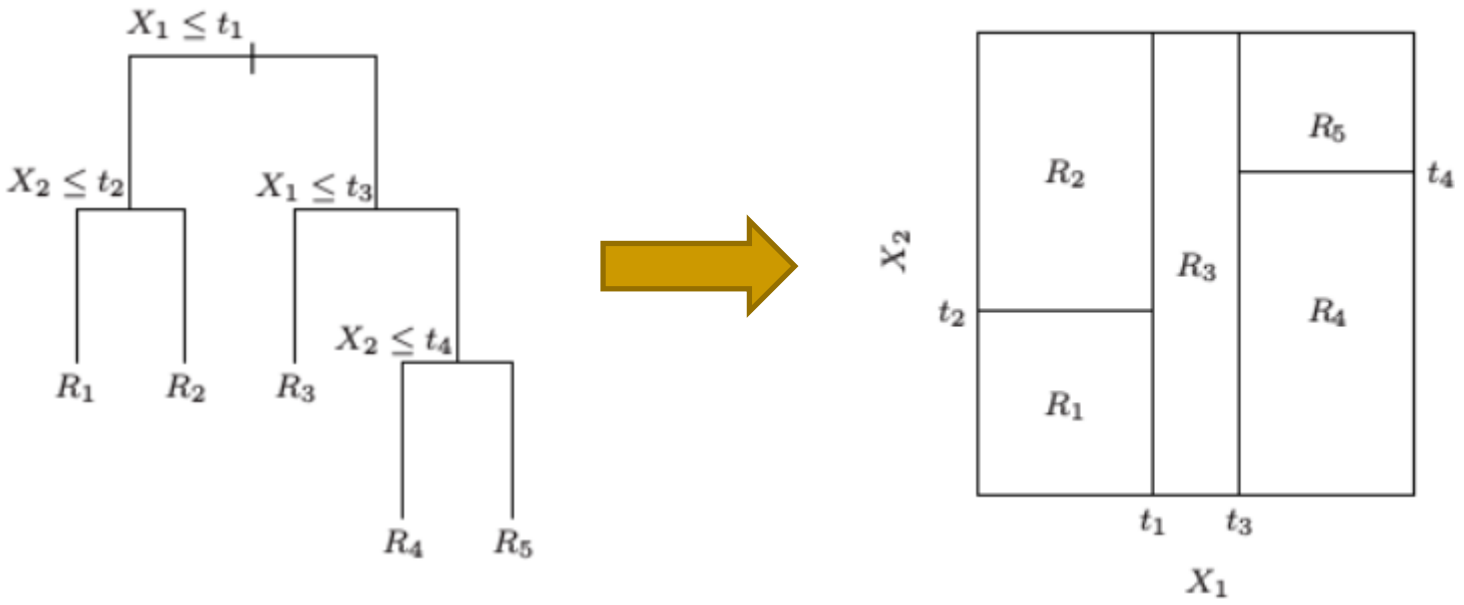
支持向量机
Support Vector Machine

Support Vector Machine

- 引例
- SVM概述
- SVM的基本原理
  - 线性可分：硬间隔SVM
  - 线性不可分：软间隔SVM
  - 非线性：核函数
- 应用实例

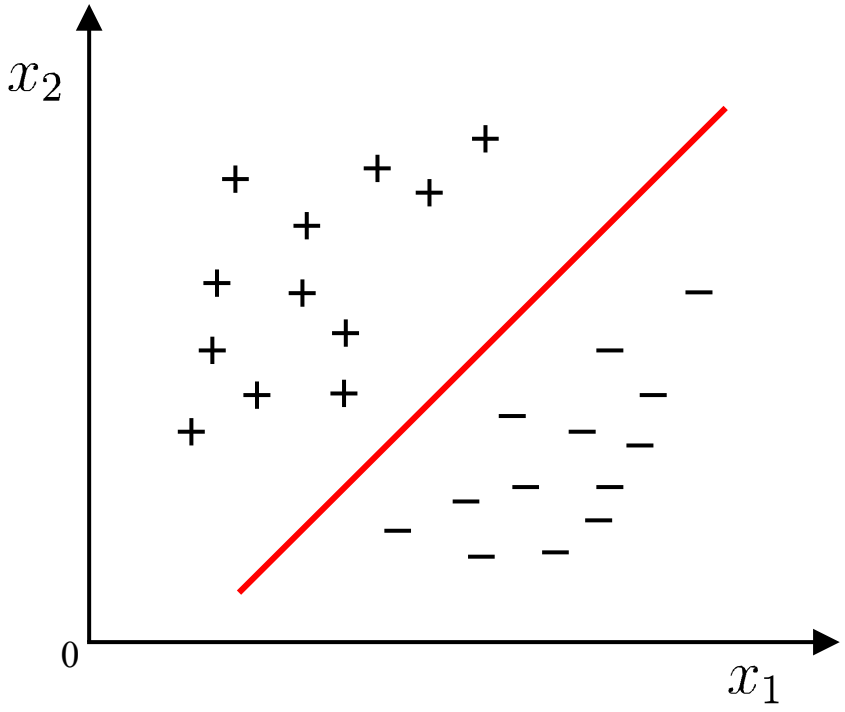
引例

- 分类问题等价于寻找输入空间的一组最优划分
- 以决策树为例



引例

- 有一类二分类问题
- 输入空间只有两类点（正例和负例）
- 目标是找到一种线性分类方式将其区分开



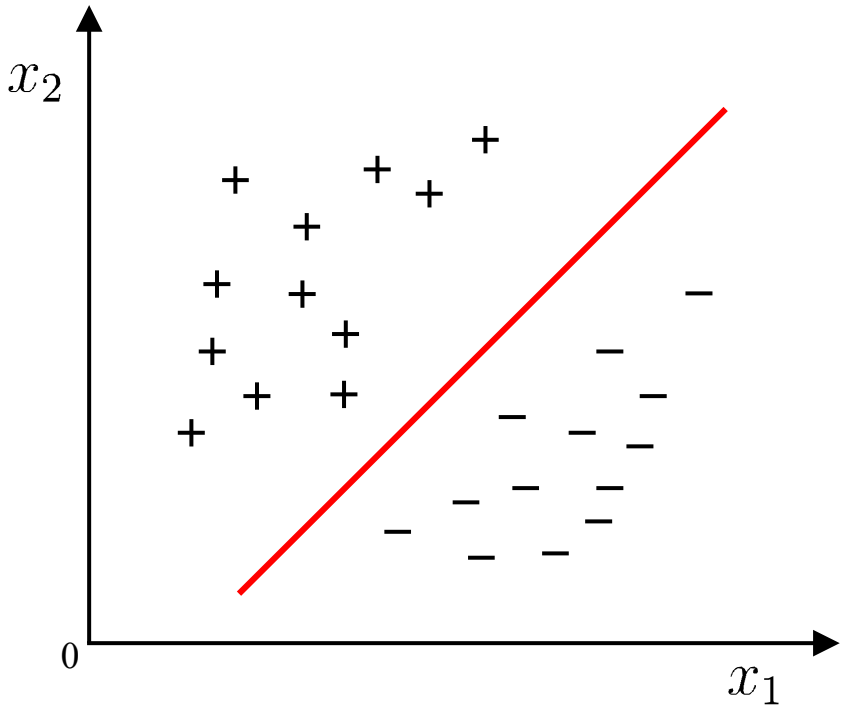
引例

- 对于二维空间来说，就是找到一条直线

$$Ax + By + C = 0$$

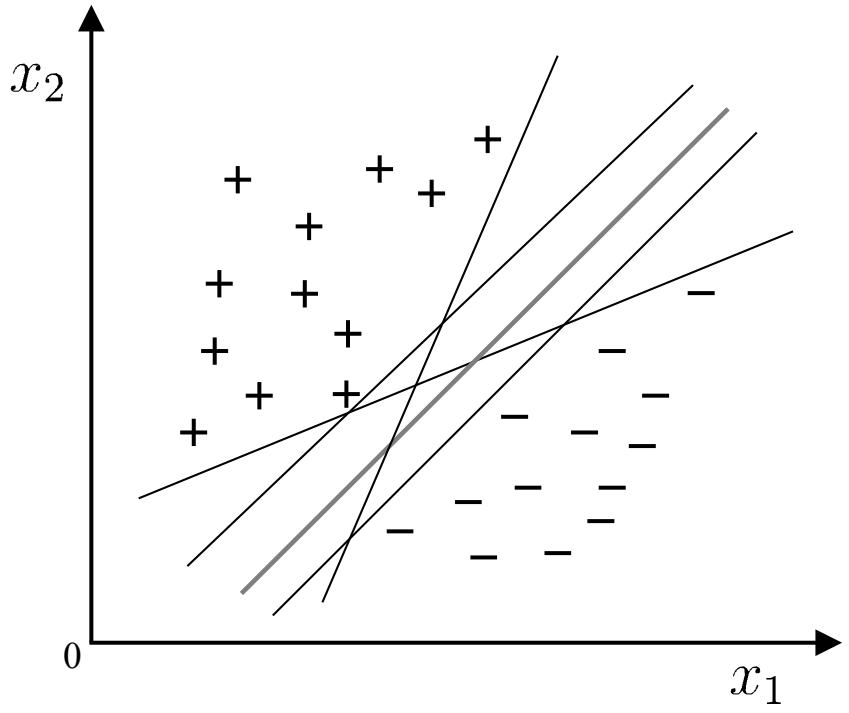
- 对于n维空间来说，就是找到一个超平面

$$w^T x + b = 0$$



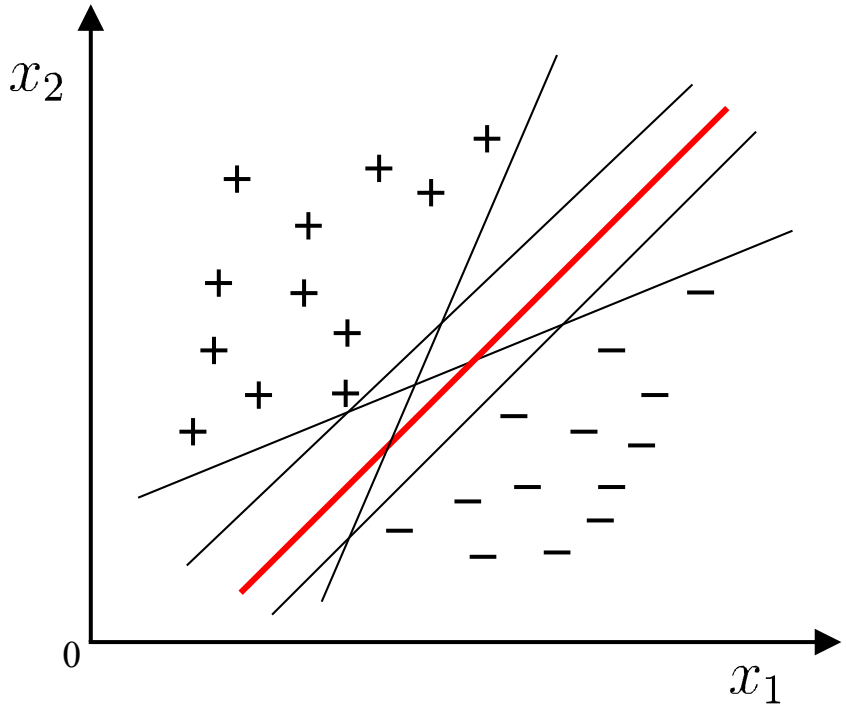
引例

- 将训练样本分开的超平面可能有很多, 哪一个好呢?



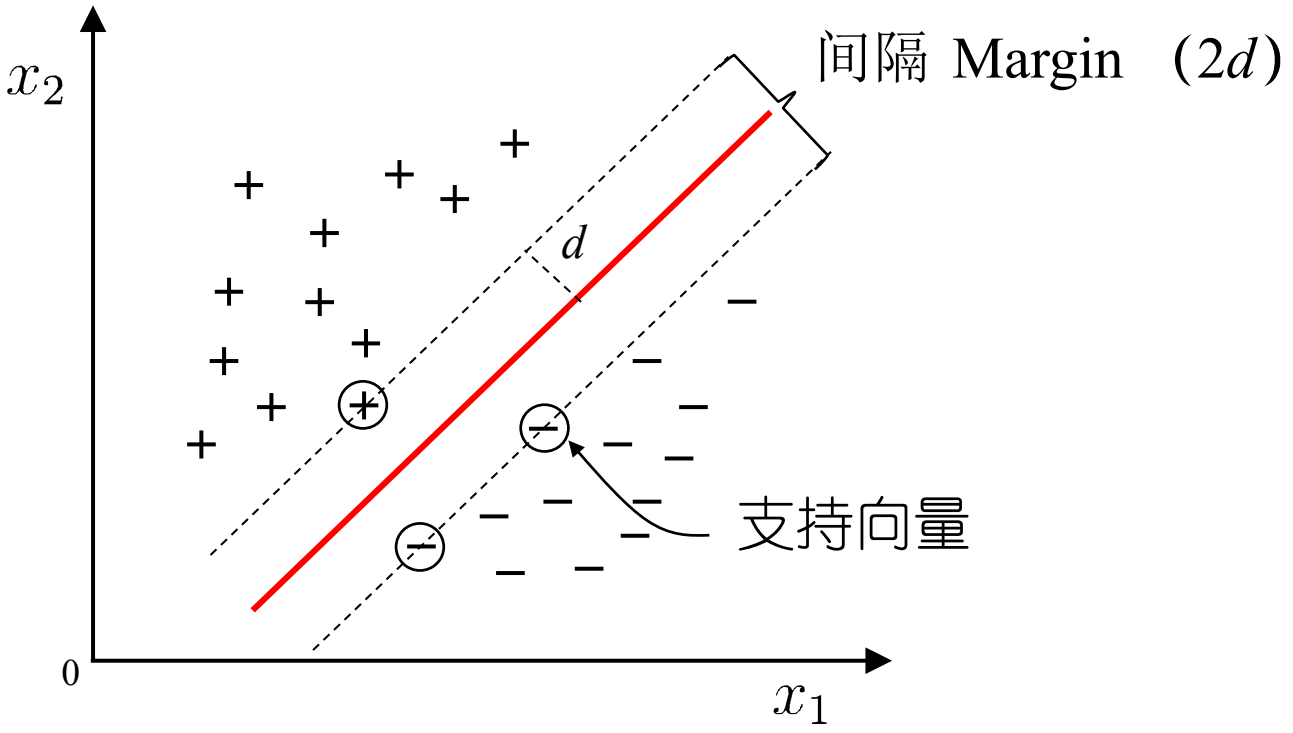
引例

- 应选择“正中间”，容忍性好，鲁棒性高，泛化能力最强



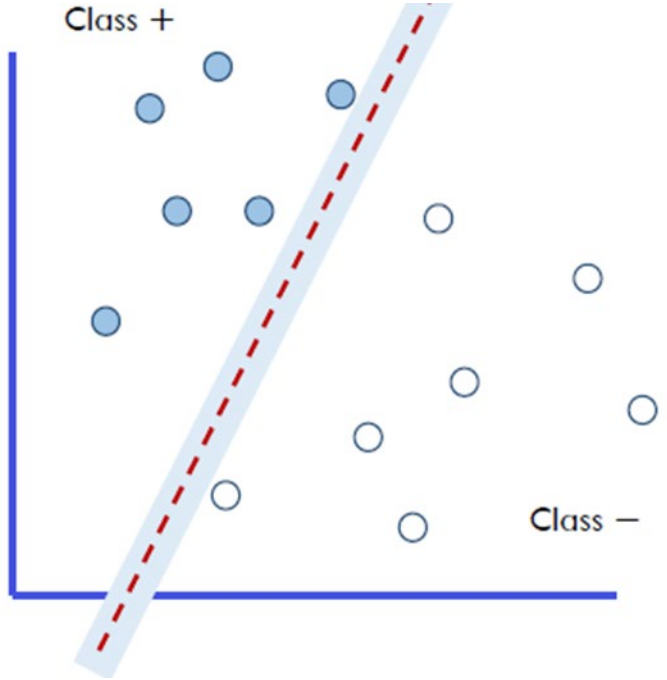
引例

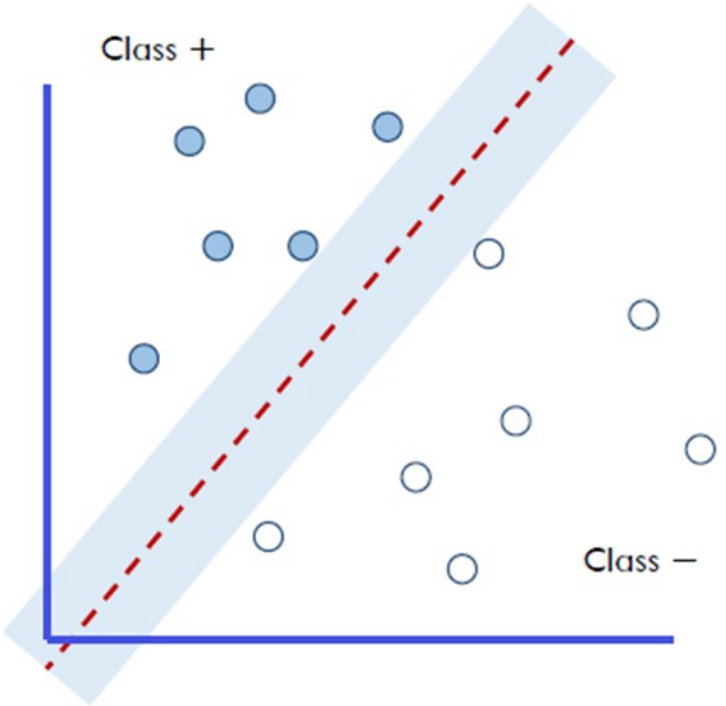
- 如何确定“正中间”的超平面？
- 支持向量机：通过求解间隔最大化来确定最佳超平面



引例

- 支持向量机：通过求解间隔最大化来确定最佳超平面　　why？





右边的模型错误容忍性更好，鲁棒性更高，泛化能力更强

SVM概述

- 传统的统计模式识别方法只有在样本趋向无穷大时，其性能才有理论的保证。**统计学习理论（STL）研究有限样本情况下的机器学习问题。SVM的理论基础就是统计学习理论。**
- 传统的统计模式识别方法在进行机器学习时，强调经验风险最小化。而单纯的经验风险最小化会产生“过学习问题”，其推广能力较差。
- **推广能力：**将学习机器(即预测函数，或称学习函数、学习模型)对未来输出进行正确预测的能力。

SVM概述

- **“过学习问题”**：某些情况下，当训练误差过小反而会导致推广能力的下降。
- 例如：对一组训练样本(x, y)，x分布在实数范围内，y取值在[0，1]之间。无论这些样本是由什么模型产生的，我们总可以用$y=sin(w*x)$去拟合，使得训练误差为0。

SVM：支持向量机

- 根据统计学习理论，机器学习的实际风险由经验风险值和置信范围值两部分组成。而基于经验风险最小化准则的学习方法只强调了训练样本的经验风险最小误差，没有最小化置信范围值，因此其推广能力较差。
- 支持向量机（Support Vector Machine, SVM）以训练误差作为优化问题的约束条件，以置信范围值最小化作为优化目标，即SVM是一种基于结构风险最小化准则的学习方法，其推广能力明显优于一些传统的学习方法。

SVM：支持向量机

- 结构风险最小原理：“经验风险”与“置信风险”的和最小。
- 风险：机器学习本质上就是一种对问题真实模型的逼近（选择一个我们认为比较好的近似模型），但真实模型一定是未知的。既然真实模型不知道，那么我们选择的近似模型与问题真实解之间究竟有多大差距，我们就没法得知。这个与问题真实解之间的误差，就叫做风险（更严格的说，误差的累积即风险）。

SVM：支持向量机

- **经验风险：**选择了一个近似模型（分类器）后，真实误差无从得知，但我们可以用某些可以掌握的量来逼近它。最直观的想法就是使用分类器在样本数据上的分类的结果与真实结果（因为样本是已经标注过的数据，是准确的数据）之间的差值来表示，这个差值叫做经验风险。
- **置信风险：**代表了我们在多大程度上可以信任分类器在未知样本上分类的结果。很显然，它是没有办法精确计算的，因此只能给出一个估计的区间，也使得整个误差只能计算上界，而无法计算准确值。

SVM概述

- **支持向量机**(Support Vector Machine, SVM)是由Cortes和Vapnik于1995年首先提出。
- 基于统计学习理论的一种机器学习方法

SVM概述

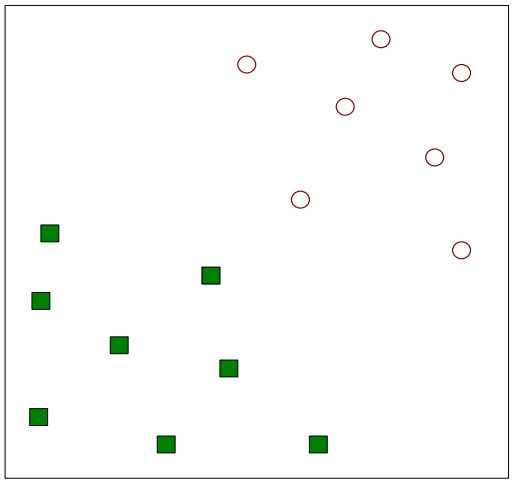
- 一种二值分类模型
- 基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化
- 也叫做 max-Margin Classifer
- 最终可转化为一个凸二次规划问题的求解
- 既可以用作分类也可以用作回归

SVM概述

- SVM在解决**小样本**、**非线性**等分类问题中表现出许多特有的优势，并能够推广到**函数拟合**等有关数据预测的应用中。
  - 手写数字识别
  - 人脸识别
  - 语音识别
  - 文本分类
  - ……

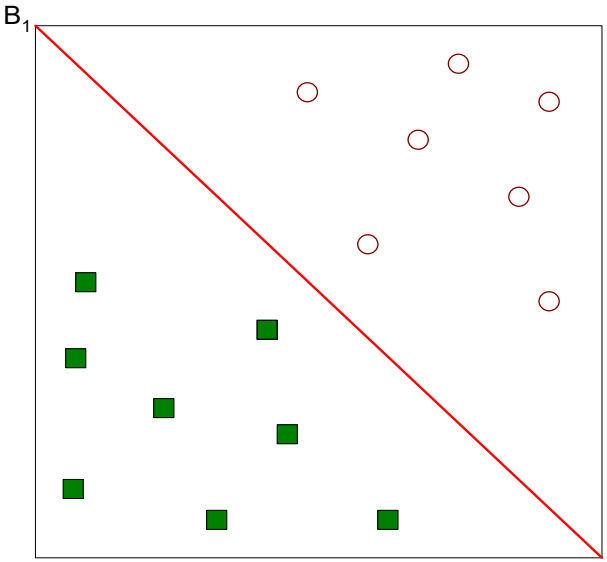
SVM的基本原理

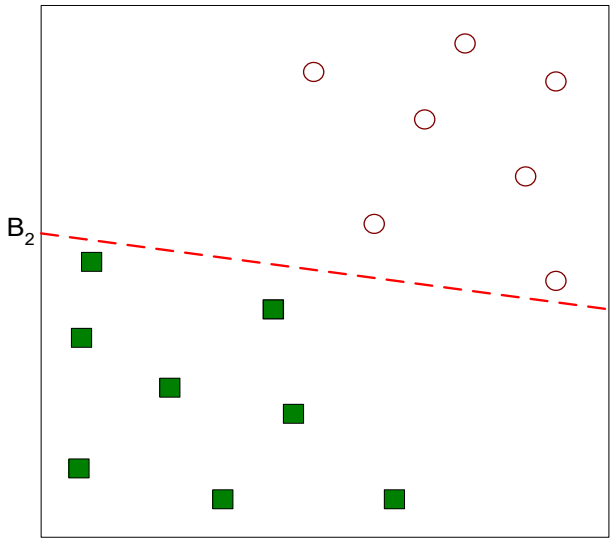
- SVM是在**两类线性可分**情况下，从获得**最优分类面**问题中提出的。
  - 例如：一个两分类问题，其中“红色空心圆圈”表示一类，“绿色实心正方形”表示另一类。
  - 问题：如何在二维平面上寻找一条直线，将这两类分开。



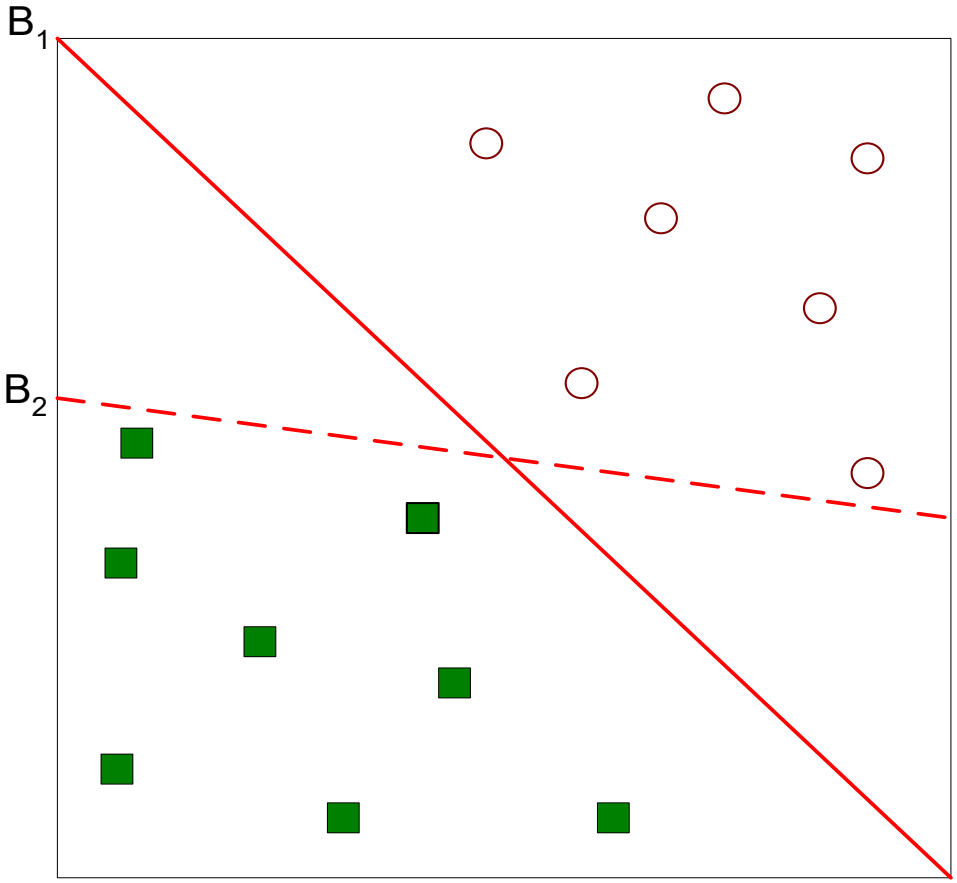
SVM的基本原理

- SVM是在**两类线性可分**情况下，从获得**最优分类面**问题中提出的。
  - **最优分类面**：要求分类面(二维情况下是分类线、高维情况下是超平面)不但能将两类正确分开，而且应使**分类间隔最大**。





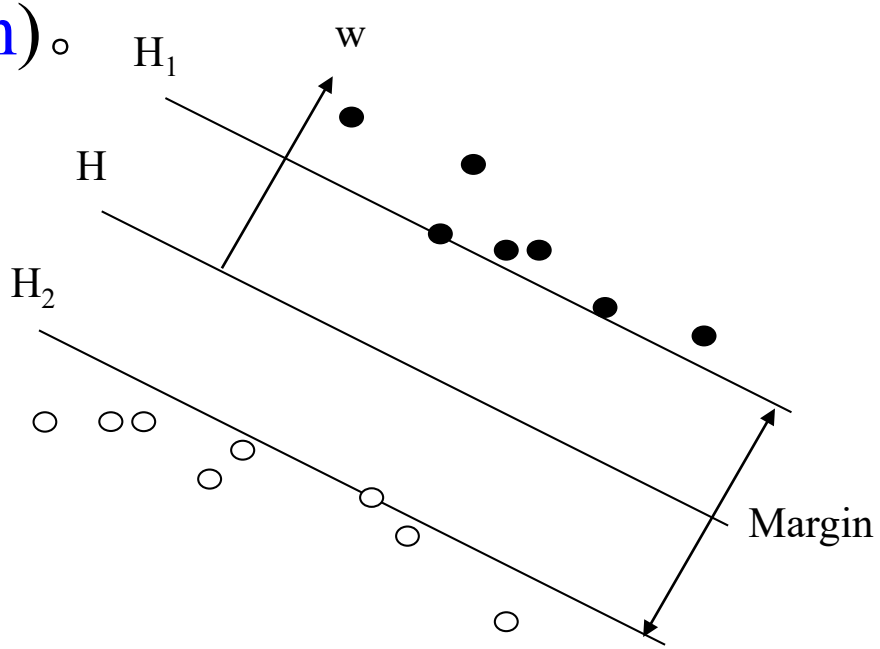
SVM的基本原理



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

SVM的基本原理

- SVM是在**两类线性可分**情况下，从获得**最优分类面**问题中提出的。
  - **分类间隔**：假设H代表分类线，H_1和H_2是两条平行于分类线H的直线，并且它们分别过每类中离分类线H最近的样本，H_1和H_2之间的距离叫做**分类间隔**(margin)。

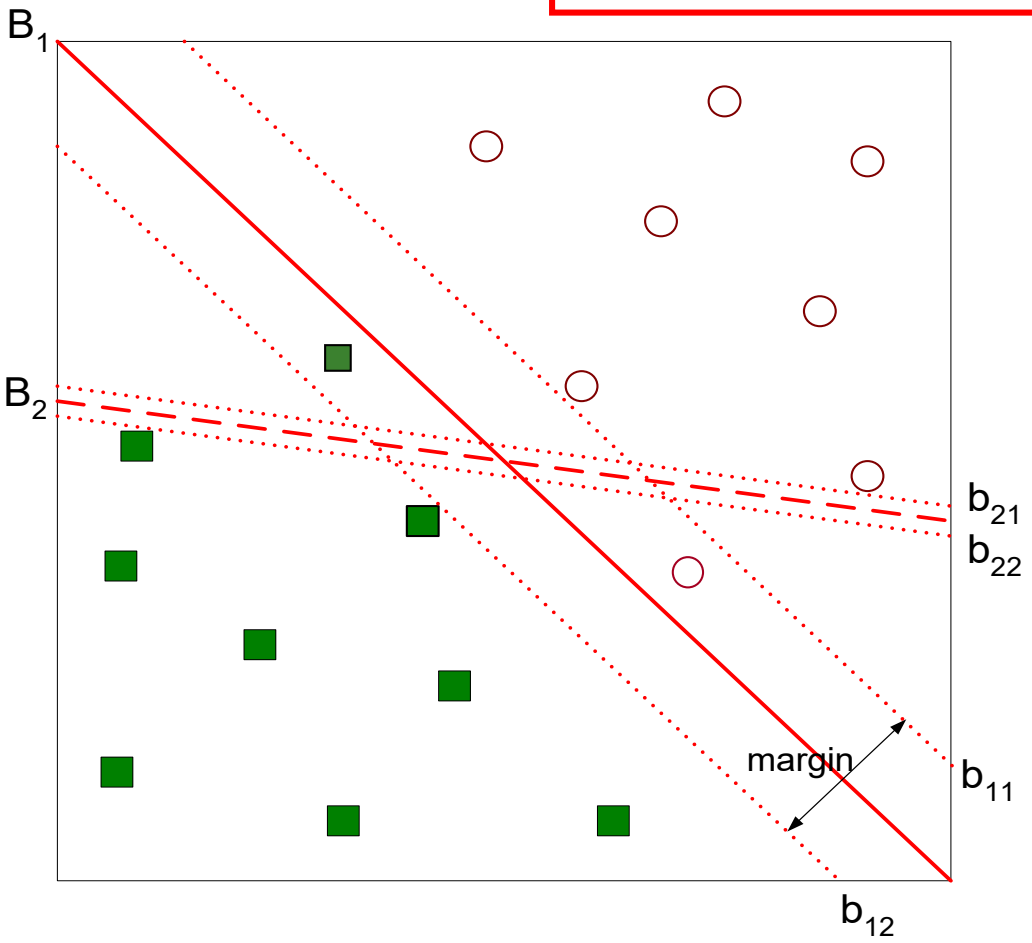


SVM的基本原理

- SVM是在**两类线性可分**情况下，从获得**最优分类面**问题中提出的。
  - SVM就是要在满足条件的众多分类面中，寻找一个能使**分类间隔达到最大**的那个分类面(二维情况下是分类线、高维情况下是超平面)。

SVM的基本原理

Margin越大，对新样本的分类(抗干扰)能力越强。



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

SVM的基本原理

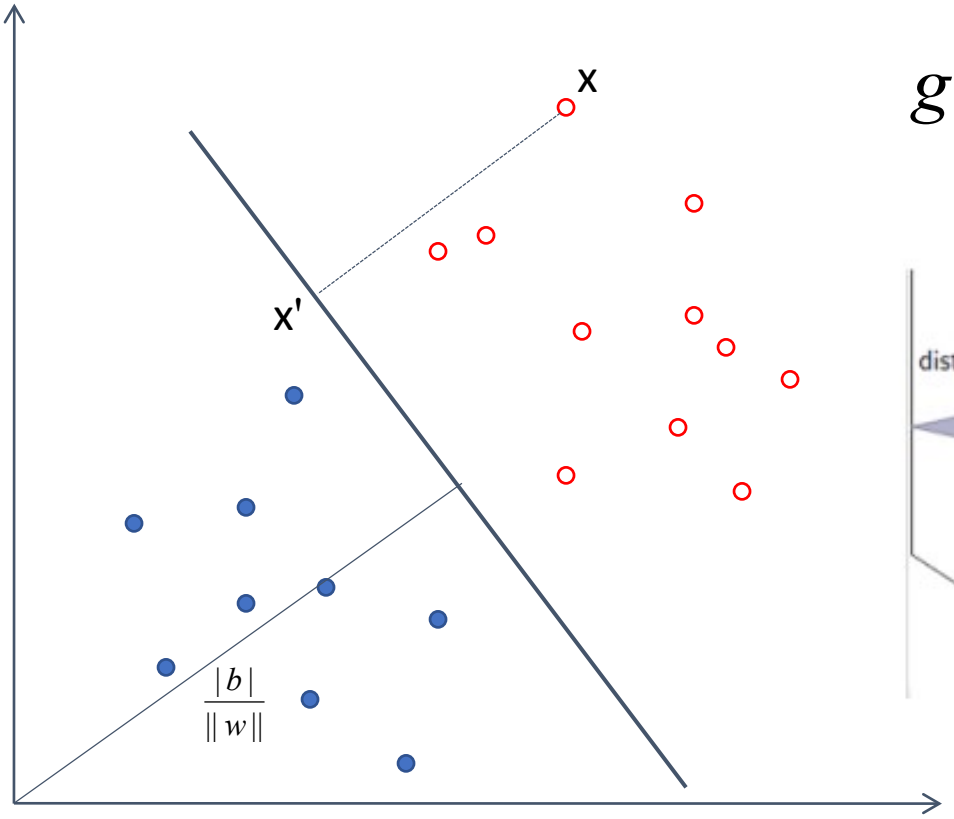
- 问题：在给定的训练数据集上，如何求得具有最大分类间隔的分类面？
- 设：两类线性可分样本集$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，其中：$x_i \in R^d$，$y_i \in \{+1, -1\}$是类别标号，$i=1, 2, \ldots, n$。
  - 对于线性可分问题，分类超平面的定义如下：

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

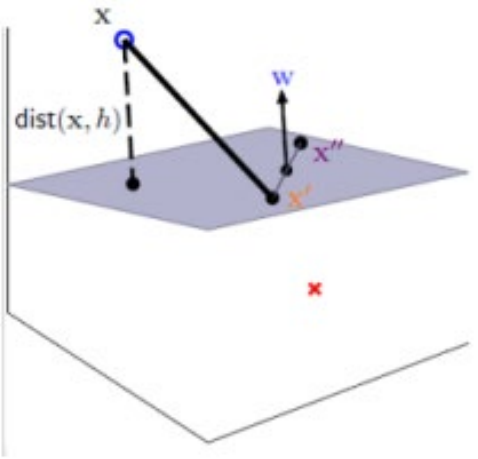
  - 其中，w和b是分类超平面的参数，且$w=\{w_1, w_2, \ldots, w_d\}$是分类超平面的法向量，$b$是偏差。

SVM的基本原理

- 其中，w和b是分类超平面的参数，且$w=\{w_1, w_2, \ldots, w_d\}$是分类超平面的法向量，$b$是偏差。



$$g(x) = w \cdot x + b$$



SVM的基本原理

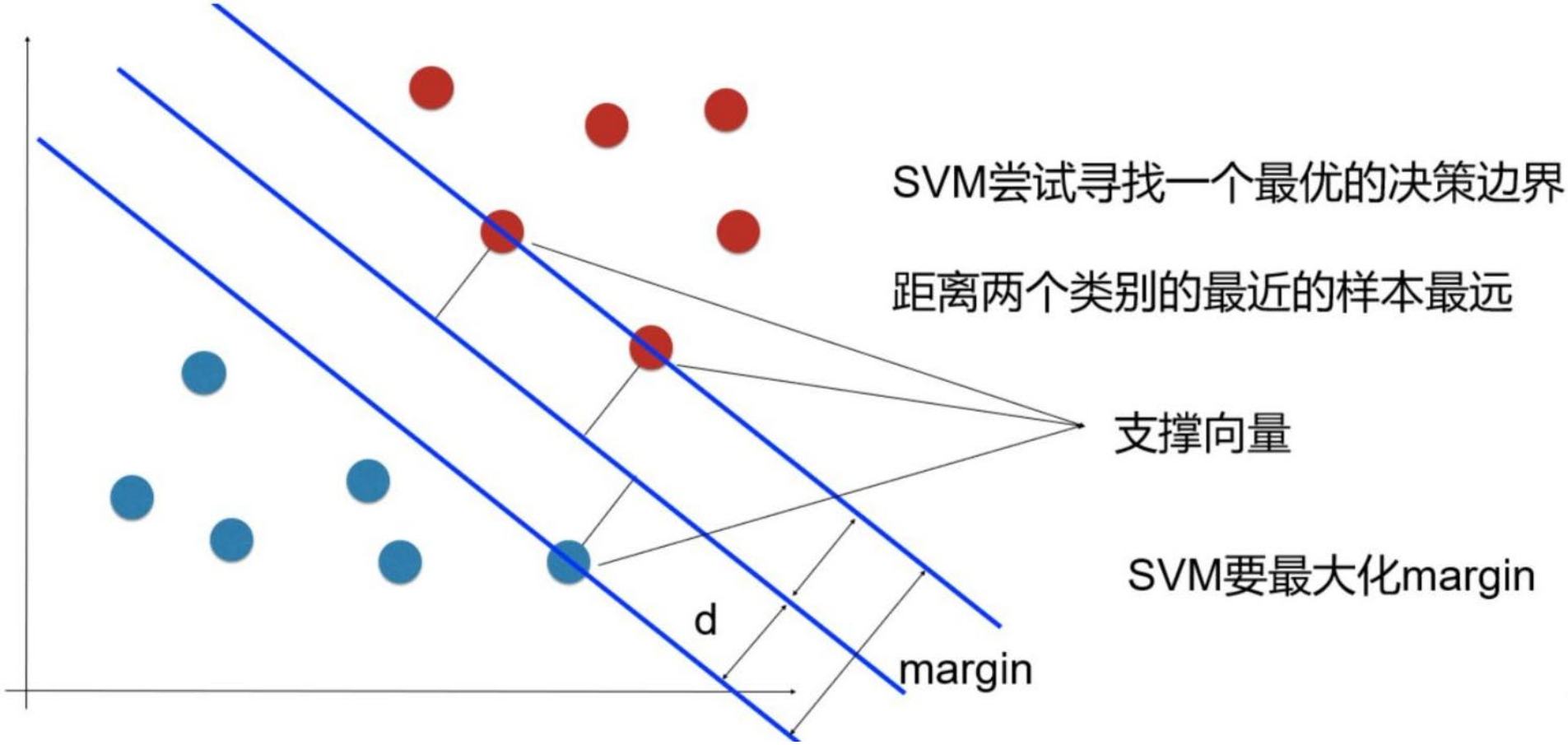
- 设：两类线性可分样本集$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$，其中：$x_i \in R^d$，$y_i \in \{+1, -1\}$是类别标号，$i=1, 2, \ldots, n$。
  - 在分类超平面上方的样本，满足如下条件：

$$w \cdot x_i + b > 0, \quad \text{for } y_i = +1$$

  - 在分类超平面下方的样本，满足如下条件：

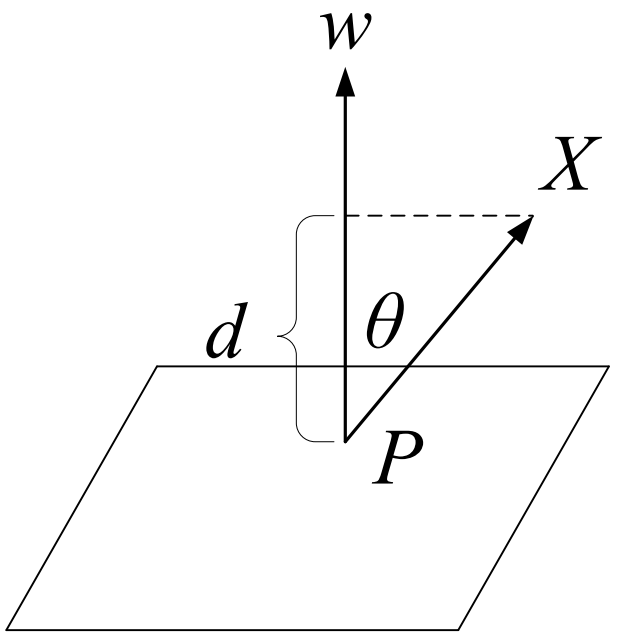
$$w \cdot x_i + b < 0, \quad \text{for } y_i = -1$$

SVM的基本原理



如图所示，根据支持向量的定义我们知道，支持向量到超平面的距离为 d，其他点到超平面的距离大于 d。

SVM的基本原理



- ***w***：法向量

$d = ||\boldsymbol{PX}||*cos\theta$

$= ||\boldsymbol{w}||*||\boldsymbol{PX}||*cos\theta / ||\boldsymbol{w}||$

$= \boldsymbol{PX} \bullet \boldsymbol{w} / ||\boldsymbol{w}||$ (几何间隔)

$\boldsymbol{PX} = (x_0\text{-}p_0, x_1\text{-}p_1, \ldots, x_m\text{-}p_m)$

$\boldsymbol{PX} \bullet \boldsymbol{w} = (x_0\text{-}p_0)w_0+(x_1\text{-}p_1)w_1+ \ldots +(x_m\text{-}p_m)\, w_m$

$=x_0w_0+x_1w_1+\ldots+x_mw_m-(p_0w_0+p_1w_1+\ldots+p_mw_m)$

$=x_0w_0+x_1w_1+\ldots+x_mw_m-(-b)$

$= x_0w_0+x_1w_1+\ldots+x_mw_m+b$

$= \boldsymbol{y_i*(w*x_i+b)}$（也叫函数间隔）

SVM的基本原理

- 二维空间上点（x , y）到直线 $Ax + By + C = 0$ 的距离公式：

$$\frac{|Ax + By + C|}{\sqrt{A^2 + B^2}}$$

- 推广到到 n 维空间后，点 $x = (x_1, x_2 \ldots x_n)$ 到超平面 $w \cdot x + b = 0$ 的距离

$$\frac{|w^T x + b|}{||w||}$$

其中 $||w|| = \sqrt{w_1^2 + \ldots w_n^2}$

SVM的基本原理

- 根据支持向量的定义，支持向量到超平面的距离为 d，其他点到超平面的距离大于 d。等价于

$$\begin{cases} \dfrac{w^T x + b}{||w||} \geq d & y = 1 \\ \dfrac{w^T x + b}{||w||} \leq -d & y = -1 \end{cases}$$

  - 分母 $||w||d$ 可以看作是一种缩放，为了方便计算和优化我们暂且令它为 1，这样做对目标函数的优化没有影响，于是：

$$\begin{cases} w^T x + b \geq 1 & y = 1 \\ w^T x + b \leq -1 & y = -1 \end{cases} \quad \Longleftrightarrow \quad y(w^T x + b) \geq 1$$

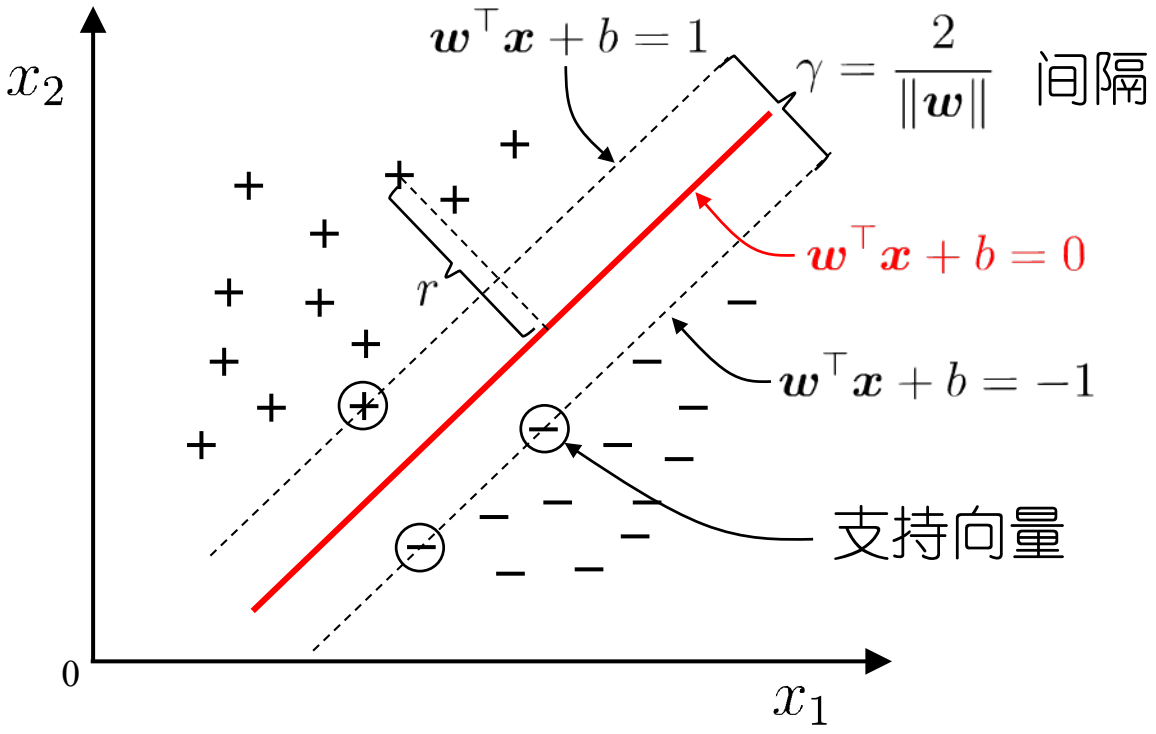
SVM的基本原理

- 根据支持向量的定义，支持向量到超平面的距离d，其他点到超平面的距离大于 d。
- 至此我们就可以得到最大间隔超平面的上下两个超平面：

每个支持向量到超平面的距离可以写为：

$$d = \frac{|w^T x + b|}{||w||}$$

$$d = \frac{y(w^T x + b)}{||w||}$$



SVM的基本原理

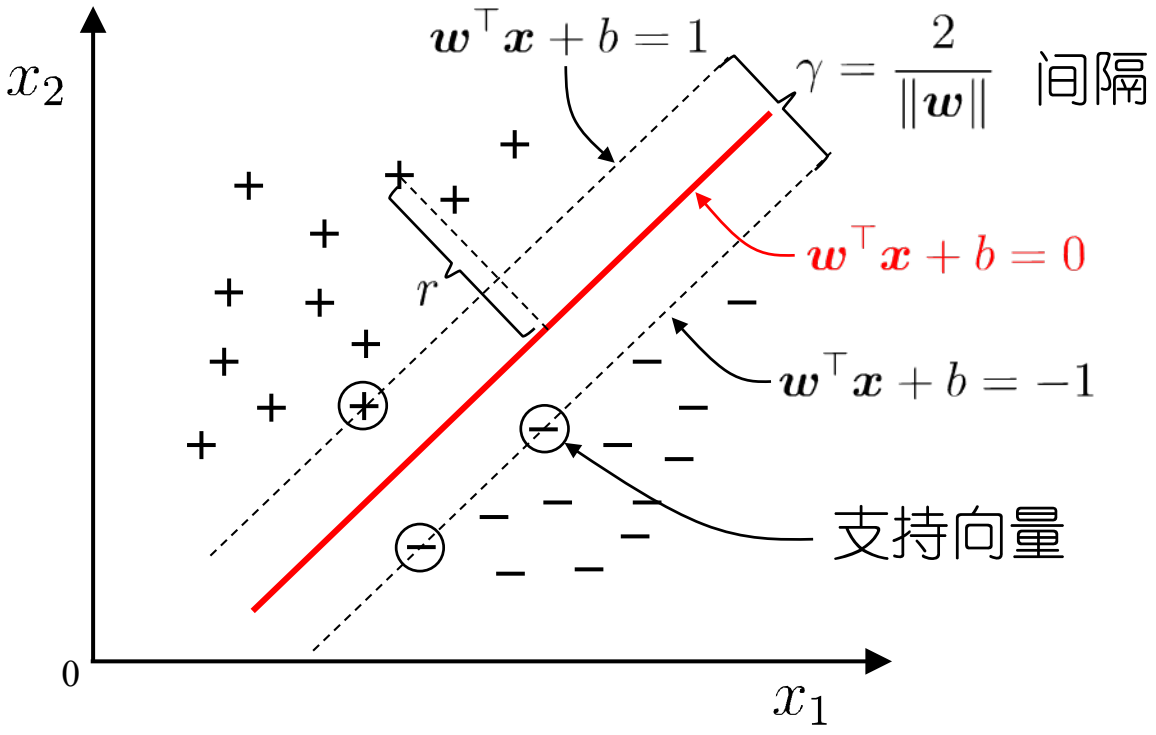
最大化间隔γ：　$\max 2 * \frac{y(w^T x+b)}{||w||}$

支持向量满足：　$y(w^T x + b) = 1$

等价于：

$$\max \frac{2}{||w||}$$

$$\min \frac{1}{2}||w||^2$$



SVM的基本原理

- 至此我们得到最终的优化问题是：

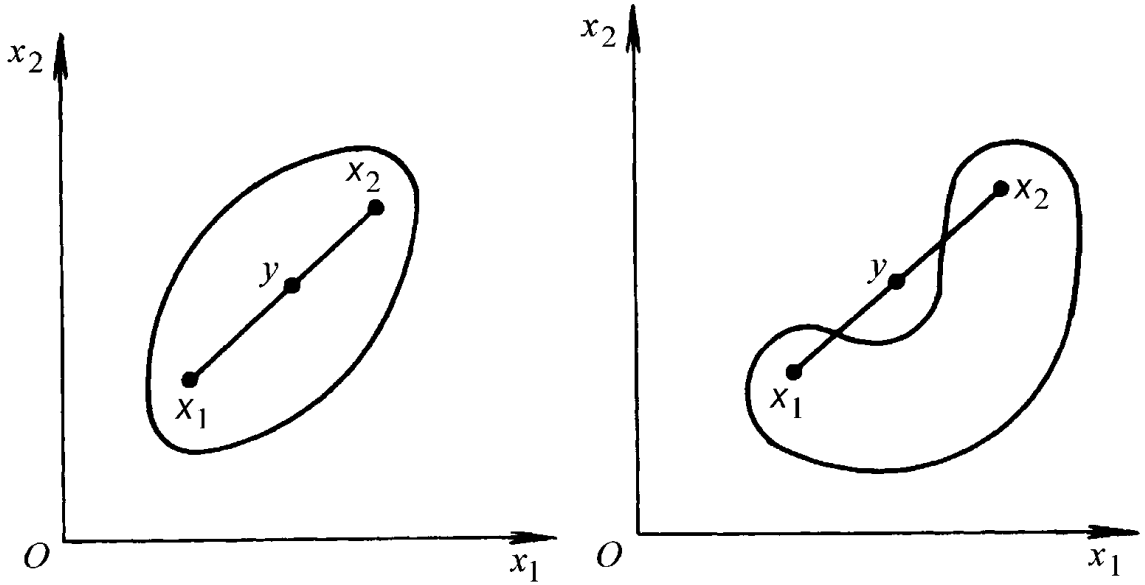
$$\min \frac{1}{2}||w||^2 \; s.t. \quad y_i\left(w^T x_i + b\right) \geq 1$$

- 把寻求分类函数 $f(x) = w.x + b$ 的问题转化为对w，b的最优化问题
- 这是一个凸优化问题，也是一个凸二次规划问题

SVM的基本原理

- 凸集：

  一个点集（或区域），如果连接其中任意两点 x_1，x_2 的线段都全部包含在该集合内，就称该点集为凸集，否则为非凸集



SVM的基本原理

- 凸函数：

  设f(x)为定义在凸集R上，且具有连续的一阶导数的函数，则f(x)在R上为凸函数的充要条件是对凸集R内任意不同两点 x_1 x_2 ，不等式

$$f(x_2) \geq f(x_1) + (x_2 - x_1)^T \nabla f(x_1)$$

  恒成立

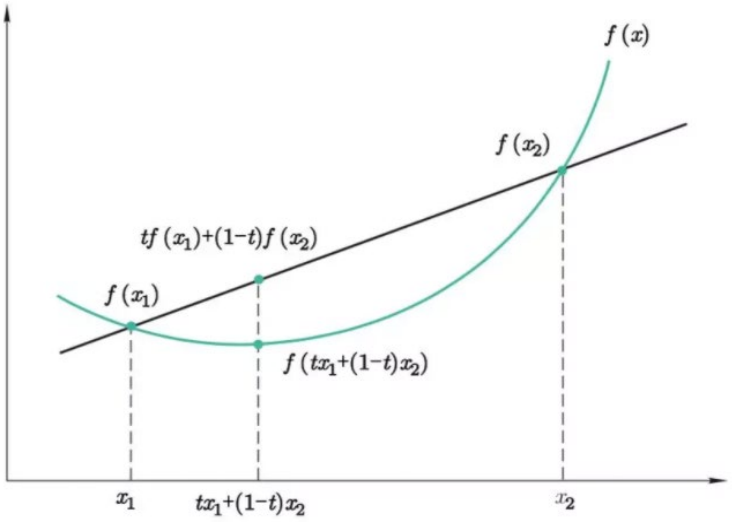


图1： 凸函数的割线在函数曲线上方

SVM的基本原理

- 凸规划：

  对于约束优化问题

$$\min f(x)$$

$$s.t. \quad g_j(x) \le 0 \quad j = 1, 2, ..., m$$

若 $f(x)$ $g_j(x)$都为凸函数，则此问题为凸规划问题。

凸规划的任何局部最优解就是全局最优解

- 求解：利用拉格朗日乘数法转换成无约束优化问题，然后转换成对偶问题求解

SVM的基本原理

- 拉格朗日乘数法：

本科高等数学学的拉格朗日程数法✦是等式约束优化问题：

$$\min f(x_1, x_2, \ldots, x_n)$$
$$s.t. \quad h_k(x_1, x_2, \ldots, x_n) = 0 \quad k = 1, 2, \ldots, l$$

我们令 $L(x, \lambda) = f(x) + \sum_{k=1}^{l} \lambda_k h_k(x)$ ，函数 $L(x, y)$ 称为 Lagrange 函数，参数 λ 称为 Lagrange 乘子**没有非负要求**。

利用必要条件找到可能的极值点✦：

$$\begin{cases} \dfrac{\partial L}{\partial x_i} = 0 \quad i = 1, 2, \ldots, n \\ \dfrac{\partial L}{\partial \lambda_k} = 0 \quad k = 1, 2, \ldots, l \end{cases}$$

具体是否为极值点需根据问题本身的具体情况检验。这个方程组称为等式约束的极值必要条件。

等式约束下的 Lagrange 乘数法引入了 l 个 Lagrange 乘子，我们将 x_i 与 λ_k 一视同仁，把 λ_k 也看作优化变量，共有 $(n + l)$ 个优化变量。

SVM的基本原理

- 现在面对的是不等式优化问题，这种情况仍是利用拉格朗日乘数法转换成无约束优化问题，然后转换成对偶问题（拉格朗日对偶问题）求解（为了更容易求解），解是最优解的前提是满足KKT条件。

不等式约束优化：

$$\begin{aligned} & min f(x) \\ s.t.\ & g_i(x) = 0 \quad i = 1, 2, , , m \\ & h_j(x) \leqslant 0 \quad j = 1, 2, , , n \end{aligned}$$

拉格朗日函数：

$$\iota(x, \alpha, \theta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{j=1}^{n} \theta_i h_i(x)$$

转化为拉格朗日对偶问题：

$$max_{\alpha,\theta}\, min_x\, \iota(x, \alpha, \theta)$$

根据KKT条件求解：

$$\begin{cases} \frac{\partial \iota(x,\alpha,\theta)}{\partial x}|_{x=x^*} = 0 \\ \frac{\partial \iota(x,\alpha,\theta)}{\partial \alpha}|_{x=x^*} = 0 \\ \frac{\partial \iota(x,\alpha,\theta)}{\partial \theta}|_{x=x^*} = 0 \\ \alpha_i \neq 0 \\ \theta_j \geqslant 0 \\ \theta_j h_j(x)|_{x=x^*} = 0 \quad \text{（互补松驰性）} \\ h_j(x)|_{x=x^*} \leq 0 \\ g_i(x)|_{x=x^*} = 0 \end{cases}$$

SVM的基本原理

我们已知 SVM 优化的主问题是：

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

$$s.t. \quad g_i(w,b) = 1 - y_i（w^T x_i + b） \leq 0, \quad i = 1, 2, \ldots, n$$

那么求解线性可分的 SVM 的步骤为：

步骤 1：

构造拉格朗日函数✦：

$$\min_{w,b} \max_{\lambda} L(w,b,\lambda) = \frac{1}{2}||w||^2 + \sum_{i=1}^{n} \lambda_i [1 - y_i (w^T x_i + b)]$$

$$s.t. \quad \lambda_i \geq 0$$

SVM的基本原理

步骤 2:

利用强对偶性转化:

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

现对参数 w 和 b 求偏导数✦:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \lambda_i x_i y_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \lambda_i y_i = 0$$

得到:

$$\sum_{i=1}^{n} \lambda_i x_i y_i = w$$

$$\sum_{i=1}^{n} \lambda_i y_i = 0$$

SVM的基本原理

我们将这个结果带回到函数中可得：

$$
\begin{aligned}
L(w,b,\lambda) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\lambda_i y_i (\sum_{j=1}^{n}\lambda_j y_j (x_i \cdot x_j) + b) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{n}\lambda_i - \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{n}\lambda_i y_i b \\
&= \sum_{j=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j (x_i \cdot x_j)
\end{aligned}
$$

也就是说：

$$
\min_{w,b} L(w,b,\lambda) = \sum_{j=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j (x_i \cdot x_j)
$$

SVM的基本原理

步骤 3:

由步骤 2 得:

$$\max_{\lambda}[\sum_{j=1}^{n}\lambda_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j(x_i \cdot x_j)]$$

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

我们可以看出来这是一个二次规划问题，问题规模正比于训练样本数，我们常用 SMO(Sequential Minimal Optimization) 算法求解。

SMO(Sequential Minimal Optimization)，序列最小优化算法，其核心思想非常简单：每次只优化一个参数，其他参数先固定住，仅求当前这个优化参数的极值。

SVM的基本原理

1. 选择两个需要更新的参数 λ_i 和 λ_j ，固定其他参数。于是我们有以下约束：

这样约束就变成了：

$$\lambda_i y_i + \lambda_j y_j = c \quad \lambda_i \geq 0, \lambda_j \geq 0$$

其中 $c = -\sum_{k \neq i,j} \lambda_k y_k$ ，由此可以得出 $\lambda_j = \frac{c - \lambda_i y_i}{y_j}$ ，也就是说我们可以用 λ_i 的表达式代替 λ_j 。这样就相当于把目标问题转化成了仅有一个约束条件的最优化问题，仅有的约束是 $\lambda_i \geq 0$ 。

2. 对于仅有一个约束条件的最优化问题，我们完全可以在 λ_i 上对优化目标求偏导，令导数为零，从而求出变量值 $\lambda_{i_{new}}$ ，然后根据 $\lambda_{i_{new}}$ 求出 $\lambda_{j_{new}}$ 。

3. 多次迭代直至收敛。

通过 SMO 求得最优解 λ^* 。

SVM的基本原理

步骤 4：

我们求偏导数时得到：

$$w = \sum_{i=1}^{m} \lambda_i y_i x_i$$

由上式可求得 w。

我们知道所有 $\lambda_i > 0$ 对应的点都是支持向量，我们可以随便找个支持向量，然后带入：$y_s(wx_s + b) = 1$，求出 b 即可，

SVM的基本原理

步骤 5：w 和 b 都求出来了，我们就能构造出最大分割超平面：$w^T x + b = 0$

分类决策函数✦：$f(x) = sign(w^T x + b)$

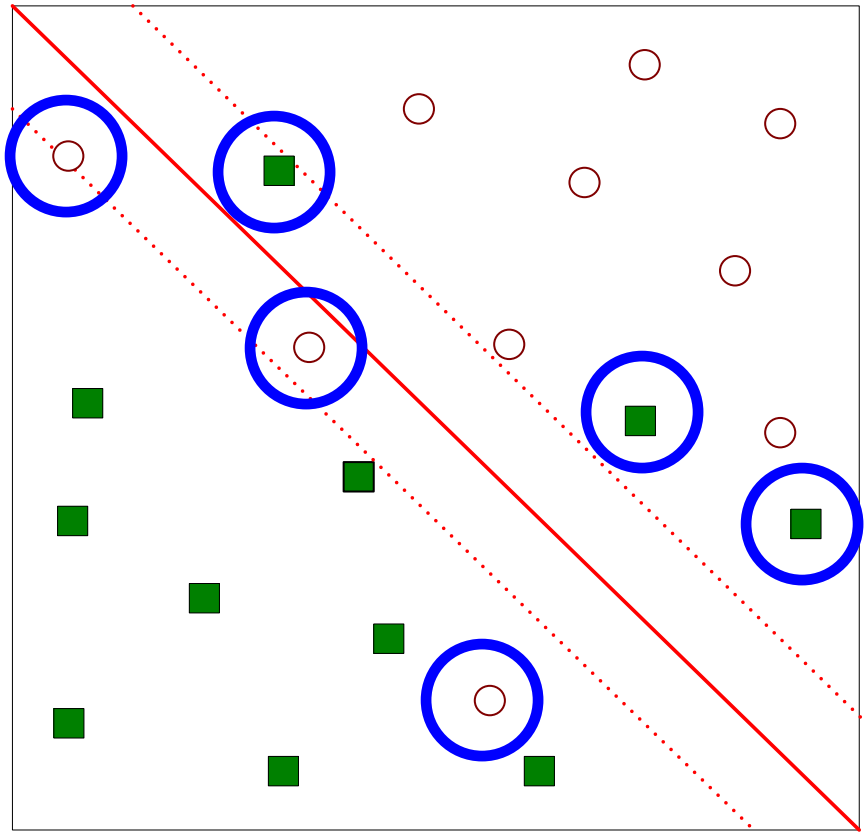
其中 $sign(\cdot)$ 为阶跃函数✦：

$$sign(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

将新样本点导入到决策函数中既可得到样本的分类。

SVM的基本原理

- 样本数据是线性不可分时，该怎么办？



SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法1：通过引入**松弛变量**(slack variables)，来构建软间隔SVM，使函数间隔加上松弛变量大于等于1，则目标函数带约束条件的最优化问题形式如下：

$$\min_{w,\,b}\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\}$$

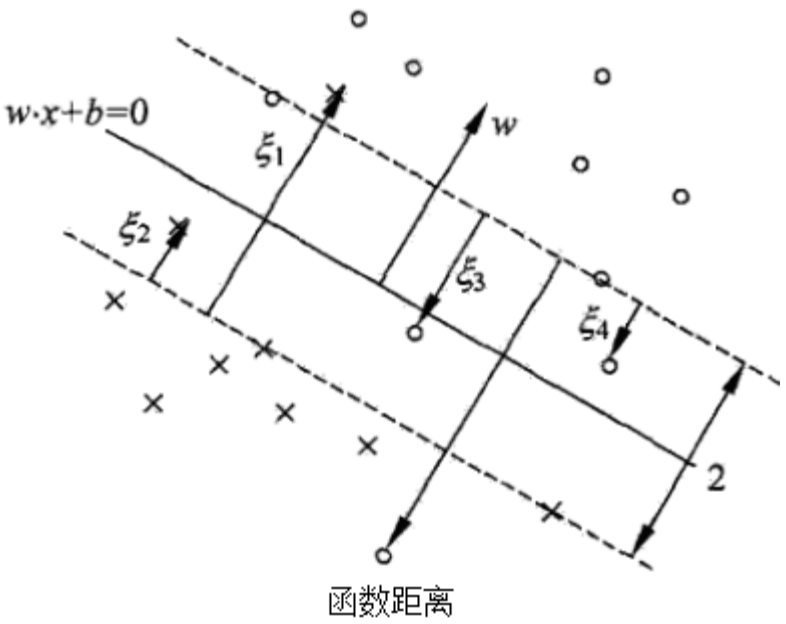
$$s.t.\ \ y_i\cdot(x_i\cdot w+b)\geq 1-\xi_i,\ \ i=1,\ldots,n$$

$$\xi_i\geq 0$$

惩罚因子，*C*通常取值为大于0的常数

SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法1：通过引入**松弛变量**(slack variables)，来构建软间隔SVM，带约束条件的最优化问题形式如下：



- 以某些点不能正确划分为代价，来换取更大的分隔间隔。
- 变量ξ_i($\xi_i \geq 0$)记录了对(x_i, y_i)分类的错误代价。

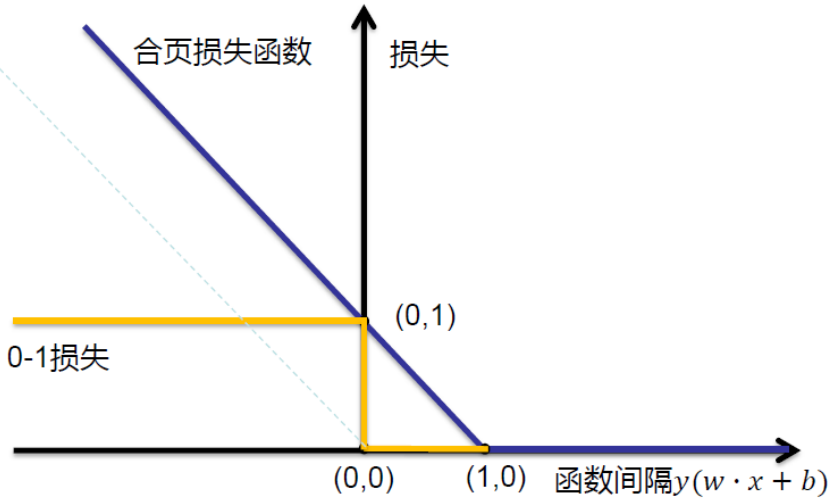
SVM的基本原理

ξ 为"松弛变量"

$$\xi_i = \max(0, 1 - y_i(w^{\mathrm{T}} x_i + b))$$

即hinge损失函数。每一个样本都有一个对应的松弛变量，表征该样本不满足约束的程度。

蓝色线代表hinge损失函数，黄色线代表0-1损失函数，可以认为它是二类分类问题的真正的损失函数，而合页损失函数是0-1损失函数的上界。



损失函数hinge loss

SVM的基本原理

- 因为松弛变量是非负的，因此样本的函数间隔可以比1小。
- 函数间隔比1小的样本被叫做离群点，放弃了对离群点的精确分类，这对分类器来说是种损失。
- 并非所有的样本点都有一个松弛变量与其对应。只有“离群点”才有。
- 松弛变量的值实际上标示出了对应的点到底离群有多远，值越大，点就越远。
- 放弃这些点也带来了好处，那就是超平面不必向这些点的方向移动，因而可以得到更大的几何间隔（在低维空间看来，分类边界也更平滑）。

SVM的基本原理

- 惩罚因子*C*决定了对离群点带来损失的重视程度。当所有离群点的松弛变量的和一定时，*C*越大，对目标函数的损失也越大，此时就暗示着非常不愿意放弃这些离群点，最极端的情况是把*C*定为无限大，这样只要稍有一个点离群，目标函数的值就变成无限大，马上让问题变成无解，这就退化成了硬间隔问题。
- 惩罚因子*C*不是一个变量，整个优化问题在解的时候，*C*是一个必须事先指定的值，指定这个值以后，解一下，就得到一个分类器，然后用测试数据看看结果好不好，不好再换一个*C*的值。如此就是一个参数寻优的过程。

SVM的基本原理

原始问题的对偶问题是

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & 0 \leqslant \alpha_i \leqslant C,\ i = 1, 2, ..., N \end{aligned}$$

下面我们来具体说明上述对偶问题是怎么得出的：

原始最优化问题的**拉格朗日函数**是

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i$$

其中，

$$\alpha_i \geqslant 0, \quad \mu_i \geqslant 0$$

对偶问题是拉格朗日函数的极大极小问题。

SVM的基本原理

首先求$L(w,b,\xi,\alpha,\mu)$对w,b,ξ的极小，由

$$\bigtriangledown_w L(w,b,\xi,\alpha,\mu) = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0$$

$$\bigtriangledown_b L(w,b,\xi,\alpha,\mu) = -\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\bigtriangledown_{\xi_i} L(w,b,\xi,\alpha,\mu) = C - \alpha_i - \mu_i = 0$$

得

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

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

$$C - \alpha_i - \mu_i = 0$$

将上式代入上面的拉格朗日函数，得

$$\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) + \sum_{i=1}^{N} \alpha_i$$

SVM的基本原理

再对极小（上式）求α**的极大**，即得**对偶问题**：

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

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

$$C - \alpha_i - \mu_i = 0$$

$$\alpha_i \geqslant 0$$

$$\mu_i \geqslant 0,\ i = 1, 2, ..., N$$

拉格朗日对偶的重要作用是将w的计算提前并消除w，使得优化函数变为拉格朗日乘子的单一参数优化问题。

将对偶最优化问题（上式）进行变换：利用等式约束（上式第二项约束）消去μ_i，从而只留下变量α_i，并将约束（上式后三项约束）写成

$$0 \leqslant \alpha_i \leqslant C$$

SVM的基本原理

再通过取负，将对目标函数求极大转换为求极小，于是得到对偶问题

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^{N}\alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^{N}\alpha_i y_i = 0 \\ & 0 \leqslant \alpha_i \leqslant C,\ i = 1, 2, ..., N \end{aligned}$$

可以**通过求解对偶问题而得到原始问题的解**(w^*, b^*)，**进而确定分离超平面和决策函数。**

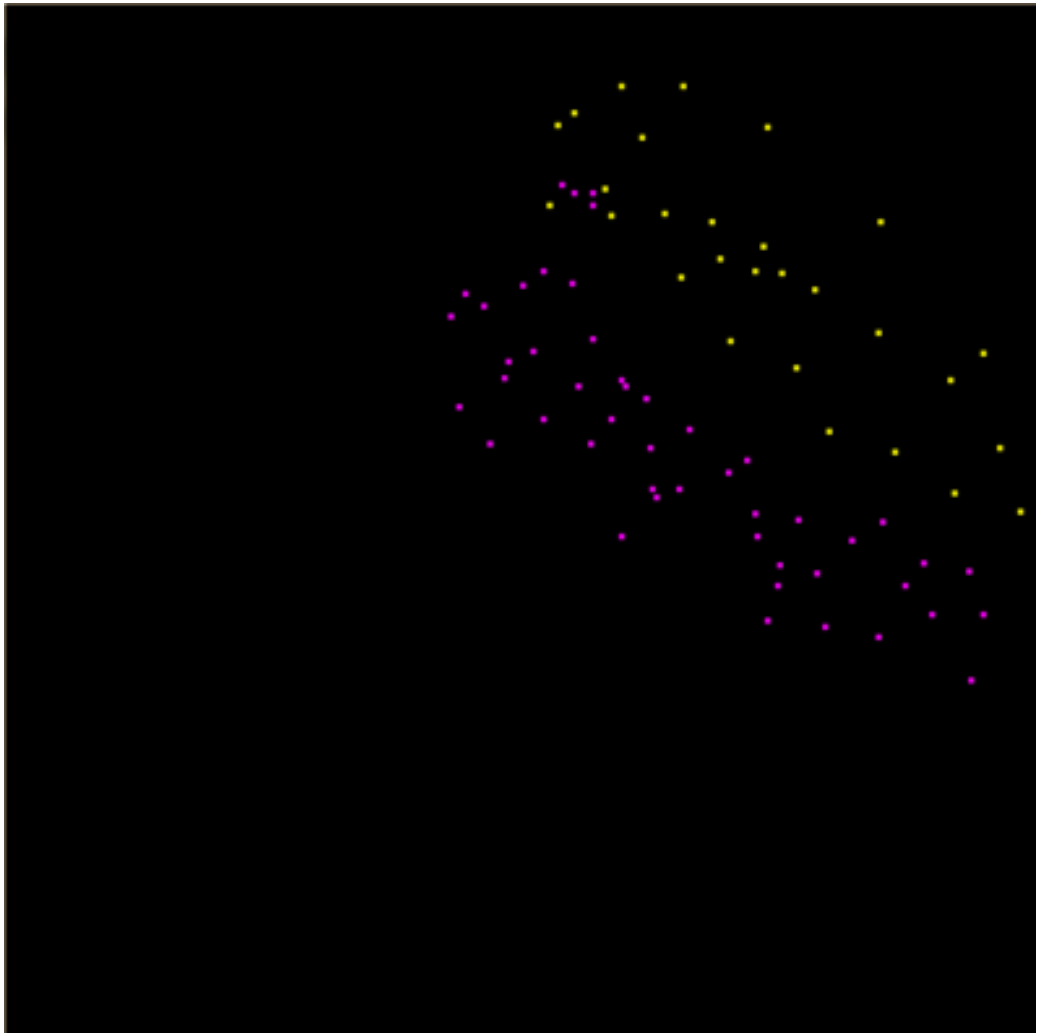
求解原始最优化问题的解w^*和b^*，得到线性支持向量机，其分离超平面为

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

分类决策函数为：$f(x) = \mathrm{sign}\left(w^{*\mathrm{T}}x + b^*\right)$

SVM的基本原理

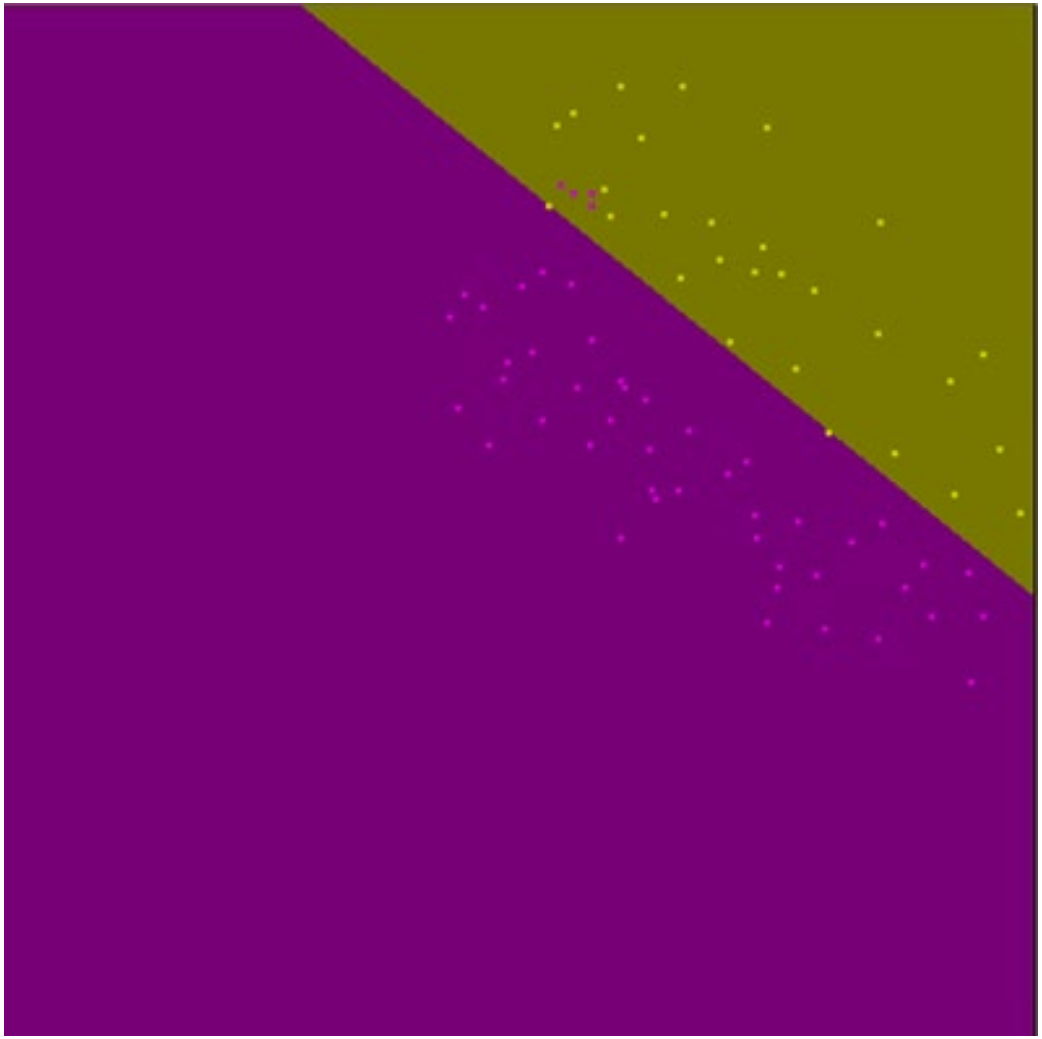
观察惩罚因子*C*的不同取值对分类的影响



原始数据

SVM的基本原理

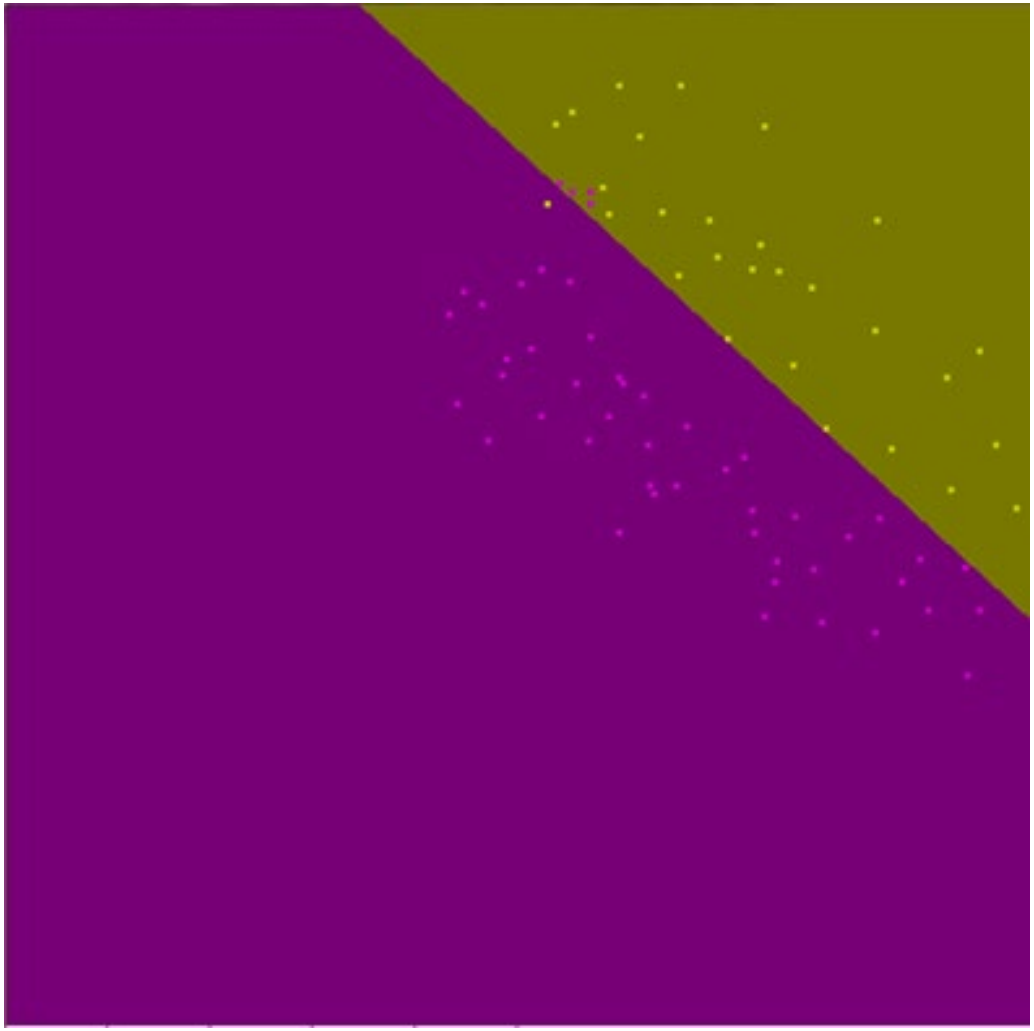
观察惩罚因子C的不同取值对分类的影响



C=100

SVM的基本原理

观察惩罚因子C的不同取值对分类的影响

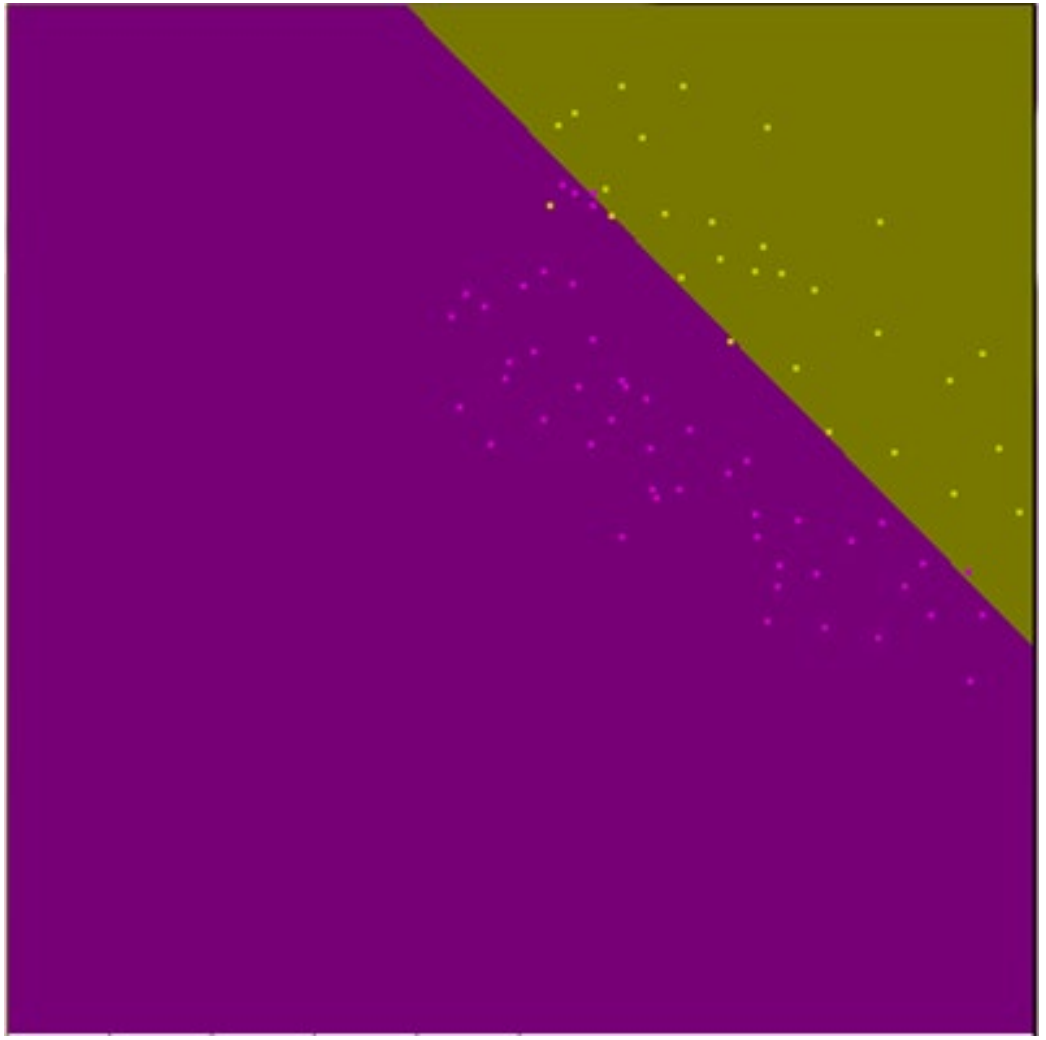


C=1,000

SVM的基本原理

观察惩罚因子*C*的不同取值对分类的影响

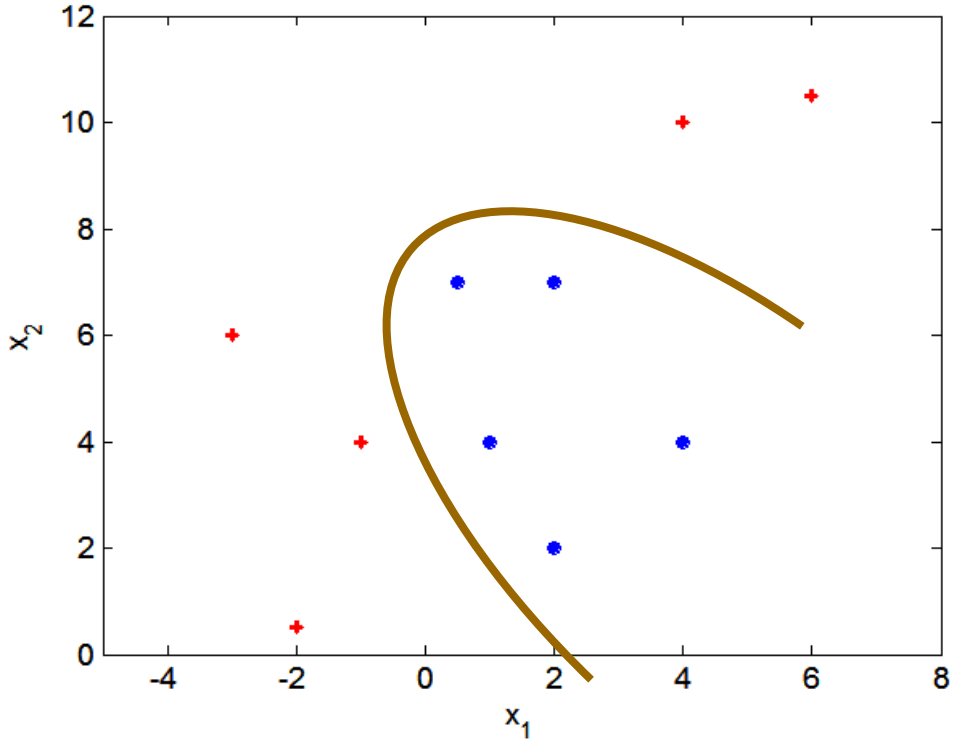
***C*越大分类超平面越向离群点移动，最终的分类超平面由离群点决定。**



C=100,000

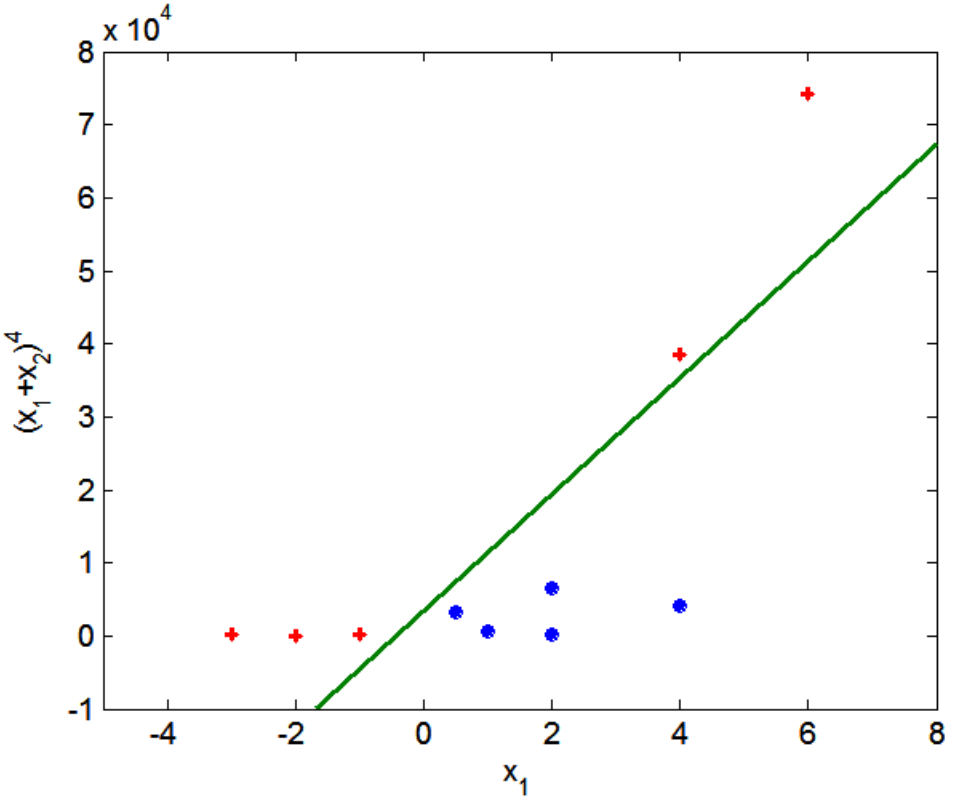
SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法2：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。



SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。

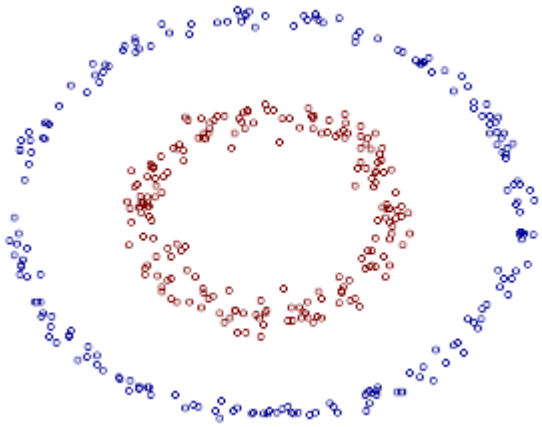


SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法1：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。
  - **数据变换到高维空间可分的理由：当维度增加到无限维的时候，一定可以让任意两个物体可分。**
    - 举一个哲学的例子：世界上本来没有两个完全一样的物体，对于所有的两个物体，可通过增加维度来让他们最终有所区别。
    - 比如：两本书，从(**颜色**，**内容**)两个维度来说，可能是一样的，可以加上**作者**这个维度，实在不行还可以加入**页码**，**拥有者**，**购买地点** ……

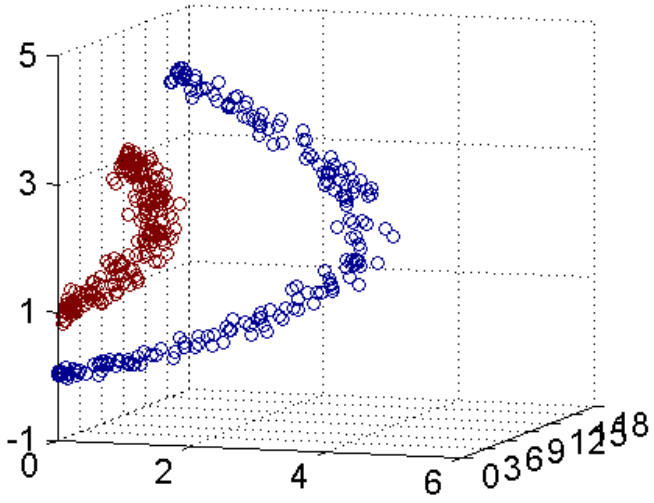
SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法1：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。
  - 使用一种**非线性变换**，可将原数据映射到高维空间中。



$$z_1 = x_1^2, z_2 = x_2^2, z_3 = x_2$$

左图中的点可被映射成三维空间中的某个点



SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法2：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。
  - 使用一种**非线性变换**，可将原数据映射到高维空间中。
  - 非线性变换的形式是什么样的？
  - 核函数:它可以将样本从原始空间映射到一个更高维的特质空间中，使得样本在新的空间中线性可分这样我们就可以使用原来的推导来进行计算，只是所有的推导是在新的空间，而不是在原来的空间中进行，即用核函数来替换当中的内积。

SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法2：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。

$$\min_{w,\,b}\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\}$$

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

$$\xi_i \geq 0$$

通过拉格朗日乘子，
可得到其对偶问题。

$$\max_{\alpha}\sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

两个样本数据的点积

$$s.t., C\geq\alpha_i\geq 0, i=1,\cdots,n$$

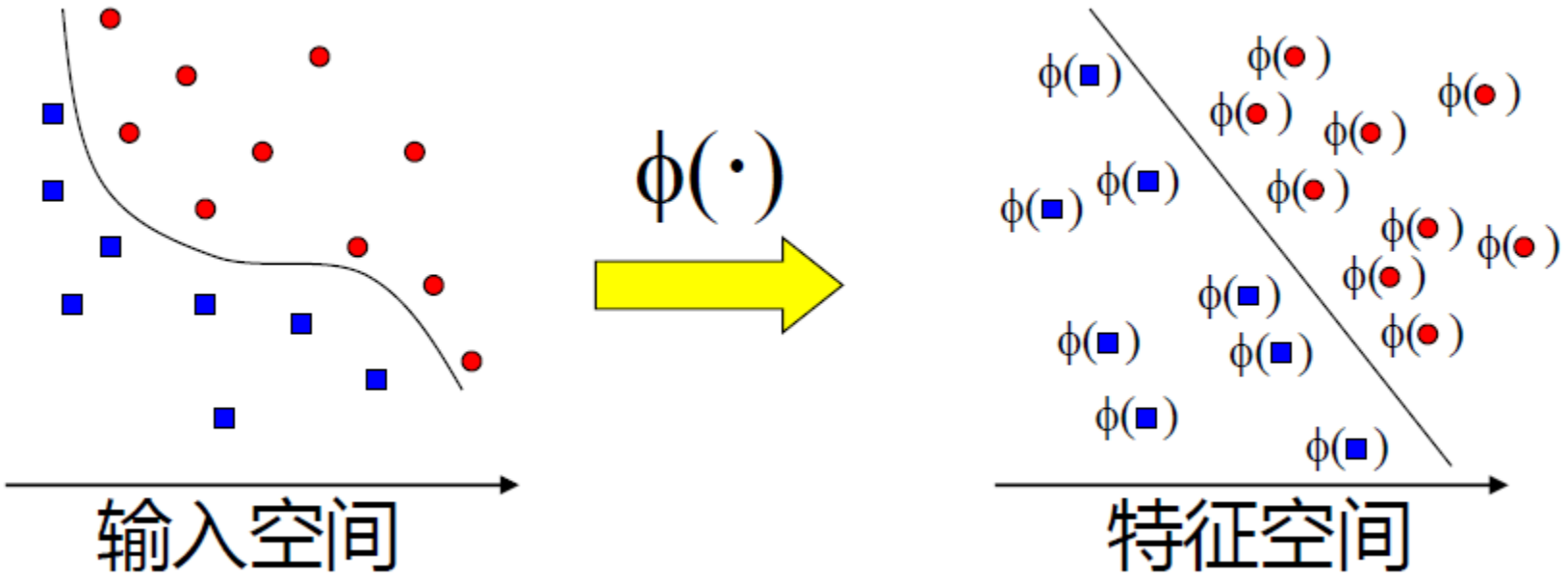
$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法1：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。
  - 使用一种**非线性变换**，可将原数据映射到高维空间中。
  - **非线性变换的形式是什么样的？**
  - 根据泛函理论，数据的点积等价于使用一个核函数$K(X_i, X_j)$，即：$K(X_i, X_j) = \Phi(X_i)\Phi(X_j)$。

SVM的基本原理

- 核技巧：根据泛函理论，数据的点积等价于使用一个核函数$K(X_i, X_j)$，即：$K(X_i, X_j) = \Phi(X_i)\Phi(X_j)$。用户核函数替换原来的内积



SVM的基本原理

- 核技巧：根据泛函理论，数据的点积等价于使用一个核函数$K(X_i, X_j)$，即：$K(X_i, X_j) = \Phi(X_i)\Phi(X_j)$。用户核函数替换原来的内积

在线性支持向量机学习的对偶问题中，用核函数$K(x, z)$替代内积，求解得到的就是非线性支持向量机

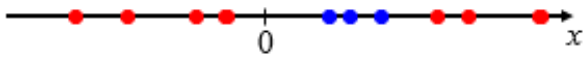
$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*\right)$$

SVM的基本原理

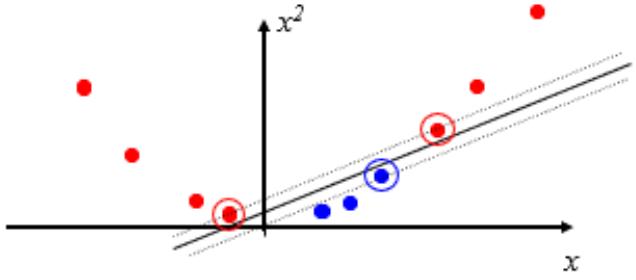
- 核技巧：根据泛函理论，数据的点积等价于使用一个核函数$K(X_i, X_j)$，即：$K(X_i, X_j) = \Phi(X_i)\Phi(X_j)$。用户核函数替换原来的内积

  - **一维空间向二维空间映射**

    下面我们考虑一维空间的二分类问题：

    我们将它进行一个二次变换，换到二维空间，这里的变换为 x -> x^2。

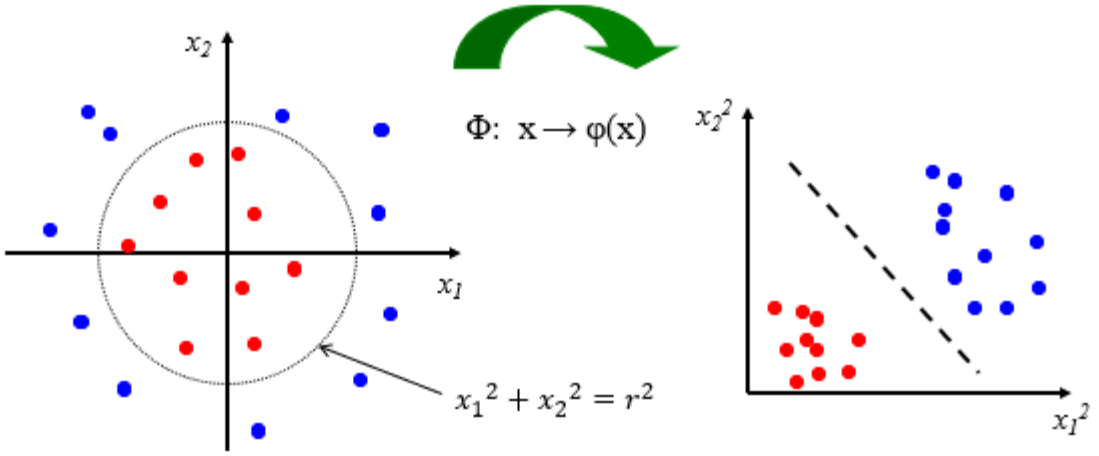
    

    从上面的例子，我们知道变换的核心思想就是：**将原始输入空间的数据集映射到高维特征空间中，从而使得数据集可分。**

SVM的基本原理

- 核技巧：根据泛函理论，数据的点积等价于使用一个核函数$K(X_i, X_j)$，即：$K(X_i, X_j) = \Phi(X_i)\Phi(X_j)$。用户核函数替换原来的内积

- 二维空间向二维特征空间映射



上图中二维空间不可分，但是变换一下坐标空间，也能实现线性可分。

SVM的基本原理

- 样本数据是线性不可分时，该怎么办？
  - 解决方法1：将样本数据转换到高维空间中，在高维空间中寻找分类超平面。
  - 常用的核函数形式如下：

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j)^d$	$d \geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh\left(\beta \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j + \theta\right)$	tanh 为双曲 正切函数, $\beta > 0, \theta < 0$

如果 Feature 的数量很大，跟样本数量差不多，这时候选用 LR 或者是 Linear Kernel 的 SVM

如果 Feature 的数量比较小，样本数量一般，不算大也不算小，选用 SVM+Gaussian Kernel

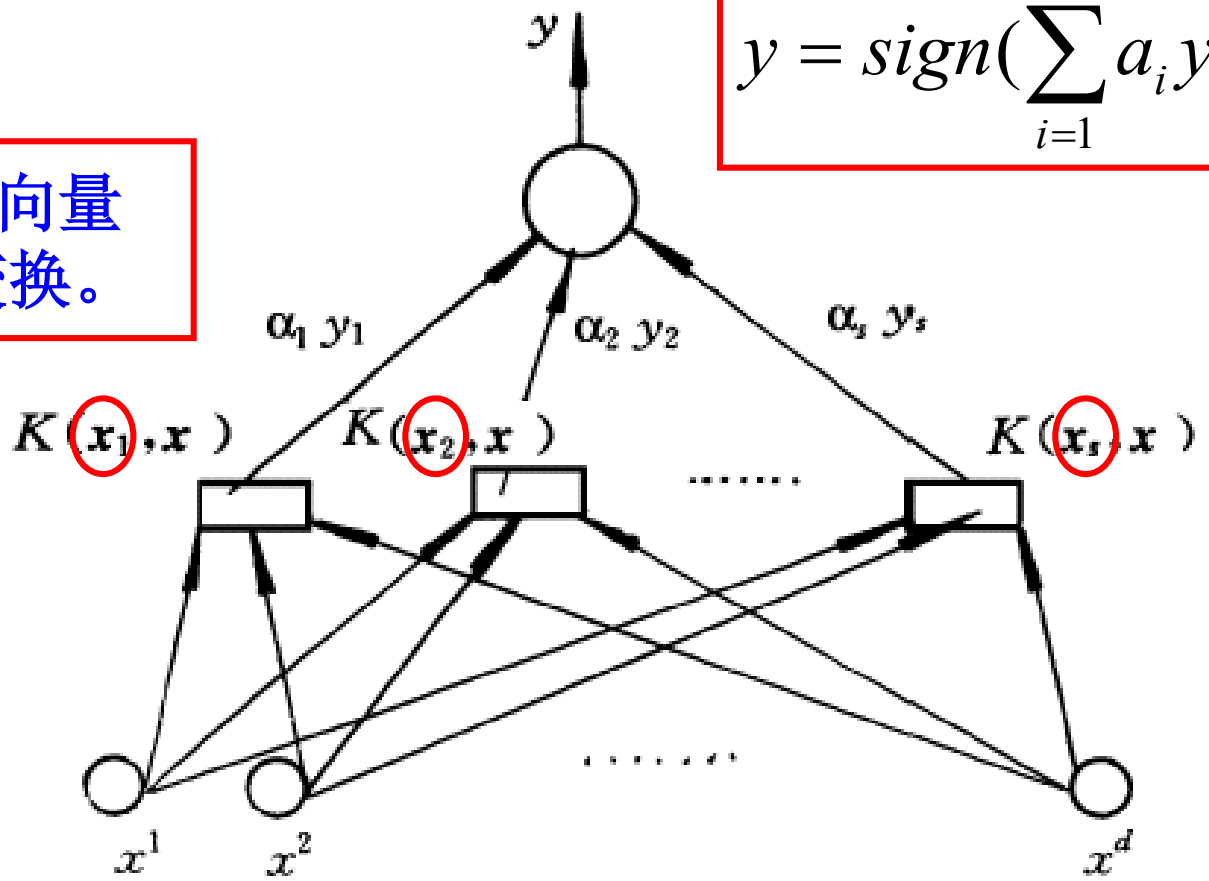
如果 Feature 的数量比较小，而样本数量很多，需要手工添加一些 feature 变成第一种情况

SVM的基本原理

- 在核函数的作用下，SVM相当于如下形式的网络结构：

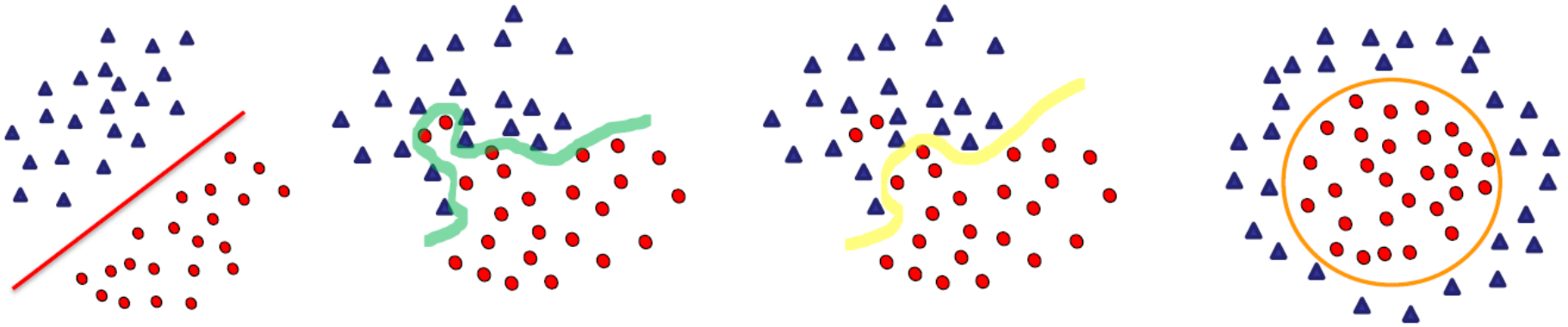
$$y = sign(\sum_{i=1}^{s} a_i y_i K(x_i, x))$$

S个支持向量参与核变换。



SVM的基本原理

硬间隔、软间隔和非线性 SVM



线性可分　　硬间隔　　软间隔　　线性不可分

假如数据是完全的线性可分的，那么学习到的模型可以称为硬间隔支持向量机。换个说法，硬间隔指的就是完全分类准确，不能存在分类错误的情况。软间隔，就是允许一定量的样本分类错误。

SVM总结

- 优点：
  - 有严格的数学推理；
  - 小样本分类器；
  - 特别适合处理复杂的非线性分类问题。
- 缺点：
  - 训练时间非常长；
  - 无法直接处理多分类问题。

SVM总结

下面是一些SVM普遍使用的准则：

- n为特征数，m为训练样本数。
- (1)如果相较于m而言，n要大许多，即训练集数据量不够支持我们训练一个复杂的非线性模型，我们选用逻辑回归模型或者不带核函数的支持向量机。
- (2)如果n较小，而且m大小中等，例如n在1-1000之间，而m在10-10000之间，使用高斯核函数的支持向量机。
- (3)如果n较小，而m较大，例如n在1-1000之间，而m大于50000，则使用支持向量机会非常慢，解决方案是创造、增加更多的特征，然后使用逻辑回归或不带核函数的支持向量机。

SVM多分类问题

- 对于$N(N>2)$ 类分类问题，有两种解决办法：
  - 1 *vs* (N–1)：需要训练N个分类器，第i个分类器用于判断样本数据是否属于第i类；
  - 1 *vs* 1：需要训练$N*(N-1)/2$个分类器，分类器(i,j)能够判断样本数据是属于第i类，还是第j类 (如，(i, j)-classifier 如果是i win,则$i = i+1$;otherwise, $j = j+1$;)。当对一个未知样本进行分类时，最后得票最多的类别即为该未知样本的类别。
- 在实际中，通常采用1 *vs* (N-1)方式解决多分类问题。