第四章 统计机器学习方法

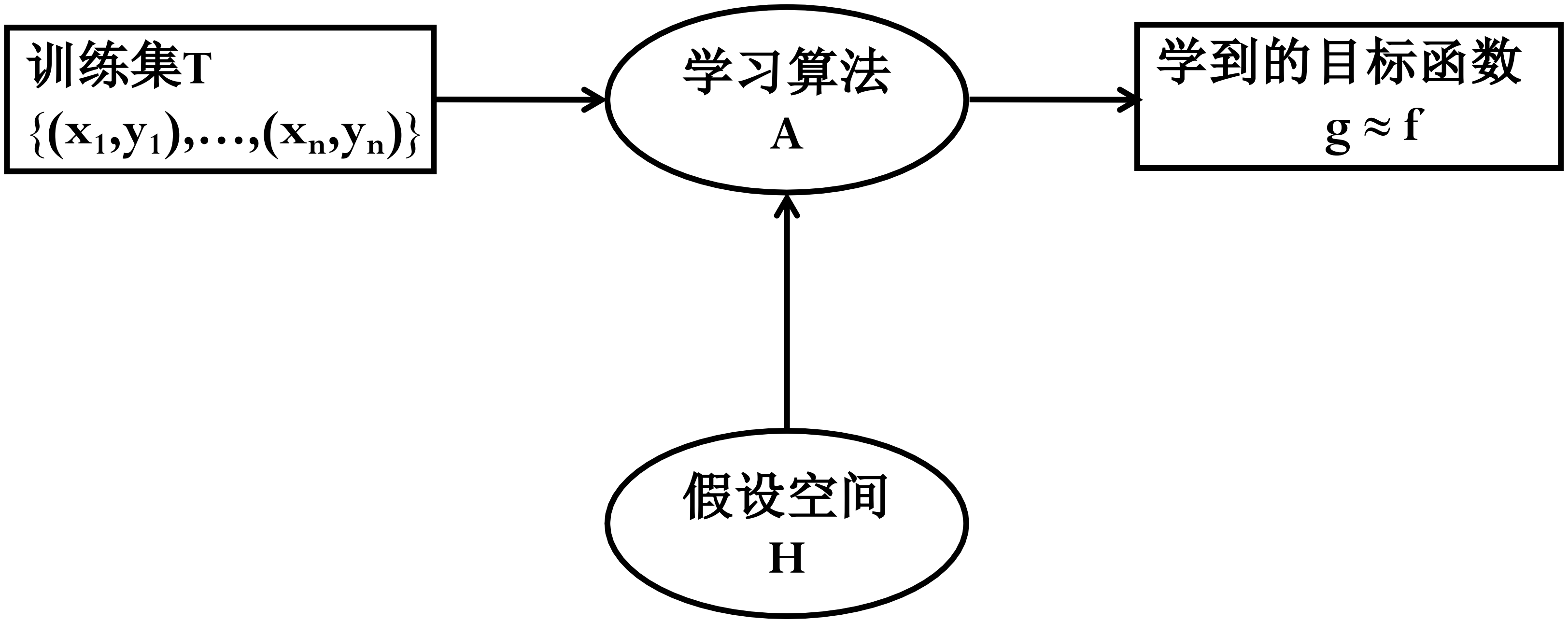
- 什么是机器学习？
  - “如果一个系统能够通过执行某个过程改进它的性能，这就是学习”
    ——赫伯特·西蒙（司马贺）

- 统计学习就是计算机系统通过运用数据及统计方法提高系统性能的机器学习

- 统计学习从数据出发，提取数据的特征，抽象出数据的模型，发现数据中的知识，又回到对数据的分析与预测中。
- 统计学习的目标就是考虑学习什么样的模型和如何学习，以使模型能对数据进行准确的预测和分析

什么是统计机器学习?

- 输入：$x \in X$
- 输出：$y \in Y$
- 未知的目标函数：$f: X \to Y$
- 训练集：$T=\{(x_1,y_1),\ldots,(x_n,y_n)\}$，
  T由f产生（可能具有噪音）
- 假设空间：$H=\{h_k\}$
- 学到的目标函数：$g \in H$
- 学习算法：A



- 学习算法A根据训练集T从假设空间H中选择一个最好的$g \approx f$

统计学习三要素：

- 模型：学习什么样的模型
  - 条件概率分布、决策函数
- 策略：模型选择的准则
  - 经验风险最小化、结构风险最小化
- 算法：模型学习的算法
  - 一般归结为一个最优化问题

- 统计机器学习分类：
  - 监督学习
  - 无监督学习
  - 半监督学习
  - 弱监督

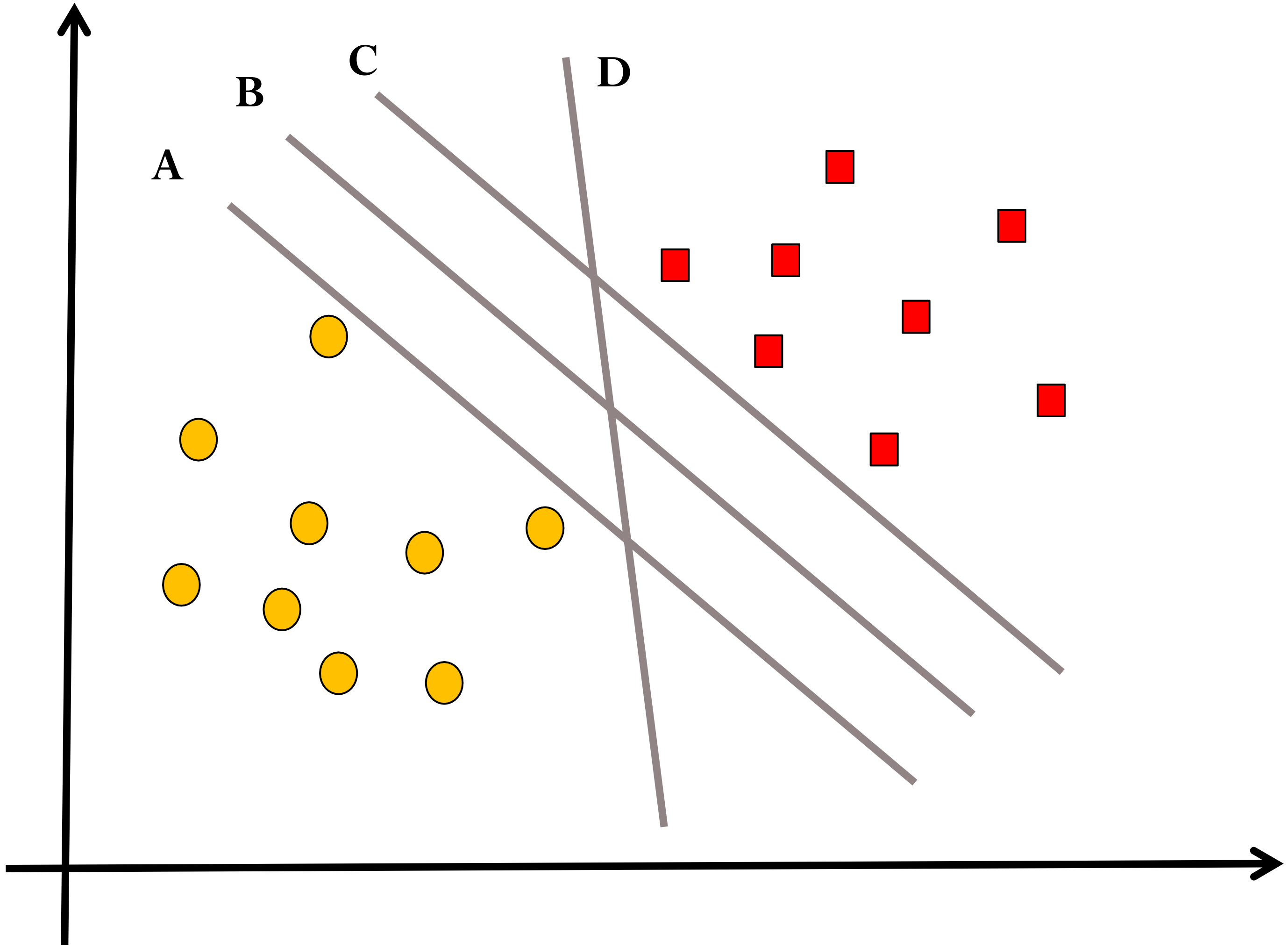
统计机器学习的应用

- 应用广泛，信息处理的各个方面几乎都要用到机器学习
  - 文字、语音识别，输入法
  - 搜索引擎
  - 推荐、广告
  - 文本处理、机器翻译
  - 图像、视频处理
  - ……

4.1 支持向量机（SVM）

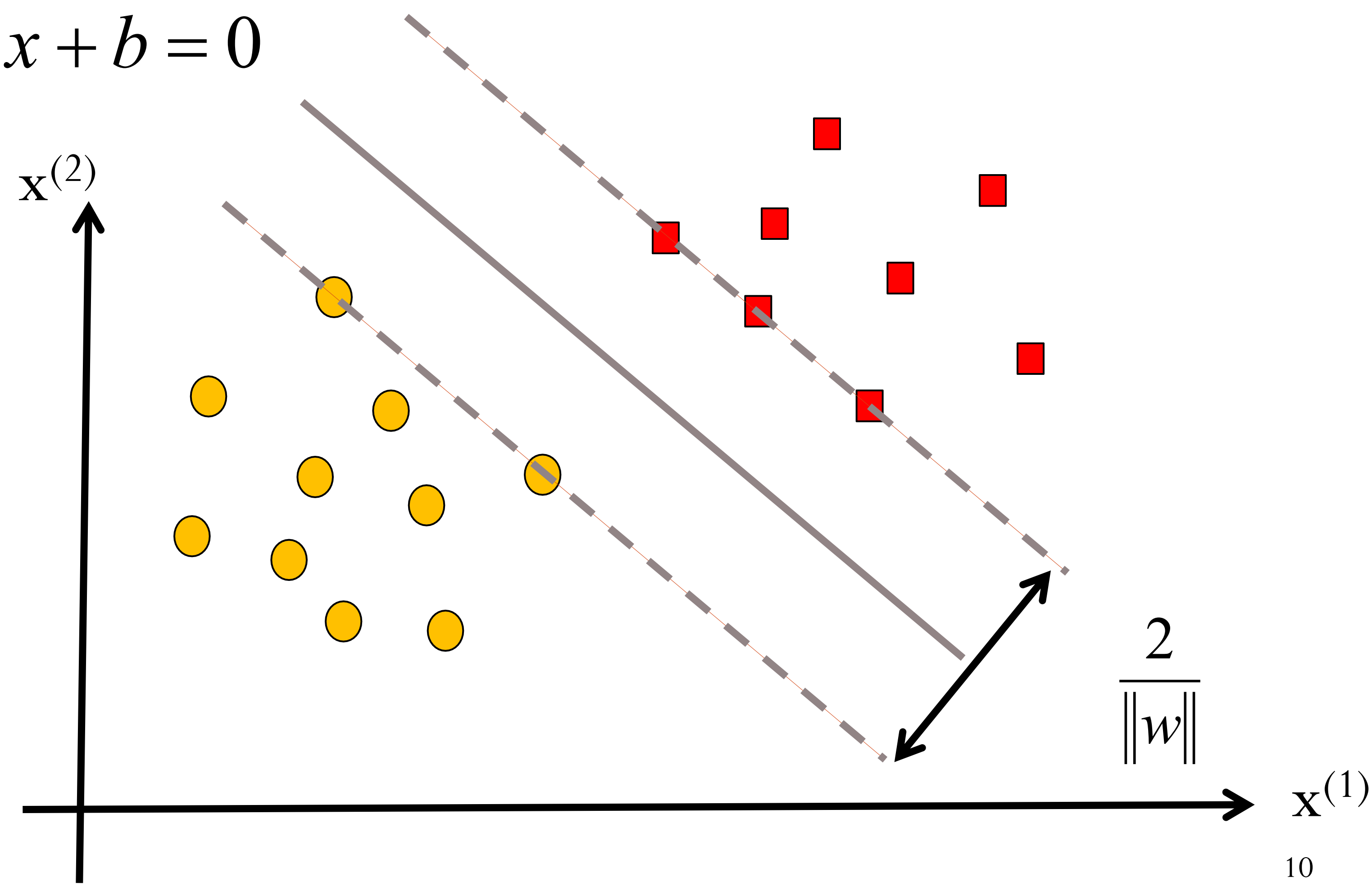
- Support Vector Machines, SVM
- 二类分类器
- 特征空间上的间隔最大化线性分类器
- 通过核技巧可实现非线性分类
- 根据模型的复杂程度可划分为：
  - 线性可分支持向量机
  - 线性支持向量机
  - 非线性支持向量机

线性可分支持向量机



最优分界面

$$w \cdot x + b = 0$$



定义4.1：给定线性可分训练集：

其中：$$T=\{(x_1,y_1),\dots,(x_N,y_N)\}$$

$$x_i\in X=\mathbf{R}^n, y_i\in Y=\{+1,-1\}, i=1,2,\dots,N$$

这里x_i为第i个特征向量，y_i为x_i的类标记，+1表示正类，-1表示负类

通过间隔最大化得到分类超平面：

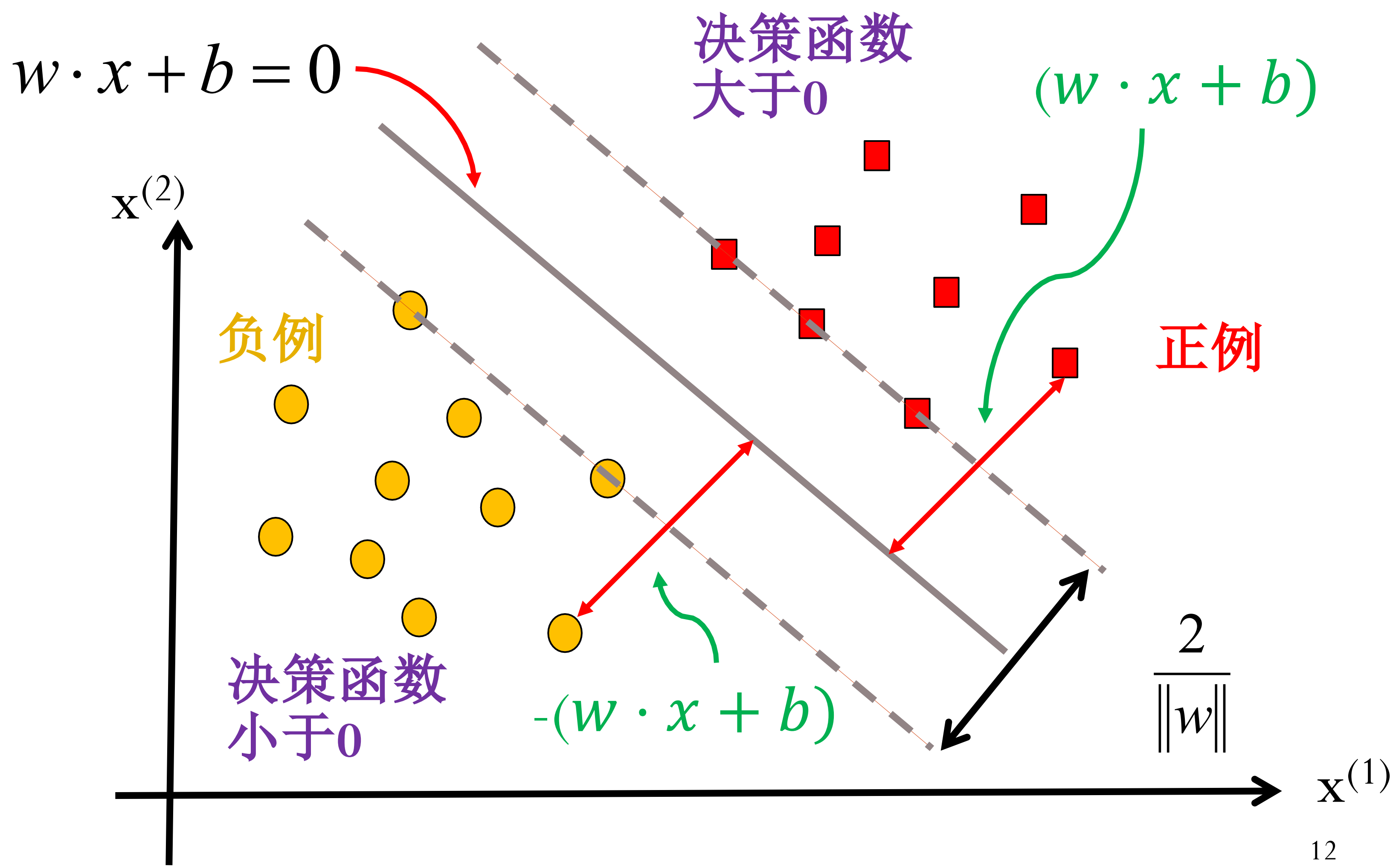
$$w^*\cdot x+b^*=0$$

相应的决策函数：

$$f(x)=sign(w^*\cdot x+b^*)$$

称为线性可分支持向量机

最优分界面



$w \cdot x + b = 0$
决策函数
大于0
$(w \cdot x + b)$
$x^{(2)}$
负例
正例
$-(w \cdot x + b)$
决策函数
小于0
$\frac{2}{\|w\|}$
$x^{(1)}$

函数间隔

- 设训练集T和超平面(w,b)，定义超平面(w,b)关于样本点(x_i,y_i)的函数间隔为：

$$\hat{\gamma}_i = y_i(w \cdot x_i + b)$$

- 定义超平面关于T的函数间隔为：

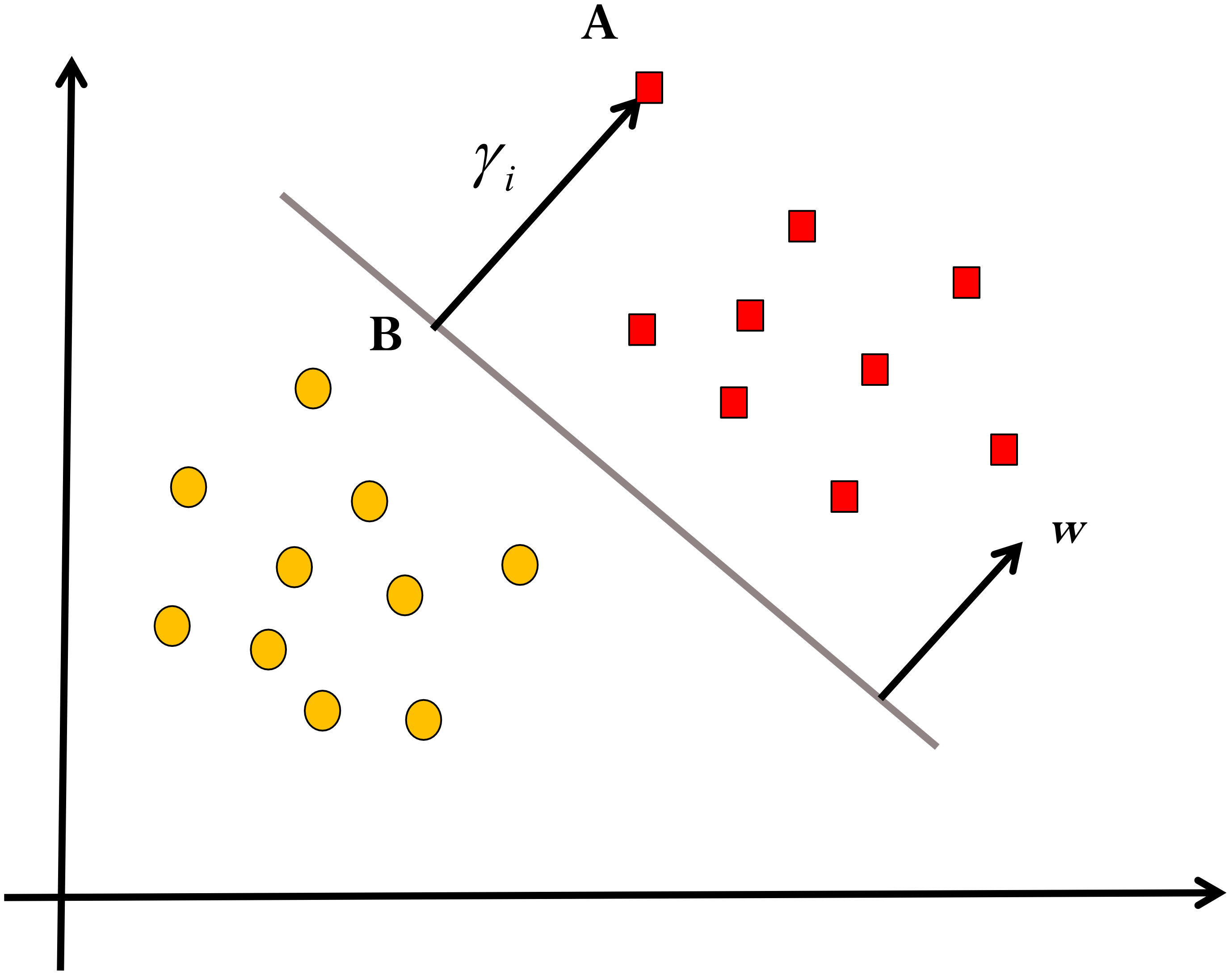
$$\hat{\gamma} = \min_i \hat{\gamma}_i$$

几何间隔

$$\gamma_{\mathrm{i}} = y_i \left(\frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|} \right)$$

$$\gamma = \min_i \gamma_i$$

其中$\|w\|$为w的L_2范数


A
γ_i
B
w

函数间隔与几何间隔的关系

$$\gamma_i = \frac{\hat{\gamma}_i}{\|w\|}$$

$$\gamma = \frac{\hat{\gamma}}{\|w\|}$$

间隔最大化

$$\max_{w,b} \gamma$$

$$s.t. \quad y_i\left(\frac{w}{\|w\|}\cdot x_i+\frac{b}{\|w\|}\right)\geq \gamma, \quad i=1,2,\ldots,N$$

用函数间隔表示为：

$$\max_{w,b}\left(\frac{\hat{\gamma}}{\|w\|}\right)$$

$$s.t. \quad y_i(w\cdot x_i+b)\geq \hat{\gamma}, \quad i=1,2,\ldots,N$$

- 由于函数间隔是可缩放的，成比例变化不影响最优化问题，所以可取 $\hat{\gamma}=1$

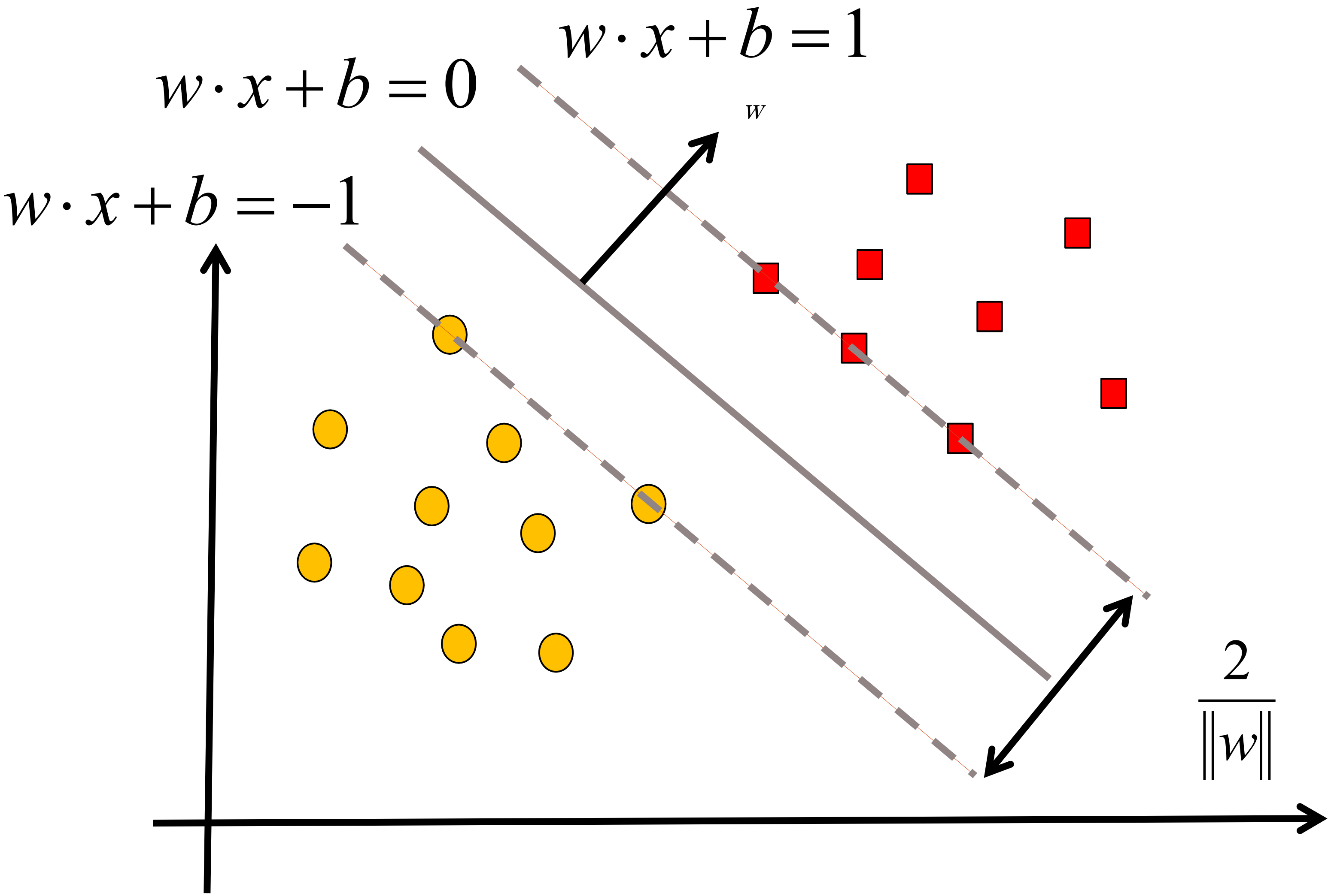
- 同时，最大化 $\frac{1}{\|w\|}$ 与最小化 $\frac{1}{2}\|w\|^2$ 是等价的，于是问题转化为如下的凸二次规划问题：

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

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1, \qquad i = 1,2,\dots,N$$

- 使上式等式成立的点构成了支持向量。

最优分界面



学习的对偶算法

◆原始问题：

$$1 - y_i(w \cdot x_i + b) \le 0$$

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

$$s.t. \quad y_i(w \cdot x_i + b) \ge 1, \qquad i = 1,2,\ldots,N$$

◆定义拉格朗日函数：

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

其中$\alpha_i \ge 0, \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_N)^T$为拉格朗日乘子向量

◆拉格朗日函数与原始优化问题的关系

约束：≤0

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i}^{N}\boxed{\alpha_i[1-y_i(w\cdot x_i+b)]}$$

$$\therefore \max_{\alpha} L(w,b,\alpha)=\begin{cases}\frac{1}{2}\|w\|^2, & \text{当满足约束条件时}\\ \infty\end{cases}$$

$\therefore \min_{w,b}\max_{\alpha} L(w,b,\alpha)$ 与原始优化问题等价

- 对偶问题

$$\min_{w,b}\max_{\alpha}\left(L(w,b,\alpha)\right) \Rightarrow \max_{\alpha}\min_{w,b}(L(w,b,\alpha))$$

$$\min_{w,b} L(w,b,\alpha) \le L(w,b,\alpha) \le \max_{\alpha} L(w,b,\alpha)$$

$$\max_{\alpha}\min_{w,b} L(w,b,\alpha) \le \min_{w,b}\max_{\alpha} L(w,b,\alpha)$$

- 当满足KKT条件时，上式等号成立
- 所以可以通过求解对偶优化问题求解原始优化问题

https://en.wikipedia.org/wiki/Minimax_theorem

https://en.wikipedia.org/wiki/Max–min_inequality

原始问题 $$\min_{w,b} \max_{\alpha} \left[\frac{1}{2}\|w\|^2 + \sum_{i}^{N} \alpha_i[1 - y_i(w \cdot x_i + b)]\right]$$

对偶问题 $$\max_{\alpha} \min_{w,b} \left[\frac{1}{2}\|w\|^2 + \sum_{i}^{N} \alpha_i[1 - y_i(w \cdot x_i + b)]\right]$$

KKT条件：

$$\nabla_{w,b} L(w, b, \alpha) = 0$$

$$\alpha_i[1 - y_i(w \cdot x_i + b)] = 0$$

$$[1 - y_i(w \cdot x_i + b)] \le 0$$

$$\alpha_i \ge 0$$

$$i = 1,2, \dots, N$$

◆对偶问题的求解：

$$\max_{\alpha}\min_{w,b}\left[\frac{1}{2}\|w\|^2+\sum_{i}^{N}\alpha_i[1-y_i(w\cdot x_i+b)]\right]$$

◆对w,b求偏导令其为0求解并代入，得到对偶问题：

$$\max_{\alpha}\left(-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i\right)$$

$$s.t.\ \ \sum_{i=1}^{N}\alpha_i y_i=0,\ \ \alpha_i\geq 0,\ i=1,2,\ldots,N$$

- **目标函数加个负号，由求极大转换成求极小，得到等价的对偶问题：**

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

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

$$\alpha_i\geq 0,\ i=1,2,\ldots,N$$

解出的最优解为α^*

如何求得w*、b*?

KKT条件：

$$\nabla_{w,b} L(w,b,\alpha) = 0$$

$$\alpha_i[1 - y_i(w \cdot x_i + b)] = 0$$

$$\nabla_w L(w,b,\alpha) = 0$$

$$\Rightarrow w^* - \sum_{i=1}^{N} \alpha_i^* y_i x_i = 0$$

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

$$\alpha_i[1 - y_i(w \cdot x_i + b)] = 0, i = 1 \ldots N$$

$$\Rightarrow b^* = y_j - w \cdot x_j, \qquad \text{选择一个}\alpha_j \neq 0$$

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

- 分类超平面：

$$w^* \cdot x + b^* = 0$$

- 分类决策函数为：

$$f(x) = sign(w^* \cdot x + b^*)$$

- 其中：

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

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

- 与$\alpha_i > 0$对应的实例x_i就是支持向量。???

KKT条件：

$$\nabla_{w,b} L(w,b,\alpha) = 0$$

$$\alpha_i[1 - y_i(w \cdot x_i + b)] = 0$$

$$[1 - y_i(w \cdot x_i + b)] \le 0$$

$$\alpha_i \ge 0$$

$$i = 1,2,\ldots,N$$

◆因此线性可分支持向量机就是求解如下的优化问题：

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

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

$$\alpha_i \geq 0,\ i = 1,2,\ldots,N$$

- 例：设正例：$\mathbf{x}_1=(3,3)^T$, $\mathbf{x}_2=(4,3)^T$,

  负例：$\mathbf{x}_3=(1,1)^T$，

- 用对偶问题求线性可分支持向量机。

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

$$= \min_{\alpha} \frac{1}{2}(18\alpha_1^2 + 25\alpha_2^2 + 2\alpha_3^2 + 42\alpha_1\alpha_2 - 12\alpha_1\alpha_3$$

$$-14\alpha_2\alpha_3) - \alpha_1 - \alpha_2 - \alpha_3$$

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

$$\alpha_i \ge 0, i = 1,2,3$$

将$\alpha_3=\alpha_1+\alpha_2$代入，并记为：

$$s(\alpha_1,\alpha_2)=4\alpha_1^2+\frac{13}{2}\alpha_2^2+10\alpha_1\alpha_2-2\alpha_1-2\alpha_2$$

通过求偏导并令其为0，易知$s(\alpha_1,\alpha_2)$在点

$\left(\frac{3}{2},-1\right)^T$取极值，但该点不满足约束$\alpha_2\geq 0$

所以最小值应该在边界上。

当$\alpha_1=0$时，最小值$s\left(0,\frac{2}{13}\right)=-\frac{2}{13}$,

当$\alpha_2=0$时，最小值$s\left(\frac{1}{4},0\right)=-\frac{1}{4}$,

于是$s(\alpha_1,\alpha_2)$在$\alpha_1=\frac{1}{4},\alpha_2=0$时达到最小,

此时$\alpha_3=\alpha_1+\alpha_2=\frac{1}{4}$

这样α_1,α_3对应的实例点x_1,x_3是支持向量

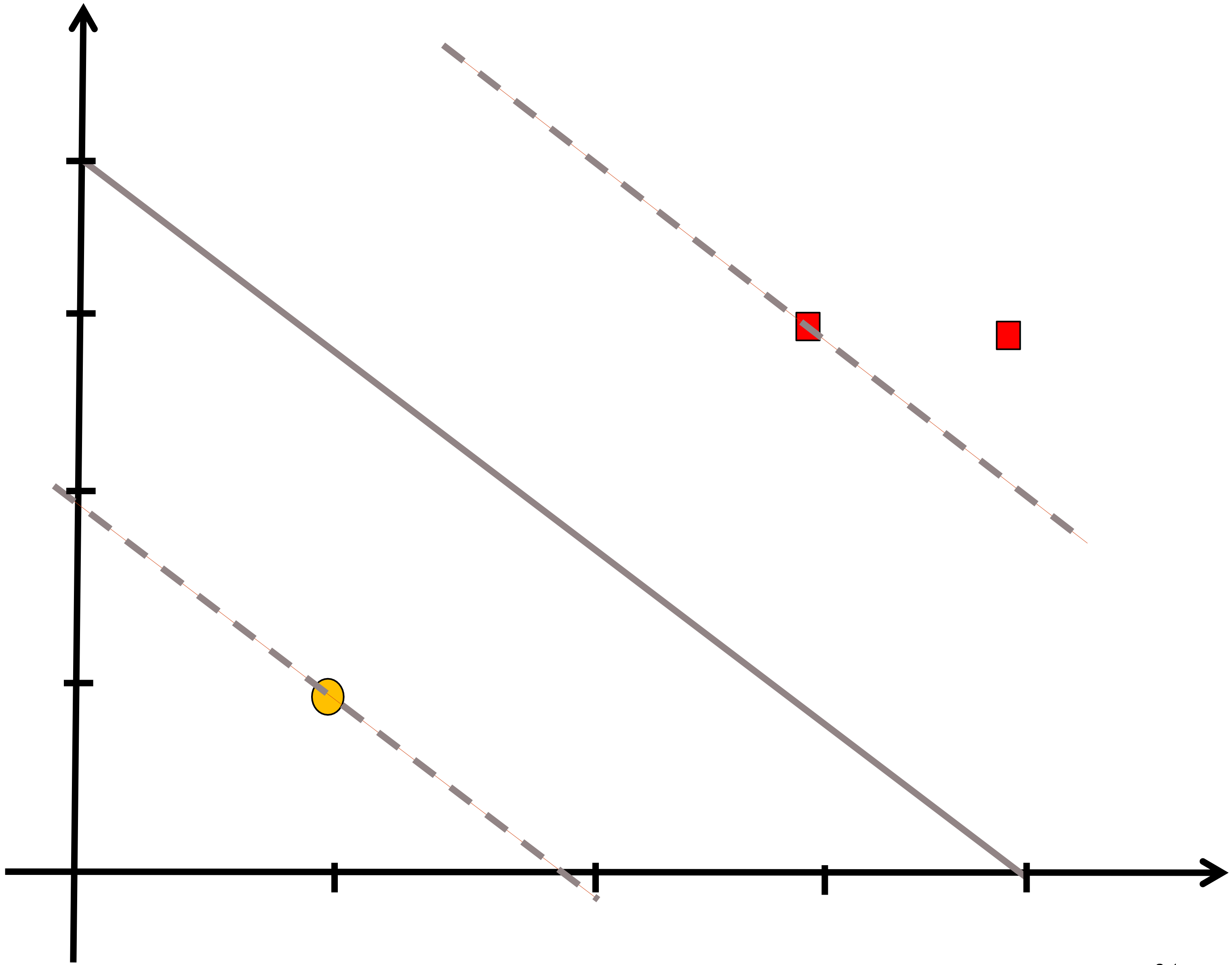
根据前面的公式得到：

$$w_1^* = w_2^* = \frac{1}{2}$$

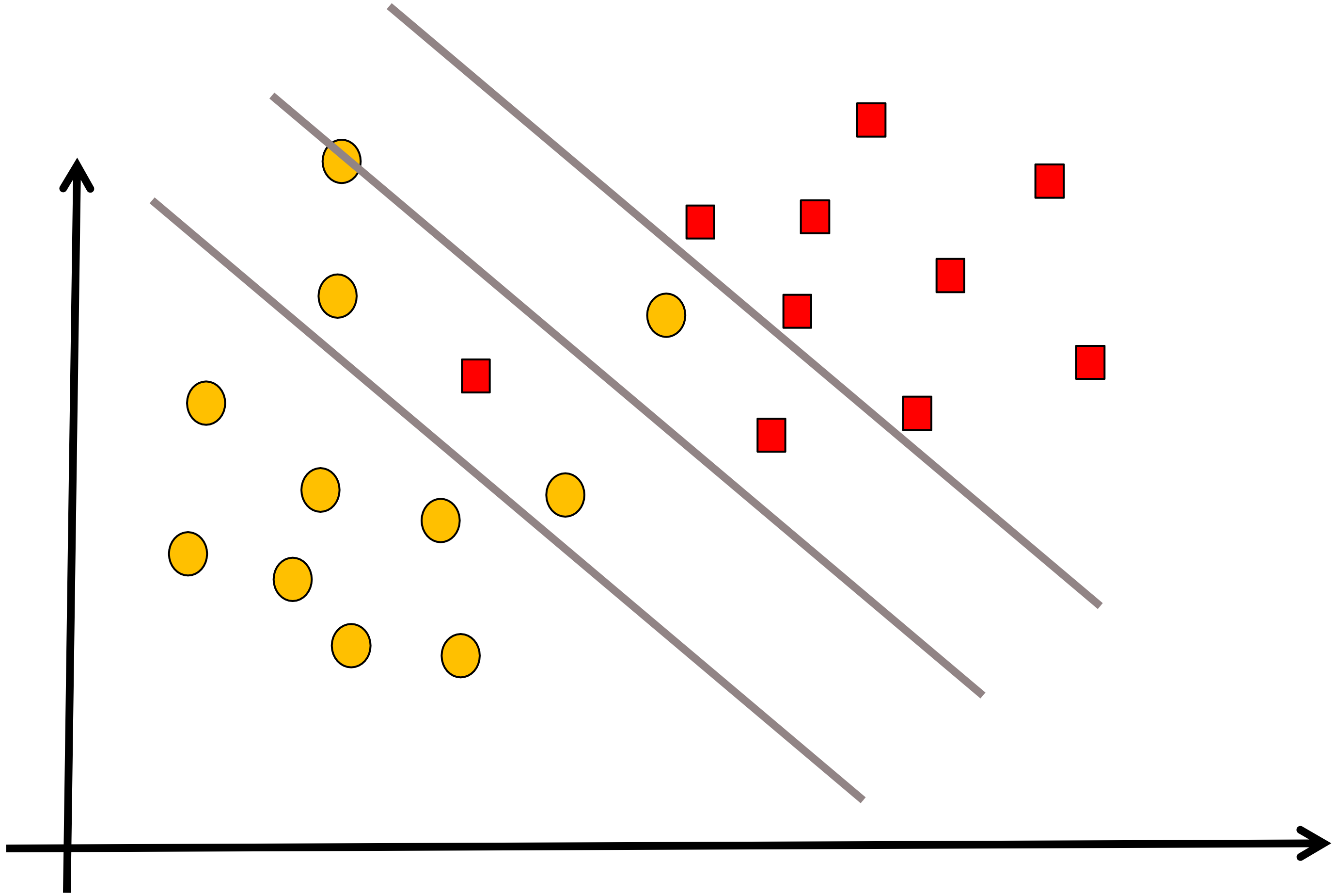
$$b^* = -2$$

分离超平面为：$\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2 = 0$

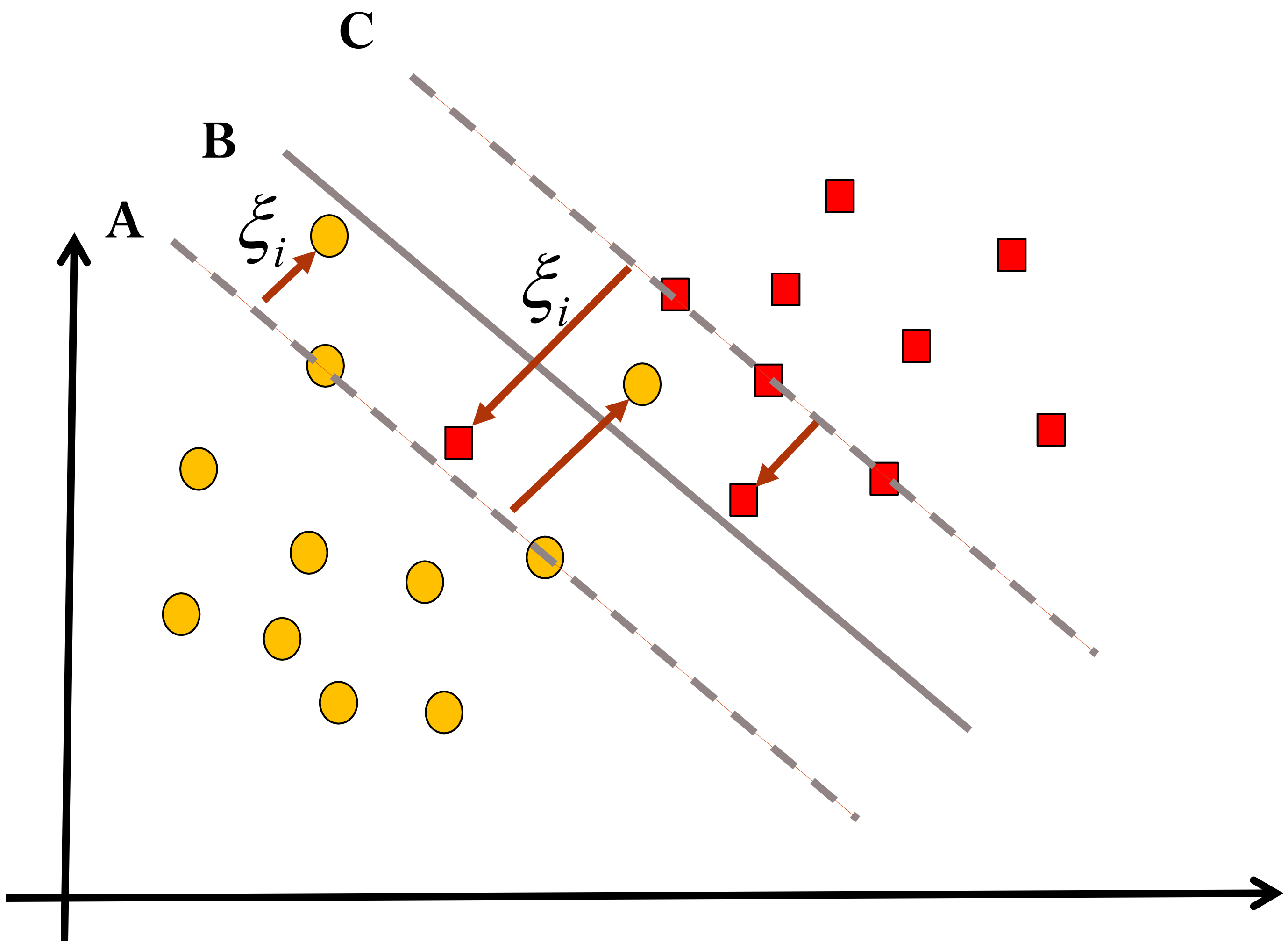
分类决策函数为：$f(x) = sign\left(\frac{1}{2}x^{(1)} + \frac{1}{2}x^{(2)} - 2\right)$



线性支持向量机



线性支持向量机



回顾：线性可分支持向量机

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

$$s.t. \quad y_i(w \cdot x_i + b) \geq 1, \qquad i = 1,2,\dots,N$$

- 某些点线性不可分，意味着这些点不满足函数间隔大于等于1的条件。
- 为此引入松弛变量 ξ_i ，使得：

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \ i = 1,2,\dots,N, \ \xi_i \geq 0$$

- 为使ξ_i尽可能的小，优化目标增加惩罚项，变为：

$$\min_{w,b,\xi}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i\right)$$

- 称为软间隔最大化
- 其中C>0为惩罚参数，C大时对误分类的惩罚增加，C小时对误分类的惩罚减少。
- 上式的含义：间隔尽量最大，同时误分类的点数尽可能小，二者由C调和。

线性支持向量机就转化为如下的优化问题（原始问题）：

$$\min_{w,b,\xi}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i\right)$$

$$s.t.\quad y_i(w\cdot x_i + b) \geq 1 - \xi_i,\qquad i = 1,2,\ldots,N$$

$$\xi_i \geq 0,\quad i = 1,2,\ldots,N$$

- **同样，通过求解对偶问题求解原始问题**

线性支持向量机的对偶问题：

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

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

$$0 \le \alpha_i \le \mathrm{C}, i = 1,2,\ldots,N$$

求得最优解$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$

计算：$w^* = \sum_{i=1}^{N} \alpha_i^* y_i x_i$

选择一个$0 < \alpha_j^* < C$,计算：

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

- 分类超平面：

$$w^* \cdot x + b^* = 0$$

- 分类决策函数为：

$$f(x) = sign(w^* \cdot x + b^*)$$

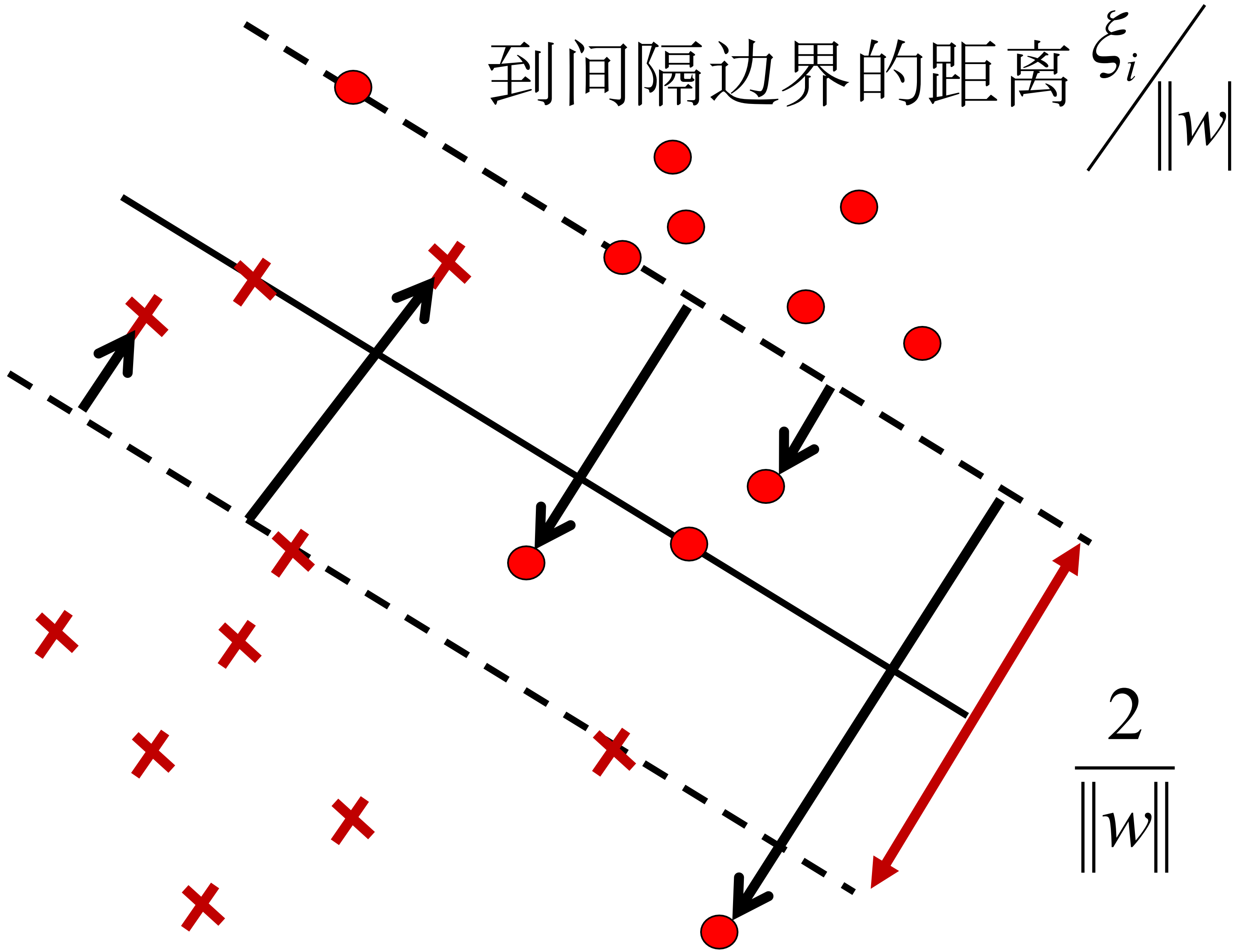
$\alpha_i^* > 0$所对应的样本x_i称为支持向量（软间隔的支持向量）

若$\alpha_i^* < C$,则$\xi_i = 0$, x_i在间隔边界上

若$\alpha_i^* = C$,$0 < \xi_i < 1$,则分类正确,x_i在间隔边界与分离超平面之间

若$\alpha_i^* = C$,$\xi_i = 1$,则x_i在超平面上

若$\alpha_i^* = C$,$\xi_i > 1$,则x_i位于误分一侧


到间隔边界的距离 $\xi_i / \|w\|$
$\frac{2}{\|w\|}$

多选题 1分

设置

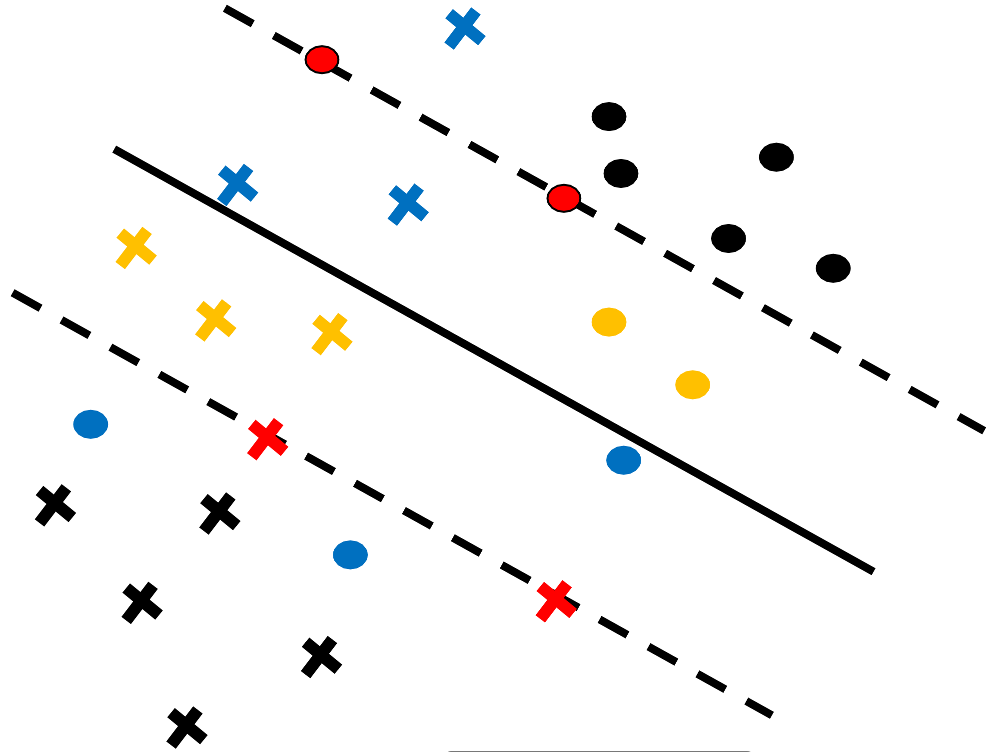
设训练样本如图所示，请选择哪些是支持向量。

A 红颜色样本

B 黄颜色样本

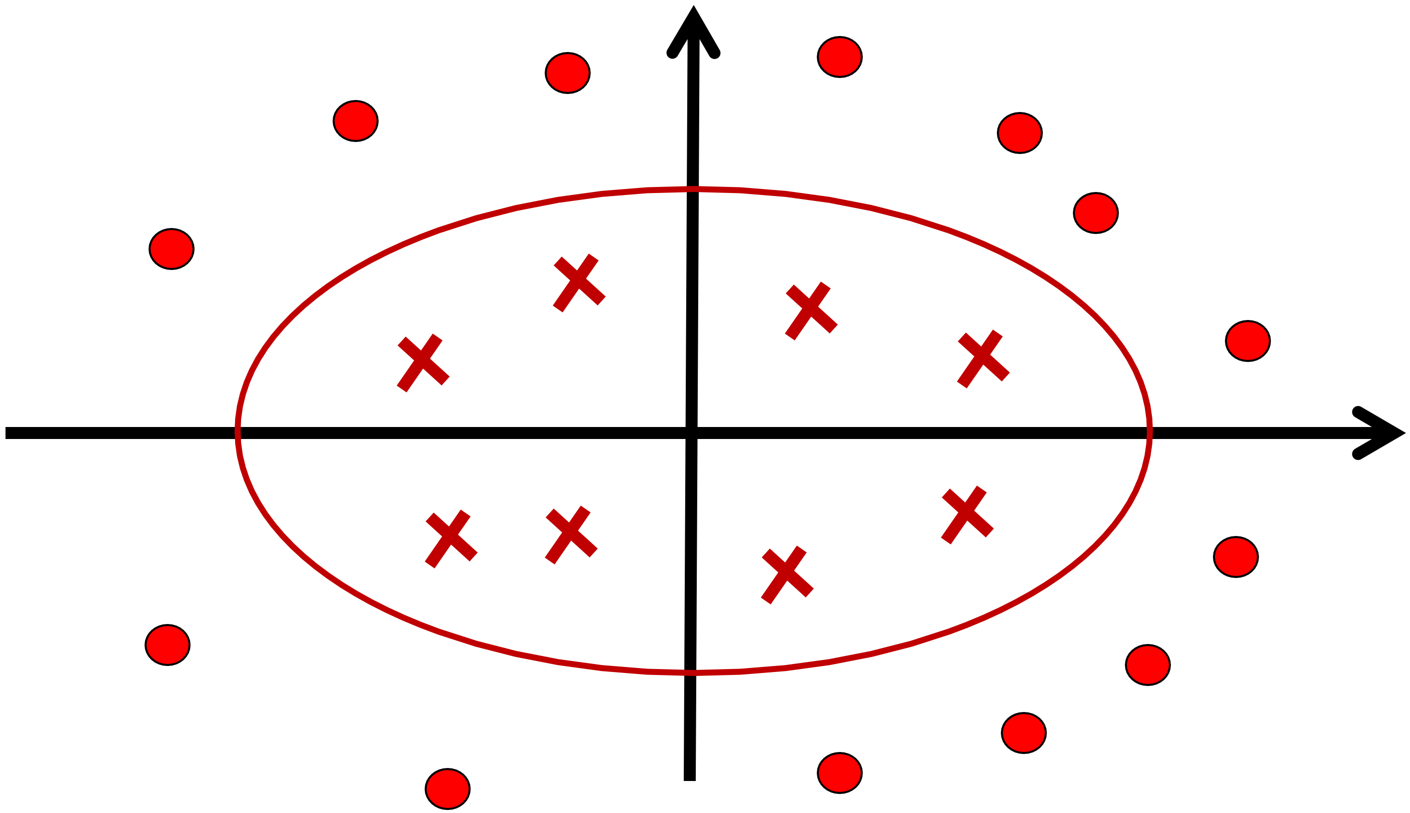
C 蓝颜色样本

D 黑颜色样本



提交

非线性支持向量机



设变换：

$$z = \phi(x) = ((x^{(1)})^2, (x^{(2)})^2)^T$$

原空间的椭圆：

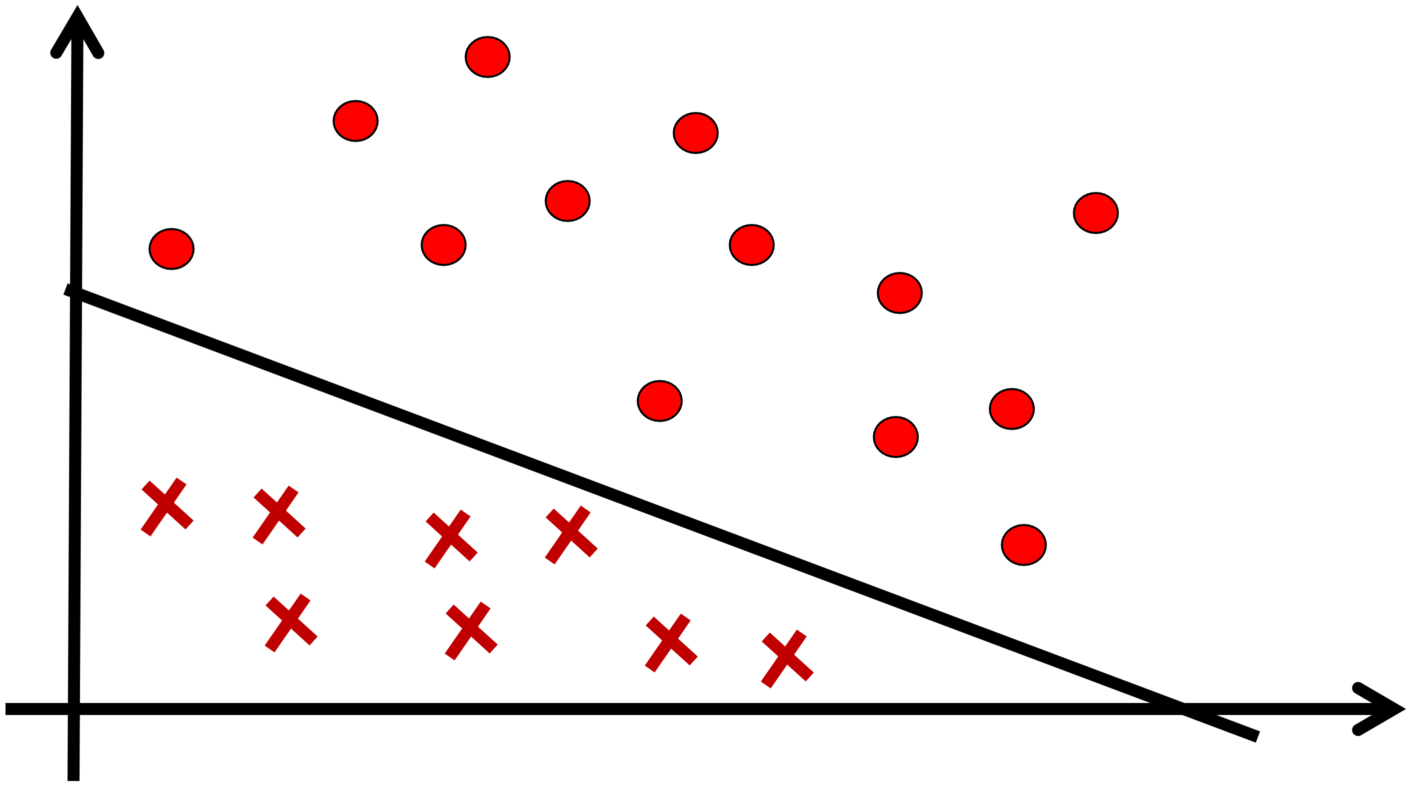
$$w_1 (x^{(1)})^2 + w_2 (x^{(2)})^2 + b = 0$$

变换为新空间中的直线：

$$w_1 z^{(1)} + w_2 z^{(2)} + b = 0$$

这样原空间的非线性可分问题

变为了新空间线性可分问题。



- 用线性分类的方法求解非线性分类问题

（1）使用一个变换，将原空间数据映射到新空间；

（2）在新空间用线性分类方法从训练数据中学习分类模型

线性支持向量机的对偶问题：

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

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

$$0 \le \alpha_i \le C, i = 1,2,\ldots,N$$

非线性支持向量机的对偶问题：

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

$$\mathrm{s.}t. \sum_{i=1}^{N}\alpha_i y_i = 0 \qquad 0 \le \alpha_i \le C, \qquad i = 1,2,\ldots,N$$

核函数

设**X**是输入空间，**H**是特征空间，如果存在

$$\phi(x): \mathbf{X} \to \mathbf{H}$$

使得对所有$x, z \in \mathbf{X}$,函数$K(x,z)$满足：

$$K(x,z) = \phi(x) \cdot \phi(z), \quad ("\cdot"\text{表示内积})$$

则称$K(x,z)$为核函数，$\phi(x)$为映射函数

核函数举例：

设核函数是$K(x,z)=(x\cdot z)^2$,试找出映射$\phi(x)$

记$x=(x^{(1)},x^{(2)}), z=(z^{(1)},z^{(2)})$,由于

$$
\begin{aligned}
(x\cdot z)^2 &= (x^{(1)}z^{(1)}+x^{(2)}z^{(2)})^2 \\
&= (x^{(1)}z^{(1)})^2+2x^{(1)}z^{(1)}x^{(2)}z^{(2)}+(x^{(2)}z^{(2)})^2
\end{aligned}
$$

所以可取映射：$\phi(x)=((x^{(1)})^2,\sqrt{2}x^{(1)}x^{(2)},(x^{(2)})^2)^T$

也可以取映射：$\phi(x)=((x^{(1)})^2,x^{(1)}x^{(2)},x^{(1)}x^{(2)},(x^{(2)})^2)^T$

可见，映射$\phi(x)$并不唯一

非线性支持向量机算法：

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

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

$$0 \le \alpha_i \le C, \ i = 1,2,\ldots,N$$

求得最优解：$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$

选择一个α^*的分量$0 < \alpha_j^* < C$，计算：

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x_j)$$

$$w^* = \sum_{i=1}^{N} \alpha_i^* y_i \emptyset(x_i)$$

分界超平面：

$$w^* \cdot \emptyset(x) + b^* = 0$$

$$\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^* = 0$$

决策函数：$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^*\right)$

常用的核函数

多项式核函数：

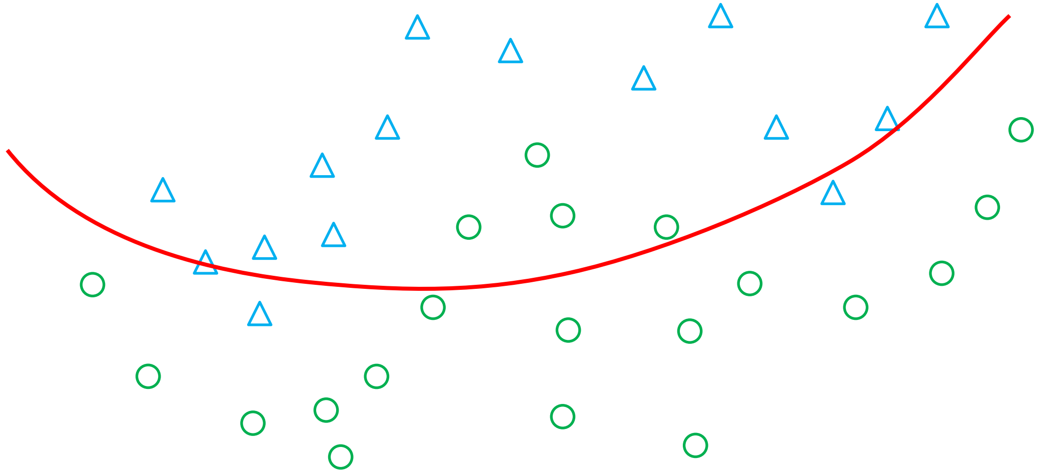
$$K(x,z)=(x\cdot z+1)^p$$

高斯核函数：

$$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right)$$

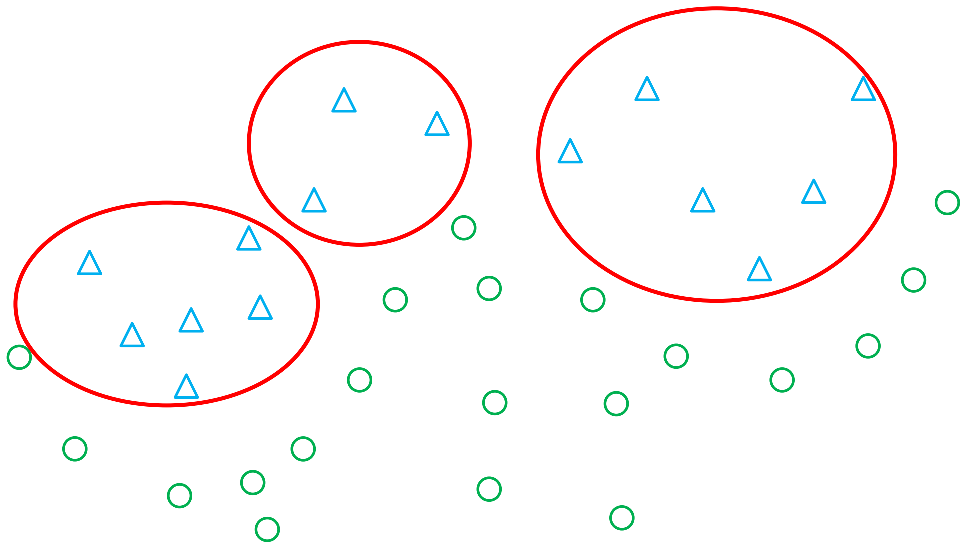
不同σ值下的分界面示意图

- σ值比较大时——欠拟合



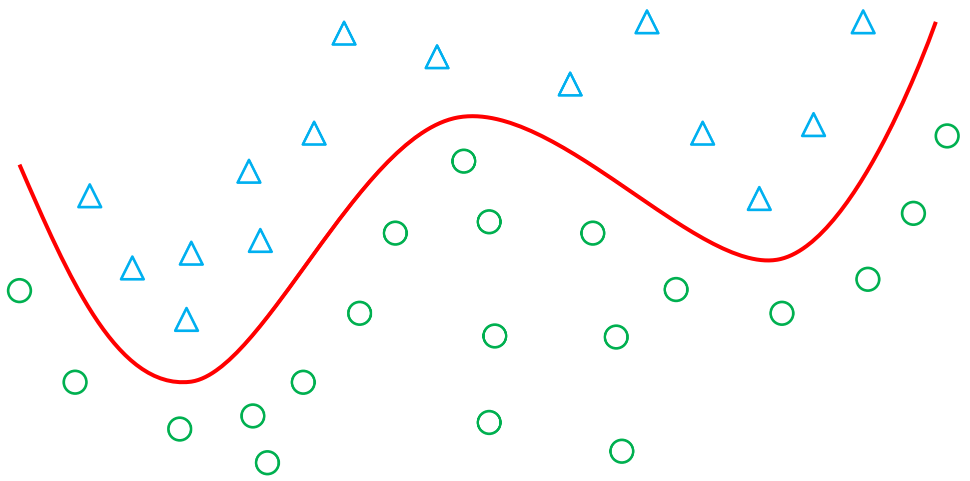
不同σ值下的分界面示意图

- σ值比较小时——过拟合

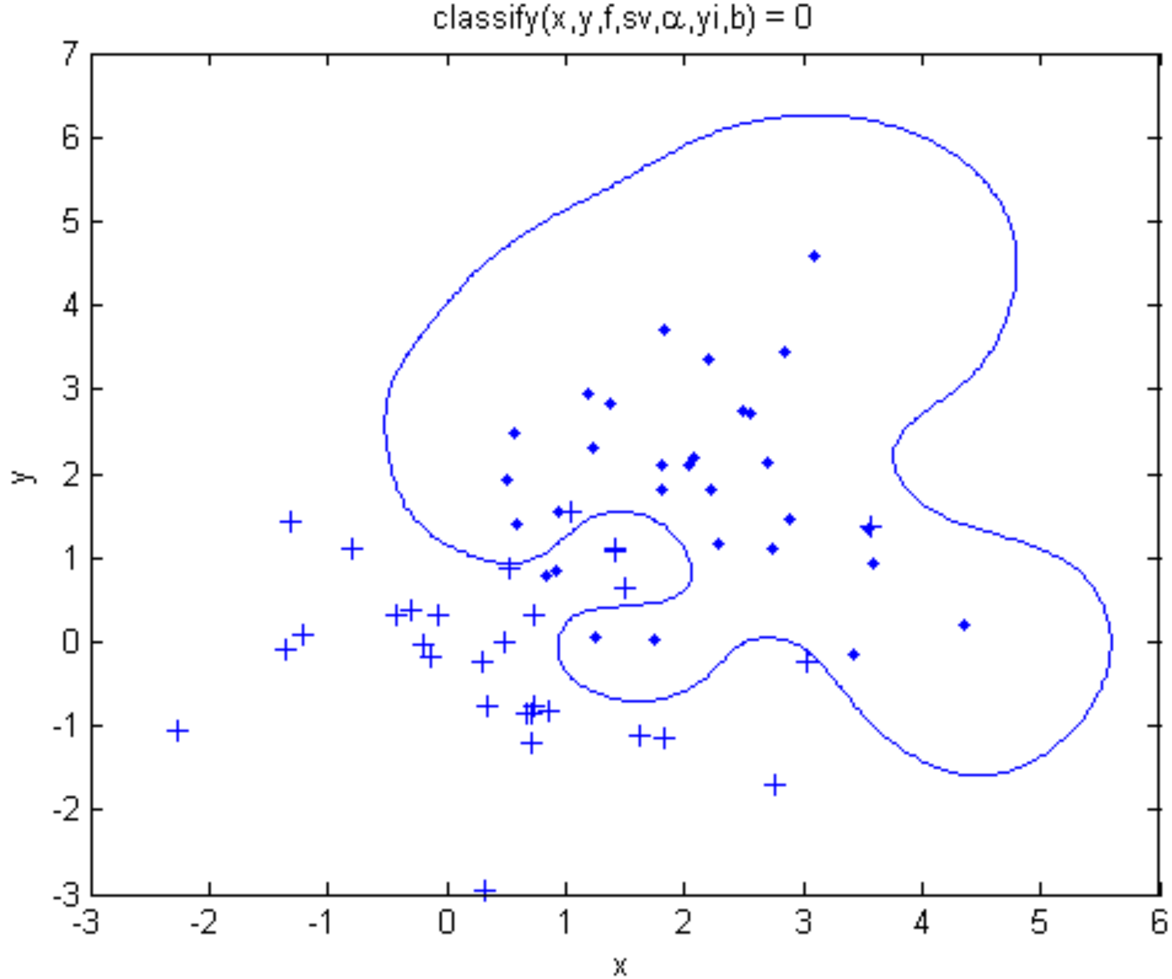


不同σ值下的分界面示意图

- σ值合适时——恰拟合



一个非线性分类的例子



序列最小最优化算法SMO

- 支持向量机的学习问题是一个凸二次规划问题，具有全局最优解。
- 有很多算法可以求解这类问题，但是当样本数多时，往往非常低效，以致无法使用。
- 为此，提出了许多快速算法，SMO 由微软的Platt与1998年提出，当时最快。

SVM用于求解多类问题

- 一对多
  - 某类为正例，其余类为负例。分类时将未知样本分类为具有最大分类函数值的那类
- 一对一
  - 任意两类构造一个SVM，分类时采取投票法决定类别
- 层次法
  - 所有类先分成两类，每类再分为两类……

SVM应用举例：文本分类

- 文本的向量空间模型
  - 文本表达为一个向量
  - $(w_{1,j}, w_{2,j}, \ldots\ldots w_{n,j})^T$
  - w_{ij}表示词项i在文档j中的权重
- 词项频率tf_{ij}权重
- tf-idf权重

- tf_{ij}权重
  - $w_{ij} = tf_{ij}$
  - tf_{ij}表示第i个词项在第j个文档中的词频

- **tf-idf权重**
  - 文档频率：df_i=出现词项i的文档数/N
    - N为训练集的文档总数
  - 逆文档频率：$idf_i=\log(1/df_i)$,
- (1) $w_{ij}= tf_{ij}*idf_i$
- 此外还有很多变形

实际中的问题

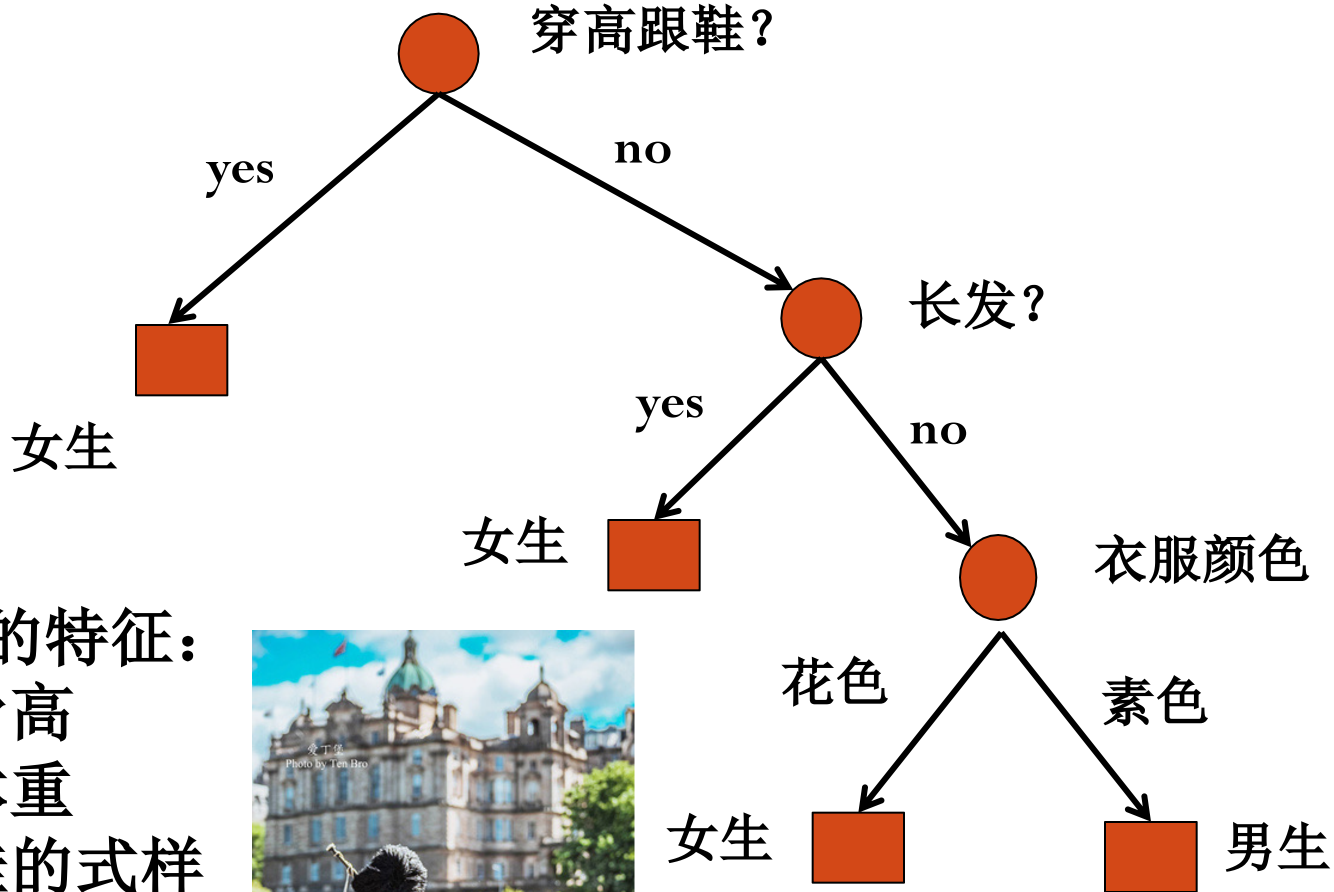
- 分类体系的建立
- 数据的收集
- 预处理
  - 分词
  - 停用词（Stop word）处理
  - 词干化（Stemming）
  - 特征选择

练习题

- 使用工具系统实现基于支持向量机方法的文本分类，并对比采用不同的特征、不同的超参数时，分类性能的优劣。

4.2 决策树

- 决策树模型是一种描述对实例进行分类的树形结构，由节点和有向边组成。
- 节点有两种类型：内部节点和叶节点。
- 内部节点表示一个特征或者属性
- 叶节点表示一个类。



可用的特征：

身高

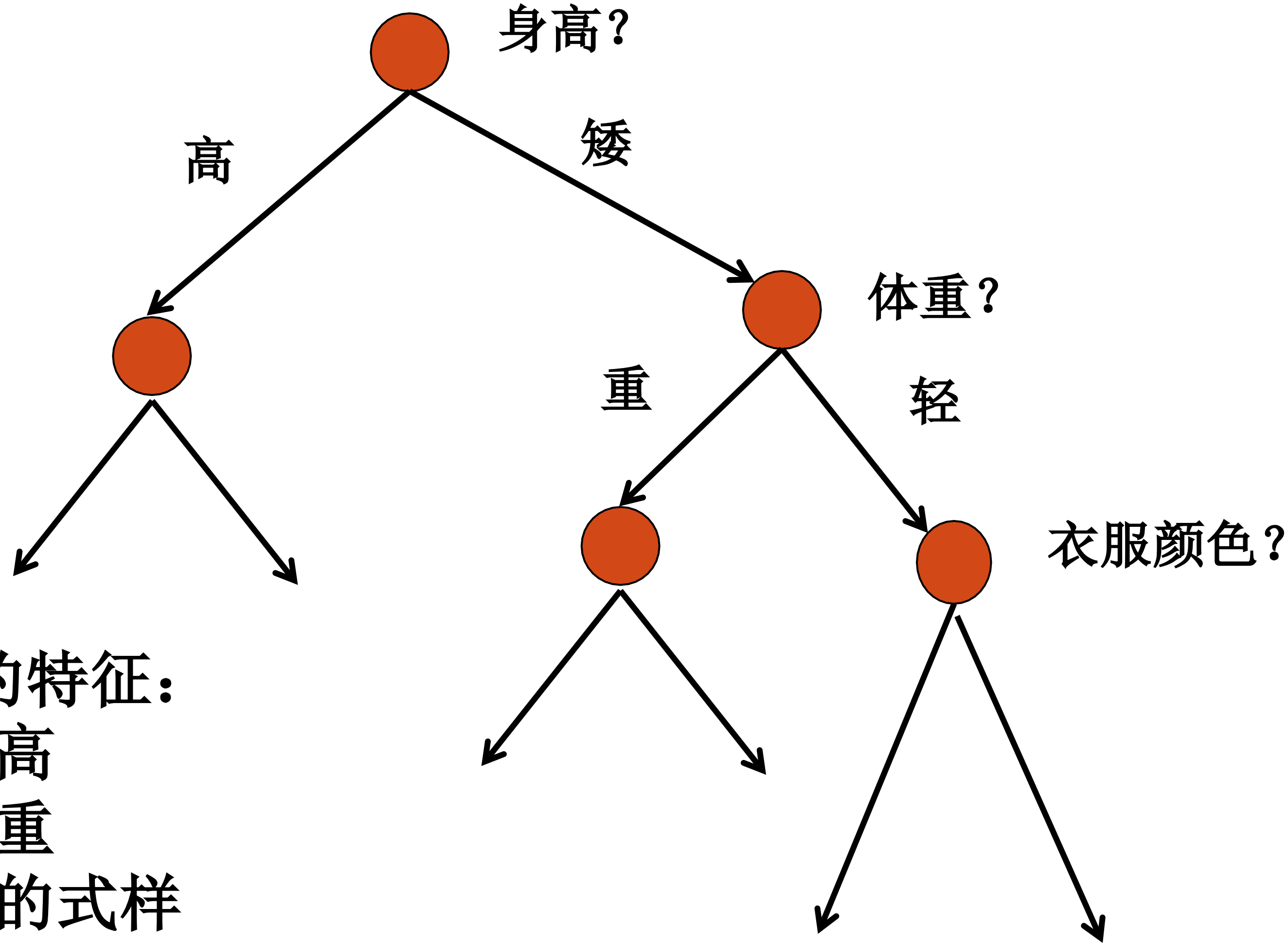
体重

鞋的式样

头发长度

衣服颜色

……



可用的特征：

身高
体重
鞋的式样
头发长度
衣服颜色
……

决策树学习

给定训练集：$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$

其中$x_i=(x_i^{(1)},\ldots,x_i^{(n)})$为输入实例，$n$为特征个数，

$y_i\in\{1,2,\ldots,K\}$为类标记

$i=1,2,\ldots,N$，N为样本容量

- 决策树学习就是从训练集中归纳出一组分类规则，得到一个与训练集矛盾较小的决策树

- 对于给定的训练集，可以构造出多个决策树，一般以损失函数最小化作为优化目标
- 从所有决策树中选取最优决策树是一个NPC问题，所以一般采用启发式方法，得到一个近似解

- 决策树学习包括
  - 特征选择
  - 决策树生成
  - 决策树剪枝

特征选择

- 一个问题中可能有不同的特征，不同的特征具有不同的分类能力，特征选择就是如何选取出那些分类能力强的特征。
- 决策树中一般按照**信息增益**选择特征
- 所谓的信息增益就是某个特征A对数据集D进行分类的不确定性减少的程度

信息增益

随机变量X的熵：

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i,\quad \text{其中} p_i = P(X = x_i), \text{也记作} H(p).$$

当概率由数据集D估计得到时，记作$H(D)$

条件熵：

$$H(Y \mid X) = \sum_{i=1}^{n} p_i H(Y \mid X = x_i)$$

表示已知X时Y的不确定性

- **特征A对数据集D的信息增益定义为：**

$$g(D,A)=H(D)-H(D\mid A)$$

- **表示特征A对数据集D的分类的不确定性减少的程度**
- **信息增益大的特征具有更强的分类能力**

- 设训练集D，K个类C_k，特征A有n个不同的取值$\{a_i,\ldots,a_n\}$，A的不同取值将D划分为n个子集$D_1\ldots D_n$，D_i中属于类C_k的样本的集合为D_{ik}，$|\cdot|$表示样本个数。
- 信息增益计算如下：

$$H(D)=-\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log_2\frac{|C_k|}{|D|}$$

$$H(D\,|\,A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i)=-\sum_{i=1}^{n}\frac{|D_i|}{|D|}\sum_{k=1}^{K}\frac{|D_{ik}|}{|D_i|}\log_2\frac{|D_{ik}|}{|D_i|}$$

$$g(D,A)=H(D)-H(D\,|\,A)$$

决策树的生成

- 两个常用的算法

- ID3
  - 一个基本的决策树生成算法
- C4.5
  - 对ID3的改进

ID3算法

- 输入：训练集D，特征集A，阈值$\varepsilon > 0$
- 输出：决策树T
- 1，若D中所有实例属于同一类C_k，则T为单节点树，将C_k作为该节点的类标记，返回T
- 2，若A为空，则T为单节点树，将D中实例数最多的类C_k作为该节点的类标记，返回T
- 3，否则计算A中各特征对D的信息增益，选择信增益息最大的特征A_g
- 4，如果A_g的信息增益小于阈值ε，则置T为单节点树，将D中实例数最大的类C_k作为该节点的类标记，返回T

- 5，否则对A_g的每一可能值a_i，依$A_g=a_i$将D分割为若干子集D_i，作为D的子节点
- 6，对于D的每个子节点D_i，如果D_i为空，则将D中实例最多的类作为标记，构建子节点
- 7，否则以D_i为训练集，以A-{A_g}为特征集，递归地调用步1~步6，得到子树T_i，返回T_i

- 例：贷款申请样本如下表所示，试用ID3算法构建决策树。

ID	年龄 A1	有工作 A2	有房子 A3	信贷情况 A4	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

$$H(D) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} = 0.971$$

$$g(D, A_1) = H(D) - \left[\frac{5}{15}H(D_1) + \frac{5}{15}H(D_2) + \frac{5}{15}H(D_3)\right]$$

$$= 0.971 - \frac{5}{15}\left[\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) + \left(-\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5}\right)\right] = 0.083$$

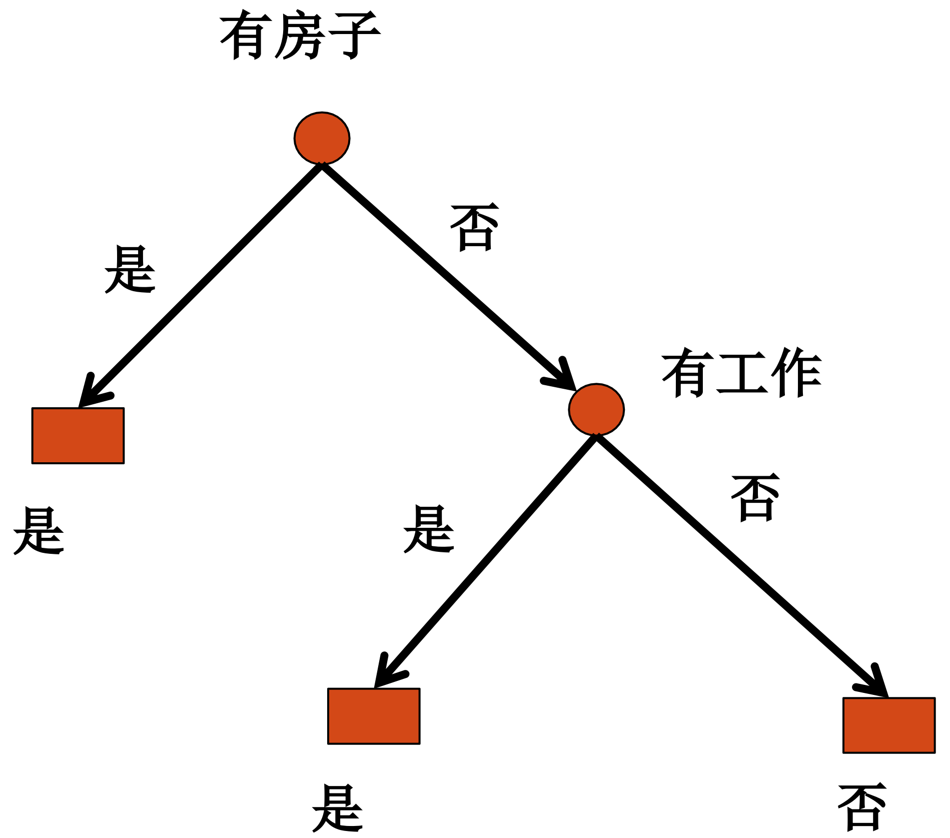
$$g(D, A_2) = 0.324$$

$$g(D, A_3) = 0.420$$ 该信息增益最大

$$g(D, A_4) = 0.363$$

A_3作为根节点，将D划分为$D_1(A_3=是)$和
$D_2(A_3=否)$,D_1成为叶结点
对D_2从特征A_1、A_2、A_4中选择特征
$g(D_2,A_1)=0.251$
$g(D_2,A_2)=0.918$ 信息增益最大
$g(D_2,A_4)=0.474$
选取A_2作为节点的特征，根据其两个取值，可以得到两个子节点，一个对应“有工作”，并且是一个叶节点，标记类别为“是”。另一个节点对应“无工作”，且样本属于同一类，也是一个叶节点，标记类别为“否”

◆生成的决策树如下：



多选题 1分

设置

关于ID3算法，请选择以下正确的说法

A 在生成决策树过程中必须使用所有的特征

B 允许生成的决策树叶节点实例类别不一样

C 同一个特征只能用在一个节点上

D 得到的决策树并不能保证最优

提交

ID3存在的问题

- 信息增益倾向于选择分枝比较多的属性

- 比如前面贷款的例子中，如果用ID做属性，将获得最大的信息增益值

信息增益比

$$g_R(D,A)=\frac{g(D,A)}{H_A(D)}$$

$$H_A(D)=-\sum_{k=1}^{n}\frac{|D_k|}{|D|}\log_2\frac{|D_k|}{|D|}$$

- 其中A为属性，A的不同取值将D划分为n个子集$D_1 \ldots D_n$

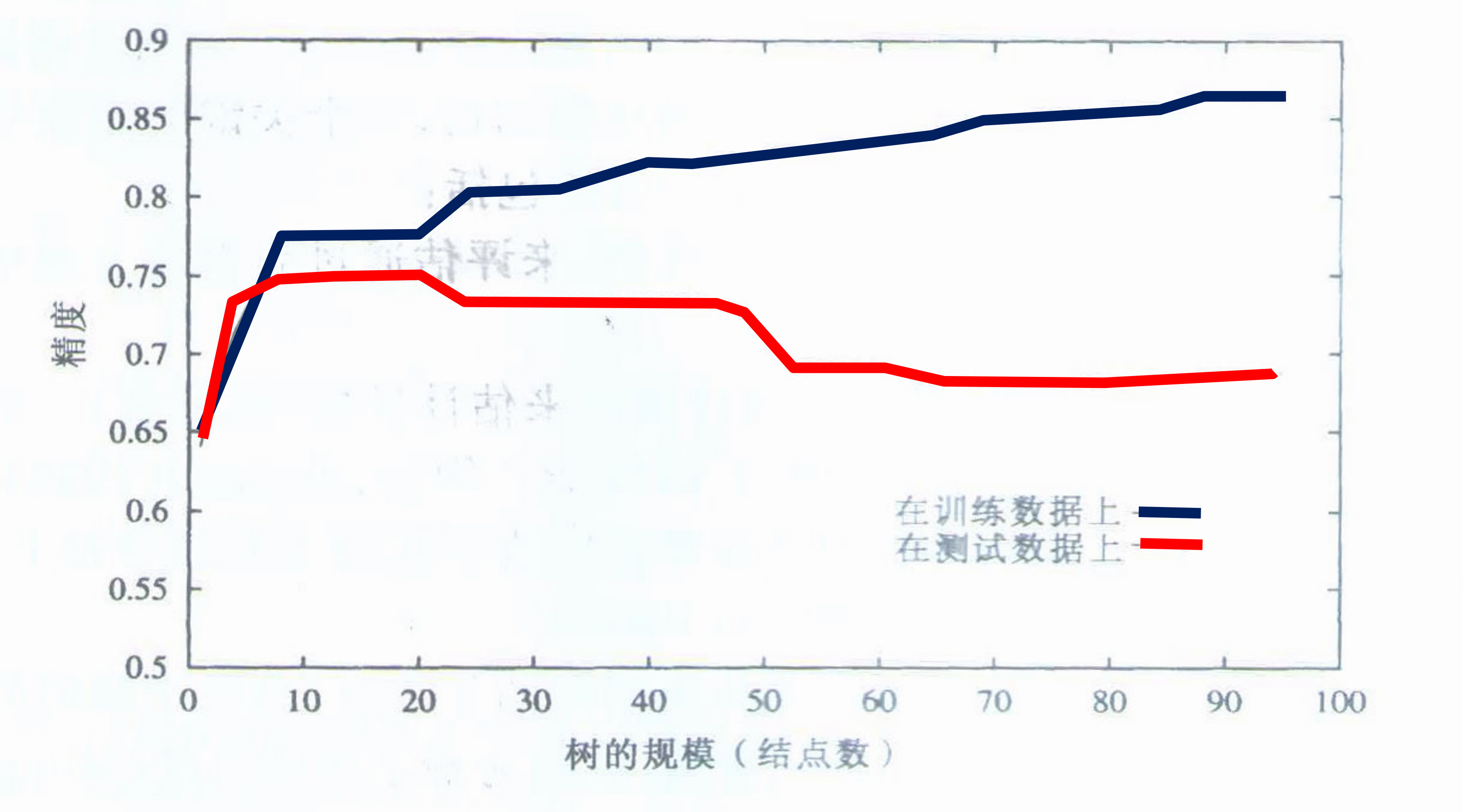
C4.5的生成算法

- 除了根据信息增益比选择特征外，C4.5算法与ID3基本一样。

- 同时C4.5增加了对连续值属性的处理，对于连续值属性A，找到一个属性值a_0，将$\leq a_0$的划分到左子树，＞a_0的划分到右子树

- 信息增益比的问题：
  - 倾向于选择分割不均匀的特征
- 解决办法
  - 先选择n个信息增益大的特征，再从这n个特征中选择信息增益比最大的特征

过拟合问题



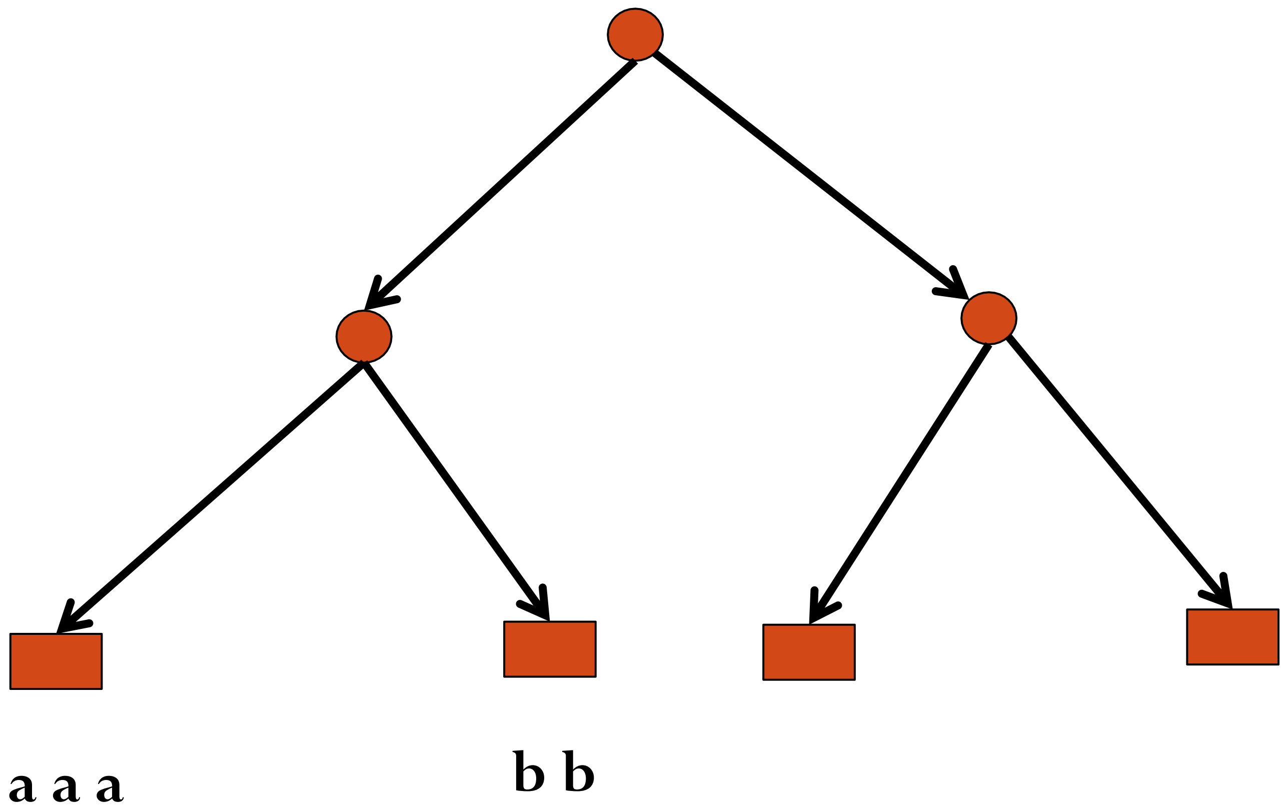
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
精度
0
10
20
30
40
50
60
70
80
90
100
树的规模（结点数）
在训练数据上
在测试数据上

决策树的剪枝

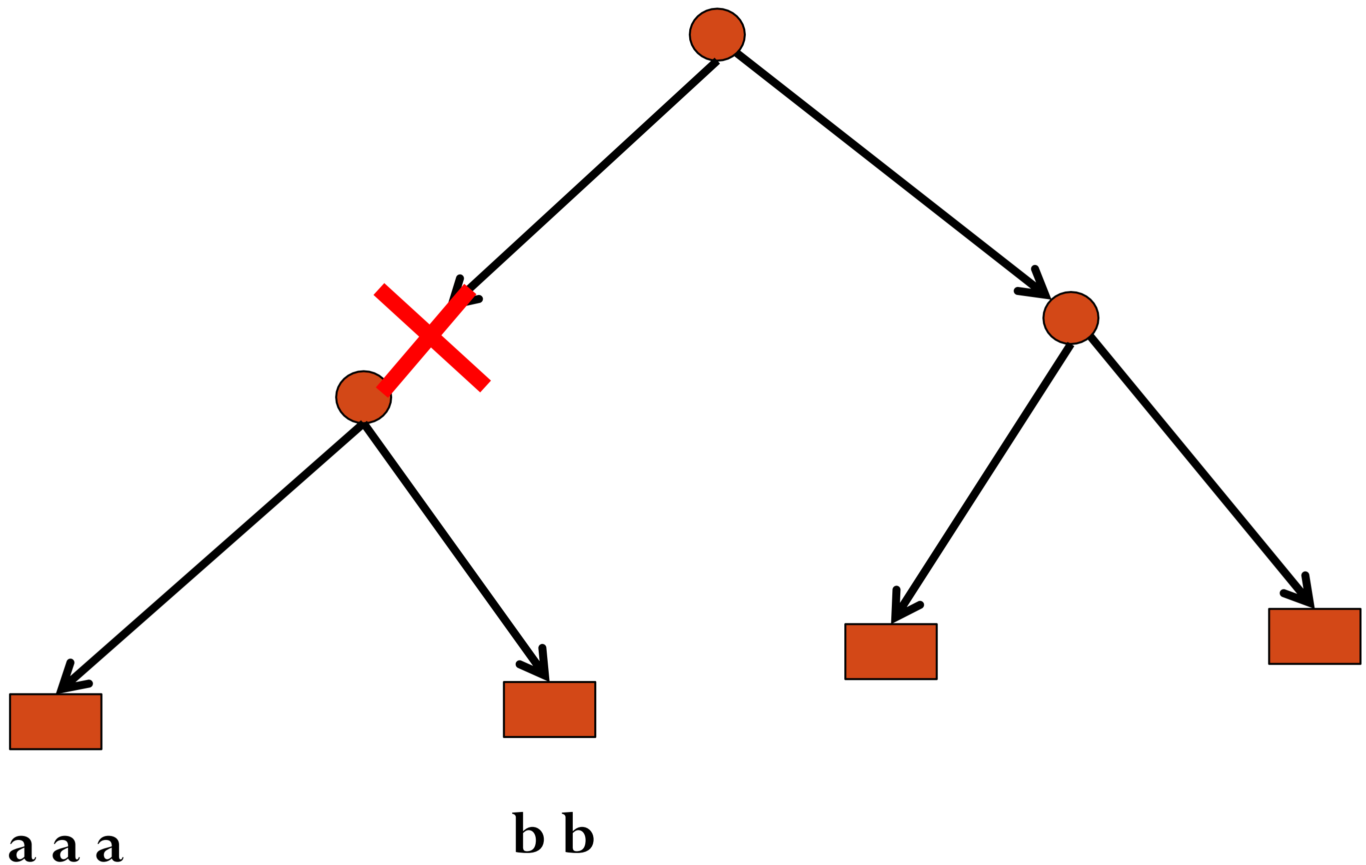
- 为了防止出现过拟合，对生成的决策树进行简化的过程称为剪枝。也就是从已经生成的树上裁掉一些子树或者叶节点，将其父节点作为新的页节点，用其实例数最大的类别作为标记。

- 这种先生成树再剪枝的方法称为后剪枝。

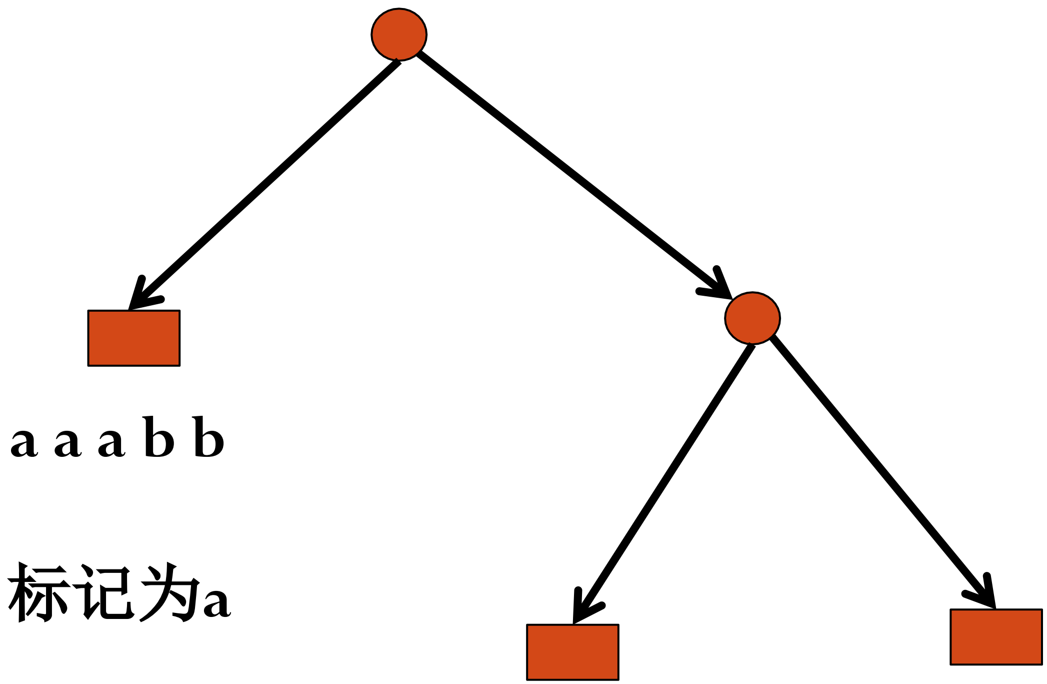
后剪枝方法示意



后剪枝方法示意



后剪枝方法示意

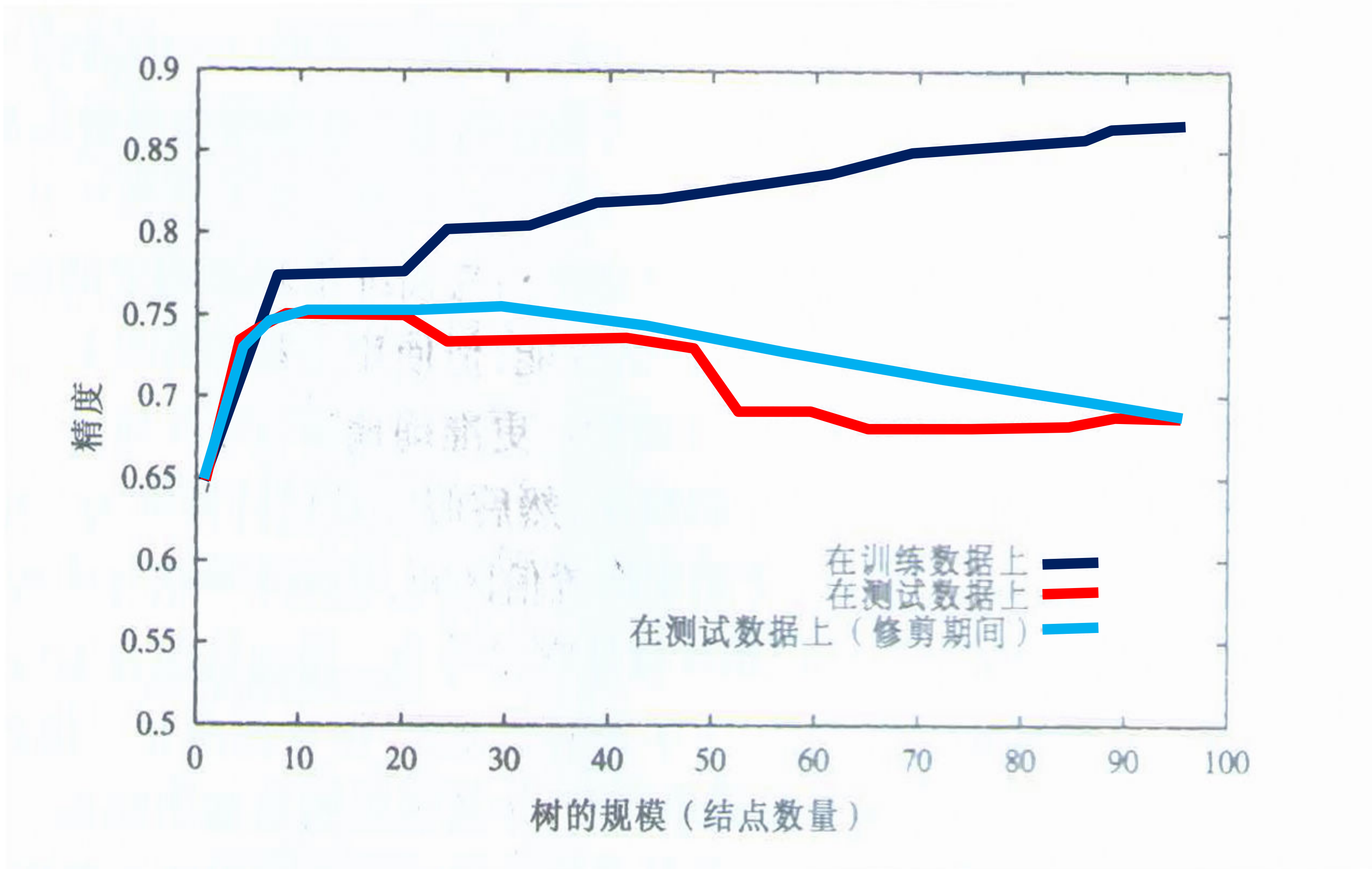


决策树的剪枝

- 当数据量大时：
  - 将数据划分为训练集、验证集和测试集

- 用训练集训练得到决策树
- 从下向上逐步剪枝
- 在验证集上测试性能，直到性能下降为止
- 最后在测试集上的性能作为系统的性能

剪枝的效果



0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
精度
0
10
20
30
40
50
60
70
80
90
100
树的规模（结点数量）
在训练数据上
在测试数据上
在测试数据上（修剪期间）

决策树的剪枝

- 当数据量小时：
  - 直接利用训练集进行剪枝

- 树T的叶节点个数为|T|，t是树T的叶节点，该节点有N_t个样本，其中k类的样本点有N_{tk}个（k=1,...,K），$H_t(T)$为叶节点t上的经验熵，a≥0为参数

◆定义损失函数：

$$C_a(T)=\sum_{t=1}^{|T|}N_tH_t(T)+a|T|$$

其中经验熵为：$H_t(T)=-\sum_k \frac{N_{tk}}{N_t}\log\frac{N_{tk}}{N_t}$

记：$C(T)=\sum_{i=1}^{|T|}N_tH_t(T)=-\sum_{t=1}^{|T|}\sum_{k=1}^{K}N_{tk}\log\frac{N_{tk}}{N_t}$

有：$C_a(T)=C(T)+a|T|$

$C(T)$表示模型对训练数据的预测误差，$|T|$表示模型的复杂程度

- 剪枝，就是当α确定时，选择损失函数最小的模型。

决策树的剪枝算法

- 输入：生成算法产生的整个树T，参数a
- 输出：修剪后的子树T_a
- (1)计算每个节点的经验熵
- (2)递归地从树的叶节点向上回缩，如果回缩后的损失函数小于等于回缩前，则剪枝，将父节点变为新的叶节点
- (3)返回2，直至不能继续为止，得到损失函数最小的子树T_a

多选题 1分

设置

请选择以下正确的说法：

A 剪枝后在训练集上性能提高

B 通过剪枝可以提高在测试集上的性能

C 决策树只能处理离散的特征值

D 决策树在训练集上的性能越高越好

提交

练习题

- 使用工具系统实现基于决策树方法的文本分类，并对ID3、C4.5算法进行比较。

随机森林

- 决策树容易过拟合
- 随机森林是由多个决策树组成的分类器
- 通过投票机制改善决策树
- 单个决策树的生成
  - 有放回的数据采样
  - 属性（特征）的采样
- 集外数据的使用
  - 单个决策树未用到的数据

小结

- 什么是统计机器学习方法?
- 朴素贝叶斯方法
- 支持向量机
  - 线性可分支持向量机
  - 线性支持向量机
  - 非线性支持向量机
- 决策树
  - ID3算法
  - C4.5算法