大数据导论

Introduction to Big Data

第10讲: 支持向量机

叶允明

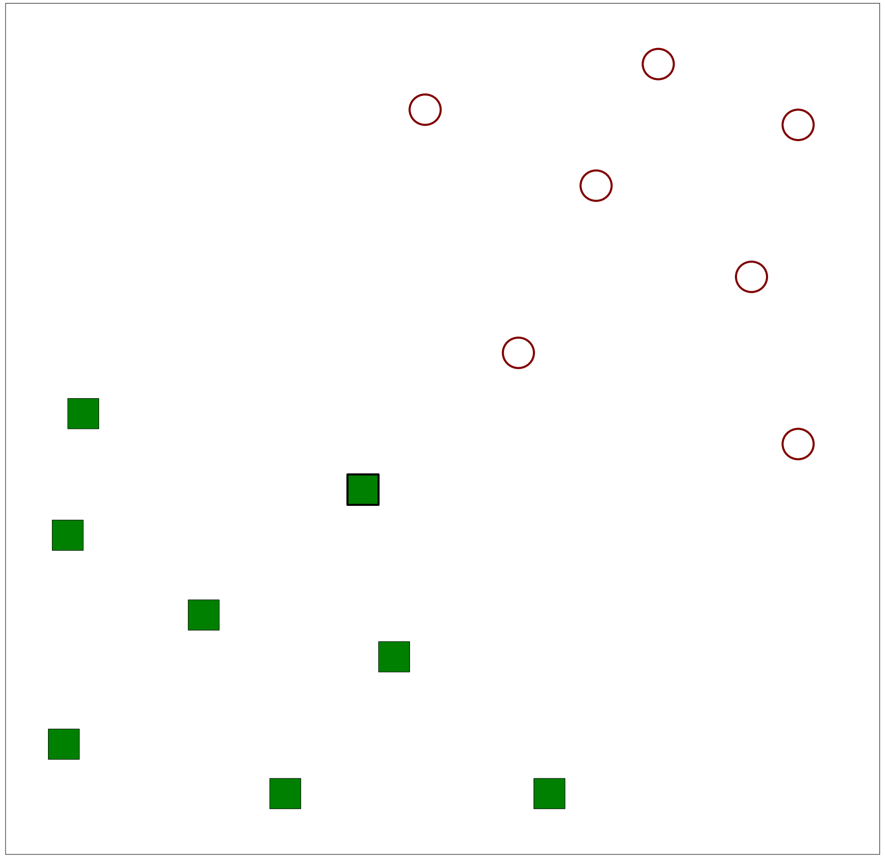
计算机科学与技术学院

哈尔滨工业大学（深圳）

目录

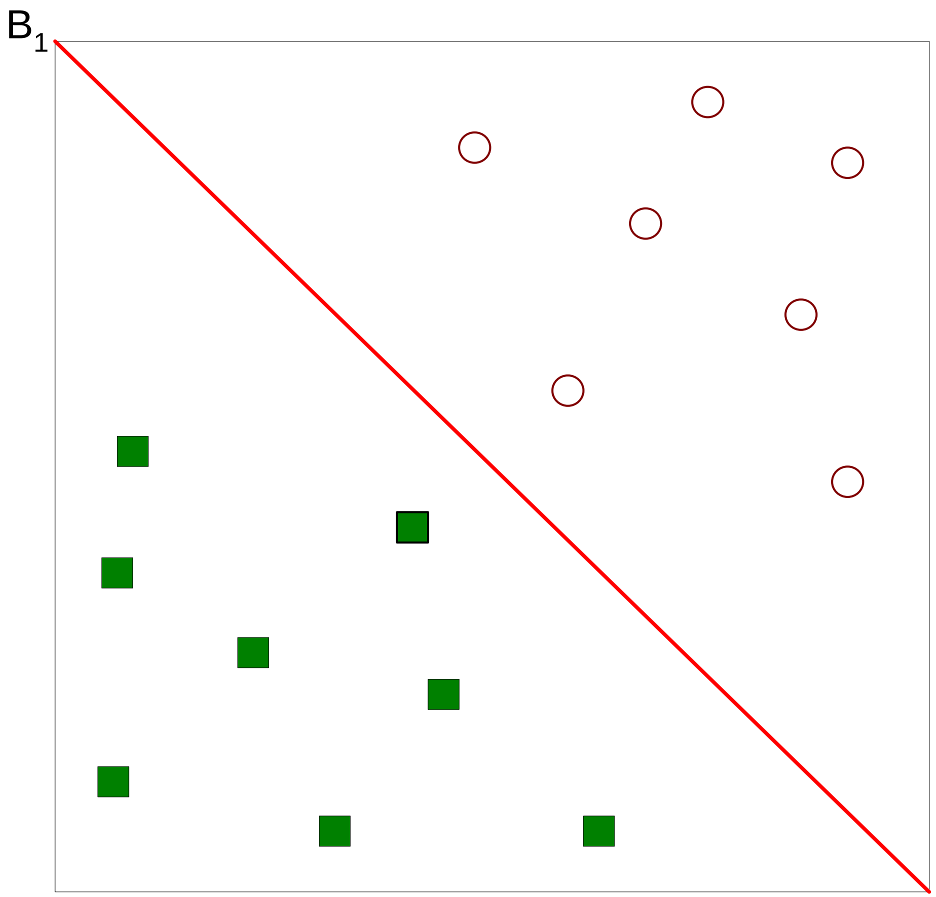
支持向量机的基本思想

Support Vector Machine

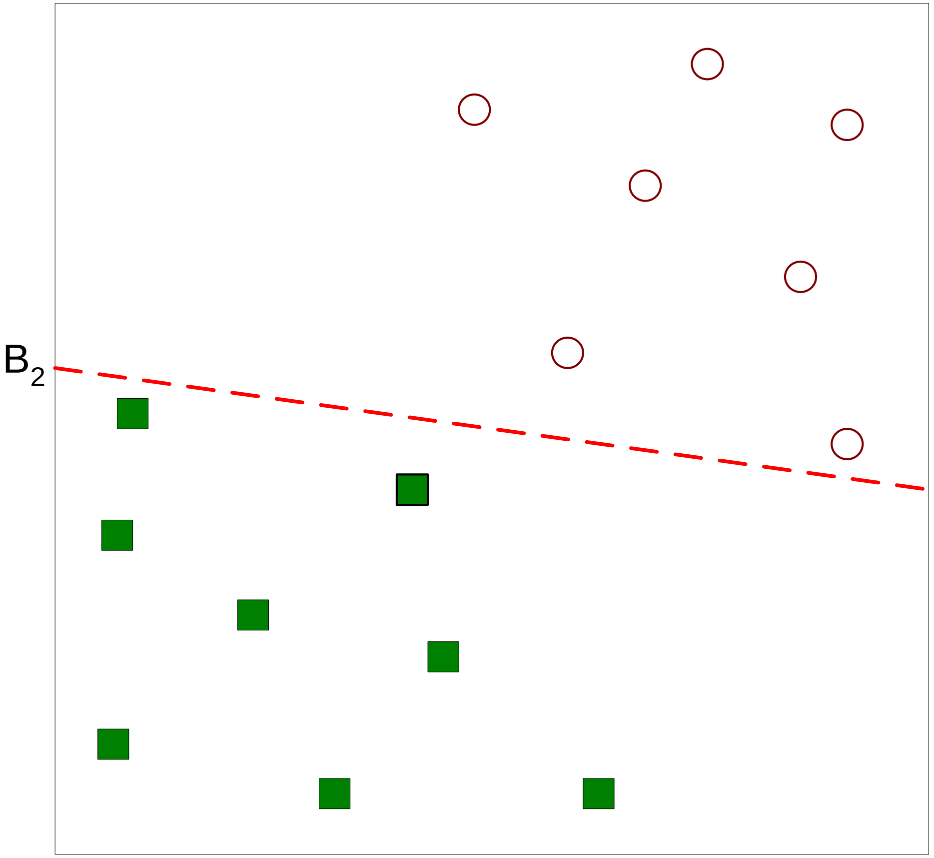


- Find a linear hyperplane (decision boundary) that will separate the data

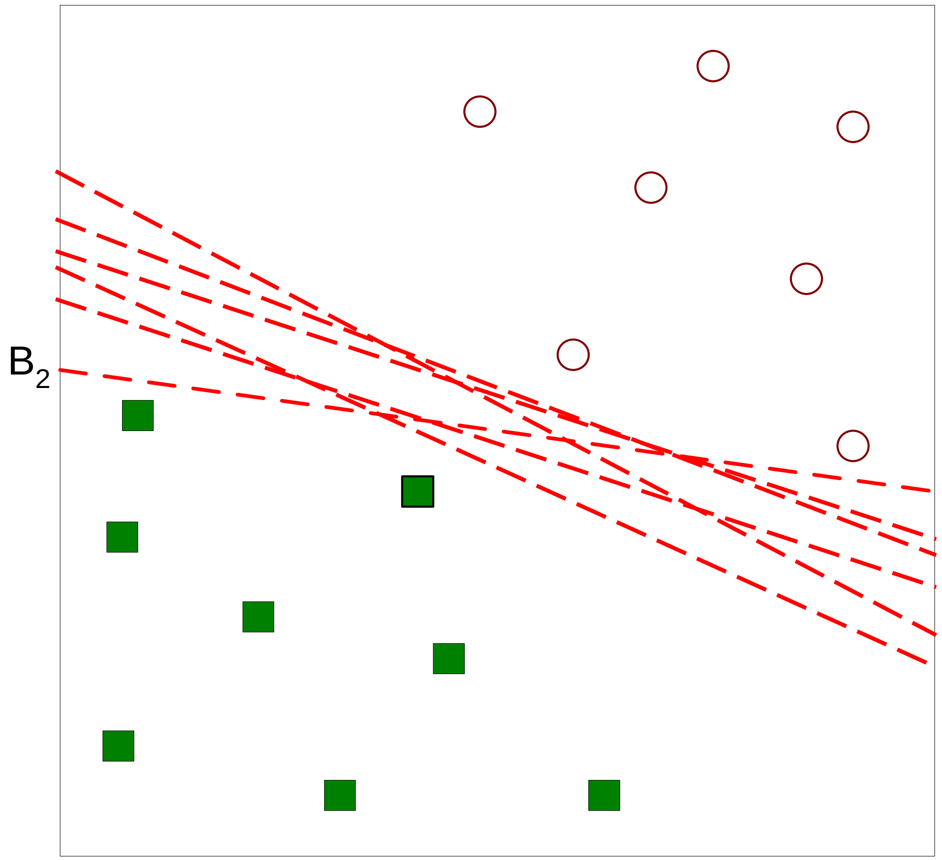
- One Possible Solution



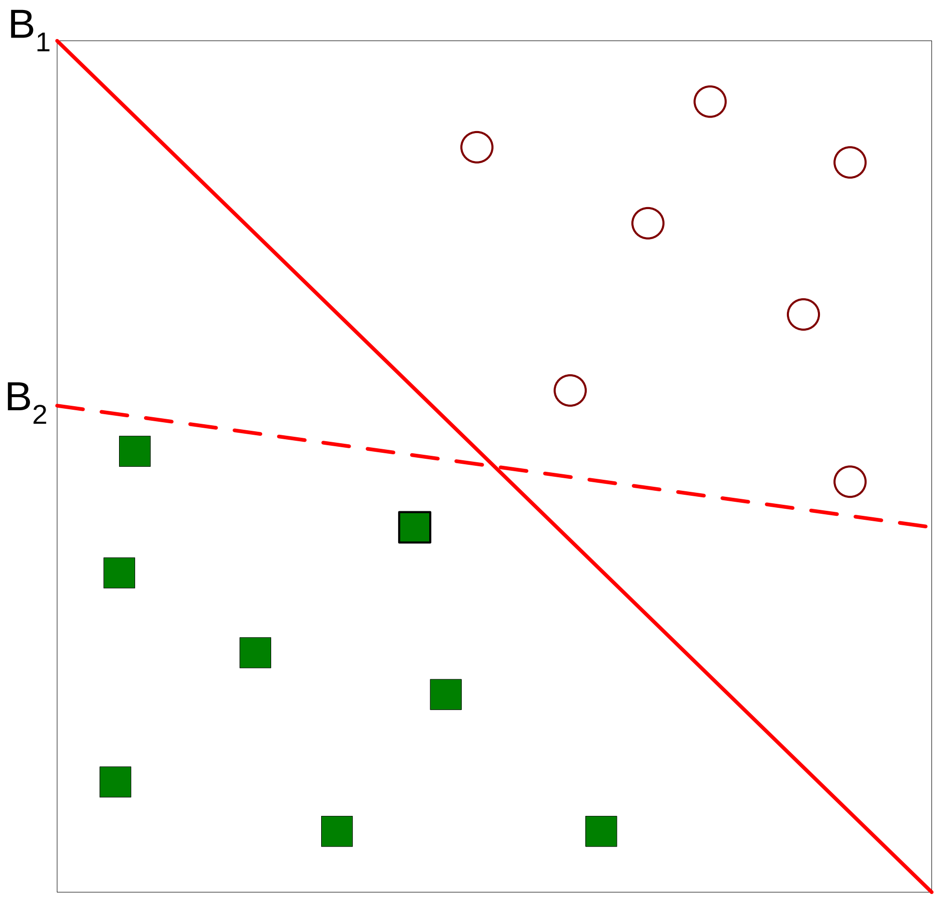
- Another possible solution



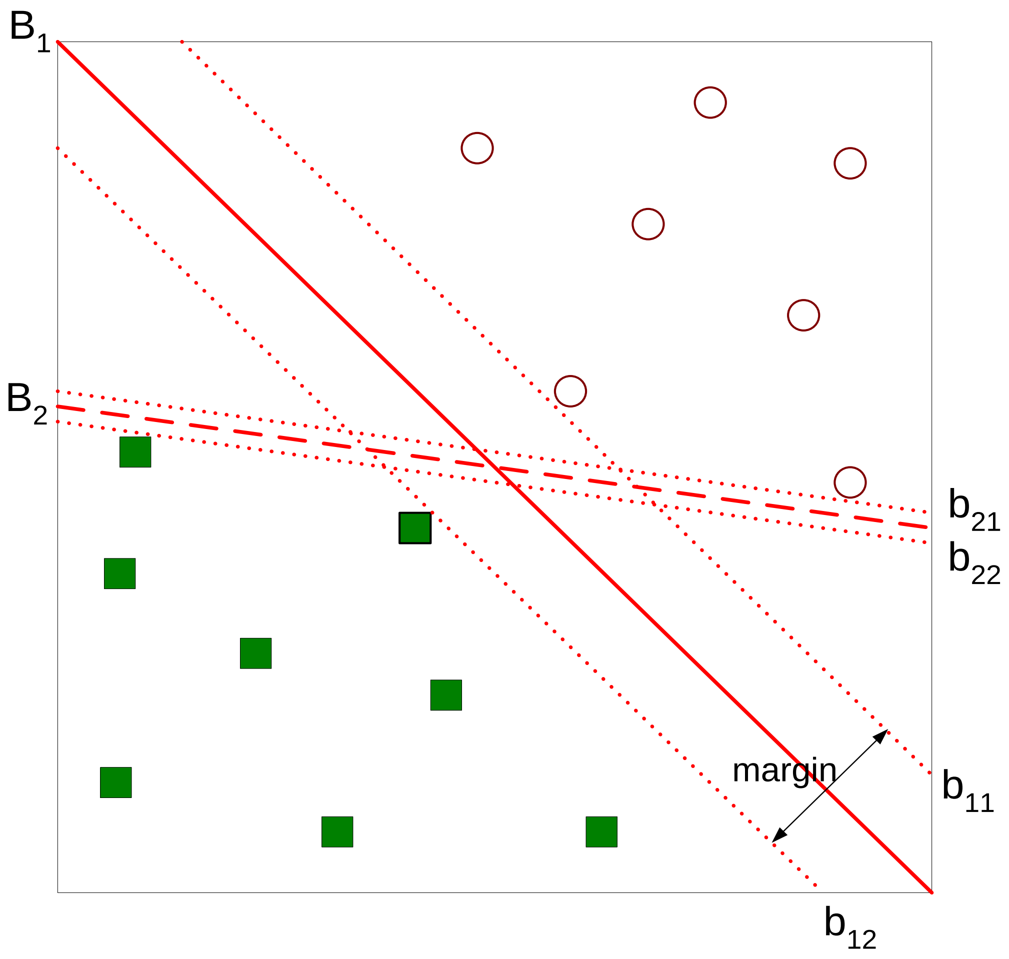
- Many possible solutions



- **Which one is better? B1 or B2?**
- **How do you define better?**

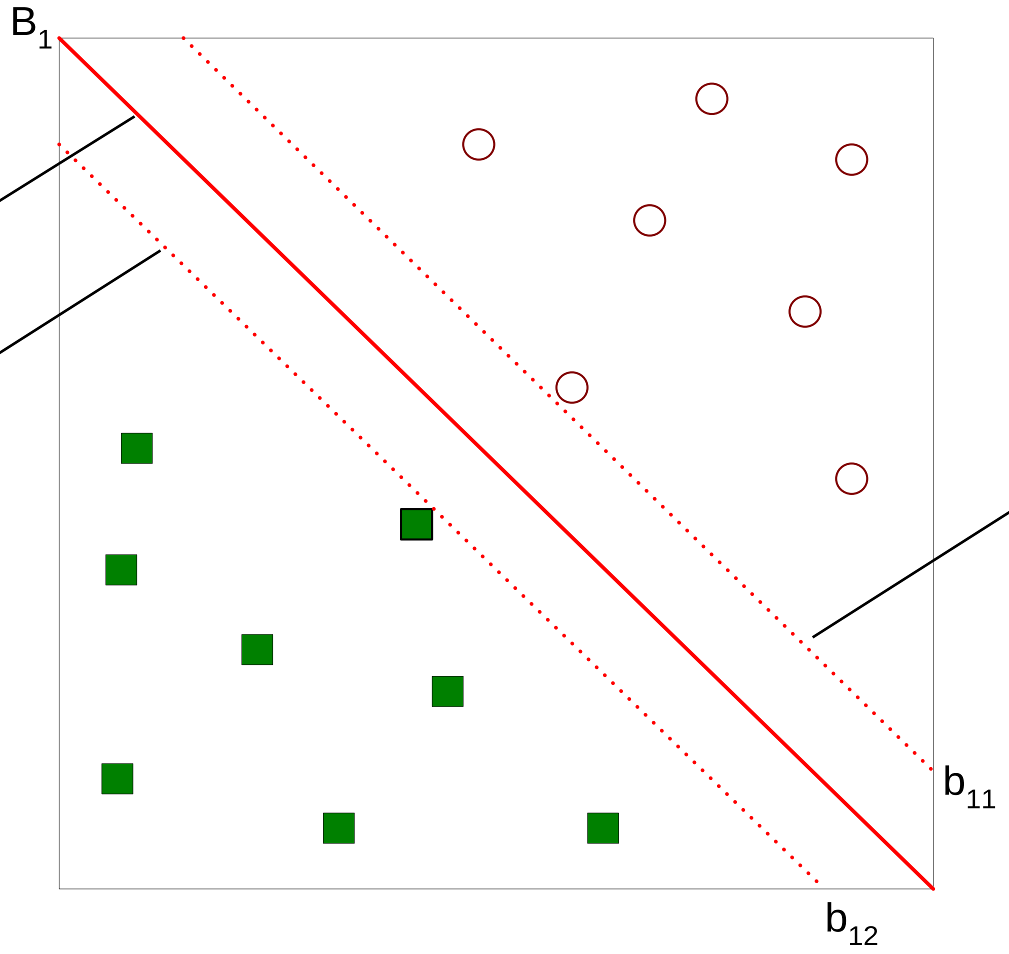


- Find hyperplane maximizes the margin => B1 is better than B2



More robust decision with larger margin!

Support Vector Machine



$$\vec{w} \bullet \vec{x} + b = 0$$

$$\vec{w} \bullet \vec{x} + b = -1$$

$$\vec{w} \bullet \vec{x} + b = +1$$

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

$$\text{Margin} = \frac{2}{\|\vec{w}\|}$$

硬间隔线性支持向量机

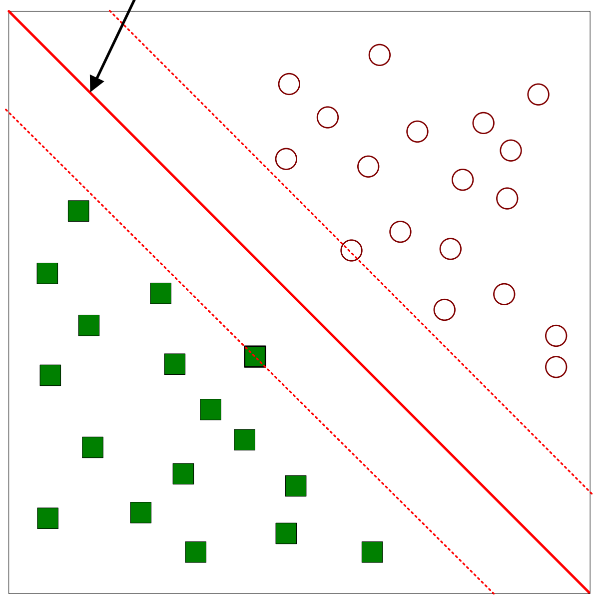
定义

- **超平面（hyperplane）**

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

$\boldsymbol{w} = (w_1, w_2, \dots, w_n)$代表超平面的法向量

b 代表原点到超平面的距离



- **线性可分数据集**

  - 数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_m, y_m)\}$，$\boldsymbol{x}_i \in \mathbb{R}^n$，$y_i\ \{-\mathbf{1}, \mathbf{1}\}$
  - 存在一个**超平面**能将D中的正负样本**严格地划分**到两侧
  - 这样的超平面称为：**分隔超平面**（separating hyperplane）

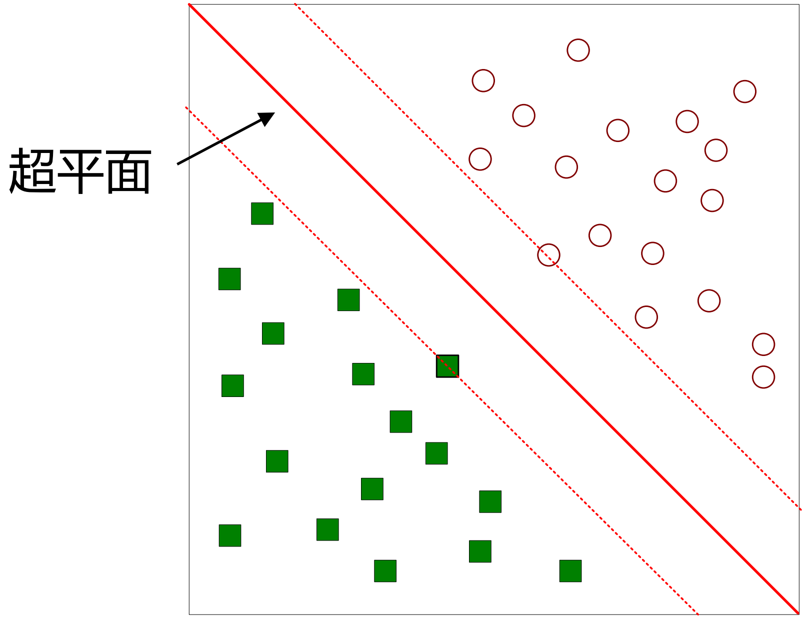
模型学习问题

对于线性可分训练数据集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}, \boldsymbol{x}_i \in \mathbb{R}^n, y_i\ \{-1,1\}$

学习目标是找到一个分隔超平面 $\boldsymbol{w}^T\boldsymbol{x} + b = 0$使其满足：

- 正确划分所有数据：

$$\left.\begin{array}{l} \boldsymbol{w}^T\boldsymbol{x}_i + b > 0,\ \ y_i = \mathbf{1} \\ \boldsymbol{w}^T\boldsymbol{x}_i + b < 0,\ \ y_i = \mathbf{-1} \end{array}\right\} \qquad y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) > 0$$



- 间隔（margin）最大:

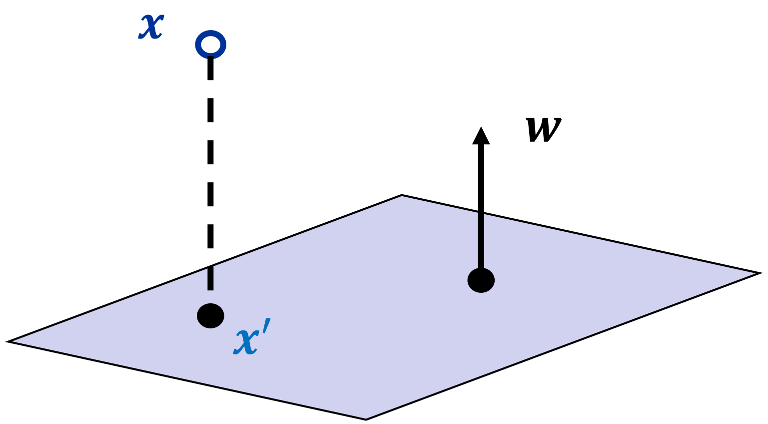
$$< \boldsymbol{w}^*, \boldsymbol{b}^* > = \max_{\boldsymbol{w},b}\ \mathrm{margin}(\boldsymbol{w}, b)$$

$$\mathrm{margin}(\boldsymbol{w}, b) = \min_{i=1,\ldots,m}\ \mathrm{distance}(\boldsymbol{x}_i, < \boldsymbol{w}, b >)$$

点到超平面的距离

- **定理** n维欧几里得空间中，点 $\boldsymbol{x} \in \mathbb{R}^n$ 到超平面 $\boldsymbol{w}^T\boldsymbol{x} + b = 0$的距离为：

$$d = \frac{|\boldsymbol{w}^T\boldsymbol{x} + b|}{\|\boldsymbol{w}\|} \qquad \|\boldsymbol{w}\| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$$



模型学习问题（续）

- 对于线性可分的训练数据集$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\dots,(\boldsymbol{x}_m,y_m)\},\boldsymbol{x}_i\in\mathbb{R}^n,y_i\ \{-1,1\}$;
- 超平面$\boldsymbol{w}^T\boldsymbol{x}+b=0$ 的间隔可以定义为：

$$\text{margin}\,(\boldsymbol{w},b)=\min_{i=1,2,\dots,m}\frac{|\boldsymbol{w}^T\boldsymbol{x}_i+b|}{\|\boldsymbol{w}\|}=\min_{i=1,2,\dots,m}\frac{1}{\|\boldsymbol{w}\|}y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)$$

原问题可改写：

$$\underbrace{<\boldsymbol{w}^*,\boldsymbol{b}^*>=\max_{\boldsymbol{w},b}\ \text{margin}(\boldsymbol{w},b)\qquad y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)>0}$$

$$<\boldsymbol{w}^*,\boldsymbol{b}^*>=\max_{\boldsymbol{w},b}\ \min_{i=1,2,\dots,m}\frac{1}{\|\boldsymbol{w}\|}y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)$$

有约束的极值条件
拉格朗日乘子法

$$\text{s.t.}\quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)>0,\qquad i=1,2,\dots,m$$

问题化简

原问题：

$$\max_{w,b} \boxed{\min_{i=1,2,\dots,m} \frac{1}{\|\boldsymbol{w}\|} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) > 0, \qquad i = 1,2,\dots,m$$

- 按相同比例变化$\boldsymbol{w}$ 和 b 后，所得间隔不变，例如 $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ 和 $6\boldsymbol{w}^T\boldsymbol{x} + 6b = 0$是同一超平面
- 不妨设：

$$\min_{i=1,2,\dots,m} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1$$

- 则原问题可以**化简**为以下有约束优化问题：

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

$$\text{s.t.} \quad \min_{i=1,2,\dots,m} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1$$

问题: $$\max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \quad \min_{i=1,2,\dots,m} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1$$

二次规划问题:

$$\min_{\boldsymbol{u}} \quad \frac{1}{2}\boldsymbol{u}^T\boldsymbol{Q}\boldsymbol{u} + \boldsymbol{p}^T\boldsymbol{u}$$

$$\text{s.t.} \quad \boldsymbol{a}_i^T\boldsymbol{u} \geq c_i, \quad i = 1,2,\dots,M$$

- 将条件改写:

$$\min_{i=1,2,\dots,m} y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1 \quad \Rightarrow \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i = 1,2,\dots,m$$

问题：

$$\max_{\boldsymbol{w}\ b} \quad \frac{1}{\|\boldsymbol{w}\|}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + b\right) \geq 1, \qquad i = 1,2,\ldots,m$$

- 可进一步转化为：

$$\max_{\boldsymbol{w},b} \frac{1}{\|\boldsymbol{w}\|} \quad \rightarrow \quad \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\| \quad \rightarrow \quad \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \rightarrow \quad \min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + b\right) \geq 1, \qquad i = 1,2,\ldots,m$$

最终可得：

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i + b\right) \geq 1, \qquad i = 1,2,\ldots,m$$

问题求解

标准问题

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i + b\right) \geq 1, \qquad i = 1,2,\ldots,m$$

凸二次规划问题

$$\min_{\boldsymbol{u}} \quad \frac{1}{2}\boldsymbol{u}^T\boldsymbol{Q}\boldsymbol{u} + \boldsymbol{p}^T\boldsymbol{u}$$

$$\text{s.t.} \quad \boldsymbol{a}_i^T\boldsymbol{u} \geq c_i, \qquad i = 1,2,\ldots,M$$

$$\boldsymbol{Q} \in \mathbb{R}^{N\times N}, \quad \boldsymbol{u} \in \mathbb{R}^N, \quad \boldsymbol{a_i} \in \mathbb{R}^N, \quad \boldsymbol{p} \in \mathbb{R}^N, \quad c_i \in \mathbb{R}$$

- 标准问题是标准凸二次规划问题。代入凸二次规划形式，可得：

$$\boldsymbol{u} = \begin{bmatrix} b \\ \boldsymbol{w} \end{bmatrix}; \quad \boldsymbol{Q} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0}_n^T \\ \boldsymbol{0}_n & \boldsymbol{I}_{n\times n} \end{bmatrix}; \quad \boldsymbol{p} = \boldsymbol{0}_{n+1}; \quad \boldsymbol{a}_i^T = y_i\begin{bmatrix} 1 & \boldsymbol{x}_i^T \end{bmatrix}; \quad c_i = 1; \quad M = m$$

- 对于凸二次规划问题，可以使用如内点法、椭球法和梯度投影法等方法求解

- 在求出最优解$\boldsymbol{w}^*$和b^*之后，我们可以得到最终的分类决策函数为：

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*)$$

硬间隔线性支持向量机标准问题的学习算法

输入：线性可分数据集$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\},\boldsymbol{x}_i\in\mathbb{R}^n,y_i\ \{-1,1\}$

输出：硬间隔最大化分离超平面和分类决策函数

- 第一步，构造并求解约束最优化问题

$$\min_{\boldsymbol{w},b}\quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$
$$\text{s.t.}\quad y_i\left(\boldsymbol{w}^T\boldsymbol{x}_i+b\right)\geq 1,\qquad i=1,2,\ldots,m$$

- 第二步，返回硬间隔最大化的分隔超平面$\boldsymbol{w}^{*T}\boldsymbol{x}+b^*=0$和分类决策函数$f(\boldsymbol{x})=\mathrm{sign}(\boldsymbol{w}^{*T}\boldsymbol{x}+b^*)$

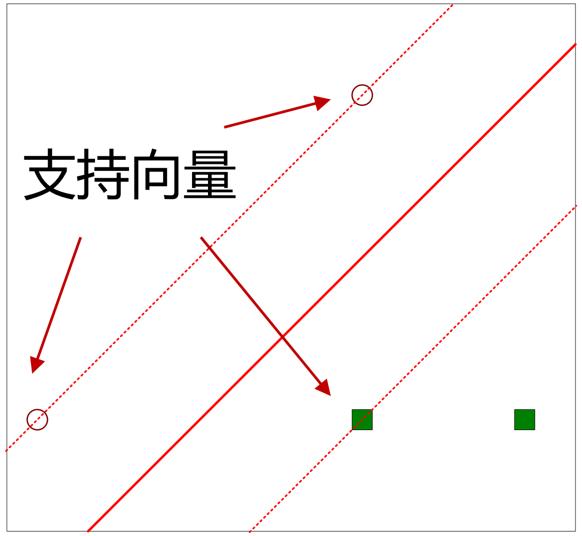
- 从线性可分数据集当中学习到**硬间隔（hard margin）最大化**的分离超平面及相应分类决策函数的过程称为**硬间隔线性支持向量机**，这里的**硬间隔**是指所有数据点都可以严格分开。
- 根据间隔最大化思想，硬间隔最大化分离超平面是唯一存在的。

问题求解举例

标准问题： $\min\limits_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$

$\text{s.t.} \quad y_i(\boldsymbol{w}^T\boldsymbol{x_i}+b) \geq 1, \qquad i=1,2,\dots,m$

使等号成立的点即为：**支持向量！**



$$\boldsymbol{X}=\begin{bmatrix}0 & 0\\ 2 & 2\\ 2 & 0\\ 3 & 0\end{bmatrix} \quad \boldsymbol{y}=\begin{bmatrix}+1\\ +1\\ -1\\ -1\end{bmatrix}$$

构造有约束优化问题

$$\min_{w,b} \quad \frac{1}{2}(w_1^2+w_2^2)$$

$$\text{s.t.} \quad \begin{aligned} b &\geq 1 \quad \text{(i)}\\ 2w_1+2w_2+b &\geq 1 \quad \text{(ii)}\\ -2w_1 \qquad\quad -b &\geq 1 \quad \text{(iii)}\\ -3w_1 \qquad\quad -b &\geq 1 \quad \text{(iv)}\end{aligned}$$

$$\begin{Bmatrix}\text{(i)}\ +\text{(iii)} & \Rightarrow & w_1 \leq -1\\ \text{(ii)}+\text{(iii)} & \Rightarrow & w_2 \geq +1\end{Bmatrix} \Longrightarrow \frac{1}{2}(w_1^2+w_2^2) \geq 1$$

进一步求解可得该最优化问题的解为： $w_1=-1,\ w_2=1,\ b=1$

故硬间隔最大化分割超平面为： $-x_1+x_2+1=0$，可以利用函数 $\text{sign}(-x_1+x_2+1)$ 来分类

存在问题

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x_i}+b\right) \geq 1, \qquad i=1,2,\ldots,m$$

凸二次规划问题：

- $n+1$个变量 $(b \in \mathbb{R},\ \boldsymbol{w} \in \mathbb{R}^n)$
- m个约束条件

- 算法存在的问题
  - 计算复杂度依赖于参数$\boldsymbol{w}$的维度，也就是特征空间的维度n。
  - 因为特征工程或非线性映射的缘故，特征空间的维度往往非常大，特别是在特征维度大于样本数量 $(n > m)$时，这会大大增加计算量。
  - 解决方法：将原问题转换成**对偶问题**，可以让计算复杂度不依赖于特征空间维度

原问题：

- $n+1$个变量
- m个约束条件

对偶问题：

- m个变量
- m个约束条件

问题转换

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{w}^T\boldsymbol{x_i} + b\right) \geq 1, \qquad i = 1,2,\dots,m$$

利用拉格朗日乘子向量 $\boldsymbol{\lambda} = [\lambda_i], \lambda_i \geq 0, i = 1,2,\dots,m$

构建原问题的拉格朗日函数为：

$$\boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \underbrace{\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}}_{\text{优化目标}} + \sum_{i=1}^{m} \lambda_i \underbrace{\left(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right)}_{\text{约束条件}}$$

利用拉格朗日函数，原问题就等价于以下min-max问题：

$$\min_{\boldsymbol{w},b}\left(\max_{\boldsymbol{\lambda}:\lambda_i \geq 0} \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda})\right) = \min_{\boldsymbol{w},b}(\underbrace{\infty}_{\text{不符合约束条件}}, \underbrace{\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}}_{\text{符合约束条件}}) = \underbrace{\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}}_{\text{符合约束条件}}$$

在$(\boldsymbol{w}, b)$给定的情况下，对于任何$\boldsymbol{\lambda}: \lambda_i \geq 0$，都有

$$\max_{\boldsymbol{\lambda}:\lambda_i \geq 0} \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) \geq \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda})$$

故对于任何$\boldsymbol{w}, b, \boldsymbol{\lambda}: \lambda_i \geq 0$，都有

$$\min_{\boldsymbol{w},b} \left(\max_{\boldsymbol{\lambda}:\lambda_i \geq 0} \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) \right) \geq \min_{\boldsymbol{w},b} \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda})$$

$$\underbrace{\min_{\boldsymbol{w},b} \left(\max_{\boldsymbol{\lambda}:\lambda_i \geq 0} \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) \right)}^{p^*} \geq \boxed{\overbrace{\max_{\boldsymbol{\lambda}:\lambda_i \geq 0} \left(\min_{\boldsymbol{w},b} \boldsymbol{L}(\boldsymbol{w}, b, \boldsymbol{\lambda}) \right)}^{d^*}}$$

拉格朗日对偶问题

- 设原问题的最优值为p^*，对偶问题的最优值为d^*
  - 若$p^* \geq d^*$，则两个问题的关系是**弱**对偶性
  - 若$p^* = d^*$，则两个问题的关系是**强**对偶性
- 若原问题满足以下条件（Slater条件），则原问题与对偶问题有强对偶性
  - ✓ 原问题是凸问题
  - ✓ 原问题有解
  - ✓ 原问题的约束条件是线性约束条件

$$\min_{\boldsymbol{w},b} \left(\max_{\boldsymbol{\lambda}:\lambda_i \geq 0} \frac{1}{2} \boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{w} + \sum_{i=1}^{m} \lambda_i (1 - y_i (\boldsymbol{w}^{\boldsymbol{T}} \boldsymbol{x}_i + b)) \right)$$

- SVM原问题满足上述条件，故$\boldsymbol{p}^* = \boldsymbol{d}^*$，**对偶问题的解就是原问题的解，**我们解决对偶问题即可

对偶问题简化

$$\max_{\boldsymbol{\lambda}:\lambda_i\geq 0}\left(\min_{\boldsymbol{w},b}\boxed{\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\sum_{i=1}^{m}\lambda_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))}\right)\qquad \boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{\lambda})$$

- $(\boldsymbol{w},b)$要满足$\boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{\lambda})$最小，则对应梯度应为0，即：

➢ $\frac{\partial \boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{\lambda})}{\partial b}=0=-\sum_{i=1}^{m}\lambda_i y_i$，我们将$\sum_{i=1}^{m}\lambda_i y_i=0$ 加到对偶问题的条件中不会影响问题的最优解，故

$$\max_{\boldsymbol{\lambda}:\lambda_i\geq 0;\sum\lambda_i y_i=0}\left(\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\sum_{i=1}^{m}\lambda_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i))-\cancel{\sum_{i=1}^{m}\lambda_i y_i\cdot b}\right)$$

$$\Rightarrow \max_{\boldsymbol{\lambda}:\lambda_i\geq 0;\sum\lambda_i y_i=0}\left(\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\sum_{i=1}^{m}\lambda_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i))\right)$$

➢ $\frac{\partial \boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{\lambda})}{\partial \boldsymbol{w}}=\boldsymbol{0}=\boldsymbol{w}-\sum_{i=1}^{m}\lambda_i y_i\boldsymbol{x}_i$，同样将$\boldsymbol{w}=\sum_{i=1}^{m}\lambda_i y_i\boldsymbol{x}_i$ 加到对偶问题的条件中，

$$\max_{\boldsymbol{\lambda}:\lambda_i\geq 0;\sum\lambda_i y_i=0;\boldsymbol{w}=\sum\lambda_i y_i\boldsymbol{x}_i}\left(\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\sum_{i=1}^{m}\lambda_i(1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i))\right)$$

$$\Rightarrow \max_{\boldsymbol{\lambda}:\lambda_i\geq 0;\sum\lambda_i y_i=0;\boldsymbol{w}=\sum\lambda_i y_i\boldsymbol{x}_i}\left(\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\sum_{i=1}^{m}\lambda_i-\boldsymbol{w}^T\boldsymbol{w}\right)$$

展开

$$\Rightarrow \max_{\boldsymbol{\lambda}:\lambda_i\geq 0;\sum\lambda_i y_i=0;\boldsymbol{w}=\sum\lambda_i y_i\boldsymbol{x}_i}\left(\min_{\boldsymbol{w},b}-\frac{1}{2}\left\|\sum_{i=1}^{m}\lambda_i y_i\boldsymbol{x}_i\right\|^2+\sum_{i=1}^{m}\lambda_i\right)\Rightarrow \boxed{\max_{\boldsymbol{\lambda}:\lambda_i\geq 0;\sum\lambda_i y_i=0;\boldsymbol{w}=\sum\lambda_i y_i\boldsymbol{x}_i}-\frac{1}{2}\left\|\sum_{i=1}^{m}\lambda_i y_i\boldsymbol{x}_i\right\|^2+\sum_{i=1}^{m}\lambda_i}$$

与$\boldsymbol{w},b$无关，可以去掉最小化

最终需求解的对偶问题

KKT条件

$$\boxed{\max_{\lambda:\lambda_i\geq 0;\sum\lambda_i y_i=0;\boldsymbol{w}=\sum\lambda_i y_i \boldsymbol{x}_i} -\frac{1}{2}\left\|\sum_{i=1}^{m}\lambda_i y_i \boldsymbol{x}_i\right\|^2+\sum_{i=1}^{m}\lambda_i}$$

总结一下，$(\boldsymbol{w}, b, \boldsymbol{\lambda})$ 是原问题与对偶问题最优解的充要条件是$(\boldsymbol{w}, b, \boldsymbol{\lambda})$满足下面的Karush-Kuhn-Tucker（KKT）条件：

- 满足原问题的约束条件： $y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}}+b\right)\geq 1,\ i=1,2,\ldots,m$
- 满足对偶问题约束条件： $\lambda_i\geq 0$
- 满足对偶问题优化条件：

➢ $\frac{\partial \boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{\lambda})}{\partial b}=0=-\sum_{i=1}^{m}\lambda_i y_i$，即$\sum_{i=1}^{m}\lambda_i y_i=0$

➢ $\frac{\partial \boldsymbol{L}(\boldsymbol{w},b,\boldsymbol{\lambda})}{\partial \boldsymbol{w}}=\boldsymbol{0}=\boldsymbol{w}-\sum_{i=1}^{m}\lambda_i y_i \boldsymbol{x}_i$，即$\boldsymbol{w}=\sum_{i=1}^{m}\lambda_i y_i \boldsymbol{x}_i$

- 满足原问题的优化条件：将原问题转换成 minmax问题时，$\boxed{\lambda_i\left(1-y_i\left(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_i+b\right)\right)=0}$

对偶互补条件

求解对偶问题

$$\max_{\lambda:\lambda_i\geq 0;\sum\lambda_i y_i=0;\boldsymbol{w}=\sum\lambda_i y_i \boldsymbol{x}_i} -\frac{1}{2}\left\|\sum_{i=1}^{m}\lambda_i y_i \boldsymbol{x}_i\right\|^2+\sum_{i=1}^{m}\lambda_i$$

改写上述问题：最大化转换为最小化（取负号）

$$\min_{\boldsymbol{\lambda}} \quad \frac{1}{2}\left\|\sum_{i=1}^{m}\lambda_i y_i \boldsymbol{x}_i\right\|^2-\sum_{i=1}^{m}\lambda_i$$

展开

$$\min_{\boldsymbol{\lambda}} \quad \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j-\sum_{i=1}^{m}\lambda_i$$

$$\text{s.t.} \quad \boxed{\sum_{i=1}^{m}\lambda_i y_i = 0}$$

$$\lambda_i \geq 0, \qquad i=1,2,\ldots,m$$

- 代入凸二次规划形式，可得：

$$\boldsymbol{u}=\boldsymbol{\lambda};\ \boldsymbol{Q}=\left[q_{ij}=y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j\right]_{m\times m};\ \boldsymbol{p}=-\boldsymbol{1}_m;$$

$\boldsymbol{e}_i\in\mathbb{R}^m$是第i行为1的单位向量

$$\sum_{i=1}^{m}\lambda_i y_i\geq 0$$
$$-\sum_{i=1}^{m}\lambda_i y_i\geq 0$$

$$\longrightarrow \boxed{\boldsymbol{a}_{\geq}=\boldsymbol{y}=[y_i]_m;\ c_{\geq}=0;\ \boldsymbol{a}_{\leq}=-\boldsymbol{y};\ c_{\leq}=0;}\ \boldsymbol{a}_i=\boldsymbol{e}_i;\ c_i=0;$$

- 由于$\boldsymbol{Q}$可能比较大，实际应用中需利用专为SVM设计的方法解上述问题，得最优解 $\boldsymbol{\lambda}$

求解对偶问题

- 在求得最优解$\boldsymbol{\lambda}^* = [\lambda_i^*]$后
- 可以利用KKT条件求得原问题的最优解$\boldsymbol{w}^*$和b^*

(KKT) 条件:

- $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1,\ i = 1,2,\dots,m$
- $\lambda_i \geq 0$
- $\sum_{i=1}^m \lambda_i y_i = 0$; $\boxed{\boldsymbol{w} = \sum_{i=1}^m \lambda_i y_i \boldsymbol{x}_i}$
- $\boxed{\lambda_i\left(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)\right) = 0}$ (对偶互补条件)

➢ $\boldsymbol{w}^* = \sum_{i=1}^m \lambda_i^* y_i \boldsymbol{x}_i$

➢ 利用对偶互补条件，如果某个j 使得$\lambda_j^* > 0$，则$1 - y_j(\boldsymbol{w}^{*T}\boldsymbol{x}_j + b) = 0$，左右两边乘以$y_j$后可得：

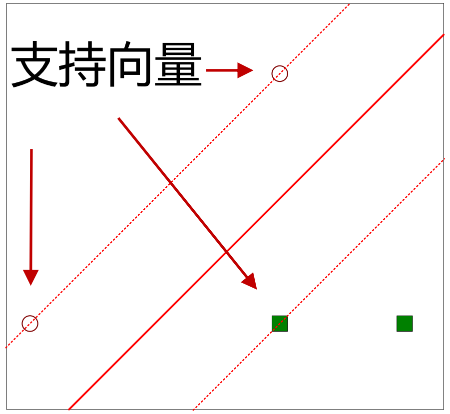
当有多个$\lambda_j^* > 0$时，算得的b可能不一样

$$b^* = y_j - \boldsymbol{w}^{*T}\boldsymbol{x}_j$$

- 最后我们可以得到最终的分类决策函数的对偶形式为

$$f(x) = \mathrm{sign}(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*)$$

- 支持向量



➢ 对偶问题的解$\boldsymbol{\lambda}^* = [\lambda_i^*](i = 1,2,\dots,n)$中对应于分量$\lambda_i^* \neq 0$的样本点$(\boldsymbol{x}_i, y_i)$的实例$\boldsymbol{x}_i$称为**支持向量。**

➢ 线性支持向量机中，支持向量可理解为经过间隔边界上的样本点，其作为向量支持着分类界限，决定决策超平面的参数取值。

硬间隔线性SVM对偶问题的学习算法

输入：线性可分数据集$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\},\boldsymbol{x}_i\in\mathbb{R}^n,y_i\ \{-1,1\}$

输出：硬间隔最大化分离超平面和分类决策函数

- 第一步，构造带约束的凸二次规划问题，并求解得到最优解$\boldsymbol{\lambda}^*=[\lambda_i^*](i=1,2,\ldots,m)$

$$\min_{\boldsymbol{\lambda}}\quad \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\lambda_i\lambda_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j-\sum_{i=1}^{m}\lambda_i$$

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

$$\lambda_i\geq 0,\qquad i=1,2,\ldots,m$$

- 第二步，计算对偶问题对应的最优解$\boldsymbol{w}^*$，并任意选择$\boldsymbol{\lambda}^*$的一个正分量$\lambda_j^*>0$，以求解b^*

$$\boldsymbol{w}^*=\textstyle\sum_{i=1}^{m}\lambda_i^* y_i\boldsymbol{x}_i$$

$$b^*=y_j-\boldsymbol{w}^{*T}\boldsymbol{x}_j$$

- 第三步，返回硬间隔最大化分离超平面$\boldsymbol{w}^{*T}\boldsymbol{x}+b^*=0$和分类决策函数$f(x)=\text{sign}(\boldsymbol{w}^{*T}\boldsymbol{x}+b^*)$

有监督学习的一些问题

混淆矩阵

- 给定m类, 混淆矩阵中的每一个表目$CM_{i,j}$表示类i的元组被分类器标记为类j的个数

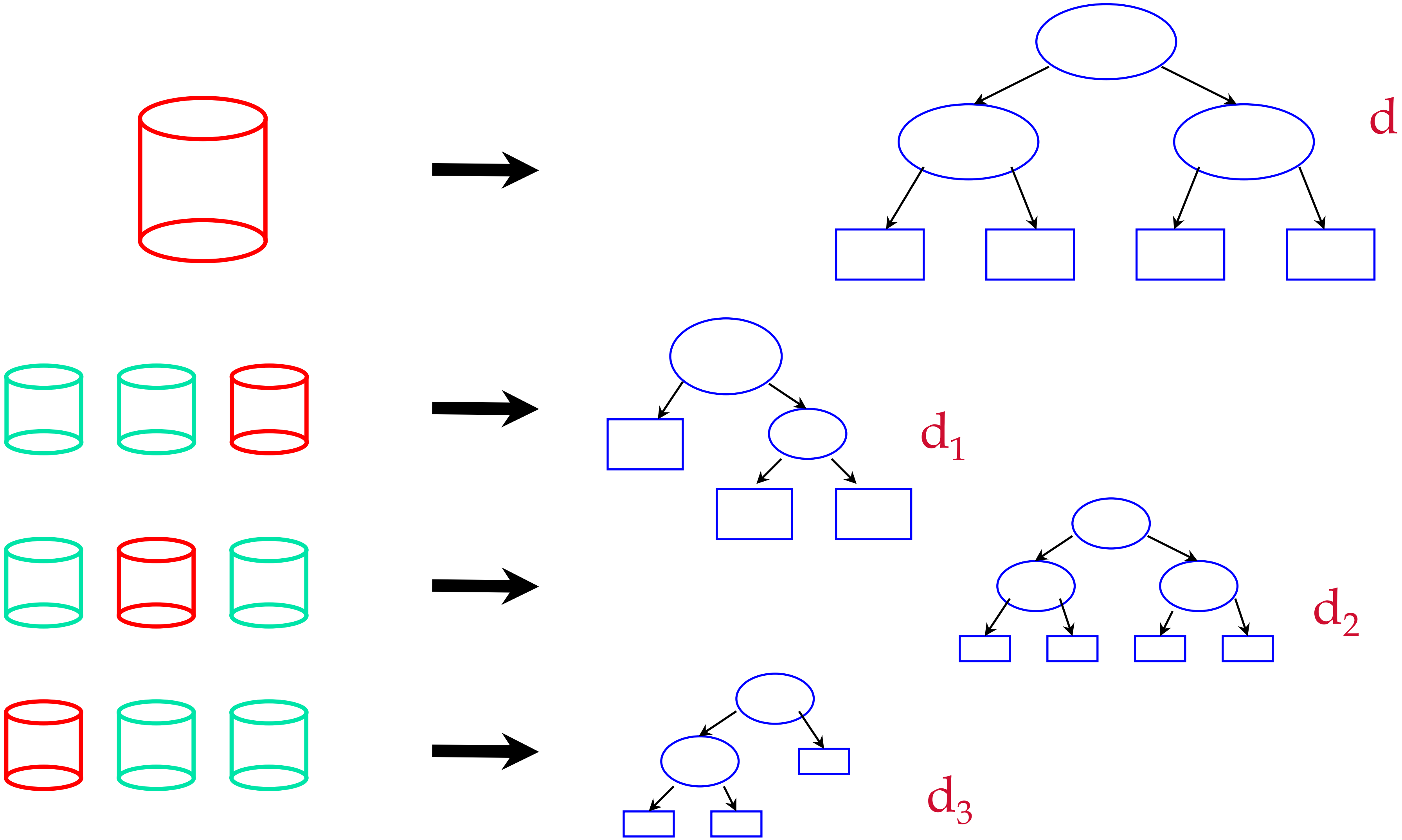
classes	buy_computer=yes	buy_computer=no	total	recognition(%)
buy_computer=yes	6954	46	7000	99.34
buy_computer=no	412	2588	3000	86.27
total	7366	2634	10000	95.52

预测类别

真实类别

	C1	C2
C1	true positives	false negatives
C2	false positives	true negatives

交叉验证

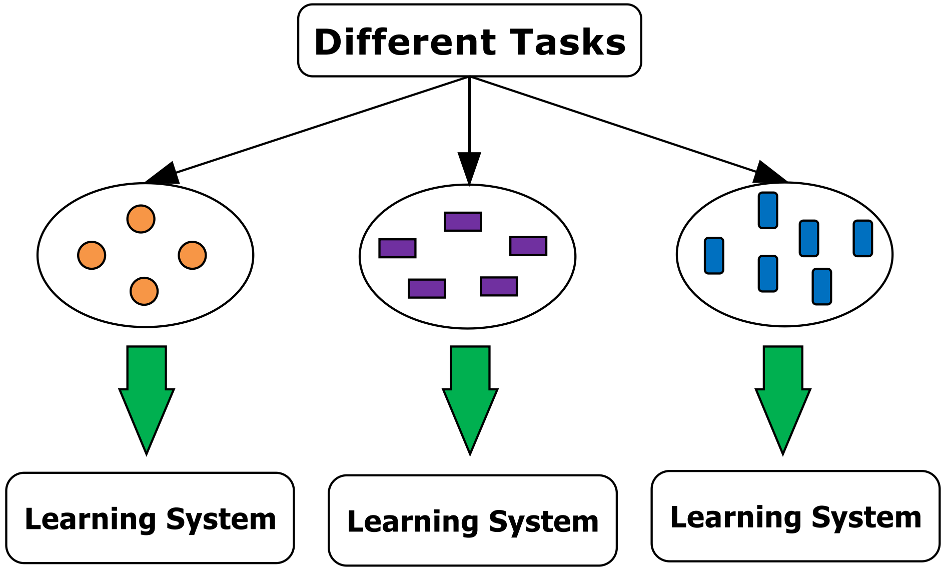


训练数据的合理利用

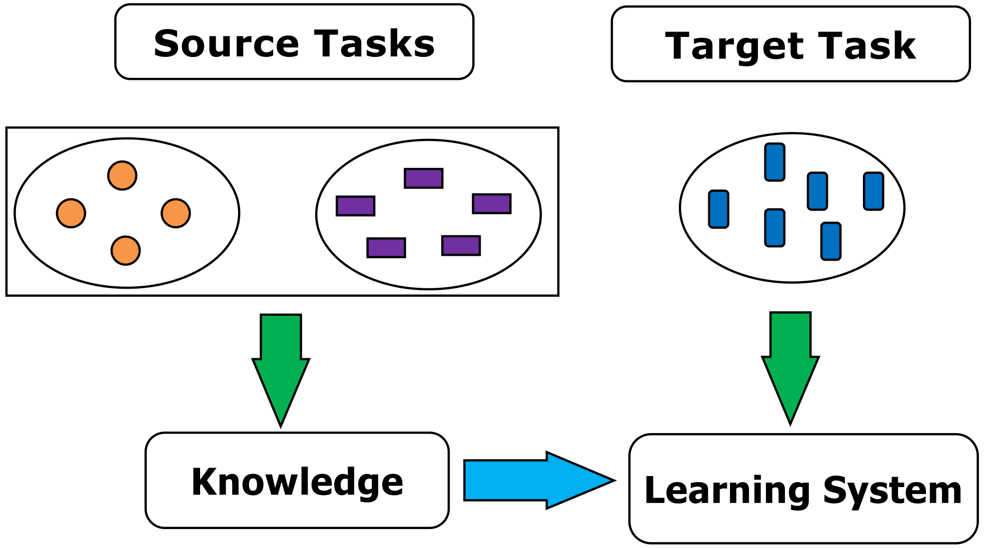
- 训练集-验证集-测试集
- 尽量少“测试”

迁移学习：概念框架

- **迁移学习目标：**从一个或多个源任务提取知识，并将这种知识用于目标任务
- **传统学习方法：**为每个新任务建立新的分类器
- **迁移学习方法：**为新任务构建分类器时，使用源任务的知识



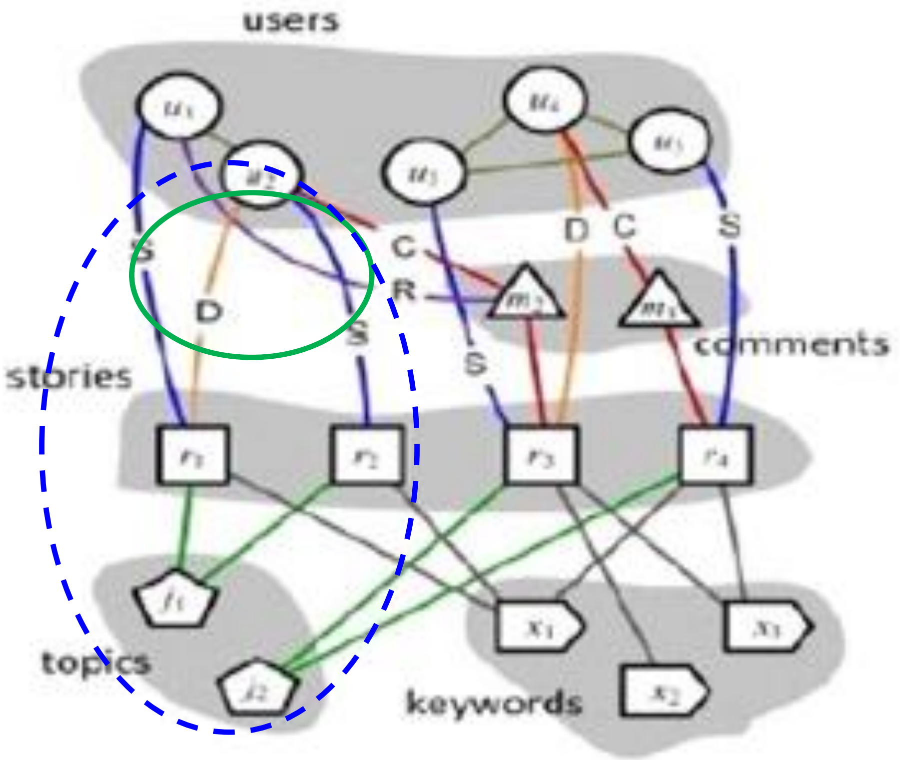
传统学习框架



迁移学习框架

其他热门研究课题

- 社交网络/关系图的分类
- 多类标分类
- 多元实例学习
- 深度学习
- ……



陈宝林著. 最优化理论与算法（第二版）.

清华大学出版社, 2005.