |t_n - y(\mathbf{x}_n)|$$

对于处于分界线 $y(\mathbf{x}_n) = 0$ 上的点，则 $\xi_n = 1$

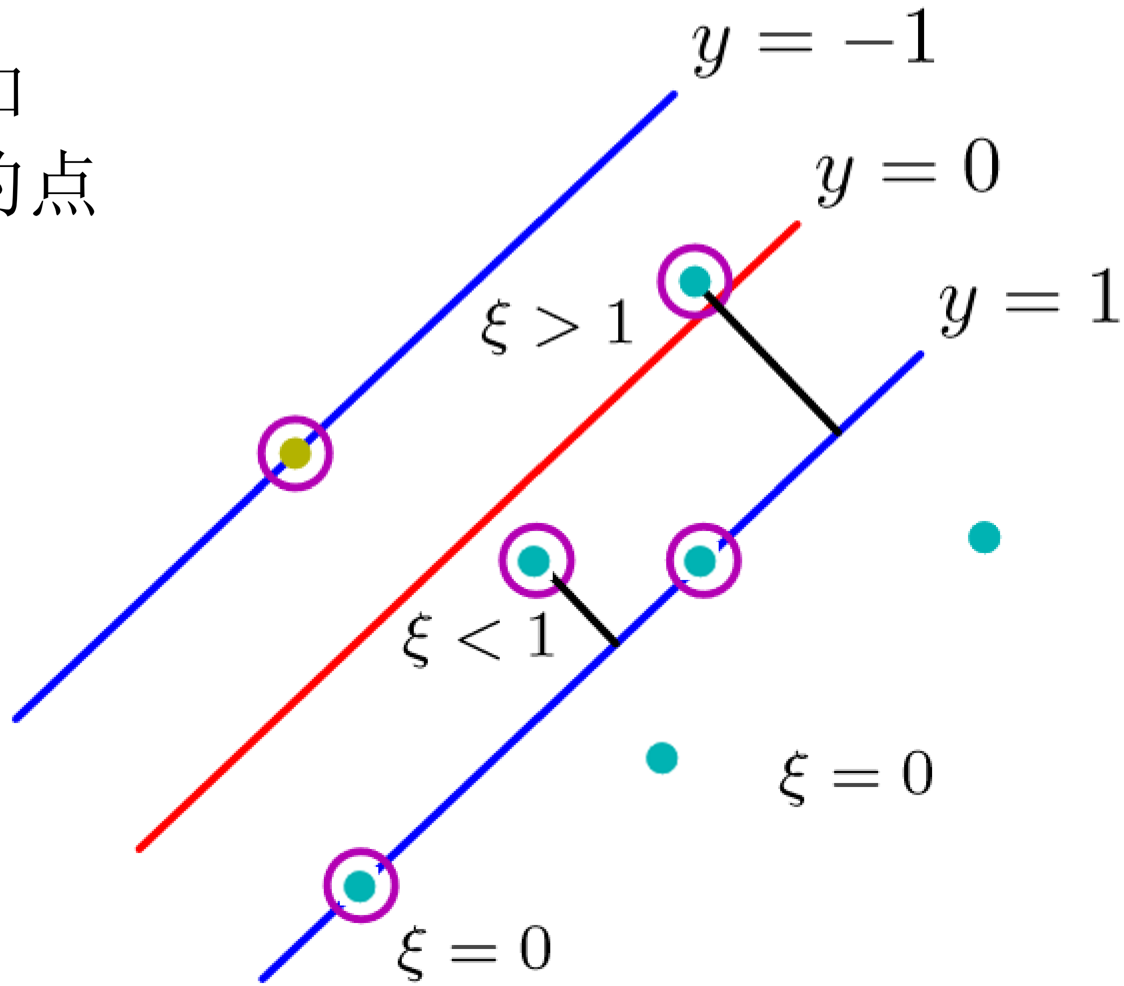
不可分情况的SVM（续1）

对于处于边界线和最小距离线之间的点

$$0 < \xi_n \leqslant 1$$

对于处于边界线错误一侧

$$\xi_n > 1$$



不可分情况的SVM（续2）

考虑松弛，将可分情况的约束

$$t_n y(\mathbf{x}_n) \geqslant 1$$

推广到带松弛约束

$$t_n y(\mathbf{x}_n) \geqslant 1 - \xi_n, \qquad n = 1, \ldots, N \qquad \text{（\#2）}$$

soft margin 推广到如下约束最小化

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{（\#3）}$$

$C > 0$

控制超参数

不可分情况的正则化解释

引入正则化项 $\xi_n \geqslant 0$ ，允许部分样本不满足

$$t_n y(\mathbf{x}_n) \geqslant 1$$

放松到满足

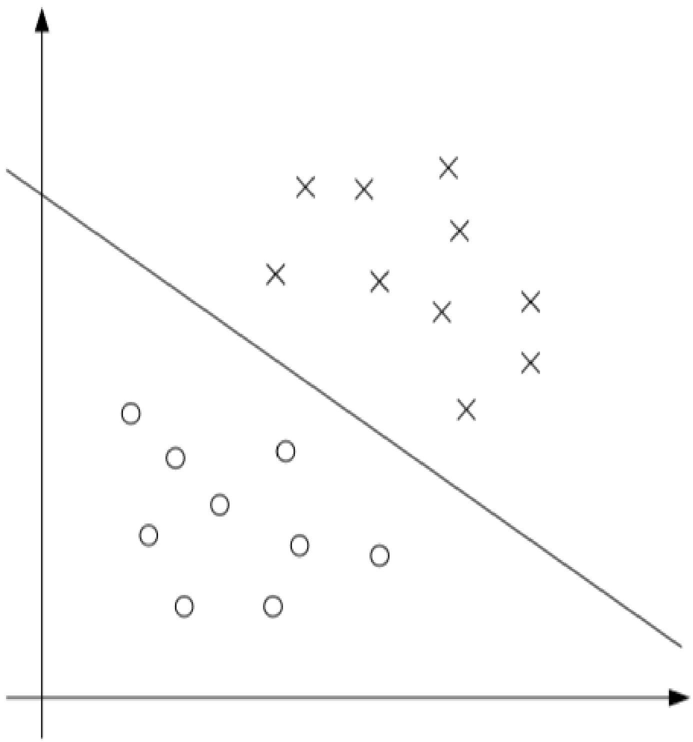
$$t_n y(\mathbf{x}_n) \geqslant 1 - \xi_n, \qquad n = 1, \ldots, N$$

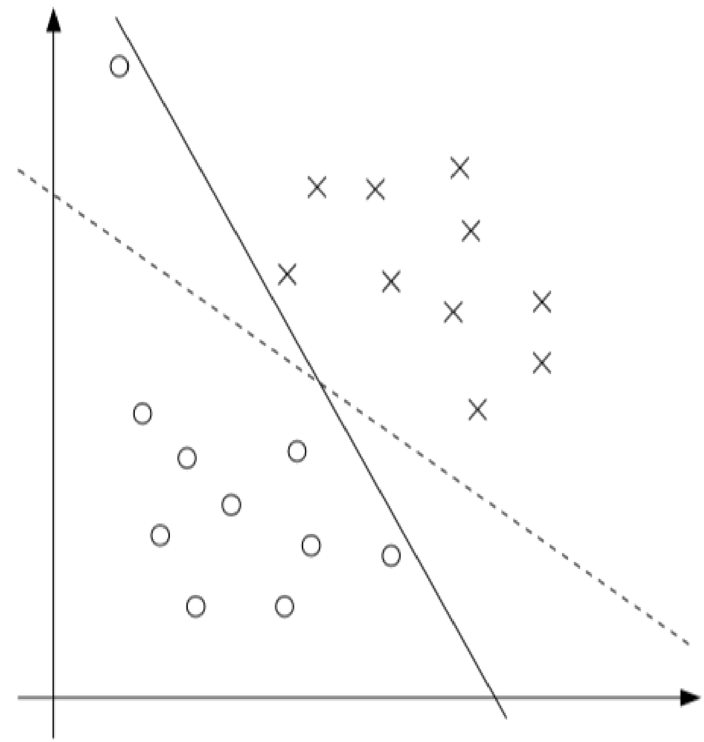
则L1正则化的目标函数为最小化：

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

不可分情况的正则化解释

- 正则化方法除了对不可分有效，即使可分情况，也可避免为了适应野值样本，而使得分界线的泛化性能变差，如下右图有一个野值样本，为了保证野值样本分类正确，泛化性明显变差。





不可分情况的SVM（续3）

考虑（#3）的目标函数和（#1）（#2）约束拉格朗日函数为

$$L(\mathbf{w}, b, \mathbf{a}) =$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N}\xi_n - \sum_{n=1}^{N}a_n\left\{t_n y(\mathbf{x}_n) - 1 + \xi_n\right\} - \sum_{n=1}^{N}\mu_n\xi_n$$

拉格朗日因子

（#4）

$\{a_n \geqslant 0\}$ and $\{\mu_n \geqslant 0\}$

不可分情况的SVM（续4）

KKT条件

$$
\begin{aligned}
a_n &\geqslant 0 \\
t_n y(\mathbf{x}_n) - 1 + \xi_n &\geqslant 0 \\
a_n \left(t_n y(\mathbf{x}_n) - 1 + \xi_n\right) &= 0 \\
\mu_n &\geqslant 0 \\
\xi_n &\geqslant 0 \\
\mu_n \xi_n &= 0
\end{aligned}
$$

不可分情况的SVM（续5）

令 $$\frac{\partial L}{\partial \boldsymbol{w}} = 0, \quad \frac{\partial L}{\partial b} = 0, \quad \frac{\partial L}{\partial \xi_{\mathrm{n}}} = 0$$

得 $$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\mathbf{x}_n)$$

$$\frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} a_n t_n = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \quad \Rightarrow \quad a_n = C - \mu_n. \qquad \Rightarrow 0 \le a_n \le C$$

（利用KKT条件第4行）

不可分情况的SVM（续6）

上页式带入拉格朗日公式，得到拉格朗日对偶问题，最大化下式

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

S. T.

$$0 \leqslant a_n \leqslant C$$

盒约束

$$\sum_{n=1}^{N} a_n t_n = 0$$

注意：此处优化问题与可分情况的优化目标式相同，约束条件不同，解二次规划问题。

不可分情况的SVM（续7）

带入 $\boldsymbol{w}$ ，得输出表达式与可分情况相同，即

$$y(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}, \mathbf{x}_n) + b.$$

分类输出

$$t(\boldsymbol{x}) = sign(y(\boldsymbol{x}))$$

不可分情况的SVM（续8）

不可分情况，SVM优化结果分析

由KKT条件

或 $a_n = 0$ 相应样本对结果无贡献

或 $a_n > 0$ 则 $t_n y(\mathbf{x}_n) = 1 - \xi_n$ **支持向量**

若 $a_n < C$ 则 $\mu_n > 0$ 故 $\xi_n = 0$ 落在最小边界线的点

若 $a_n = C$ 则 $\xi_n \leqslant 1$ 落在最小边界线外正确分类

$\xi_n > 1$ 落在分界线外错误分类

不可分情况的SVM（续9）

确定 b，对于满足如下集合的点

$0 < a_n < C$ have $\xi_n = 0$ so that $t_n y(\mathbf{x}_n) = 1$

以上点集记为 $\mathcal{M}$ ，支持向量集 $\mathcal{S}$

$$t_n \left(\sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) + b \right) = 1 \qquad \text{对 } \mathcal{M}$$

故：

$$b = \frac{1}{N_{\mathcal{M}}} \sum_{n \in \mathcal{M}} \left(t_n - \sum_{m \in \mathcal{S}} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right)$$

4.3 不可分SVM的另一种解释

定义：合页损失函数（hinge loss function）

$$L_h(z) = \max(0, 1 - z)$$

则有：

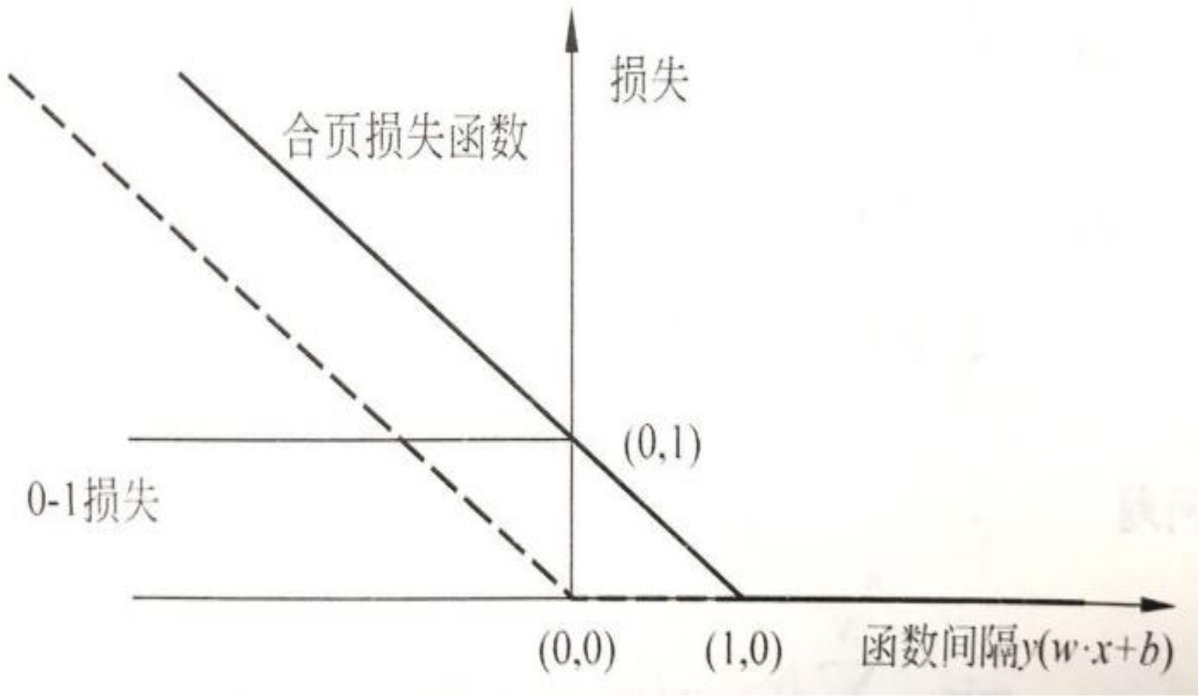
$$\sum_{n=1}^{N} \xi_n = \sum_{n=1}^{N} L_h\left(t_n\left(\boldsymbol{w}^T \boldsymbol{x}_n + b\right)\right)$$

SVM等价为正则化合页目标函数：

$$\min_{\boldsymbol{w},b}\left\{\sum_{n=1}^{N} L_h\left(t_n\left(\boldsymbol{w}^T \boldsymbol{x}_n + b\right)\right) + \lambda \|\boldsymbol{w}\|^2\right\}$$

$$= \min_{\boldsymbol{w},b}\left\{\sum_{n=1}^{N} \max\left[0, 1 - t_n\left(\boldsymbol{w}^T \boldsymbol{x}_n + b\right)\right] + \lambda \|\boldsymbol{w}\|^2\right\}$$

损失函数的例子：合页、感知机、0-1



一种变化的SVM：ν-SVM

最大化

$$\widetilde{L}(\mathbf{a}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

S. T.

$$0 \leqslant a_n \leqslant 1/N$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

$$\sum_{n=1}^{N} a_n \geqslant \nu.$$

ν 是超参数
表示 $\xi_n > 0$
的点的比例

5. 多分类SVM（Multiclass SVM)

SVM是本身针对2分类问题的，是一种决策方法多分类需要各种扩展方法。

一种K类分类器： one-versus-the-rest

$y_k(\mathbf{x})$

用SVM分两类，K作为正样，其他K-1作为负类

分类输出

$$y(\mathbf{x}) = \max_k y_k(\mathbf{x})$$

其他方法，各有一些利弊

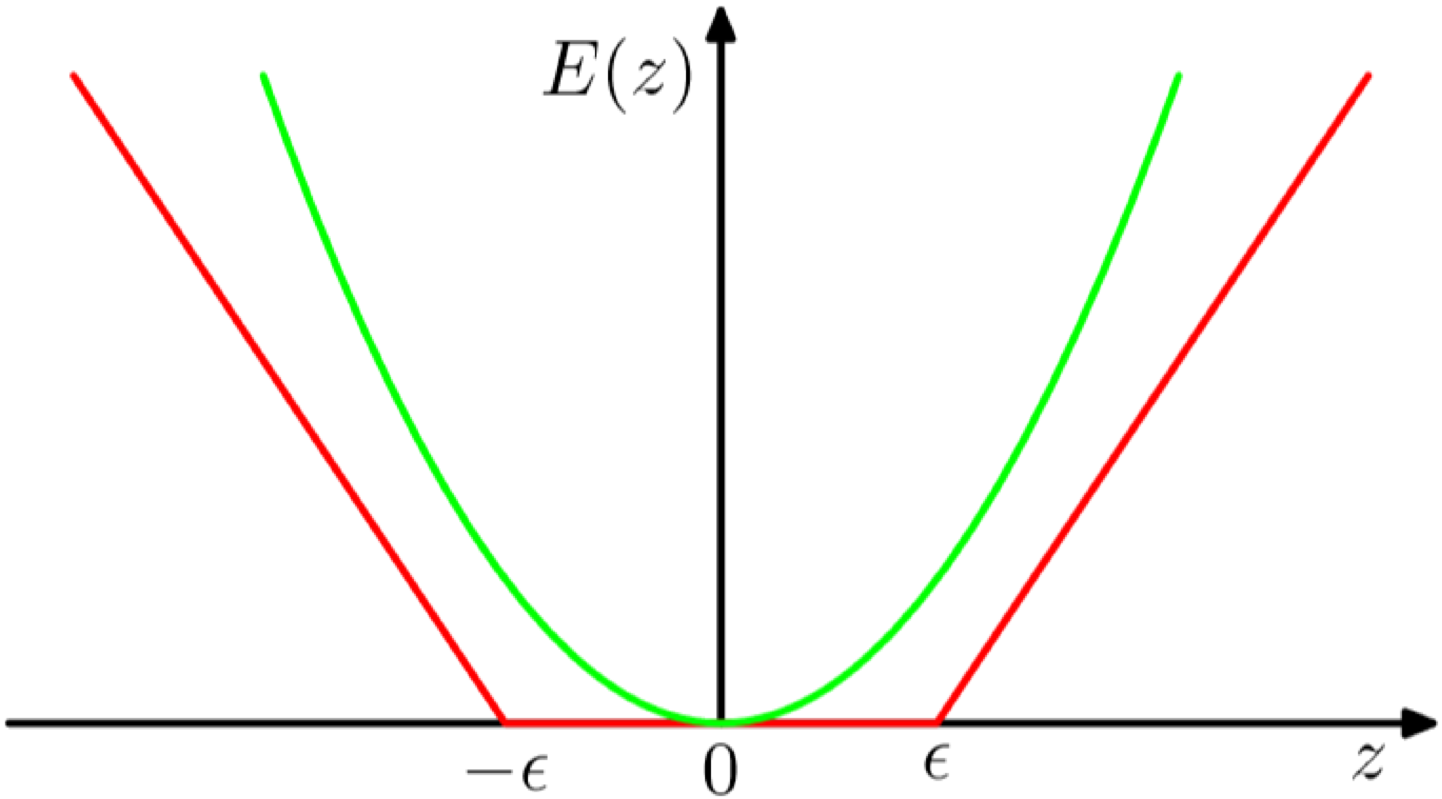
6. 回归SVM（SVMs for Regression）

为了获得稀疏表示，定义新的误差函数，取代误差平方和。ε-不敏感误差函数 ε-insensitive error function

$$E_\epsilon(y(\mathbf{x})-t)=\begin{cases}0, & \text{if } |y(\mathbf{x})-t|<\epsilon; \\ |y(\mathbf{x})-t|-\epsilon, & \text{otherwise}\end{cases}$$

正则化误差函数为（C为反正则化系数）

$$C\sum_{n=1}^{N}E_\epsilon(y(\mathbf{x}_n)-t_n)+\frac{1}{2}\|\mathbf{w}\|^2 \qquad \text{(\#1)}$$



ε-不敏误差函数与平方误差比较

回归SVM（续1）

对每一个点 $\mathbf{x}_n$ 定义两个松弛变量

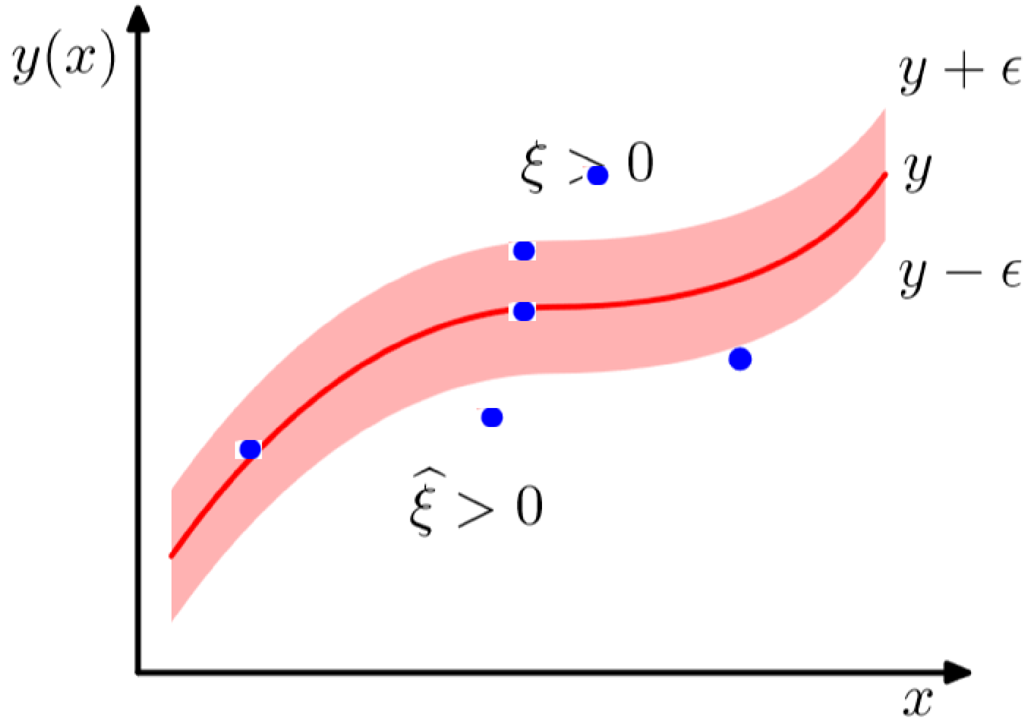
$$\xi_n \geqslant 0 \quad 和 \quad \widehat{\xi}_n \geqslant 0 \qquad （\#2）$$

$$\xi_n > 0 \quad 对应 \quad t_n > y(\mathbf{x}_n) + \epsilon,$$

$$\widehat{\xi}_n > 0 \quad 对应 \quad t_n < y(\mathbf{x}_n) - \epsilon,$$

$$y_n - \epsilon \leqslant t_n \leqslant y_n + \epsilon, \quad 对应 \quad \epsilon\text{-tube}$$

回归SVM（续2）



通过松弛变量，所有点满足下式

$$
\begin{aligned}
t_n &\leqslant y(\mathbf{x}_n) + \epsilon + \xi_n \\
t_n &\geqslant y(\mathbf{x}_n) - \epsilon - \widehat{\xi}_n.
\end{aligned} \qquad (\#3)
$$

回归SVM（续3）

由松弛变量的定义，（#1）目标函数改写为

$$C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}\|\mathbf{w}\|^2$$

结合（#2）（#3）的约束，得拉格朗日函数为

$$\begin{aligned} L \; = \; & C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n)+\frac{1}{2}\|\mathbf{w}\|^2-\sum_{n=1}^{N}(\mu_n\xi_n+\widehat{\mu}_n\widehat{\xi}_n) \\ & -\sum_{n=1}^{N}a_n(\epsilon+\xi_n+y_n-t_n)-\sum_{n=1}^{N}\widehat{a}_n(\epsilon+\widehat{\xi}_n-y_n+t_n) \end{aligned}$$

（#4）

参数为 $a_n \geqslant 0, \quad \widehat{a}_n \geqslant 0, \mu_n \geqslant 0, \quad \widehat{\mu}_n \geqslant 0$

回归SVM（续）

计算各导数

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} (a_n - \widehat{a}_n)\phi(\mathbf{x}_n)$$

$$\frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{n=1}^{N} (a_n - \widehat{a}_n) = 0$$

$$\frac{\partial L}{\partial \xi_n} = 0 \quad \Rightarrow \quad a_n + \mu_n = C$$

$$\frac{\partial L}{\partial \widehat{\xi}_n} = 0 \quad \Rightarrow \quad \widehat{a}_n + \widehat{\mu}_n = C. \qquad \text{（\#5）}$$

回归SVM（续4）

将（#5）带入拉格朗日表达式，得到对偶表示，即最大化下式

$$\widetilde{L}(\mathbf{a},\widehat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n-\widehat{a}_n)(a_m-\widehat{a}_m)k(\mathbf{x}_n,\mathbf{x}_m) - \epsilon\sum_{n=1}^{N}(a_n+\widehat{a}_n)+\sum_{n=1}^{N}(a_n-\widehat{a}_n)t_n$$

并有盒约束 （#6）

$$0 \leqslant a_n \leqslant C$$
$$0 \leqslant \widehat{a}_n \leqslant C$$

和

$$\sum_{n=1}^{N}(a_n-\widehat{a}_n)=0$$

回归SVM（续5）

解以上二次规划问题，求得 a_n $\widehat{a}_n$

由 $\mathbf{w}$ 的表达式，得回归的预测公式为

$$y(\mathbf{x}) = \sum_{n=1}^{N} (a_n - \widehat{a}_n) k(\mathbf{x}, \mathbf{x}_n) + b$$

回归SVM（续6）

KKT条件，

结合KKT条件可对回归SVM的结果做分析。

$$
\begin{aligned}
a_n(\epsilon + \xi_n + y_n - t_n) &= 0 \\
\widehat{a}_n(\epsilon + \widehat{\xi}_n - y_n + t_n) &= 0 \\
(C - a_n)\xi_n &= 0 \\
(C - \widehat{a}_n)\widehat{\xi}_n &= 0.
\end{aligned}
$$

回归SVM（续7）

回归SVM结果分析

$a_n \neq 0$ 必有 $\epsilon + \xi_n + y_n - t_n = 0$,

$(\xi_n = 0)$ 处在 ϵ-tube 的上边界上

$(\xi_n > 0)$ 处在 ϵ-tube 的上边界上侧

$\widehat{a}_n \neq 0$ 必有 $\epsilon + \widehat{\xi}_n - y_n + t_n = 0$

$\widehat{\xi}_n$ 为0，处在下边界上，>0处于下边界下侧

$\widehat{a}_n \neq 0$ 或 $a_n \neq 0$ 的点称为**支持向量**

（两者不会同时非0）

回归SVM（续8）

求偏置参数b

对于支持向量中一些点 $0 < a_n < C$

则 $\xi_n = 0$ 故 $\epsilon + y_n - t_n = 0$

得：
$$\begin{aligned} b &= t_n - \epsilon - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) \\ &= t_n - \epsilon - \sum_{m=1}^{N}(a_m - \widehat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) \end{aligned}$$

对 $0 < \hat{a}_n < C$ 的点也求得一个b，两者平均的结果更好

一种变化的回归SVM：回归ν-SVM

一种变化的回归SVM描述如下

最大化：

$$\widetilde{L}(\mathbf{a}, \widehat{\mathbf{a}}) = -\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}(a_n - \widehat{a}_n)(a_m - \widehat{a}_m)k(\mathbf{x}_n, \mathbf{x}_m) + \sum_{n=1}^{N}(a_n - \widehat{a}_n)t_n$$

S.T.

$$0 \leqslant a_n \leqslant C/N$$

$$0 \leqslant \widehat{a}_n \leqslant C/N$$

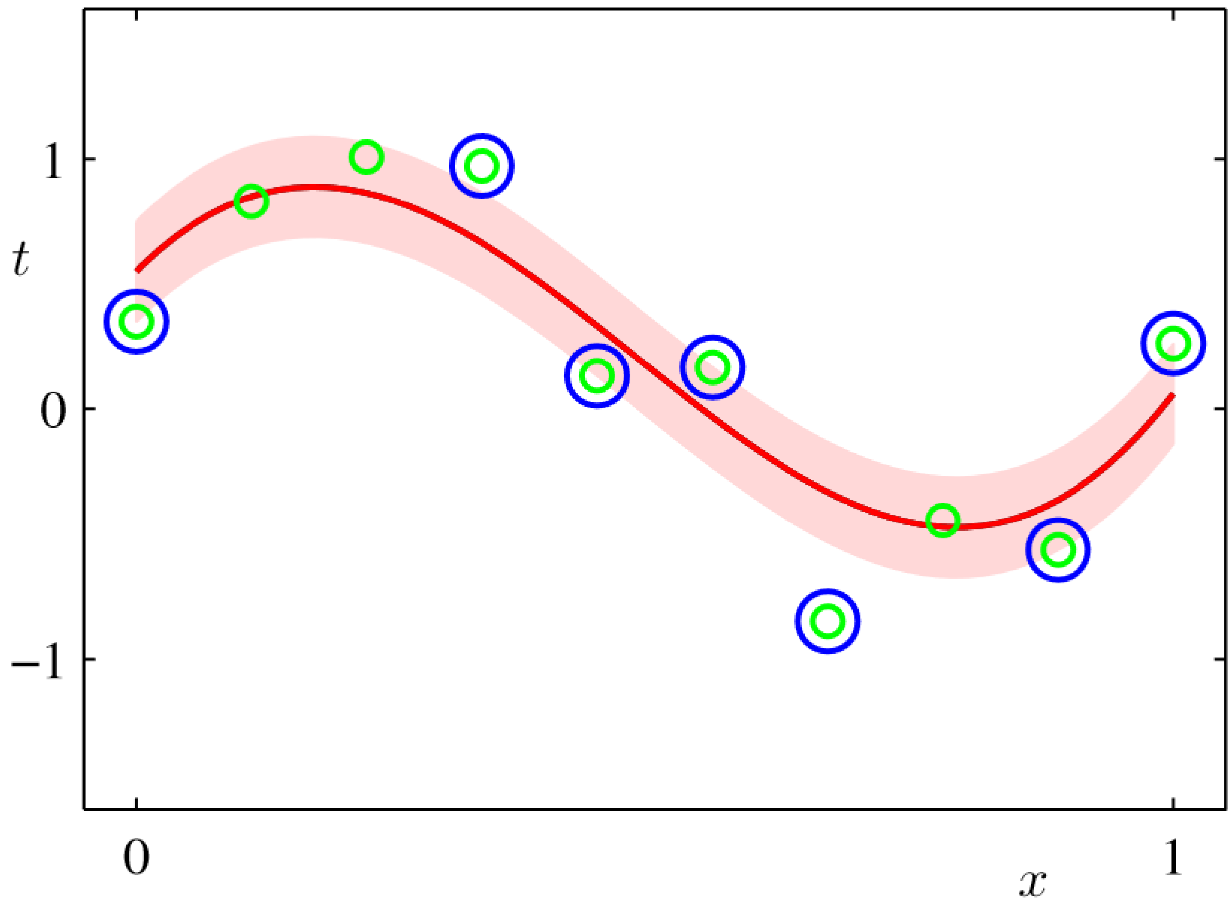
$$\sum_{n=1}^{N}(a_n - \widehat{a}_n) = 0$$

$$\sum_{n=1}^{N}(a_n + \widehat{a}_n) \leqslant \nu C.$$

用ν参数代替ε参数，有更直观性

回归ν-SVM的例子

蓝圈表示支持向量使用的是高斯核



7. SVM的一个优化算法：SMO Sequential Minimal optimization

预备知识：坐标上升算法（**Coordinate Ascent：CA**）最大化如下目标函数

$$L\left(a_1, \cdots, a_i, \cdots, a_N\right)$$

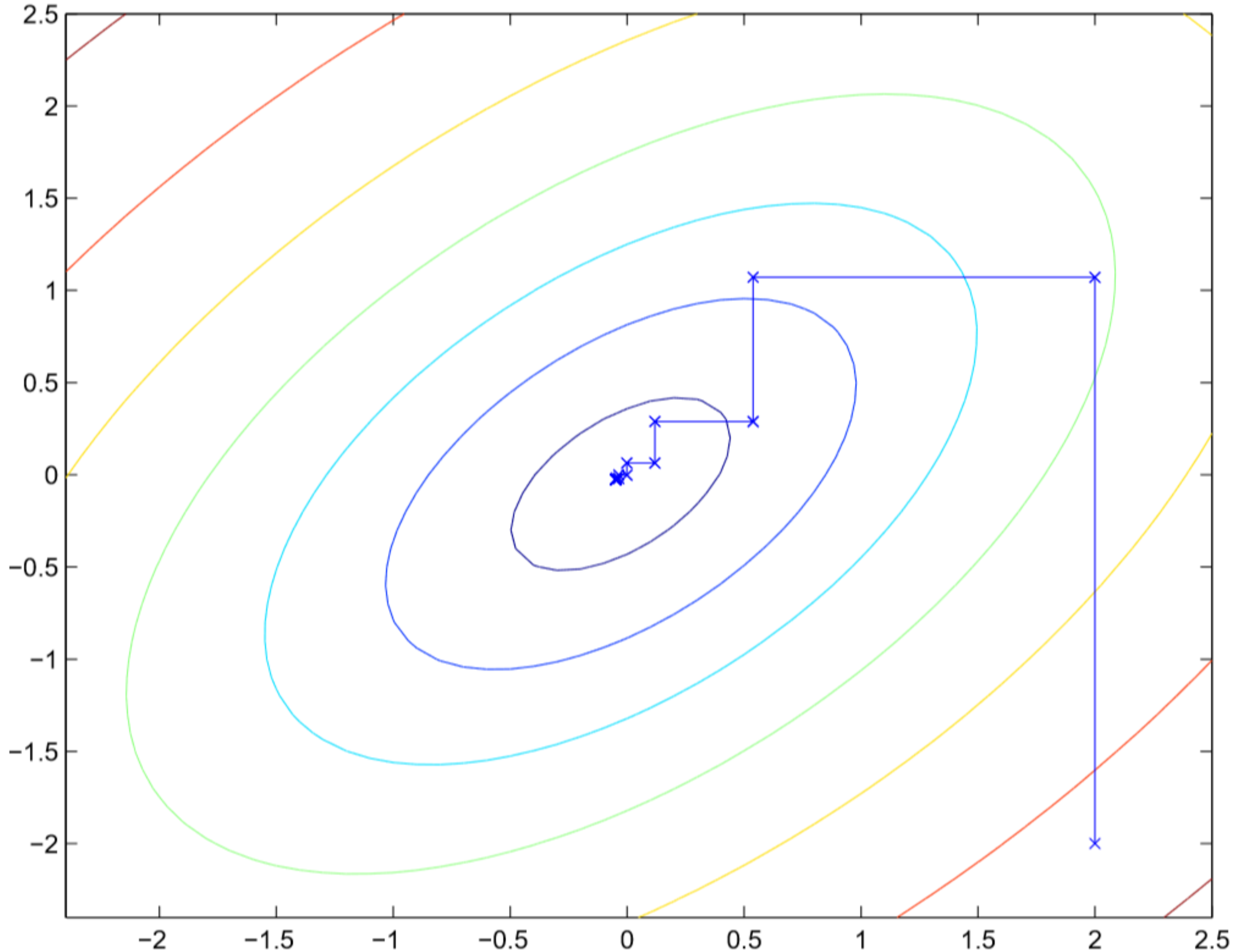
每次固定其他变量，只求对变量 a_i 的优化，即

$$a_i = \max_{\hat{a}_i}\left\{L\left(a_1, \cdots, \hat{a}_i, \cdots, a_N\right)\right\}$$

该过程序列循环执行，直到收敛！

示例CA

两维情况下
CA的示例



SVM的一个优化算法：SMO（续1）

CA优化算法，应用于SVM优化，即如下问题

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m)$$

S. T.

$$0 \leqslant a_n \leqslant C$$

$$\sum_{n=1}^{N} a_n t_n = 0$$

由于约束可改写为

$$a_i t_i = -\sum_{\substack{j=1 \\ j \neq i}}^{N} a_j t_j$$

故用CA算法对SVM优化， 一次至少优化两个系数，固定其他系数

SVM的一个优化算法：SMO（续2）

每次选两个系数，不失一般性，选 a_1, a_2

由约束条件

$$a_1 t_1 + a_2 t_2 = -\sum_{j=3}^{N} a_j t_j = \varsigma$$

ς 在算法中，由于其他系数假设是常数，故是常数

$$a_1 = (\varsigma - a_2 t_2) t_1$$

带入目标函数

$$\tilde{L}((\varsigma - a_2 t_2) t_1, a_2, \cdots, a_i, \cdots, a_N)$$

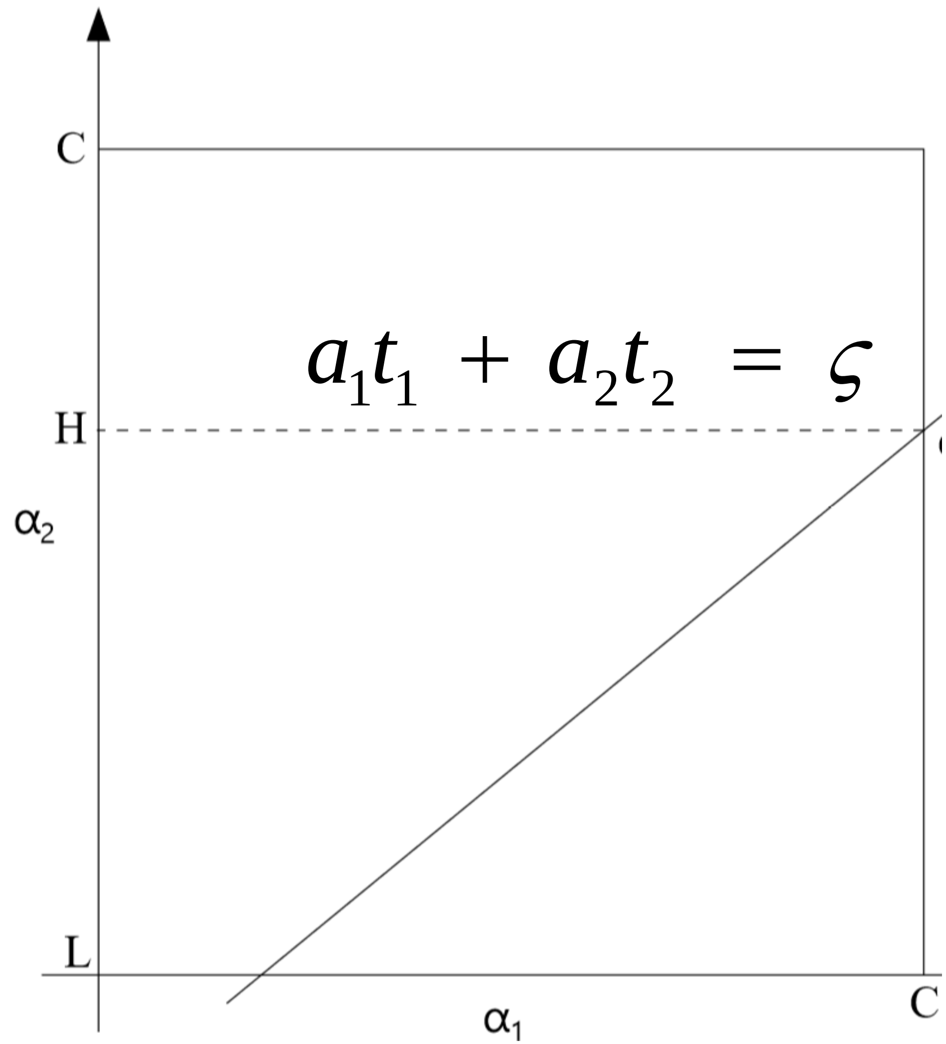
$$= A a_2^2 + B a_2 + C$$

SVM的一个优化算法：SMO（续3）

$$\frac{\partial \tilde{L}}{\partial a_2} = 0 \Rightarrow \text{求得：} a_2^{new,uc}$$

注意，由于盒约束和约束方程，得

$$L \le a_2 \le H$$



SVM的一个优化算法：SMO （续4）

综合考虑，得到解

$$a_2^{new} = \begin{cases} H & a_2^{new,uc} > H \\ a_2^{new,uc} & L \le a_2^{new,uc} \le H \\ L & a_2^{new,uc} < L \end{cases}$$

$$a_1^{new} = \left(\varsigma - a_2^{new} t_2\right) t_1$$

SVM的一个优化算法：SMO （续5）

每次的一对优化变量的选取采用启发式，
首先选择一个不满足KKT条件的点对应的系数
再选择可能引起大的改变的系数组成一对，
循环迭代，直到收敛。

算法的细节描述参考[John Platt,1999]

SVM 有SVM Light、LIBSVM等专用软件包，
主要通用平台也支持SVM

SVM的一点说明

1. 在**SVM**中，最容易应用的是线性**SVM**方法，即核函数取

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}'$$

2. 在比较复杂的分类问题中，怎样选择好的核函数，是一个没有解决的问题，人为因素和启发式方法为主。

3. 大规模问题**SVM**的优化有若干研究，**SMO**是其中之一。