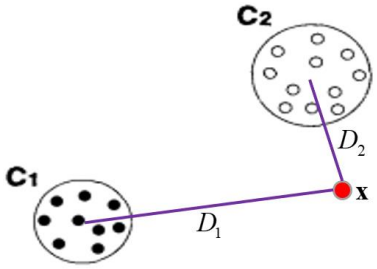
模式识别
(Pattern Recognition)

广东工业大学集成电路学院　邢延

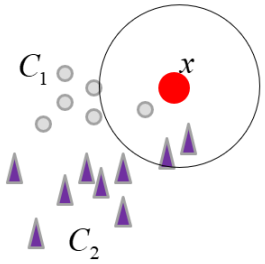
- **已经讲过的分类器**

  - **基于距离的分类器**



$\because D_2 < D_1$
$\therefore \mathbf{x} \in C_2$

  - **基本思想**



$K=3$时，$x \in C_2$

在所有样本中找到与测试样本的K个最近邻者，其中各类别所占个数表示成 K_i，i=1,2,…,m。

定义判别函数为：$g_i(\mathbf{x})=K_i$
则决策规则为：

$$j = \arg\max_i g_i(\mathbf{x}), i = 1,2...,m$$

K-近邻一般采用K为奇数，跟投票表决一样，避免因两种票数相等而难以决策。

  - **朴素Bayes分类器**

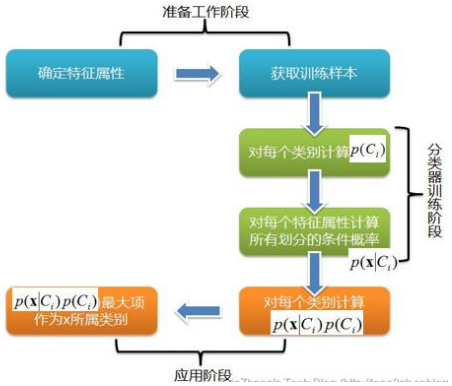
$$p(C_i|\mathbf{x}) = \frac{p(\mathbf{x}|C_i)p(C_i)}{p(\mathbf{x})}, 1 \le i \le m$$

其中，$p(\mathbf{x}) = \sum_{i=1}^{m} p(\mathbf{x}|C_i)p(C_i)$，即$\mathbf{x}$的概率分布

$$posterio = \frac{likelihood \times prior}{evidence}$$

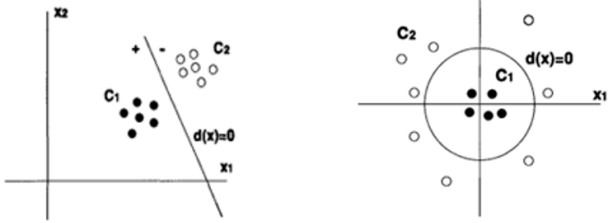
后验概率主要由先验概率和似然函数的乘积所决定

给定训练数据集，evidence是个常数



- **线性分类器与非线性分类器**

  - 判别函数
    - 定义
      - ◆分类器用于判定待分类数据所属类别的函数
    - 种类
      - ◆线性函数 ------ 线性分类器
      - ◆非线性函数 -------- 非线性分类器



- **线性 还是 非线性?**

  - **基于距离的分类器**

    一般是非线性分类器，特殊情况下当分类边界是线性时，才是线性分类器。

  - **Bayes分类器**

    朴素bayes分类器一般是非线性分类器，但某些具有特定属性的朴素贝叶斯分类器才是线性分类器。

第四讲 支持向量机

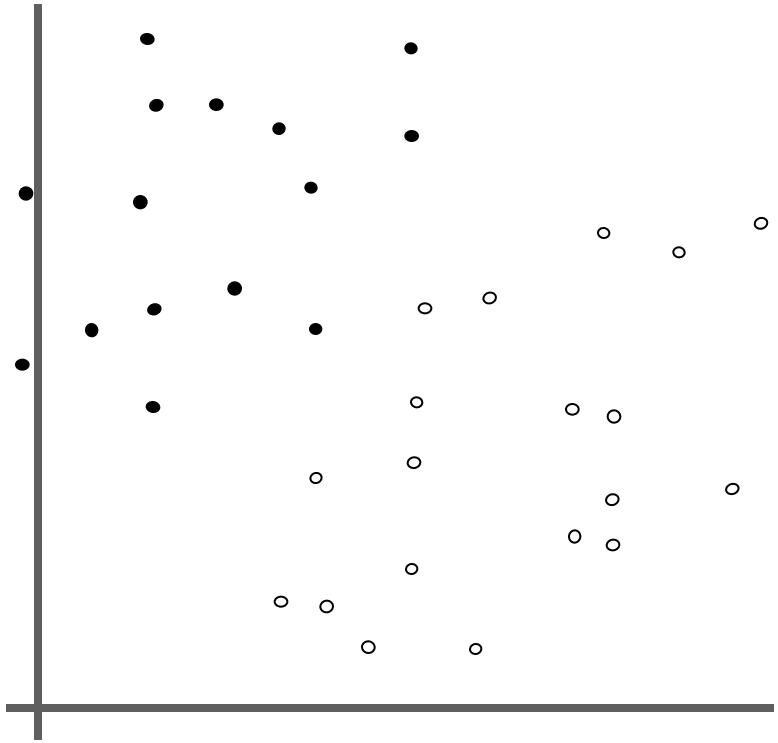
（04 Support Vector Machines）

内容提要

- 支持向量机的原理
- 线性支持向量机
- 软间隔支持向量机
- 非线性支持向量机

支持向量机的原理

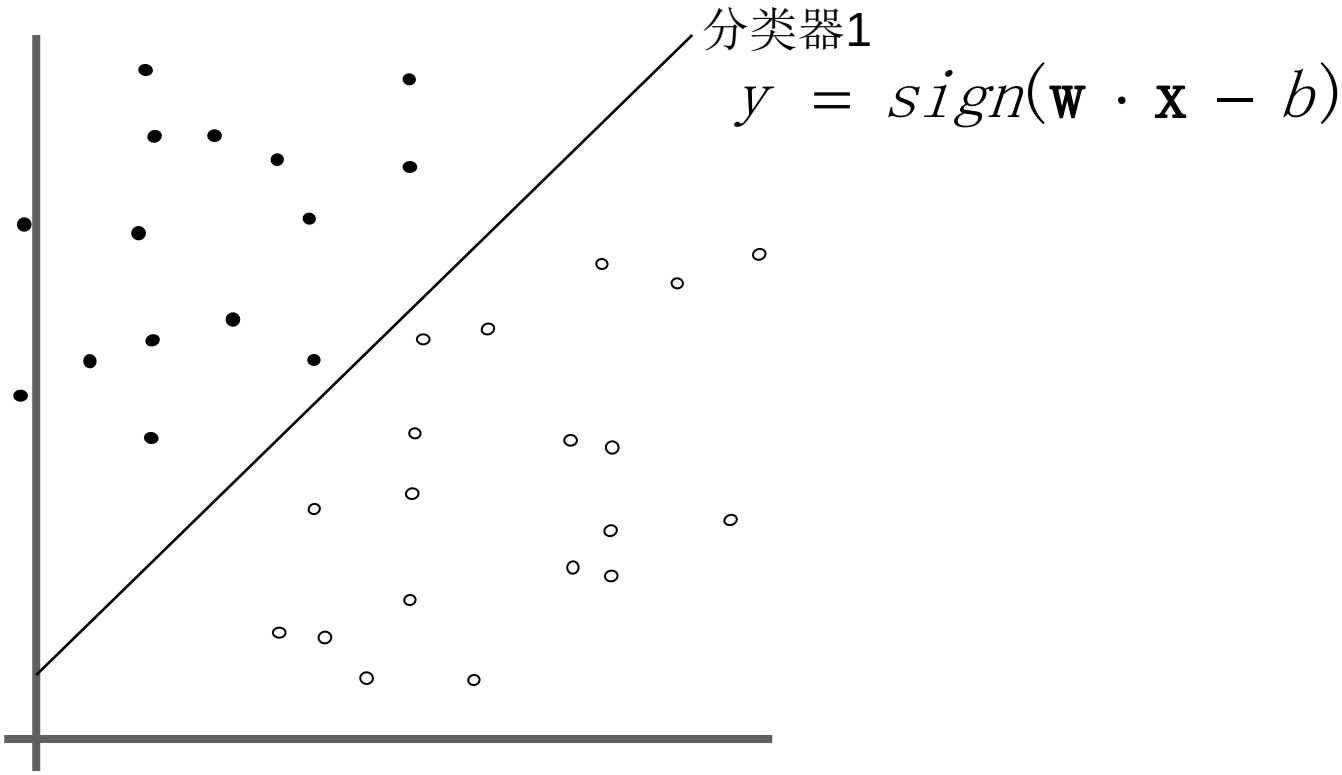
- **线性分类器的深入分析**



如何分类?

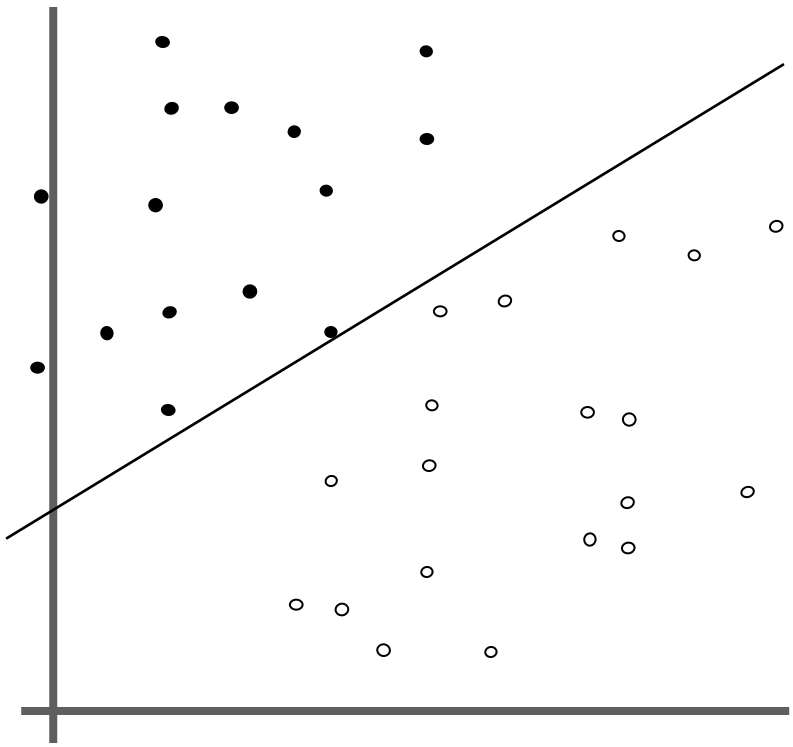
支持向量机的原理

- 线性分类器的深入分析



支持向量机的原理

- **线性分类器的深入分析**



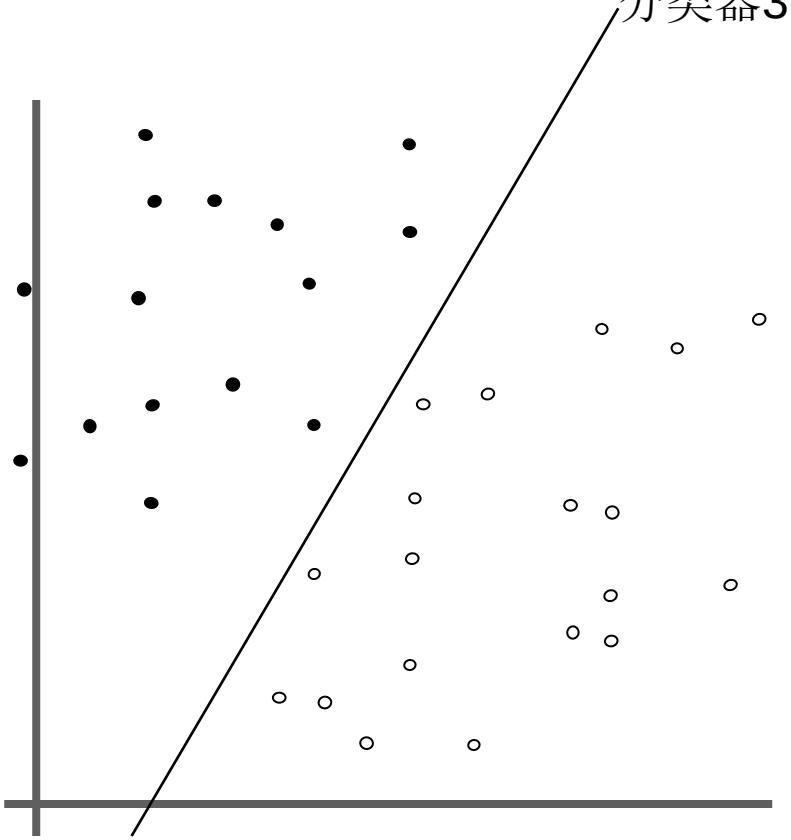
分类器2

$$y = sign(\mathbf{w} \cdot \mathbf{x} - b)$$

支持向量机的原理

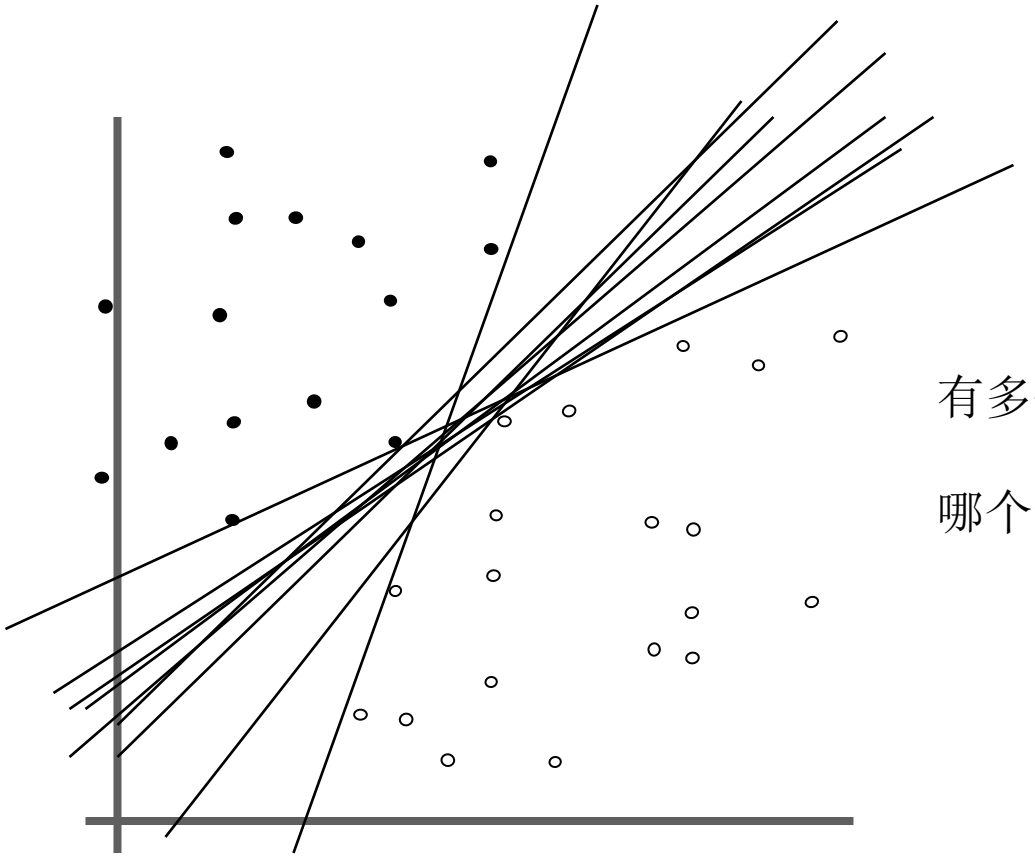
- **线性分类器的深入分析**

分类器3 $y = sign(\mathbf{w} \cdot \mathbf{x} - b)$



支持向量机的原理

- **线性分类器的深入分析**

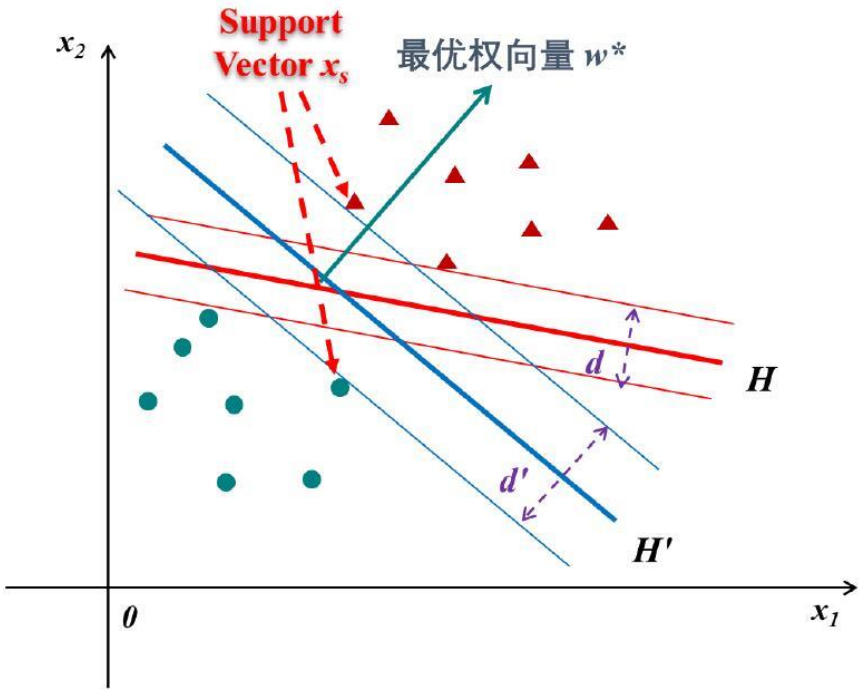


有多个分类器！

哪个最优？

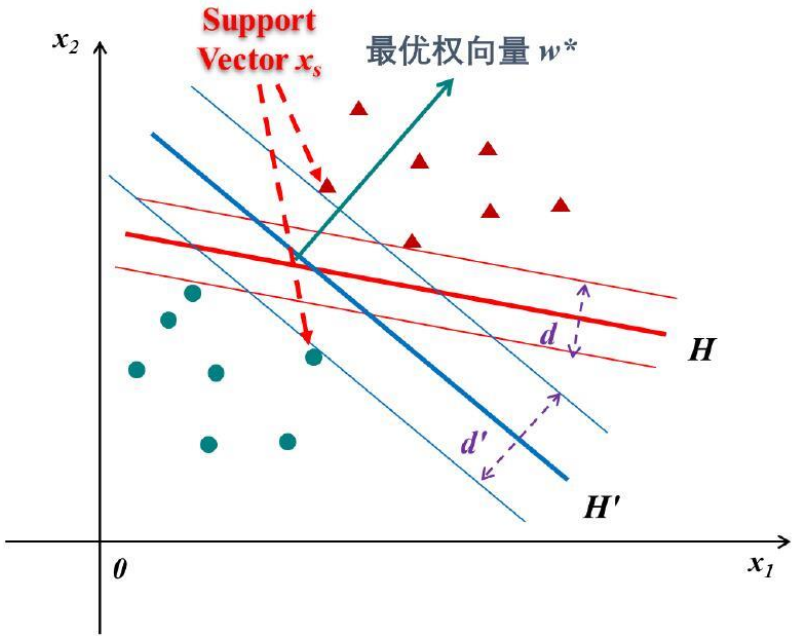
支持向量机的原理

- 分类间隔d（Margin of Classification）
  - 定义
    - 任何一个求解得到的权向量w，都会带来一系列平行的分类决策边界，其可平移的范围具有一定的宽度，即分类间隔，如图中所示的d。



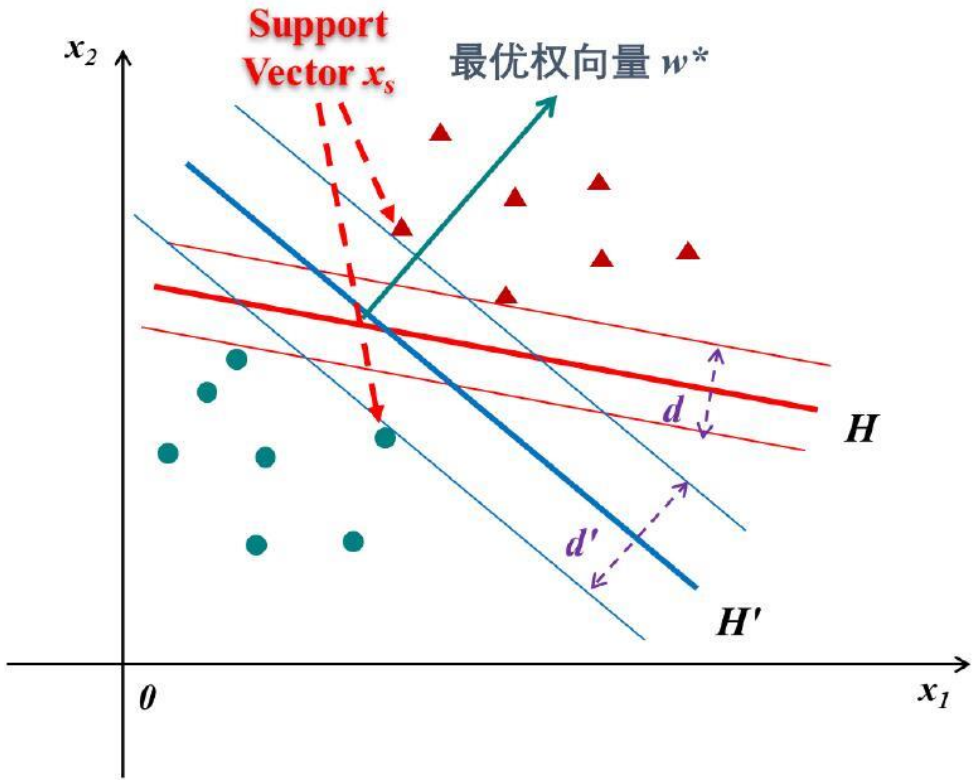
支持向量机的原理

- 最大间隔分类器
  - 分类间隔越大，两类样本做分类决策时的裕量也就越大，分类错误也就越少；
  - 最大间隔分类器是最优的。



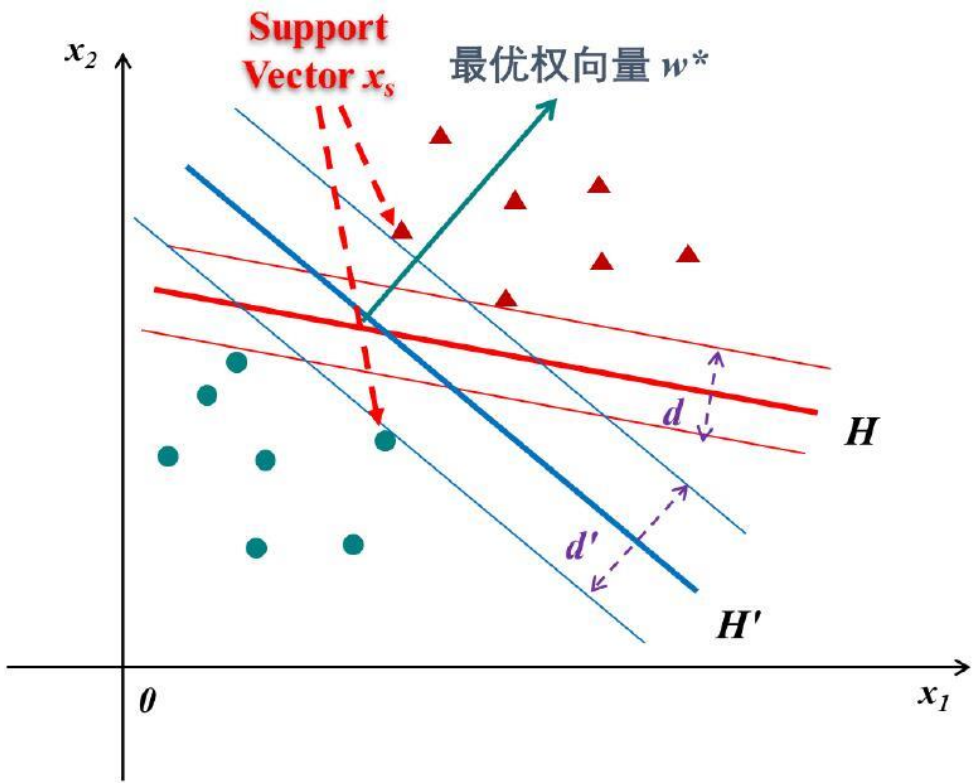
支持向量机的原理

- **支持向量(Support Vectors)**
  - **支持向量是距离分类决策边界最近的样本。**
  - **分类间隔仅由支持向量决定!**



支持向量机的原理

- 支持向量机(Support Vector Machine, SVM)



支持向量机：
是支持向量的线性组合

支持向量机的训练：
求解最优权向量，以获得最大分类间隔下的支持向量

线性支持向量机

- **SVM的训练**

求取分类间隔的最大值 $d = 2\frac{|G_{ij}(x_s)|}{\|w\|}$

↓

固定 $|G_{ij}(x_s)| = 1$ 将 max d 的问题转化成 min||w||的问题

↓

带约束的二次优化问题：$\min \frac{1}{2}\|w\|^2$，约束：

$$\begin{cases} sgn(G_{ij}(x^{(1)}))(w^T x^{(1)} + w_0) \geq 1 \\ sgn(G_{ij}(x^{(2)}))(w^T x^{(2)} + w_0) \geq 1 \\ \vdots \\ sgn(G_{ij}(x^{(l_i)}))(w^T x^{(l_i)} + w_0) \geq 1 \\ sgn(G_{ij}(x^{(l_i+1)}))(w^T x^{(l_i+1)} + w_0) \geq 1 \\ sgn(G_{ij}(x^{(l_i+2)}))(w^T x^{(l_i+2)} + w_0) \geq 1 \\ \vdots \\ sgn(G_{ij}(x^{(l)}))(w^T x^{(l)} + w_0) \geq 1 \end{cases}$$

↓

使用拉格朗日乘子法将其转化为无约束优化问题

↓

$$\min_{\|w\|} \max_{\alpha} L(w, w_0, \alpha)$$

拉格朗日乘子法

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha^{(i)}(sgn(G_{ij}(x^{(i)}))(w^T x^{(i)} + w_0) - 1)$$

↓ 分别对权向量 w 和偏置量 w_0 求偏导

唯一解：得拉格朗日乘子α*，$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T$，其中绝大部分分量的值=0，只有少数分量的值大于0

↓

线性支持向量机

- **SVM的训练**

大于0的α*分量对应的训练样本就是支持向量

⇩

权值矩阵：$\boldsymbol{w}^* = \sum_{j=1}^{J} \boldsymbol{\alpha}_j^* y_j \boldsymbol{x}_j$, 其中$\boldsymbol{\alpha}_j^* > 0$

偏置量：$w_0 = y_i - \sum_{j=1}^{J} \boldsymbol{\alpha}_j^* y_j (\boldsymbol{x}_j, x_i)$, 其中$\boldsymbol{\alpha}_j^* > 0$

⇩

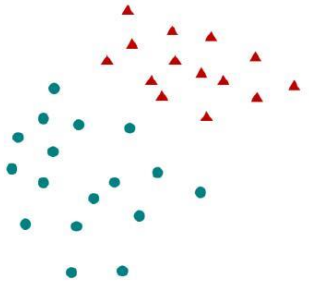
分类决策函数：$f(\boldsymbol{x}) = sign(\sum_{j=1}^{J} \boldsymbol{\alpha}_j^* y_j (\boldsymbol{x}, \boldsymbol{x}_j))$, 其中$\boldsymbol{\alpha}_j^* > 0$

线性支持向量机

- **SVM的泛化能力强**
  - **遵循结构风险最小化准则**

经验风险最小化： $\min(R_{emp})$
Empirical Risk Minimization

训练数据



经验风险最小化： $\min(R_{emp})$
Empirical Risk Minimization

结构风险最小化： $\min(R(w))$ → 未知的新数据
Structure Risk Minimization

$$R(w) \le R_{emp}(w) + \phi\left(\frac{h}{l}\right)$$ → 泛化误差上界

$$\phi\left(\frac{h}{l}\right) = \sqrt{\frac{h\left(\ln\left(\frac{2l}{h}\right)+1\right) - \ln\left(\frac{\eta}{4}\right)}{l}}$$ → 置信风险

η是随机噪声带来的样本误差

l：训练集中的样本数量
h：分类器的VC维 → 是一类函数具有的分类能力；

给定数据，分类器函数形式的阶次越低，其VC维也就越小，分类器结构风险也就越小，泛化能力也就越强。

线性支持向量机

- **SVM的泛化能力强**
  - **遵循结构风险最小化准则**

经验风险最小化： $\min(R_{emp})$
Empirical Risk Minimization

结构风险最小化： $\min(R(w))$
Structure Risk Minimization

$$R(w) \le R_{emp}(w) + \phi\left(\frac{h}{l}\right)$$

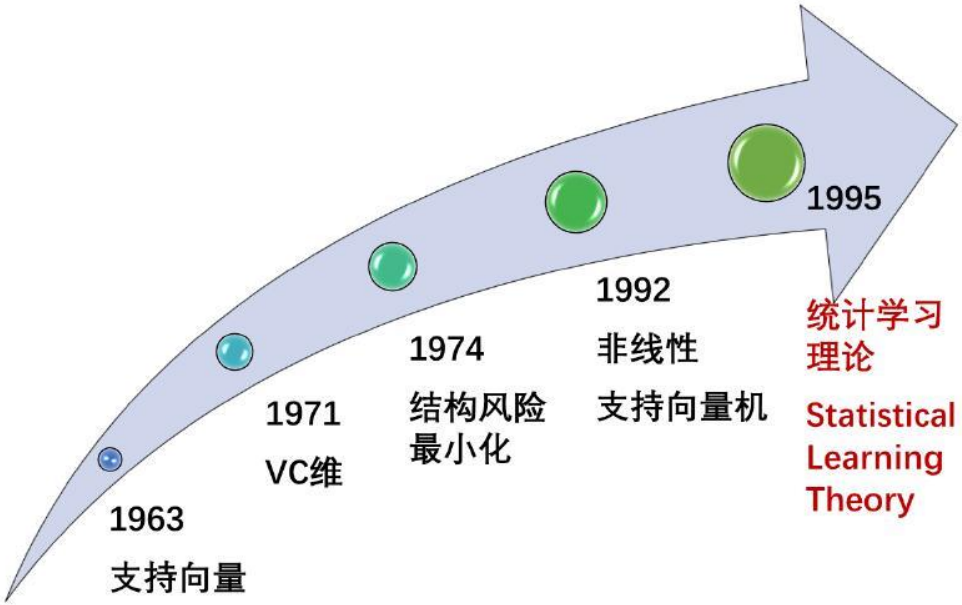
$$\phi\left(\frac{h}{l}\right) = \sqrt{\frac{h\left(\ln\left(\frac{2l}{h}\right)+1\right) - \ln\left(\frac{\eta}{4}\right)}{l}}$$

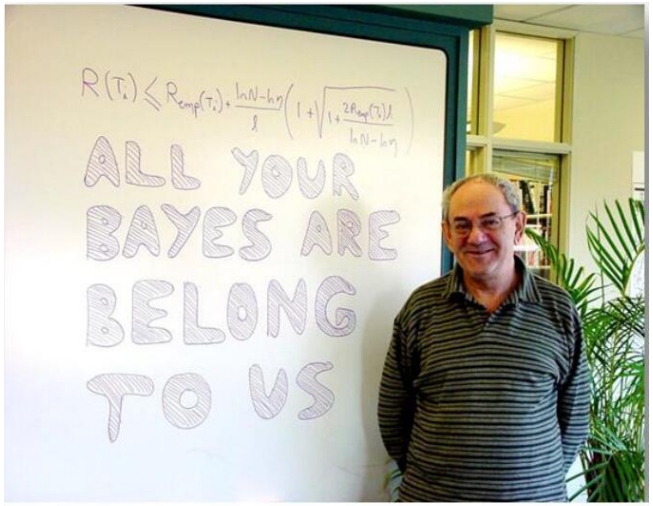
l：训练集中的样本数量
h：分类器的VC维

支持向量机是阶次最低的线性函数！
因此，在不需要大量训练集样本的情况下，也能取得泛化能力非常强的分类器训练结果。

线性支持向量机

- 统计学习理论的发展史

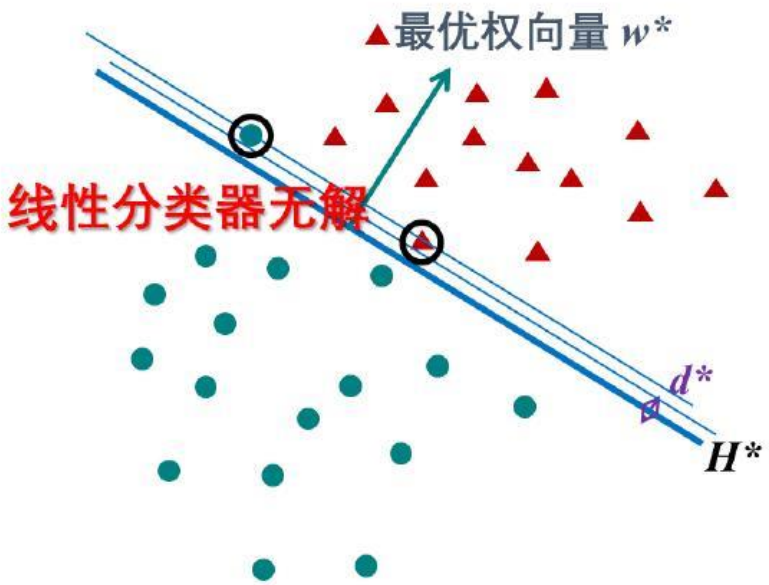




Vladimir N. Vapnik

软间隔支持向量机

- 当数据中有噪声或者异常点时，成为线性不可分问题



异常点到线性分类器决策边界的距离一定比支持向量到分类决策边界的距离更近；

异常点的判别函数值的绝对值，一定是小于1；

在约束条件中减去一项正数 ξ，使判别函数的绝对值允许小于1；

软间隔支持向量机

- 当数据中有噪声或者异常点时，成为线性不可分问题

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi^{(i)}$$

软间隔支持向量机

$$s.t. \quad sgn\left(G_{ij}\left(x^{(l_i)}\right)\right)\left(w^T x^{(l_i)} + w_0\right) \geq 1 - \xi^{(l_i)}$$

ξ 就称为松弛变量

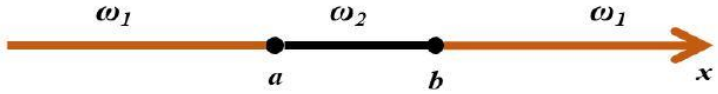
C，称为惩罚因子

软间隔支持向量机

- 本质上是线性SVM
  - 训练过程与线性SVM类似

非线性支持向量机

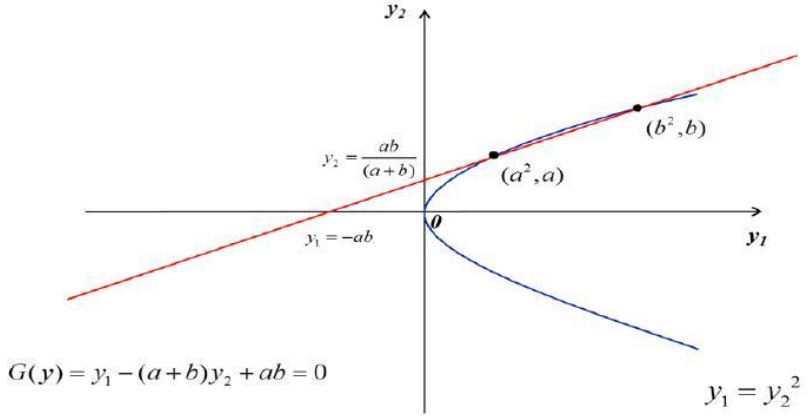
- 采用广义线性化方法
  - 将低维空间中的一个非线性分类问题往高维空间映射，从而转化为一个线性分类问题



$G(x) = (x-a)(x-b)$ ← 非线性判别函数

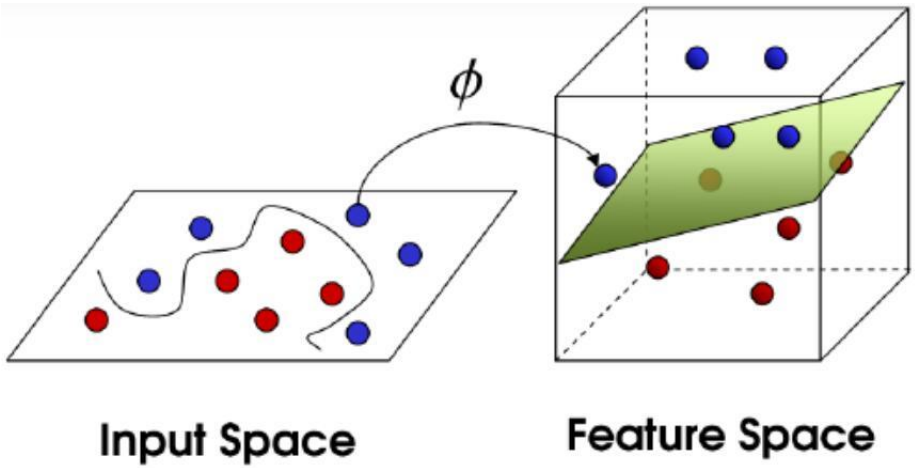
令 $y_1=x^2$, $y_2=x$ ← 广义线性化

$G(y) = y_1 - (a+b)y_2 + ab$ ← 线性判别函数



非线性支持向量机

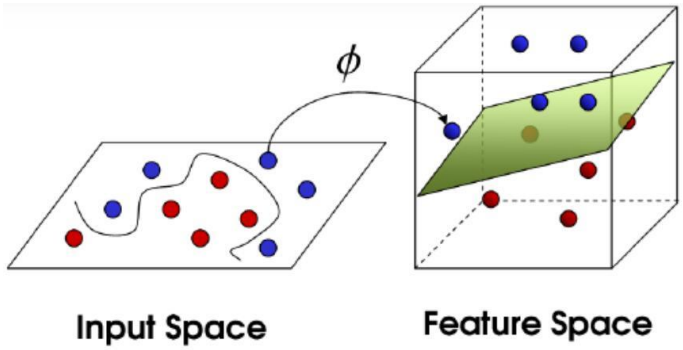
- 采用广义线性化方法
  - 难点
    - ◆怎么知道应该映射到多少维的特征空间，非线性分类问题才会转化成线性分类问题呢?
    - ◆如何找到合适的映射函数?
    - ◆将问题转化到高维空间中后，会带来巨大的计算量问题，甚至会因为维度灾难造成问题根本无法解决。



非线性支持向量机

- 采用广义线性化方法
  - 采用核函数
    - ◆在高维空间的线性支持向量机求解过程中，和在最终得到的线性判别函数中，除了类别标签以外，并没有用到原始空间中的样本x(i)映射到高维空间中的像y(i)；
    - ◆用到的只是高维空间中两个向量的内积；

分类决策函数：$f(\boldsymbol{\phi}(\boldsymbol{x})) = sign(\sum_{j=1}^{J} \boldsymbol{\alpha}_j^* y_j(\phi(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}_j))),$其中$\boldsymbol{\alpha}_j^* > 0$



非线性支持向量机

- **采用广义线性化方法**
  - **采用核函数**
    - **如果不经过原始特征空间到高维特征空间的映射过程，就能够计算出两个低维空间的向量在高维空间中的内积，就可以实现非线性支持向量机求解的目标；**
    - **核函数是这样一类函数，它的输入是低维空间中的两个向量，输出是这两个向量经过同一个映射到另一个空间以后的内积.**

核函数的定义

设 $\mathbb{X}$ 是 $\mathbb{R}^n$ 中的一个子集，称定义在 $\mathbb{X} \times \mathbb{X}$ 上的函数 $k(x,y)$ 是核函数，如果存在一个从 $\mathbb{X}$ 到希尔伯特空间(特征空间) $\mathbb{H}$ 的映射 ϕ

$$\phi : \xi \mapsto \phi(\xi) \in \mathbb{H}$$

使得对任意的 $x, y \in \mathbb{X}$ ，有

$$k(x,y) = (\phi(x), \phi(y)) = \phi(x)^T \phi(y)$$

都成立。

非线性支持向量机

- **采用广义线性化方法**
  - **采用核函数**

具体例子

假设 $A=(1,2)^T$、$B=(3,4)^T$，构造一个映射 $\phi(\cdot)=(x_1^2,\sqrt{2}x_1x_2,x_2^2)^T$，则可知

$$\phi(A)=(1,2\sqrt{2},4)^T$$
$$\phi(B)=(9,12\sqrt{2},16)^T$$

因此通过映射 $\phi(\cdot)$ 将点 A、B 从二维平面升维到三维空间。然后计算

$$\begin{aligned}\phi(A)^T\phi(B)&=1\times 9+2\sqrt{2}\times 12\sqrt{2}+4\times 16\\&=9+48+64\\&=121\end{aligned}$$

高维空间中执行了11次乘法运算、2次根号运算和2次加法运算

上述运算是在映射后的高维空间下做内积，那么是否能直接在原始的空间中进行相应的运算，使得低维情况下的运算结果等于高维情况下的运算结果呢？答案是肯定的可以通过核函数 $k(x,y)=(x^Ty)^2$ 来实现

$$\begin{aligned}k(A,B)&=(A^TB)^2\\&=(1\times 3+2\times 4)^2\\&=121\end{aligned}$$

低维空间中仅执行了3次乘法运算和1次加法运算

非线性支持向量机

- 采用广义线性化方法
  - 采用核函数
    - 低维空间和高维空间通过核函数联通起来
    - 优点是避免了维度灾难
      - 高维空间中的运算计算量很大呈指数级别复杂度，难以解决；
      - 低维空间中的运算计算量很小
      - 但是两者的最终结果是一致的。

非线性支持向量机

- **常用的核函数**

  - **多项式核函数**

$$K(x^{(i)}, x^{(j)}) = (x^{(i)T}x^{(j)} + c)^d, \quad d = 1, 2, \ldots\ldots$$

  - **Sigmoid核函数**

$$K(x^{(i)}, x^{(j)}) = \tanh(\beta x^{(i)T}x^{(j)} + \gamma)$$

  - **径向基核函数**

$$K(x^{(i)}, x^{(j)}) = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\sigma^2}\right), \quad \text{高斯型}$$

非线性支持向量机

- 如何选择合适的核函数?
  - 无论是核函数的形式还是参数，都没有确定的选择方法，只能依靠经验来试
  - 核函数方法配合软间隔方法，能够为大多数问题都找到支持向量机的解

非线性支持向量机

- **增强分类器泛化能力的方法**

经验风险最小化： $\min(R_{emp})$
Empirical Risk Minimization

结构风险最小化： $\min(R(w))$
Structure Risk Minimization

$$R(w) \le R_{emp}(w) + \phi\left(\frac{h}{l}\right)$$

$$\phi\left(\frac{h}{l}\right) = \sqrt{\frac{h\left(\ln\left(\frac{2l}{h}\right)+1\right) - \ln\left(\frac{\eta}{4}\right)}{l}}$$

l：训练集中的样本数量
h：分类器的VC维

方法一：
先选择VC维低的分类器形式，以降低置信风险，再通过分类器参数的优化来降低经验风险，如支持向量机。

方法二：
加大训练集的规模，但会带来计算量的问题，如深度学习算法。

小结

- **支持向量机的特点**
  - **解决两类分类问题**
  - **数据需要先进行规范化处理**
  - **对噪声敏感**
  - **修改支持向量能够改变分类器的性能**

模式识别算法编程实例演示

- **实例4：支持向量机**
  - **参见Jupyter Notebook文档目录：lesson04**