人工智能 I

杨杰

yangjie@seu.edu.cn

2025/11/17，五四楼-303，9:50~11:25

本周四实验课安排

周次	9	
日期	2025/11/20	
星期	星期四	
时间	16:40-18:15	19:00-20:35
机房地点	中心楼实验室D区	五五楼435软实七
实验任务	支持向量机	

注意两个时间段是不同机房！请勿跑错！

注意18:15机房电脑自动关机！请及时保存！

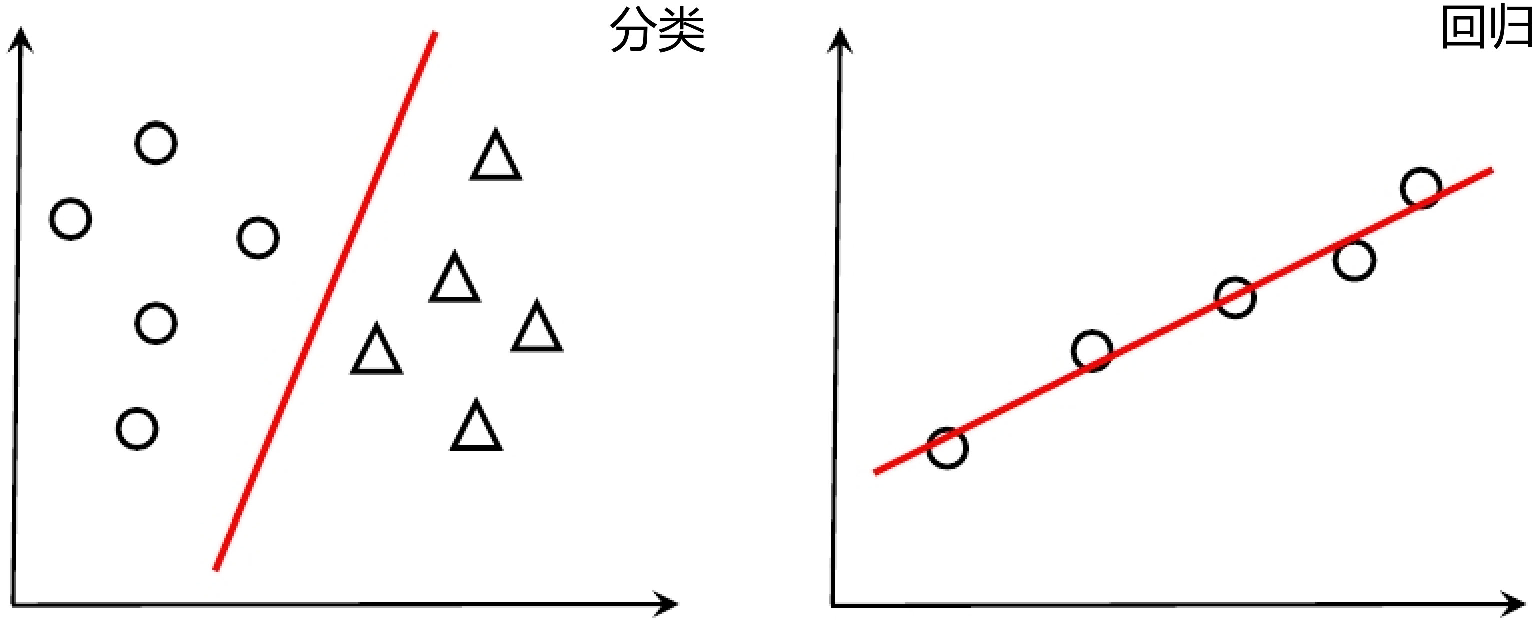
本节课安排

- 监督学习：
  - 线性回归
  - 对率回归
  - 支持向量机
  - 决策树
  - 随机森林
  - 贝叶斯分类器
  - 感知机与神经网络

本节课安排

- 监督学习：
  - 线性回归
  - 对率回归
  - 支持向量机
  - 决策树
  - 随机森林
  - 贝叶斯分类器
  - 感知机与神经网络

线性模型



线性模型(linear model)试图学得一个通过属性的**线性组合**来进行预测的函数

$$f(x) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b$$

向量形式： $f(x) = w^{\mathrm{T}} x + b$

简单、基本、可理解性好

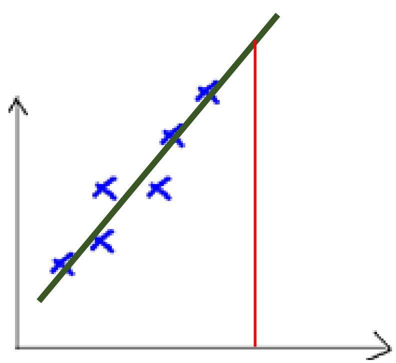
一元(uni-variate)线性回归

- 示例：根据年龄预测小孩身高

年龄（月）	身高（cm）
12	75
24	88
36	98
48	101
…	…

给定一组历史数据，如何预测未来身高？

解决方法：用一条直线去尽量准的拟合这些数据，然后如果有新的输入过来，我们可以在将直线上这个点对应的值返回。



一元(uni-variate)线性回归

- 形式化描述
  - 训练数据

$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\cdots$$
$$(x_m, y_m)$$

年龄（月）	身高（cm）
12	75
24	88
36	94
48	101
…	…

  - 预测模型

$$f(x) = wx + b, \qquad w \in \mathbb{R}, b \in \mathbb{R}$$

  - 学习算法

最小化MSE，有：$(w^*, b^*) =$

$$\arg\min_{w,b} \underbrace{\sum_{i=1}^{m} (y_i - wx_i - b)^2}_{E(w,b)}$$

一元(uni-variate)线性回归

分别对w和b求导：

$$\frac{\partial E(w,b)}{\partial w} = 2\left(w\sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m}(y_i - b)x_i\right)$$

$$\frac{\partial E(w,b)}{\partial b} = 2\left(mb - \sum_{i=1}^{m}(y_i - wx_i)\right)$$

令导数为0，得到闭式(closed-form)解：

$$w^* = \frac{\sum_{i=1}^{m} y_i(x_i - \bar{x})}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m}\left(\sum_{i=1}^{m} x_i\right)^2}$$

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

$$b^* = \frac{1}{m}\sum_{i=1}^{m}(y_i - w^* x_i)$$

多元(multi-variate)线性回归

- 数据

$$(\mathbf{x}_1, y_1)$$
$$(\mathbf{x}_2, y_2)$$
$$\cdots\cdots$$
$$(\mathbf{x}_m, y_{\mathrm{m}})$$

$$\mathbf{x}_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{di} \end{bmatrix} \in \mathbb{R}^d, y_i \in \mathbb{R}$$

- 预测模型

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b, \qquad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} \in \mathbb{R}^d, b \in \mathbb{R}$$

- 学习算法

$$(\mathbf{w}^*, b^*) = \arg\min_{(\mathbf{w},b)} \sum_{i=1}^{m} (f(\mathbf{x}_i) - y_i)^2$$

多元(multi-variate)线性回归

- 推导过程

$$X = \begin{bmatrix} \mathbf{x}_1 & \cdots & \mathbf{x}_m \\ 1 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dm} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{(d+1)\times m}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m \qquad \tilde{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\sum_{i=1}^{m}(f(\mathbf{x}_i) - y_i)^2 = \| X^T\tilde{\mathbf{w}} - \mathbf{y} \|_2^2$$

令导数为0, 得： $XX^T\tilde{\mathbf{w}} = X\mathbf{y}$

- 若XX^T满秩或正定，则$\tilde{\mathbf{w}}^* = \left(XX^T\right)^{-1}X\mathbf{y}$
- 若XX^T不满秩，则存在多个解，需引入正则化(regularization)

正则化（regularization）

- 前面所接触的学习模型：

$$\min_f \sum_{i=1}^{m} \ell(f(\mathbf{x}_i), y_i)$$

- 更一般的学习模型：

λ：超参数

$$\min_f \underbrace{\sum_{i=1}^{m} \ell(f(\mathbf{x}_i), y_i)}_{\text{经验风险}} + \lambda \underbrace{\Omega(f)}_{\text{结构风险}}$$

经验风险
(empirical risk)
描述模型与训练数据的契合程度

结构风险
(structural risk)
描述模型本身的某些性质

- 正则化可理解为"罚函数法"
  - 通过对不希望的结果施以惩罚，使得优化过程趋向于希望目标
  - 归纳偏好

带正则化的多元线性回归

- 2-范数正则化：

$$\tilde{\mathbf{w}}^* = \arg\min_{\tilde{\mathbf{w}}} \| X^T\tilde{\mathbf{w}} - \mathbf{y} \|_F^2 + \lambda \| \tilde{\mathbf{w}} \|_2^2$$

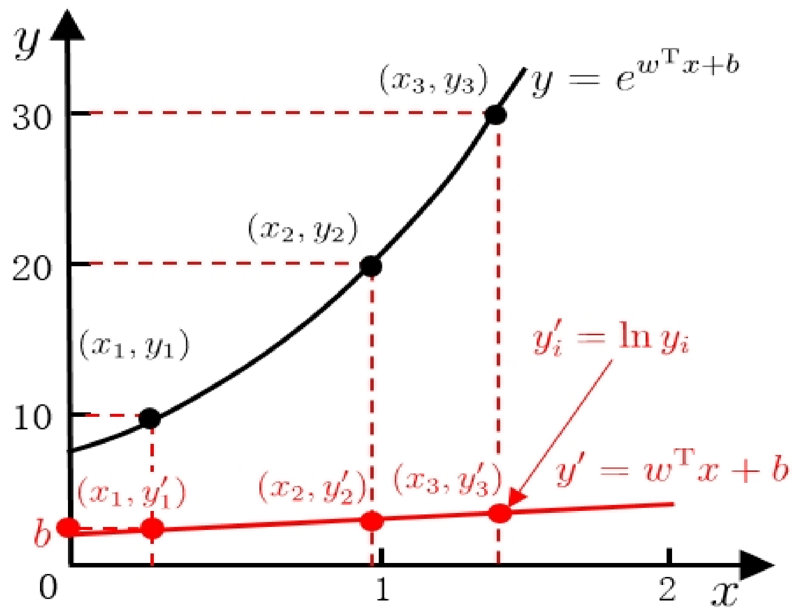
$$\tilde{\mathbf{w}}^* = \left(XX^T + \lambda I\right)^{-1} X\mathbf{y}$$

$$XX^T = U \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_{d+1} \end{bmatrix} U^T$$

$$XX^T + \lambda I = U \begin{bmatrix} \sigma_1 + \lambda & 0 & 0 & 0 \\ 0 & \sigma_2 + \lambda & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sigma_{d+1} + \lambda \end{bmatrix} U^T$$

广义线性回归

- 线性模型无法拟合非线性数据（欠拟合）



令预测值逼近y的衍生物？

对数线性回归

$$\ln y = \mathbf{w}^T\mathbf{x} + b$$

等价于用$e^{\mathbf{w}^T\mathbf{x}+b}$逼近$y$

- 广义线性回归

$$y = \underbrace{g}_{}^{-1}(\mathbf{w}^T\mathbf{x} + b)$$

单调可微的 联系函数(link function)

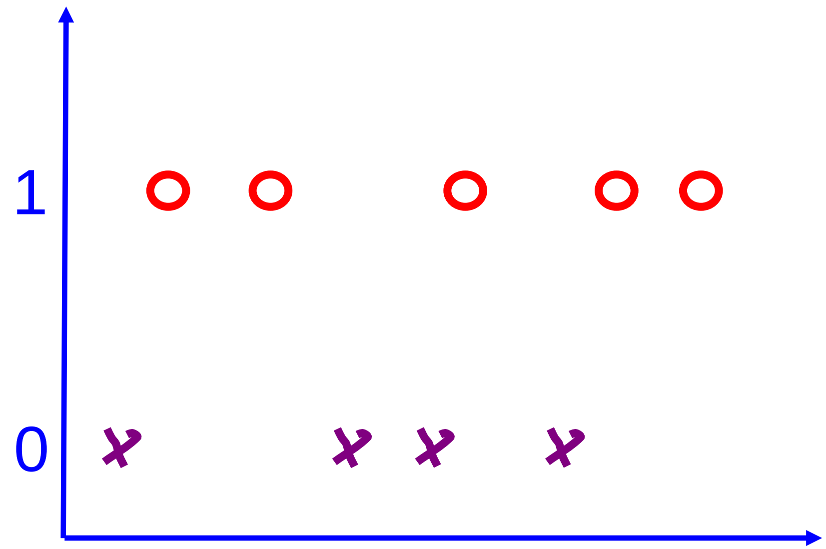
本节课安排

- 监督学习：
  - 线性回归
  - 对率回归
  - 支持向量机
  - 决策树
  - 随机森林
  - 贝叶斯分类器
  - 感知机与神经网络

二分类问题

年龄（月）	偏高 or 偏矮
12	0
24	1
36	1
48	0
...	...

直接应用线性回归模型，效果差



线性回归模型产生的输出：$z = \mathbf{w}^T\mathbf{x} + b$

期望输出：$y \in \{0,1\}$ or $\{-1,1\}$

找z和y的联系函数

$y \in \{-,+\}$

用于分类的联系函数

线性回归模型产生的输出：$z = \mathbf{w}^T\mathbf{x} + b$

期望输出：$y \in \{0,1\}$

找z和y的联系函数

单位阶跃函数
(unit-step function)：

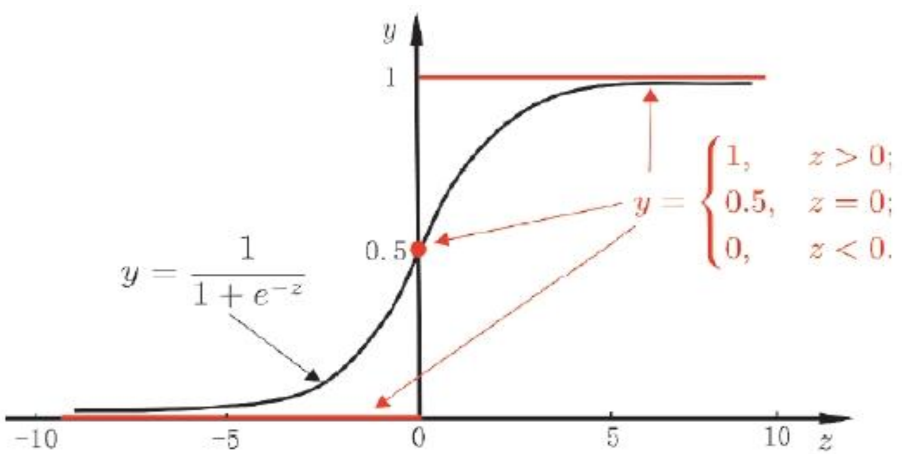
$$y = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1, & z > 0 \end{cases}$$

性质不好，
需找“替代函数”
(surrogate function)

对数几率函数
(logistic function)
简称对率函数

$$y = \frac{1}{1 + e^{-z}}$$

单调可微、任意阶可导



对率回归

对率回归(Logistic Regression)，也称逻辑回归：

以对率函数为联系函数的线性回归

$$y = \frac{1}{1+e^{-z}} \Rightarrow y = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$$

注意：对率回归是分类算法

可变形为：

$$\log \frac{y}{1-y} = \mathbf{w}^T\mathbf{x} + b$$

几率(odds), 反映了 x 作为正例的相对可能性

会出现溢出

计算仍然存在困难，直接应用线性回归算法不可取！

求解思路

将类标$y \in \{0,1\}$看作类后验概率值 $y = P(c = +|\mathbf{x})$

预测模型： $P_{\mathbf{w}}(c = +|\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{w}^T\mathbf{x}+b)}}$

于是，可使用“极大似然估计法”

(maximum likelihood method)

已知某组独立同分布样本满足某概率分布$P_{\boldsymbol{\theta}}(\mathbf{a})$，但是其中参数$\boldsymbol{\theta}$未知。极大似然估计是参数估计的方法之一：

寻找参数使这组样本出现的概率最大!

$$\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} P_{\boldsymbol{\theta}}(\mathbf{a}_i) \Longleftrightarrow \max_{\boldsymbol{\theta}} \log\left(\prod_{i=1}^{m} P_{\boldsymbol{\theta}}(\mathbf{a}_i)\right)$$

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log\left(P_{\boldsymbol{\theta}}(\mathbf{a}_i)\right)$$

求解思路

记$\mathbf{w}=[\mathbf{w};b],\mathbf{x}=[\mathbf{x};1]$, 于是

$$P_{\mathbf{w}}(c=+|\mathbf{x})=\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}=\frac{e^{\mathbf{w}^T\mathbf{x}}}{1+e^{\mathbf{w}^T\mathbf{x}}}$$

$$P_{\mathbf{w}}(c=-|\mathbf{x})=1-P(c=+|\mathbf{x})=\frac{1}{1+e^{\mathbf{w}^T\mathbf{x}}}$$

$$\max_{\mathbf{w}}\sum_{i=1}^{m}\log\big(P_{\mathbf{w}}(c_i|\mathbf{x}_i)\big) \Longleftrightarrow \min_{\mathbf{w}}-\sum_{i=1}^{m}\log\big(P_{\mathbf{w}}(c_i|\mathbf{x}_i)\big)$$

$$\min_{\mathbf{w}}\sum_{i=1}^{m}\left(-y_i\mathbf{w}^T\mathbf{x}_i+\log(1+e^{\mathbf{w}^T\mathbf{x}_i})\right)$$

注意：$y_i\in\{0,1\}$

高阶可导连续凸函数，可用经典的数值优化方法, 如梯度下降法/牛顿法

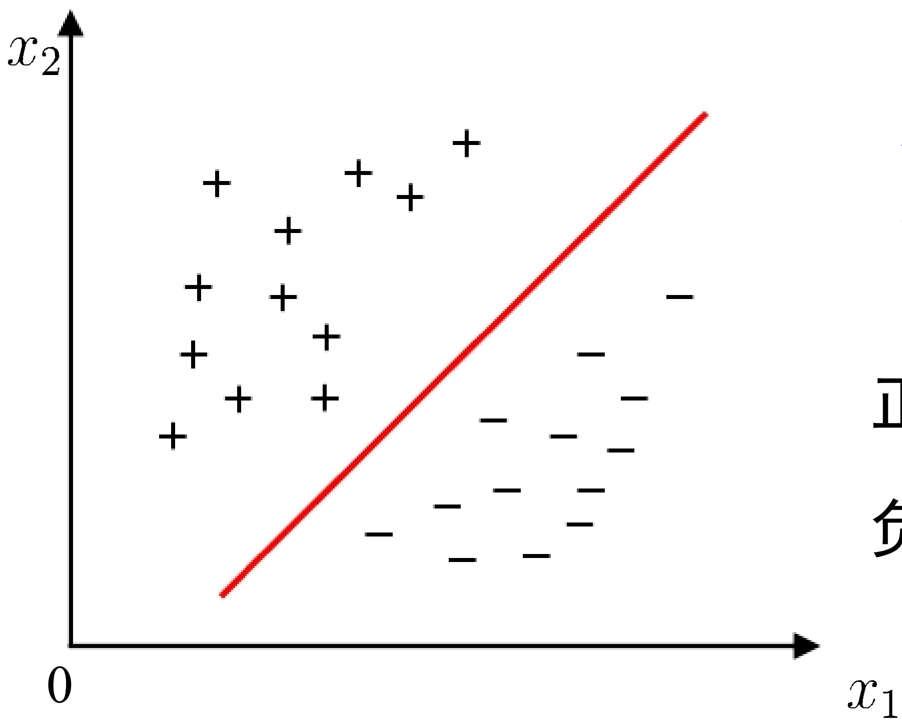
本节课安排

- 监督学习：
  - 线性回归
  - 对率回归
  - 支持向量机
  - 决策树
  - 随机森林
  - 贝叶斯分类器
  - 感知机与神经网络

线性模型做分类

- 对率回归：
  - 通过“联系函数”，把分类问题转化为回归问题
- 如何直接做分类？

在样本空间中寻找一个超平面，将不同类别的样本分开



超平面：$\mathbb{R}^d$的$d-1$维仿射子空间

$$\{\mathbf{x}|\mathbf{w}^T\mathbf{x}+b=0,\mathbf{x}\in\mathbb{R}^d\}$$

正类：$\{\mathbf{x}|\mathbf{w}^T\mathbf{x}+b>0,\mathbf{x}\in\mathbb{R}^d\}$

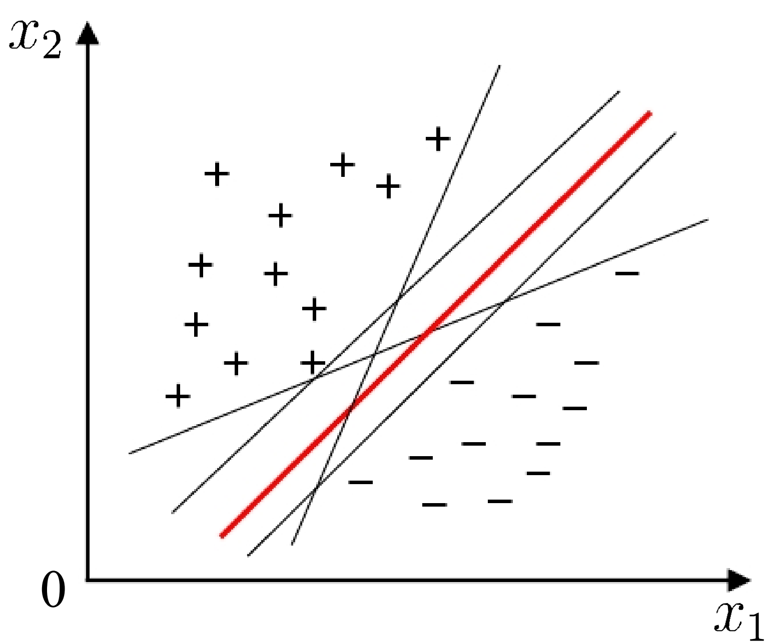
负类：$\{\mathbf{x}|\mathbf{w}^T\mathbf{x}+b<0,\mathbf{x}\in\mathbb{R}^d\}$

线性模型做分类

- 假设类标为{-1, +1}且训练数据线性可分，那么所有能正确划分训练数据的超平面可表示为：

$$\left.\begin{aligned}&\text{正样本}(y_i=1):\ \mathbf{w}^T\mathbf{x}_i+b\geq 1\\&\text{负样本}(y_i=-1):\ \mathbf{w}^T\mathbf{x}_i+b\leq -1\end{aligned}\right\}\ \begin{aligned}y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1,\\ i=1,\cdots,m\end{aligned}$$

将训练样本分开的超平面可能有很多，哪一个更好呢？



直观上应选择"正中间"，容忍性好，鲁棒性高，泛化能力最强.

支持向量与间隔

超平面方程：

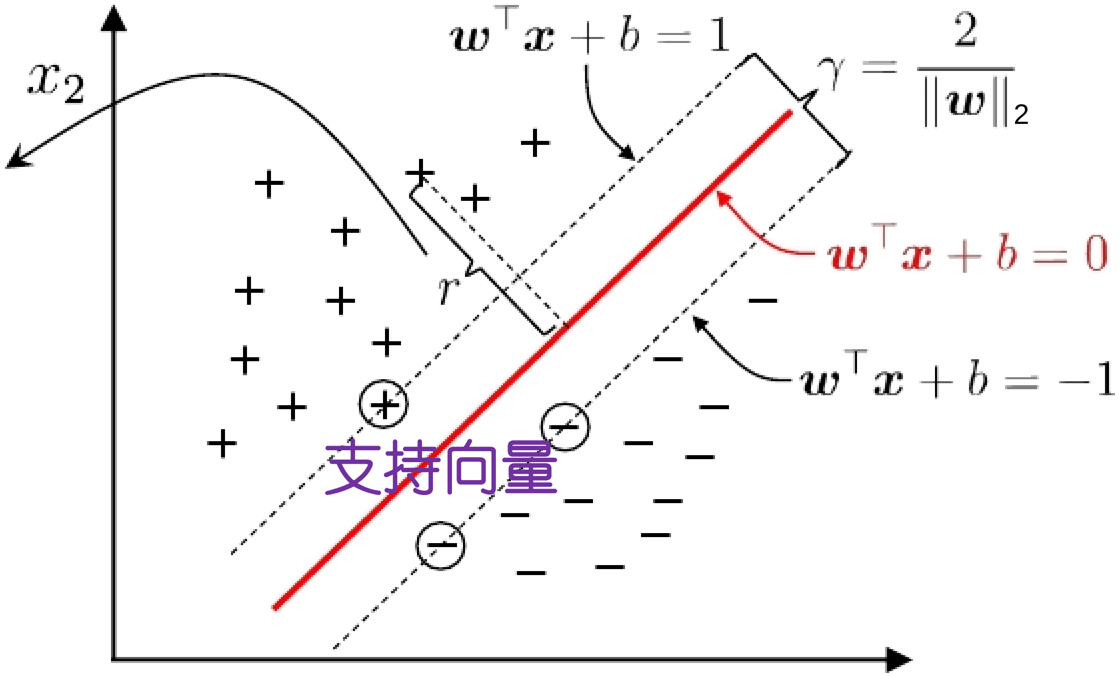
$\mathbf{w}^T\mathbf{x} + b = 0$ $\qquad$ $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, i = 1, \cdots, m$

距离超平面最近的几个训练样本使得等式成立，它们被称为"支持向量" (Support Vector)

两个异类支持向量到超平面的距离之和为

$$\gamma = \frac{2}{\| \mathbf{w} \|_2}$$

"间隔" (Margin)



SVM的基本模型

最大间隔(Large Marin)：

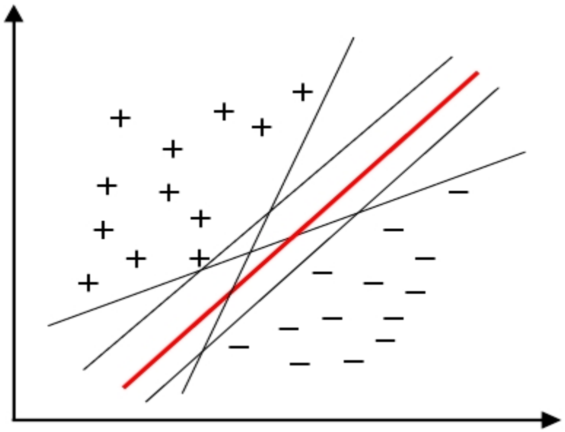
$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|_2}$$

$$\text{s.t.}, \ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \ i = 1, \cdots, m$$

$$\Updownarrow$$

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2$$

$$\text{s.t.}, \ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \ i = 1, \cdots, m$$



“正中间”的泛化能力最强

- 凸二次规划问题，能用优化计算包求解。
- 若$d > m$, 使用拉格朗日乘子法效率更高

带有约束的优化问题

$$\text{带有约束的优化问题}\begin{cases} \min_{x\in D} f(x) \\ \text{s.t. } g_i(x) \le 0, i=1,2,\ldots,q \\ h_j(x)=0, j=q+1,\ldots,m \end{cases}$$

其中f(x)是目标函数，g(x)为不等式约束，h(x)为等式约束。

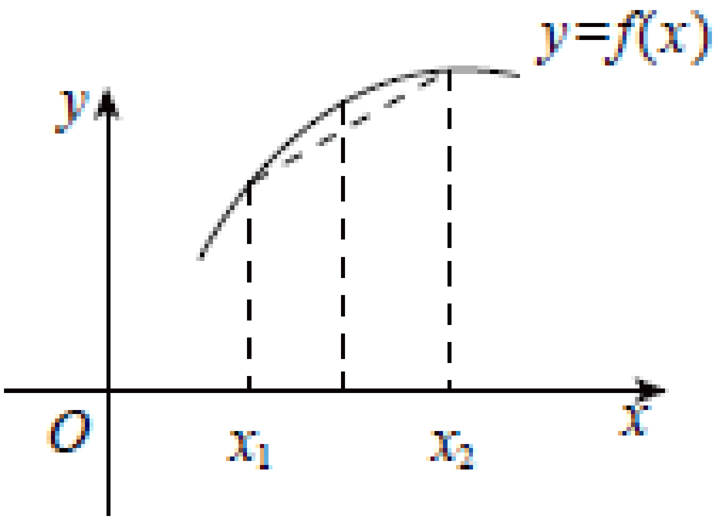
若f(x)，h(x)，g(x)三个函数都是线性函数，则该优化问题称为线性规划。若任意一个是非线性函数，则称为非线性规划。

若目标函数为二次函数，约束全为线性函数，称为二次规划。

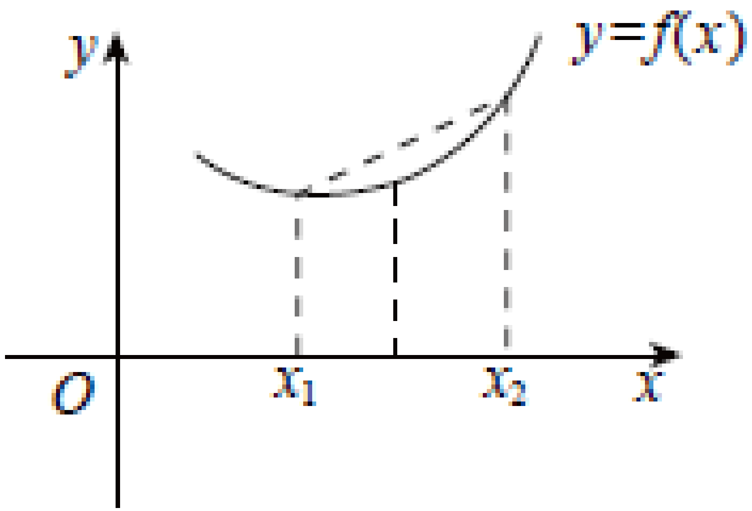
若f(x)为凸函数，g(x)为凸函数，h(x)为线性函数，则该问题称为**凸优化。**
注意这里不等式约束g(x)<=0则要求g(x)为凸函数，若g(x)>=0则要求g(x)为凹函数。

凸优化的任一局部极值点也是全局极值点，局部最优也是全局最优。

凹凸性



凹



凸

拉格朗日乘子法

- Lagrange Multiplier Method
- 用途：消除或简化优化问题中的约束

$$\min_{\mathbf{x}} f(\mathbf{x})$$
$$\text{s.t. } g_i(\mathbf{x}) = 0, i = 1, \cdots, m$$
$$h_i(\mathbf{x}) \leq 0, i = 1, \cdots, n$$

拉格朗日乘子：

$$\boldsymbol{\alpha} = [\alpha_1; \cdots; \alpha_m]$$

$$\boldsymbol{\beta} = [\beta_1; \cdots; \beta_n], \beta_i \geq \mathbf{0}$$

拉格朗日函数：$L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_{i=1}^{m} \alpha_i g_i(\mathbf{x}) + \sum_{i=1}^{n} \beta_i h_i(\mathbf{x})$

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$
$$\text{s.t., } \beta_i \geq 0, i = 1, \cdots, n$$

SVM的对偶问题

$$\min_{\mathbf{w},b} \frac{1}{2} \| \mathbf{w} \|_2^2$$
$$\text{s.t.}, \ y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1, \ i = 1, \cdots, m$$

- 第一步：引入拉格朗日乘子$\boldsymbol{\alpha} = [\alpha_1; \cdots; \alpha_m], \alpha_i \geq 0$，得到拉格朗日函数：

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

- 第二步：求解$\min\limits_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha})$，即对$\mathbf{w}$与$b$的偏导数为0，得：

$$\mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i , \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \qquad -\frac{1}{2} w^T w + \sum_{i=1}^{m} \alpha_i$$

- 第三步：回代可得对偶问题（Dual Problem）：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^{m} \alpha_i$$
$$\text{s.t.}, \ \textstyle\sum_{i=1}^{m} \alpha_i y_i = 0; \ \alpha_i \geq 0, i = 1, \cdots, m$$

b怎么求？

解的稀疏性

最终预测模型：$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$

训练完成后，预测模型只和支持向量有关

- Karush-Kuhn-Tucker (KKT)条件：

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(\mathbf{x}_i) - 1 \geq 0 \\ \alpha_i (y_i f(\mathbf{x}_i) - 1) = 0 \end{cases} \Rightarrow$$

必有$\alpha_i = 0$

或$y_i f(\mathbf{x}_i) = 1$

支持向量

参数b的求法：

设$\mathbf{x}_j$为支持向量，即$y_j(\mathbf{w}^T\mathbf{x}_j + b) = 1$

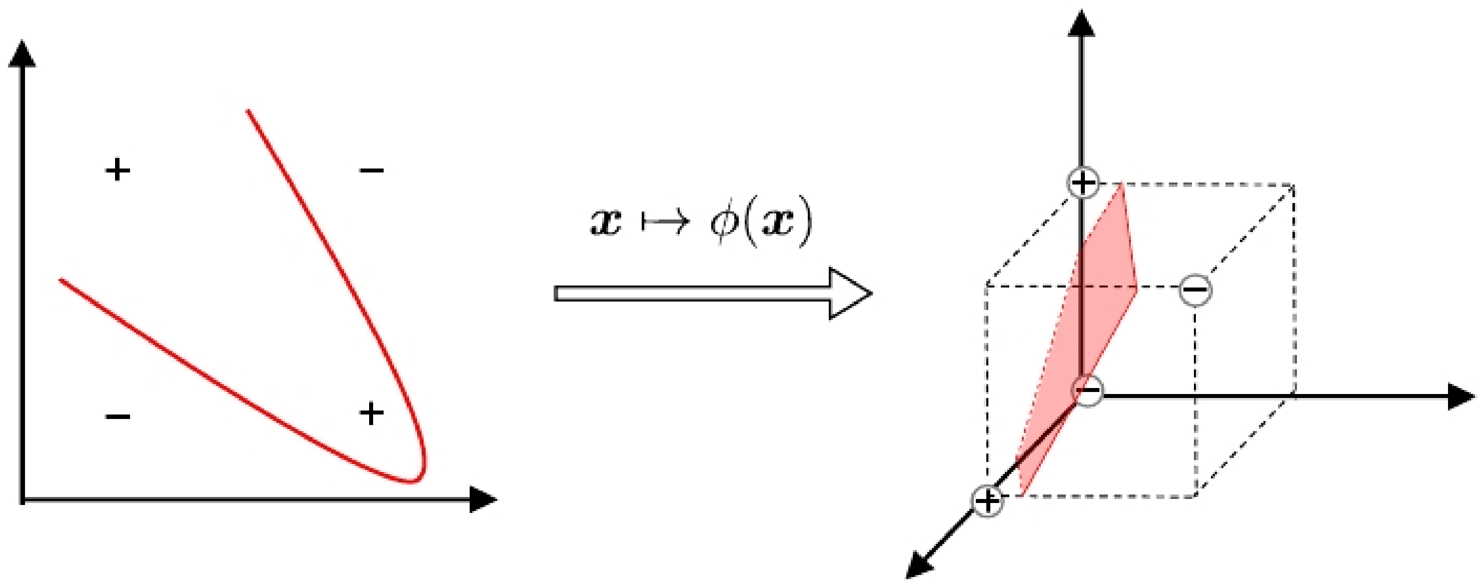
$$b = \frac{1}{y_j} - \mathbf{w}^{\mathrm{T}}\mathbf{x}_j = y_j - \mathbf{w}^{\mathrm{T}}\mathbf{x}_j$$

$$b^* = \frac{1}{|\mathrm{SV}|} \sum_{\mathbf{x}_j \in \mathrm{SV}} (y_j - \mathbf{w}^{\mathrm{T}}\mathbf{x}_j)$$

特征空间映射

- 若不存在一个能正确划分两类样本的超平面, 怎么办?

将样本从原始空间映射到一个更高维的特征空间, 使样本在这个特征空间内线性可分



如果原始空间是有限维（属性数有限），那么一定存在一个高维特征空间使样本线性可分

在特征空间中

- 预测模型：

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

- 原始问题：

$$\min_{\mathbf{w}} \frac{1}{2} \| \mathbf{w} \|_2^2$$

$$\text{s.t. } y_i(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b) \geq 1, i = 1, \cdots, m$$

- 对偶问题：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \boxed{\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)} - \sum_{i=1}^{m} \alpha_i$$

$$\text{s.t., } \sum_{i=1}^{m} \alpha_i y_i = 0; \ \alpha_i \geq 0, i = 1, \cdots, m$$

- 最终预测模型：

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i y_i \boxed{\phi(\mathbf{x}_i)^T \phi(\mathbf{x})} + b$$

ϕ只以内积形式出现！

核函数

- 核函数(kernel function)：衡量样本在特征空间中的相似度

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j)$$ 绕过得到显式特征映射的困难

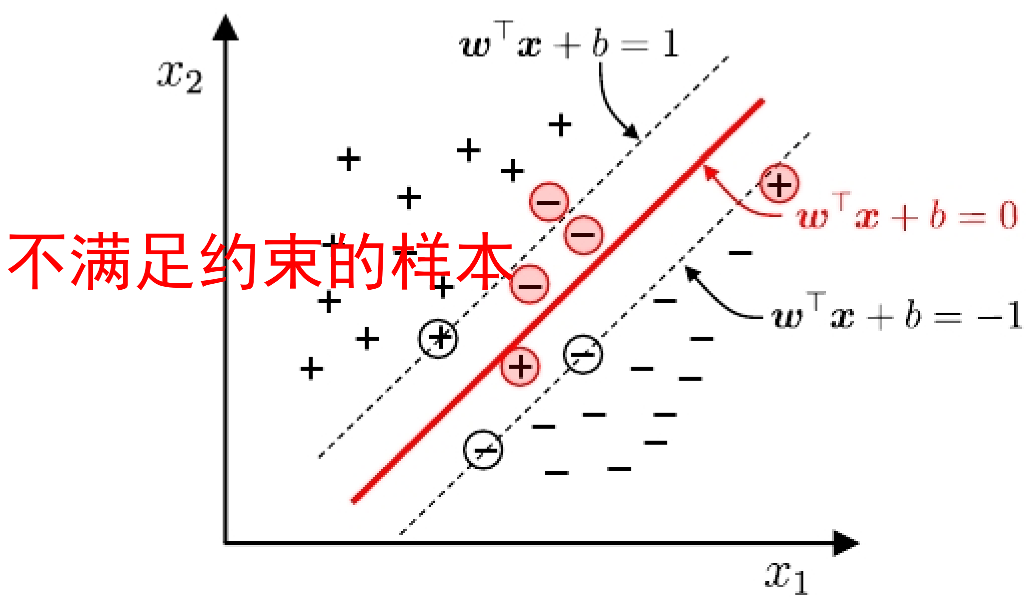
名称	表达式	参数
线性核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$	
多项式核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^d$	$d \geq 1$为多项式的次数
高斯核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\Vert \mathbf{x}_i - \mathbf{x}_j \Vert_2^2}{2\sigma^2}\right)$	$\sigma > 0$为高斯核的带宽
拉普拉斯核	$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\Vert \mathbf{x}_i - \mathbf{x}_j \Vert_2}{\sigma}\right)$	$\sigma > 0$为拉普拉斯核的带宽

- 核函数选择成为决定SVM性能的关键！情况不明时：
  - 若$d \geq m$，选用线性核
  - 若$d < m$，选用高斯核
- 可对核函数进行学习，核学习(Kernel Learning)
  - 难度和学习特征映射ϕ差不多

软间隔

- 现实中很难找到合适的核函数，使训练样本在特征空间中线性可分
- 即便线性可分，也很难断定是否会导致过拟合

引入软间隔(soft margin), 允许在一些样本上不满足约束



软间隔SVM

引入松弛变量：

$$\min_{\mathbf{w},\xi} \frac{1}{2} \parallel \mathbf{w} \parallel_2^2 + C\sum_{i=1}^{m} \xi_i$$

$$\text{s.t.}, \ y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \xi_i \geq 0$$

对偶问题：

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) - \sum_{i=1}^{m} \alpha_i$$

$$\text{s.t.}, \ \sum_{i=1}^{m} \alpha_i y_i = 0; \ 0 \leq \alpha_i \leq C, i = 1, \cdots, m$$

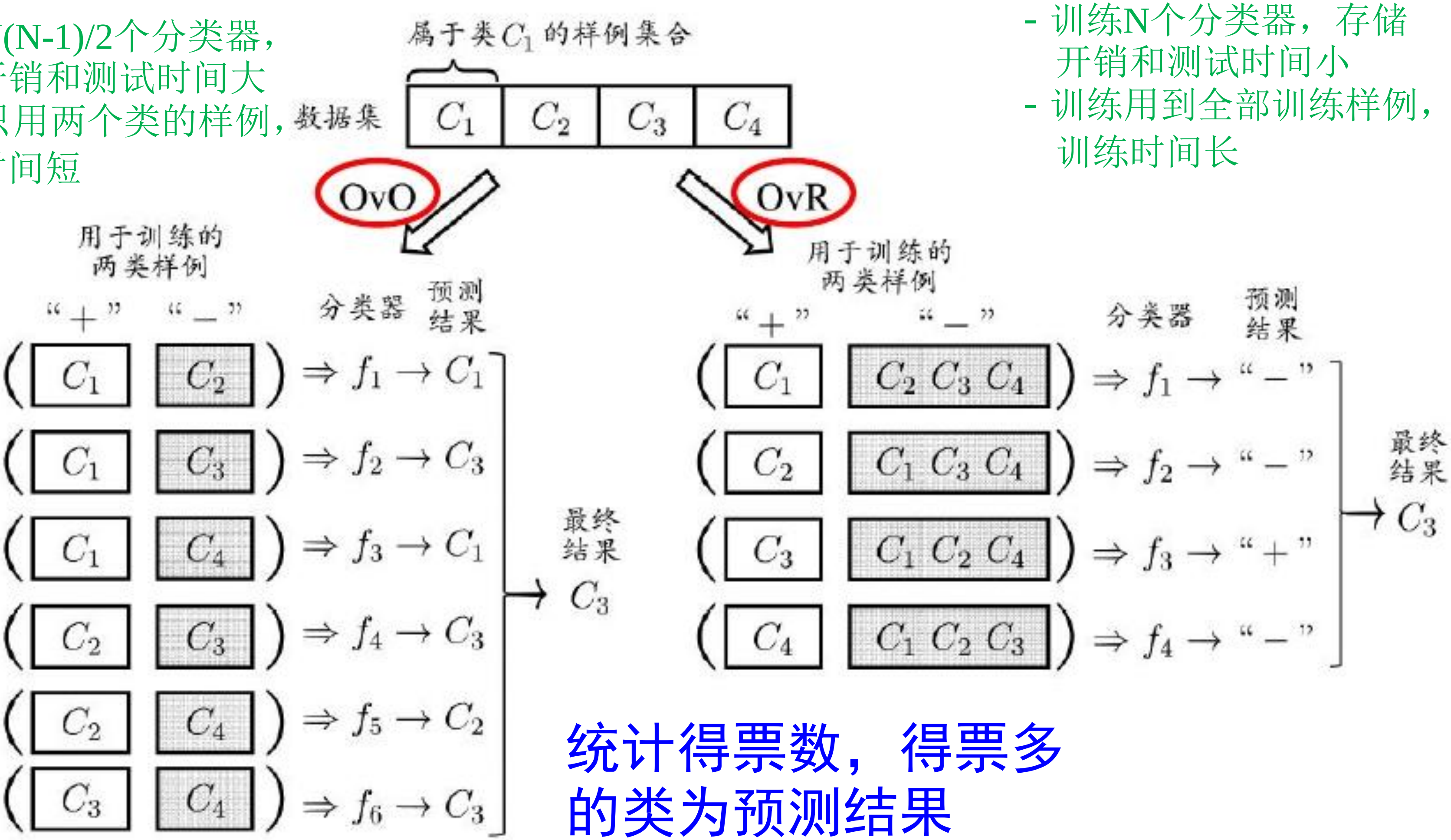
多分类

拆解法：将一个多分类任务拆分为若干个二分类任务求解

-训练N(N-1)/2个分类器，存储开销和测试时间大

-训练只用两个类的样例，训练时间短

- 训练N个分类器，存储开销和测试时间小
- 训练用到全部训练样例，训练时间长



统计得票数，得票多的类为预测结果

Take Home Message

- 支持向量机的”最大间隔”思想
- 对偶问题及其解的稀疏性
- 通过向高维空间映射解决线性不可分的问题
- 引入”软间隔”缓解特征空间中线性不可分的问题