机器学习

计算机学院

杨晓春
xcyang@bit.edu.cn

1

第四章：支持向量机

- Tom M. Mitchell, McGraw Hill, 2003
- http://www.cs.cmu.edu/~awm/tutorials
- 周志华，机器学习，2016

目标

- 理解支持向量机SVM的原理与目标
- 掌握支持向量的计算过程和算法步骤

3

概念

- 线性可分支持向量机
  - 硬间隔最大化hard margin maximization
  - 硬间隔支持向量机
  - 对偶问题
- 线性支持向量机
  - 软间隔最大化soft margin maximization
  - 软间隔支持向量机
  - 正则化
- 非线性支持向量机
  - 核函数kernel function

注：以上概念的提法，各个文献并不十分统一。

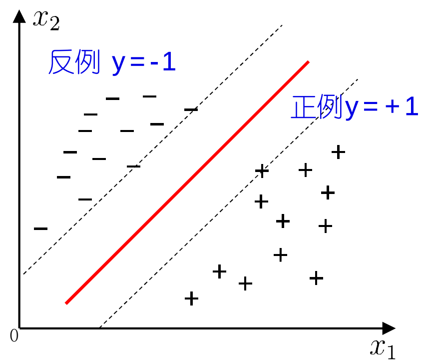
4

概念

- 线性可分支持向量机
  - 硬间隔最大化hard margin maximization
  - 硬间隔支持向量机
  - 对偶问题
- 线性支持向量机
  - 软间隔最大化soft margin maximization
  - 软间隔支持向量机
- 非线性支持向量机
  - 核函数kernel function

5

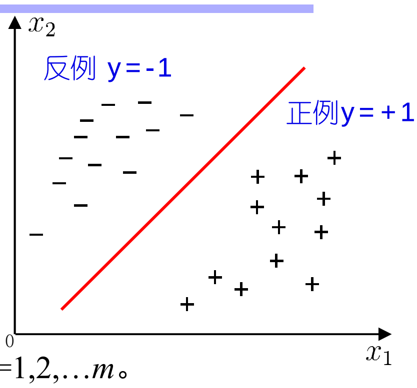
线性可分支持向量机



6

引子

- 线性模型：在样本空间中寻找一个超平面，将不同类别的样本分开
- 假设给定一个特征空间上的训练数据集 $D=\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2) \ldots (\mathbf{x}_m, y_m)\}$



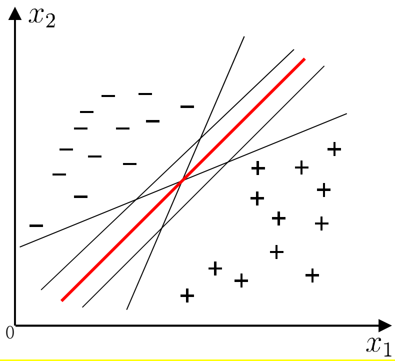
  - 其中，$\mathbf{x}_i \in R_d$，$y_i \in \{+1,-1\}$，$i=1,2,\ldots m$。
  - $\mathbf{x}_i$为第i个实例(若$d>1$，$\mathbf{x}_i$为向量)；
  - y_i为$\mathbf{x}_i$的类标记；
    - 当$y_i=+1$时，称$\mathbf{x}_i$为正例；
    - 当$y_i=-1$时，称$\mathbf{x}_i$为负例；
  - $(\mathbf{x}_i, y_i)$称为样本点。

7

线性分类问题

- 若两个集合有部分相交，如何定义超平面使两个集合“尽量”分开？

Q:将训练样本分开的超平面可能有很多, 哪一个好呢?



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

任何一个划分都可以，但哪个最好？

A:应选择"正中间", 容忍性好, 鲁棒性高, 泛化能力最强.

本页课件来源：周志华《机器学习》及其课件，致谢：张腾

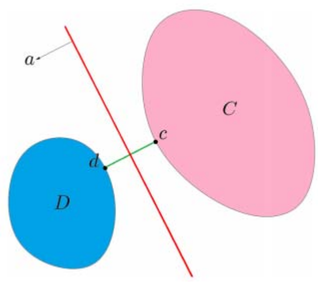
8

分割超平面

- 设C和D为两不相交的凸集，则存在超平面P，P可以将C和D分离。

$$\forall x \in C, a^T x \le b \text{且} \forall x \in D, a^T x \ge b$$

- 两个集合的距离，定义为两个集合间元素的最短距离。
- 做集合C和集合D最短线段的垂直平分线。

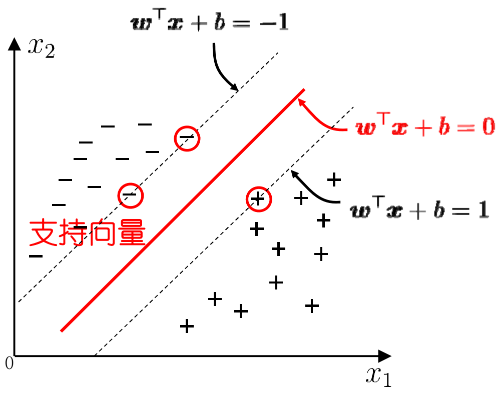


本页课件致谢：普开数据 邹伟

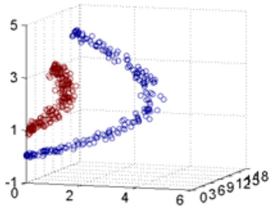
(硬)间隔与支持向量

超平面方程(线性方程): $\boldsymbol{w}^\top \boldsymbol{x} + b = 0$

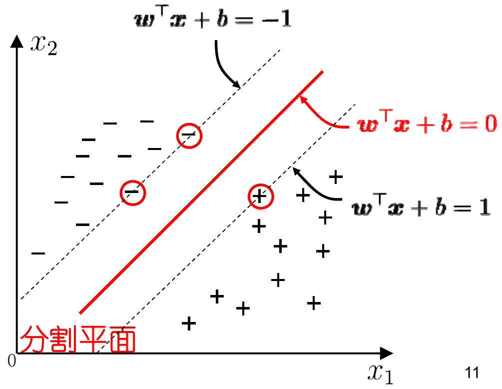
- 最优分割超平面
  - 支持向量
  - 集合边界的若干点为“基础”计算超平面的方向
  - 两个集合边界上这些点的平均作为超平面的截距



线性可分支持向量机



- 给定线性可分训练数据集，通过间隔最大化
  平面为 $y(x) = \boldsymbol{w}^{\top}\boldsymbol{x} + b$
- 相应的决策分类函数 $f(x) = \mathrm{sign}(\boldsymbol{w}^{\top}\boldsymbol{x} + b)$
  - 该决策分类函数称为：线性可分支持向量机
- 整理符号
  - 训练集：$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n$
  - 目标值：$y_1, y_2, \ldots y_n$
  - 新数据分类：
    - $y(x) = \boldsymbol{w}^{\top}\boldsymbol{x} + b$
    - $\mathrm{sign}(y(x))$

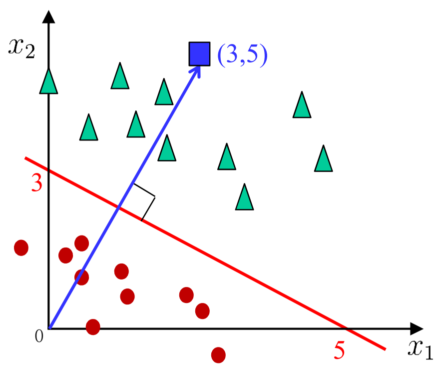


线性可分支持向量机

- 参数**w**表示的是超平面的法线方向（垂直方向）
- 当**w**确定了，分割线的斜率就确定了

➔一组与法线垂直的超平面

$w_1x_1+w_2x_2+b=0$



$3x_1+5x_2-15=0$

$\boldsymbol{w}^{\top}\boldsymbol{x} + b$

$\boxed{\mathbf{w} = (3;5)}$

$\mathbf{x} = (x_1; x_2)$

$b = -15$

目的：求**w**, b

▲ > 0

● < 0

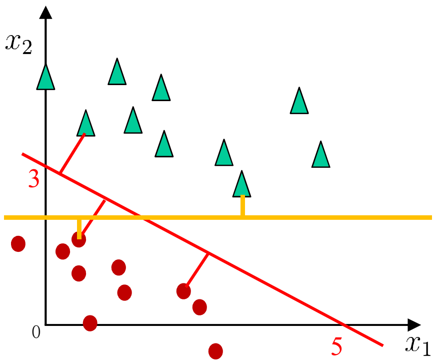
线性可分支持向量机

- 点到直线的距离：

$$dist = \frac{|w_1x_1 + w_2x_2 + b|}{\sqrt{w_1^2 + w_2^2}} = \frac{|w^Tx + b|}{||w||}$$

- 不可导，所以变形得到点到超平面的距离公式

$$dist = \frac{(w^Tx + b) * y}{||w||}$$



Q1: 如何确定**w**?
Q2: 为何是选中间？

推导目标函数

- 根据题设 $y(x) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$
- 得到：

$$\begin{cases} y(x_i) > 0 \Leftrightarrow y_i = +1 \\ y(x_i) < 0 \Leftrightarrow y_i = -1 \end{cases} \Rightarrow y_i \cdot y(x_i) > 0$$

- **w** 和 b等比缩放，

$$\frac{y_i \cdot y(x_i)}{\|w\|} = \frac{y_i \cdot (w^T \cdot x_i + b)}{\|w\|}$$

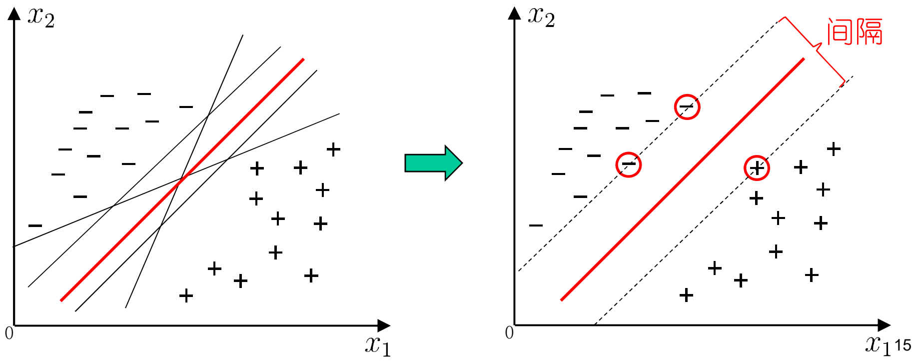
本页课件致谢：普开数据 PUKAI DATA 邹伟

支持向量机基本型

$$\frac{y_i \cdot y(x_i)}{\|w\|} = \frac{y_i \cdot (w^T \cdot x_i + b)}{\|w\|}$$

- **最大间隔**分离超平面: 寻找参数**w**和b, 使得间隔最大.
- 目标函数：

$$\arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|} \min_i \left[y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \right] \right\}$$



15

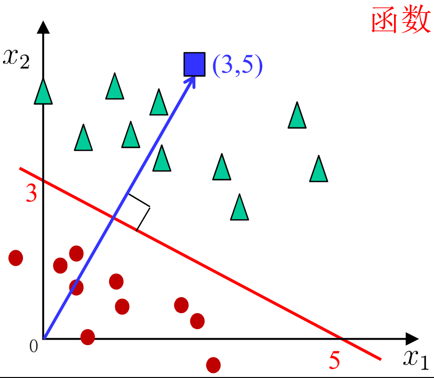
支持向量机基本型

$$\frac{y_i \cdot y(x_i)}{\|w\|} = \frac{y_i \cdot (w^T \cdot x_i + b)}{\|w\|}$$

- **最大函数间隔**分离超平面: 寻找参数**w**和b, 使得间隔γ_i最大
- 函数间隔 $\gamma_i = y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)$
- 目标函数：

$$\arg\max_{\boldsymbol{w},b} \left\{ \frac{1}{\|\boldsymbol{w}\|} \min_i \left[\underline{y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)} \right] \right\}$$

函数间隔(functional margin)可以无限小



$3x_1+5x_2-15=0$

$30x_1+50x_2-150=0$

$0.3x_1+0.5x_2-1.5=0$

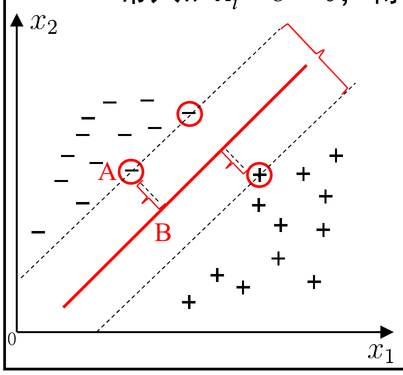
➔加约束条件，让红线部分≥1

16

函数间隔和几何间隔

$$\frac{y_i\cdot y(x_i)}{\|w\|}=\frac{y_i\cdot\left(w^T\cdot x_i+b\right)}{\|w\|}$$

- 定义全部样本上的函数间隔 $\gamma=\min\gamma_i$，$i=1,2\ldots m$
- 任意点到分割平面的距离（几何间隔**geometric margin**）
  - A点：(x_i, y_i)，B点：$x_B = x_i - \gamma_i \frac{w}{||w||}$
  - 带入$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b = 0$，得 $\gamma_i = y_i((\frac{w}{||w||})^T x_i + \frac{b}{||w||})$



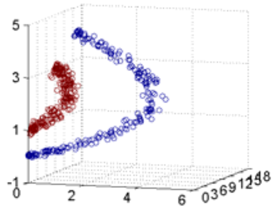
当$\|\mathbf{w}\|=1$时，就是函数间隔

函数间隔和几何间隔

$$\frac{y_i\cdot y(x_i)}{\|w\|}=\frac{y_i\cdot\left(w^T\cdot x_i+b\right)}{\|w\|}$$

- 分割平面 $y(x) = \boldsymbol{w}^\top \boldsymbol{x} + b$
- 总可以通过等比例缩放的方法，使得两类点的函数值都满足$|y(x) \geqslant 1|$
- 原目标函数 $\arg\max_{\boldsymbol{w},b}\left\{\frac{1}{\|\boldsymbol{w}\|}\min_i\left[\underbrace{y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)}_{\geqslant 1}\right]\right\}$
- 新目标函数 $\arg\max_{\boldsymbol{w},b}\frac{1}{\|\boldsymbol{w}\|}$
- 约束条件 $y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)\geq 1$

建立目标函数



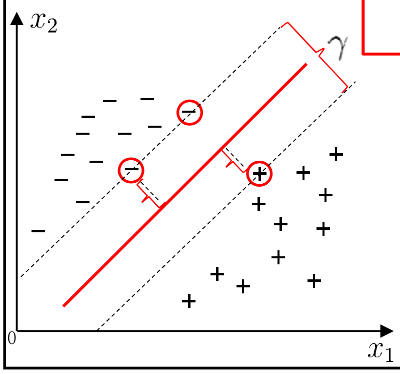
$$\arg\max_{\boldsymbol{w},b} \ \frac{2}{\|\boldsymbol{w}\|}$$

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

SVM的目标函数

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

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



$\gamma = \dfrac{2}{\|\boldsymbol{w}\|}$ 间隔

两个异类支持向量到超平面的距离之和

19

概念

- **线性可分支持向量机**
  - 硬间隔最大化hard margin maximization
  - 硬间隔支持向量机
  - 求解目标函数
- 线性支持向量机
  - 软间隔最大化soft margin maximization
  - 软间隔支持向量机
- 非线性支持向量机
  - 核函数kernel function

20

求解目标函数—拉格朗日乘子法

- 目标 $\arg\min\limits_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$

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

- 凸二次优化问题：希望可以高效求解

➔引入拉格朗日乘子$\alpha_i \geq 0$ 得到拉格朗日函数

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

对w,b求偏导得

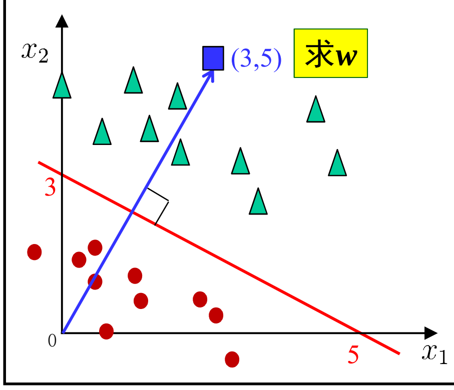
$$\frac{\partial L}{\partial w} = 0 \quad \Rightarrow \quad \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$

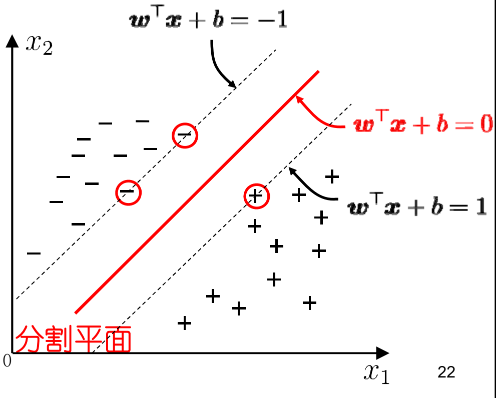
$$\frac{\partial L}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \alpha_i y_i = 0$$

理解参数间的关系

- w和样本相关

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$





求解目标函数—对偶问题

- 目标 $$\arg\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$ $$\text{s.t.}\ \ y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) \geq 1,\ i = 1, 2, \ldots, m.$$
- 拉格朗日函数 $$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1\right)$$ $$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \qquad \sum_{i=1}^{m} \alpha_i y_i = 0$$
- 原问题是极小极大问题 $$\min_{w,b} \max_{\alpha} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$$
- 该问题的对偶问题是极大极小问题 $$\max_{\alpha} \min_{w,b} L(\boldsymbol{w}, b, \boldsymbol{\alpha})$$

23

计算拉格朗日的对偶函数

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$
$$0 = \sum_{i=1}^{m} \alpha_i y_i$$

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y_i\left(w^T x_i + b\right) - 1\right)$$
$$= \frac{1}{2} w^T w - w^T \sum_{i=1}^{m} \alpha_i y_i x_i - b \sum_{i=1}^{m} \alpha_i y_i + \sum_{i=1}^{m} \alpha_i$$
$$= \frac{1}{2} w^T \sum_{i=1}^{m} \alpha_i y_i x_i - w^T \sum_{i=1}^{m} \alpha_i y_i x_i - b \cdot 0 + \sum_{i=1}^{m} \alpha_i$$
$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\left(\sum_{i=1}^{m} \alpha_i y_i x_i\right)^T \sum_{i=1}^{m} \alpha_i y_i x_i$$
$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j$$
$$a^* = \arg\max_{\alpha} \left(\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \right)$$

本页课件致谢：普开数据 PUKAI DATA 邹伟

24

整理目标函数

添加负号 ⟶
$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j - \sum_{i=1}^{m}\alpha_i$$
$$\text{s.t.} \sum_{i=1}^{m}\alpha_i y_i = 0,\ \alpha_i \geq 0,\ i = 1, 2, \ldots, m.$$

b的偏导得到的约束　　凸函数得到的约束

25

解的稀疏性

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i$$

- 最终模型： $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b = \boxed{\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^\top \boldsymbol{x}} + b$
- KKT条件 (Karush-Kuhn-Tucker)：

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

$$\begin{cases} \alpha_i \geq 0, \\ y_i f(\boldsymbol{x}_i) \geq 1, \\ \alpha_i(y_i f(\boldsymbol{x}_i) - 1) = 0. \end{cases}$$

y_i和$f(x_i)$符号相同

$$y_i f(\boldsymbol{x}_i) > 1 \Rightarrow \alpha_i = 0$$

支持向量机解的**稀疏性**：训练完成后，大部分的训练样本都不需保留，最终模型仅与支持向量有关.

本页课件来源：周志华《机器学习》及其课件，致谢：张腾

26

线性可分支持向量机学习算法

- 构造并求解约束最优化问题

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\top}\boldsymbol{x}_j - \sum_{i=1}^{m}\alpha_i$$

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

- 求最优解α*

27

线性可分支持向量机学习算法

- 计算

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

$$b^* = y_i - \sum_{i=1}^{m}\alpha_i^* y_i x_i x_j$$

- 求分离超平面

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

- 分类决策函数

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

28

求解方法 – SMO (Sequential Minimal Optimization)

- 求a^* $$a^* = \arg\max_{\alpha}\left(\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j\right)$$
- 基本思路：不断执行如下两个步骤直至收敛.
  - 第一步：选取一对需更新的变量α_i和α_j
  - 第二步：固定α_i和α_j以外的参数, 求解对偶问题更新α_i和α_j
- 仅考虑α_i和α_j时, 对偶问题的约束变为

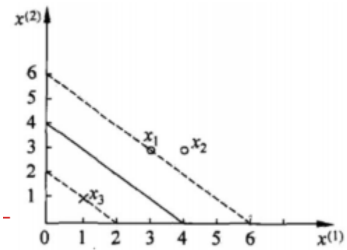
$$\alpha_i y_i + \alpha_j y_j = -\sum_{k\neq i,j}\alpha_k y_k, \quad \alpha_i \geq 0, \quad \alpha_j \geq 0.$$

  - 用一个变量, 该问题具有闭式解.表示另一个变量, 回代入对偶问题可得一个单变量的二次规划
- 偏移项b：通过支持向量来确定

将m个解问题，转换成两个变量的求解问题：并且目标函数是凸的

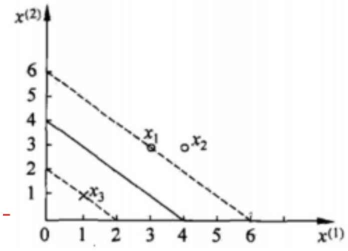
举例

- 给定3个数据点：
  - 正例点x_1=(3,3)^T， x_2= (4,3)^T
  - 负例点x_3=(1,1)^T
- 求线性可分支持向量机



举例

- 给定3个数据点：
  - 正例点 $x_1=(3,3)^T$，$x_2=(4,3)^T$
  - 负例点 $x_3=(1,1)^T$
- 求线性可分支持向量机



- 目标函数

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

$$=\frac{1}{2}\left(18\alpha_1^2+25\alpha_2^2+2\alpha_3^2+42\alpha_1\alpha_2-12\alpha_1\alpha_3-14\alpha_2\alpha_3\right)-\alpha_1-\alpha_2-\alpha_3$$

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

$$\alpha_i\geq 0,\quad i=1,2,3$$

举例：将约束带入目标函数，化简计算

- 将 $\alpha_1+\alpha_2=\alpha_3$
- 带入目标函数，得到关于 α_1，α_2 的函数：

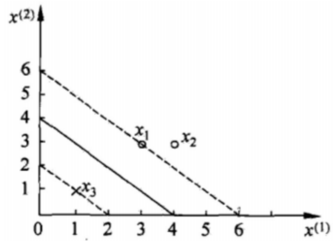
$$s(\alpha_1,\alpha_2)=4\alpha_1^2+\frac{13}{2}\alpha_2^2+10\alpha_1\alpha_2-2\alpha_1-2\alpha_2$$

- 对 α_1，α_2 求偏导并令其为0，易知 $s(\alpha_1,\alpha_2)$ 在点(1.5,-1)处取极值。而该点不满足条件 $\alpha_2\geq 0$，所以，最小值在边界上达到。
- 当 $\alpha_1=0$ 时，最小值 s(0,2/13)=-2/13=-0.1538
- 当 $\alpha_2=0$ 时，最小值 s(1/4,0)=-1/4=-0.25
- 于是，$s(\alpha_1,\alpha_2)$ 在 $\alpha_1=1/4$，$\alpha_2=0$ 时达到最小，此时，$\alpha_3=\alpha_1+\alpha_2=1/4$

举例：分离超平面

- $\alpha_1=\alpha_3=1/4$ 对应的点 x_1,x_3 是支持向量。
- 带入公式：$w^*=\sum_{i=1}^{m}\alpha_i^* y_i x_i$

$$b^*=y_i-\sum_{i=1}^{m}\alpha_i^* y_i x_i x_j$$



- 得到 $w_1=w_2=0.5$，$b=-2$
- 因此，分离超平面为 $\frac{1}{2}x_1+\frac{1}{2}x_2-2=0$
- 分离决策函数为

$$f(x)=sign\left(\frac{1}{2}x_1+\frac{1}{2}x_2-2\right)$$

本页课件致谢：普开数据 PUKAI DATA 邹伟

33

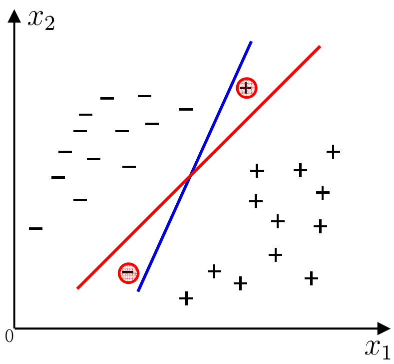
概念

- 线性可分支持向量机
  - 硬间隔最大化hard margin maximization
  - 硬间隔支持向量机
  - 对偶问题
- 线性支持向量机
  - 软间隔最大化soft margin maximization
  - 软间隔支持向量机
  - 正则化
- 非线性支持向量机
  - 核函数kernel function

34

线性支持向量机

- 不一定分类完全正确的超平面就是最好的
- 样本数据本身线性不可分

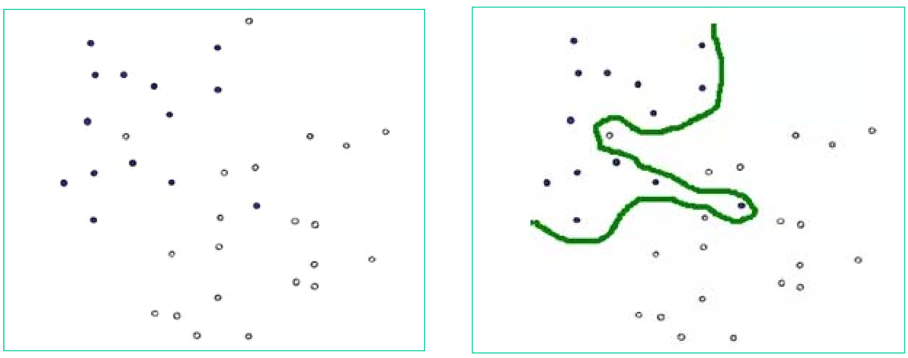


35

线性支持向量机

- 线性不可分

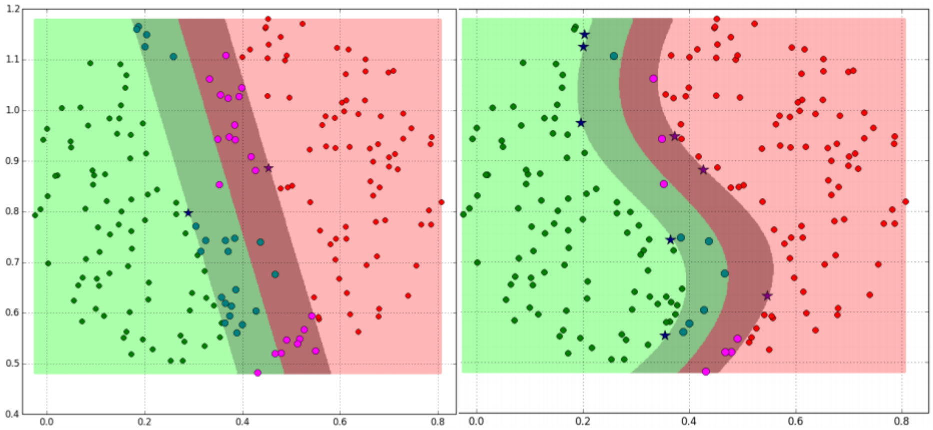
➔但有些情况：用**线性模型**可以进行进行分割



36

线性不可分的办法 – 曲线划分

- 实例



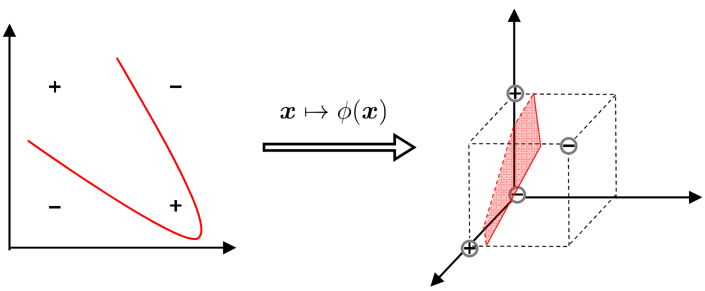
本页课件致谢：普开数据 PUKAI DATA 邹伟

37

线性不可分

Q:若不存在一个能正确划分两类样本的超平面, 怎么办?

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



本页课件来源：周志华《机器学习》及其课件，致谢：张腾

38

关于线性模型的理解

- x_i可以映射为$\phi(\boldsymbol{x}_i)$
  - 表示x_i的任意阶，但w是线性的
- 通用的SVM模型
  - 划分超平面为 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b$

只以内积的形式出现

原始问题
$$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t.} \ \ y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \geq 1, \ i = 1, 2, \ldots, m.$$

对偶问题
$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boxed{\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)} - \sum_{i=1}^{m} \alpha_i$$
$$\text{s.t.} \ \ \sum_{i=1}^{m} \alpha_i y_i = 0, \ \ \alpha_i \geq 0, \ i = 1, 2, \ldots, m.$$

预测
$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i \boxed{\phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x})} + b$$

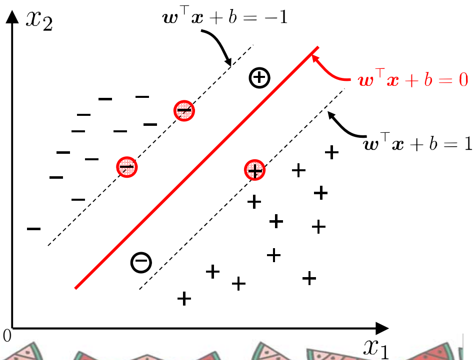
39

线性不可分的办法 – 软间隔

Q:现实中, 很难确定合适的核函数使得训练样本在特征空间中线性可分; 同时一个线性可分的结果也很难断定是否是由于过拟合造成的.

A:引入"软间隔"的概念, 允许支持向量机在一些样本上不满足约束 $y_i(w \cdot x_i + b) \geq 1$

不满足约束的样本



本页课件来源：周志华《机器学习》及其课件，致谢：张腾

40

线性不可分的办法 – 软间隔

- 若数据线性不可分，则增加松弛因子$\xi_i \geq 0$， 使函数间隔加上松弛变量大于等于1。这样， 约束条件变成

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

- 目标函数

以多大的比例去重视松弛

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

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

$$\xi_i \geq 0, \quad i = 1,2,\cdots,m$$

41

带松弛因子的SVM拉格朗日函数

- 拉格朗日函数 $\min_{w,b,\xi} L(w,b,\xi,\alpha,\mu) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i^\top \cdot x_j) + \sum_{i=1}^{m}\alpha_i$

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \qquad \sum_{i=1}^{m}\alpha_i y_i = 0$$

- 带松弛因子的拉格朗日函数

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

- 对w, b, ζ求偏导

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{m}\alpha_i y_i$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \mu_i$$

42

带入目标函数

- 将求导结果带入目标函数，得

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

- 对上式求关于α的极大，得

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

$$\begin{aligned} s.t.\ & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ & C - \alpha_i - \mu_i = 0 \\ & \alpha_i \ge 0 \\ & \mu_i \ge 0,\quad i = 1,2,\ldots,m. \end{aligned} \qquad 0 \le \alpha_i \le C$$

43

最终的目标函数

- 整理得到对偶问题

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

$$\begin{aligned} s.t.\ & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ & 0 \le \alpha_i \boxed{\le C,}\quad i = 1,2,\ldots,m. \end{aligned}$$

跟线性可分支持向量机的区分

- 当C=∞，退化成线性可分支持向量机

44

线性支持向量机学习算法

- 构造并求解约束最优化问题

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

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

$$0 \le \alpha_i \le C, \quad i = 1,2,\ldots,m.$$

- 求得最优解α*

45

线性支持向量机学习算法

- 计算

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

$$b^* = \frac{\max_{i:\ y_i=-1} w^* \cdot x_i + \min_{i:\ y_i=1} w^* \cdot x_i}{2}$$

- 注意：计算b*时，需要使用满足条件$0<\alpha_j<C$的向量
- 实践中往往取支持向量的所有值取平均，作为b*

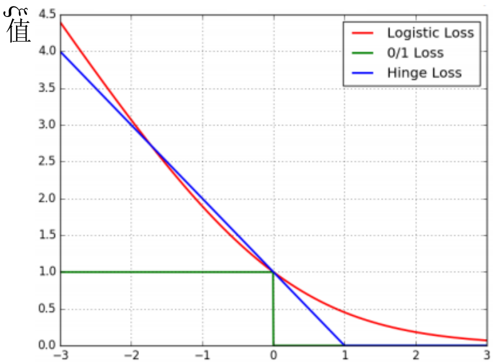
- 求得分离超平面 $\boldsymbol{w}^*\boldsymbol{x} + \mathrm{b} = 0$
- 分类决策函数 $f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^*\boldsymbol{x} + \mathrm{b})$

46

损失函数分析

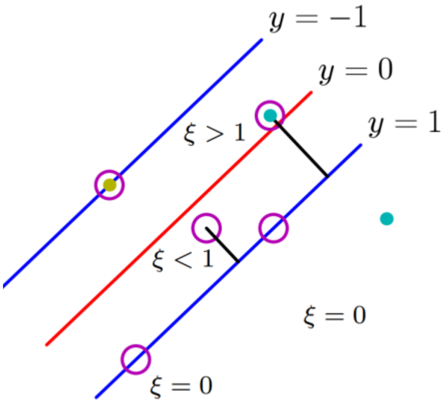
$$y_i(w\cdot x_i+b)\ge 1-\xi_i$$

- 绿色：0/1损失
- 蓝色：SVM Hinge损失函数
- 红色：Logistic损失函数



正例样本离红色线的距离

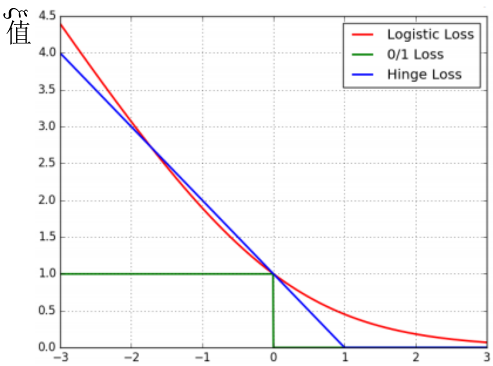
考虑一个正例被分类的情况



损失函数分析

$$y_i(w\cdot x_i+b)\ge 1-\xi_i$$

- 绿色：0/1损失
- 蓝色：SVM Hinge损失函数
- 红色：Logistic损失函数



正例样本离红色线的距离 z

$$L_{0/1}=\begin{cases}1 & z<0\\ 0 & z\geqslant 0\end{cases}$$

$$l_H=\begin{cases}0 & z>1\\ 1-z & z\leqslant 1\end{cases}$$

$$l_H=\max(0,1-z)$$

通用的表示

$l_H=\max(0, \text{margin}-g(x))$

0/1损失函数

- 基本想法
  - 最大化间隔的同时, 让不满足约束的样本应尽可能少.

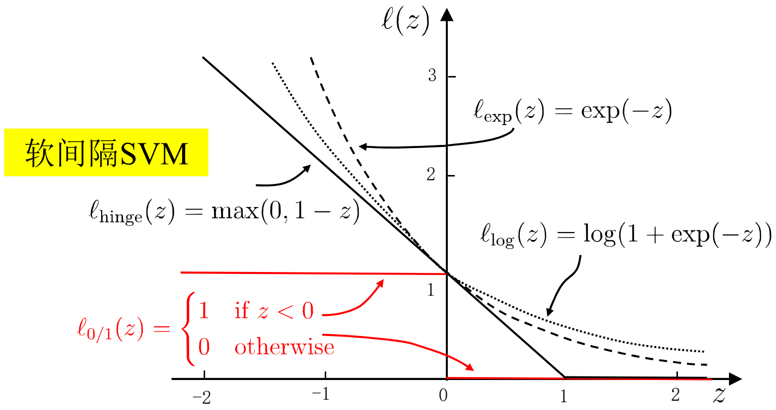
$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} l_{0/1}\left(y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)-1\right)$$

  其中 $l_{0/1}$ 是“0/1损失函数”

$$l_{0/1} = \begin{cases} 1 & z<0 \\ 0 & otherwise \end{cases}$$

- 存在的问题：0/1损失函数非凸、非连续, 不易优化！

替代损失



替代损失函数数学性质较好, 一般是0/1损失函数的上界

重新审视目标函数

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

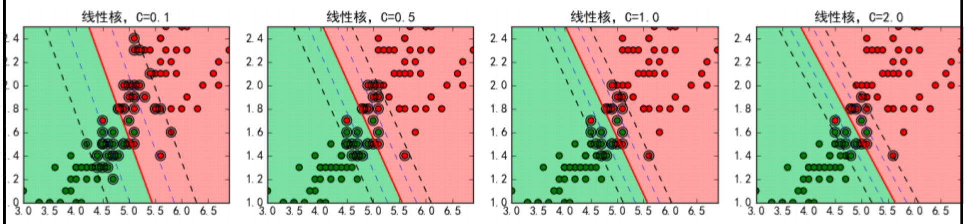
$$s.t.\quad y_i(w\cdot x_i+b)\geq 1-\xi_i,\quad i=1,2,\cdots,m$$

$$\xi_i\geq 0,\quad i=1,2,\cdots,m$$

- SVM自带L2正则项
  - 理论上既能分类，又能防止过拟合
- C越大，越重视损失；而不重视正则
- C越小，越重视模型的泛化能力（实验中带宽越宽）

51

SVM参数举例



- SVM自带L2正则项
  - 理论上既能分类，又能防止过拟合
- C越大，越重视损失；而不重视正则
- C越小，越重视模型的泛化能力（实验中带宽越宽）

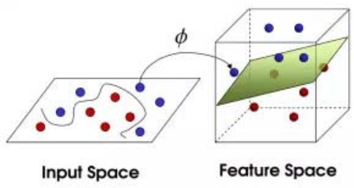
52

概念

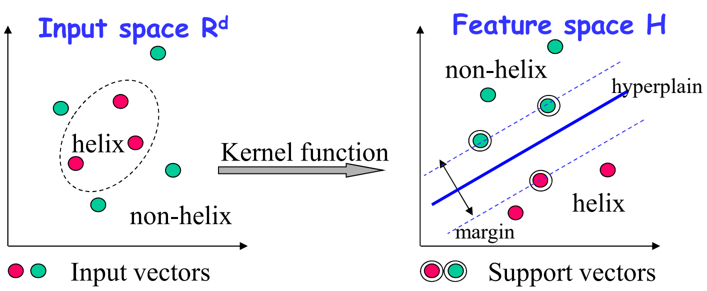
- 线性可分支持向量机
  - 硬间隔最大化hard margin maximization
  - 硬间隔支持向量机
  - 对偶问题
- 线性支持向量机
  - 软间隔最大化soft margin maximization
  - 软间隔支持向量机
  - 正则化
- 非线性支持向量机
  - 核函数kernel function

53

非线性分类



- 特征空间中的支持向量
  - 超平面 → 决策函数



54

核函数

- 线性可分支持向量机

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

$$\text{s.t.} \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \;\; \alpha_i \geq 0, \;\; i = 1, 2, \ldots, m.$$

- 线性支持向量机

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

$$s.t. \quad \sum_{i=1}^{m}\alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1,2,\ldots,m.$$

- 构造核函数
  - 度量样本间相似性 $\boxed{K(x_i, x_j) = \phi(x_i)^T\phi(x_j)}$

55

核函数

- 可以使用核函数，将原始输入空间映射到新的特征空间，从而，使得原本线性不可分的样本可能在核空间可分。
  - 线性核函数： $\kappa(x_1, x_2) = x_1 \cdot x_2$
  - 多项式核函数： $\kappa(x_1, x_2) = (x_1 \cdot x_2 + c)^d$
  - 高斯核RBF函数： $\kappa(x_1, x_2) = \exp\left(-\gamma \cdot \|x_1 - x_2\|^2\right)$
  - Sigmoid核函数： $\kappa(x_1, x_2) = tanh(x_1 \cdot x_2 + c)$
- 在实际应用中，往往依赖先验领域知识/交叉验证等方案才能选择有效的核函数。
  - 没有更多先验信息，则使用高斯核函数

56

核函数

- 基本想法：不显式地设计核映射，而是设计核函数.

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^\top \phi(\boldsymbol{x}_j)$$

- Mercer定理(充分非必要)：只要一个对称函数所对应的核矩阵半正定，则它就能作为核函数来使用.
- 常用核函数：

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

57

理解核函数

- 多项式核函数 $\kappa(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y})^2$

$$\kappa(\vec{x}, \vec{y}) = (\vec{x} \cdot \vec{y})^2$$

$$\Rightarrow \left(\sum_{i=1}^{n} x_i y_i\right)^2 \qquad \Rightarrow \Phi(\vec{x}) = vec(x_i x_j)\Big|_{i,j=1}^{n}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j y_i y_j$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} (x_i x_j)(y_i y_j)$$

特殊的，若n＝3，即：$\Phi(\vec{x}) = \begin{pmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{pmatrix}$

58

理解核函数

- 多项式核函数 $\kappa(\vec{x},\vec{y})=(\vec{x}\cdot\vec{y}+c)^2$

$$\kappa(\vec{x},\vec{y})=(\vec{x}\cdot\vec{y}+c)^2$$

$$\Rightarrow(\vec{x}\cdot\vec{y})^2+2c\vec{x}\cdot\vec{y}+c^2$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i x_j)(y_i y_j)+\sum_{i=1}^{n}\left(\sqrt{2c}x_i\cdot\sqrt{2c}x_j\right)+c^2$$

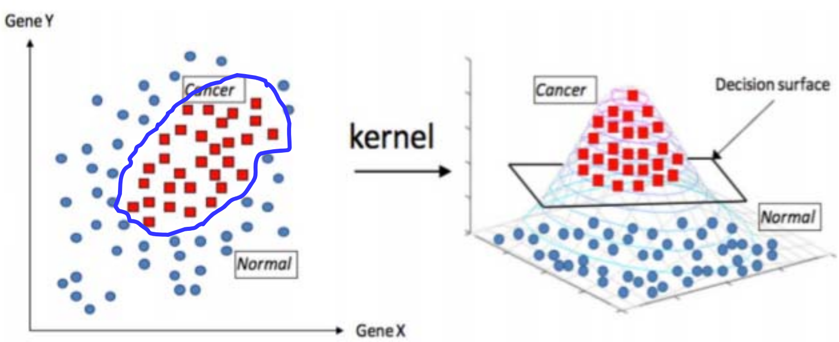
$$\Rightarrow\Phi(\vec{x})=\left(vec(x_i x_j)\Big|_{i,j=1}^{n},vec\left(\sqrt{2c}x_i\right)\Big|_{i=1}^{n},c\right)$$

特殊的，若n=3，即：$\Phi(\vec{x})=\begin{pmatrix} x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \\ \sqrt{2c}x_1 \\ \sqrt{2c}x_2 \\ \sqrt{2c}x_3 \\ c \end{pmatrix}$

59

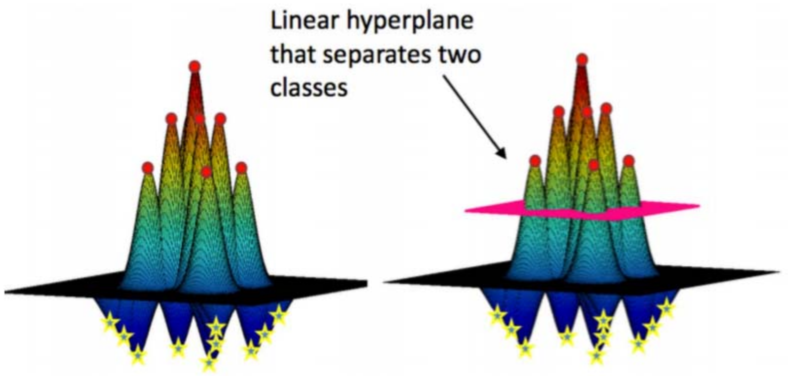
核函数映射

- 高维空间的线性可分，在低维空间是线性不可分的



60

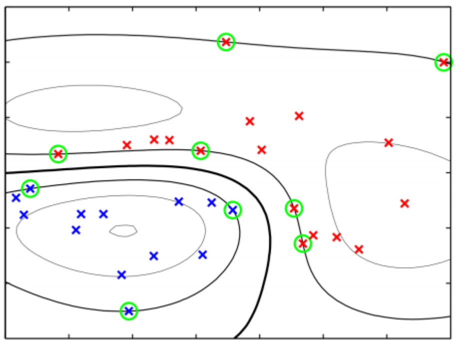
核函数映射



61

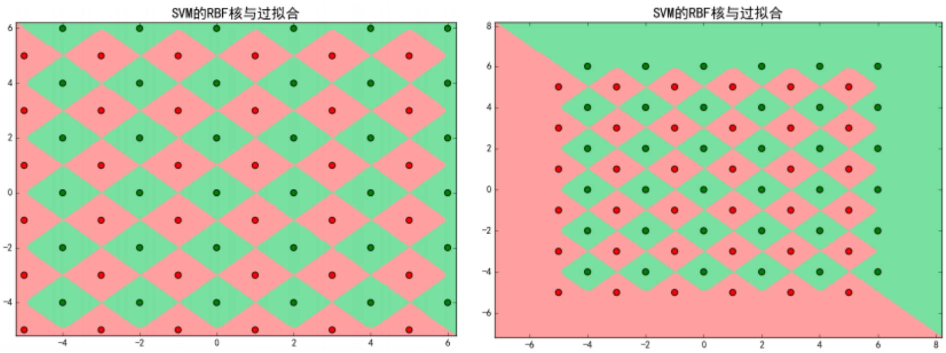
高斯核

- 粗线是分割超“平面”
- 其他线是y(x)的等高线
- 绿色圈点是支持向量点



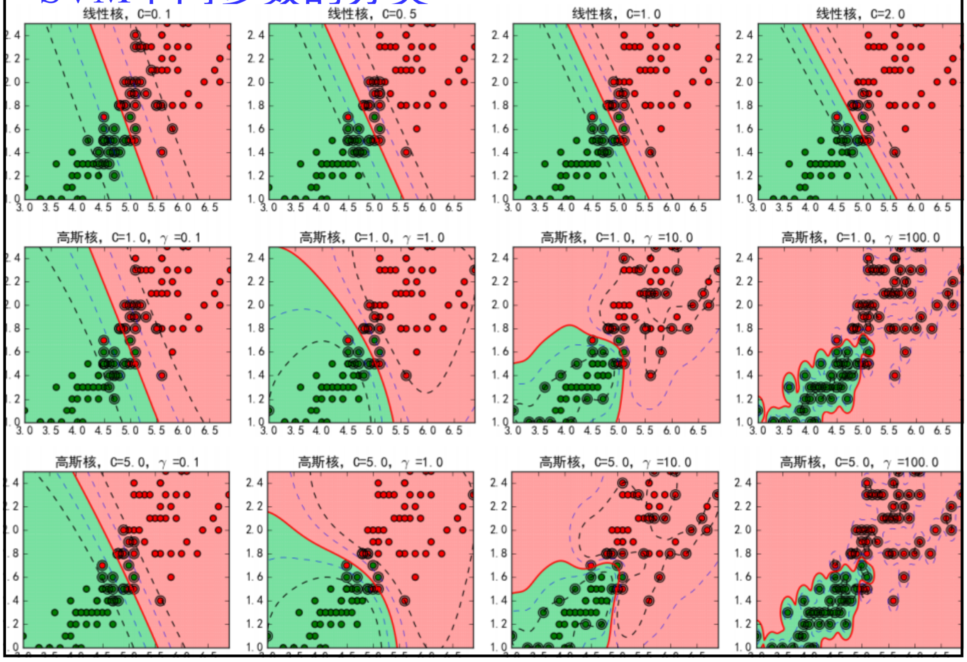
62

高斯核分类



63

SVM不同参数的分类



支持向量回归

- 给定数据集

$$D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$$

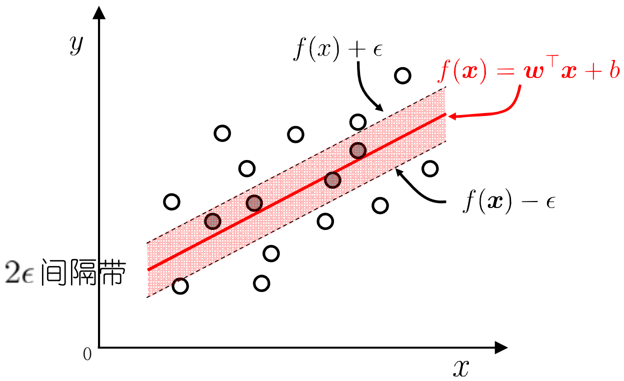
$$\boldsymbol{x}_i=(x_{i1};x_{i2};\ldots;x_{id})\quad y_i\in\mathbb{R}$$

- 支持向量回归目标

$$f(\boldsymbol{x}_i)=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b \text{ 使得 } f(\boldsymbol{x}_i)\simeq y_i$$

支持向量回归（SVR）

特点: 允许模型输出和实际输出间存在 ϵ 的偏差.



本页课件来源：周志华《机器学习》及其课件，致谢：张腾

支持向量回归（SVR）

- SVR的形式化表示

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\underbrace{\sum_{i=1}^{m} \ell_\epsilon(f(\boldsymbol{x}_i) - y_i)}_{\text{损失函数}}$$

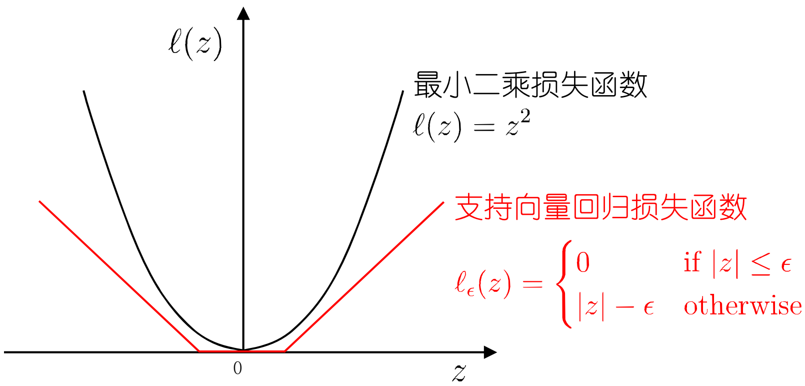
  – ϵ：正则化常数
  – 不敏感损失函数

$$\ell_\epsilon(z) = \begin{cases} 0, & \text{if } |z| \leqslant \epsilon\,; \\ |z| - \epsilon, & \text{otherwise.} \end{cases}$$

67

损失函数

落入中间2ϵ间隔带的样本不计算损失，从而使得模型获得稀疏性.



本页课件来源：周志华《机器学习》及其课件，致谢：张腾

68

支持向量回归（SVR）

- SVR的形式化表示

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\ell_\epsilon(f(\boldsymbol{x}_i)-y_i)$$

  – C：正则化常数
  – ϵ 不敏感损失函数 $\ell_\epsilon(z)=\begin{cases}0, & \text{if } |z|\leqslant\epsilon\text{ ;}\\ |z|-\epsilon, & \text{otherwise.}\end{cases}$

- 引入松弛变量 ξ_i, $\hat{\xi}_i$

$$\min_{\boldsymbol{w},b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)$$

$$\begin{aligned}\text{s.t. } & f(\boldsymbol{x}_i)-y_i\leqslant\epsilon+\xi_i\ ,\\ & y_i-f(\boldsymbol{x}_i)\leqslant\epsilon+\hat{\xi}_i\ ,\\ & \xi_i\geqslant 0,\ \hat{\xi}_i\geqslant 0\ ,\ i=1,2,\ldots,m.\end{aligned}$$

69

支持向量回归（SVR）

- 引入拉格朗日乘子 $\mu_i\geqslant 0,\ \hat{\mu}_i\geqslant 0,\ \alpha_i\geqslant 0,\ \hat{\alpha}_i\geqslant 0,$

得到拉格朗日函数

$$\begin{aligned}& L(\boldsymbol{w},b,\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},\boldsymbol{\xi},\hat{\boldsymbol{\xi}},\boldsymbol{\mu},\hat{\boldsymbol{\mu}})\\ &=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\mu_i\xi_i-\sum_{i=1}^{m}\hat{\mu}_i\hat{\xi}_i\\ &+\sum_{i=1}^{m}\alpha_i\left(f(\boldsymbol{x}_i)-y_i-\epsilon-\xi_i\right)+\sum_{i=1}^{m}\hat{\alpha}_i\left(y_i-f(\boldsymbol{x}_i)-\epsilon-\hat{\xi}_i\right)\end{aligned}$$

代入 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b$ ，另各参数的偏导为零，

得：

$$\begin{aligned}\boldsymbol{w}&=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\boldsymbol{x}_i\ ,\\ 0&=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)\ ,\\ C&=\alpha_i+\mu_i\ ,\\ C&=\hat{\alpha}_i+\hat{\mu}_i\ .\end{aligned}$$

70

支持向量回归（SVR）

- 得到SVR的对偶问题

$$\max_{\alpha,\hat{\alpha}} \sum_{i=1}^{m} y_i(\hat{\alpha}_i-\alpha_i)-\epsilon(\hat{\alpha}_i+\alpha_i) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\hat{\alpha}_i-\alpha_i)(\hat{\alpha}_j-\alpha_j)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$$

$$\text{s.t.} \sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)=0 ,\quad 0\leqslant \alpha_i,\hat{\alpha}_i\leqslant C .$$

- 上述过程需满足KTT条件，即

$$\begin{cases} \alpha_i(f(\boldsymbol{x}_i)-y_i-\epsilon-\xi_i)=0 , \\ \hat{\alpha}_i(y_i-f(\boldsymbol{x}_i)-\epsilon-\hat{\xi}_i)=0 , \\ \alpha_i\hat{\alpha}_i=0 ,\ \xi_i\hat{\xi}_i=0 , \\ (C-\alpha_i)\xi_i=0 ,\ (C-\hat{\alpha}_i)\hat{\xi}_i=0 . \end{cases}$$

则：

$$f(\boldsymbol{x})=\sum_{i=1}^{m}\underline{(\hat{\alpha}_i-\alpha_i)}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}+b$$

使 $(\hat{\alpha}_i-\alpha_i)\neq 0$ 的样本：SVR的支持向量

71

形式化

原始问题

$$\min_{\boldsymbol{w},b,\xi_i,\hat{\xi}_i} \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)$$

对偶问题

$$\min_{\alpha,\hat{\alpha}} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\hat{\alpha}_i)(\alpha_j-\hat{\alpha}_j)\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=1}^{m}(\alpha_i(\epsilon-y_i)+\hat{\alpha}_i(\epsilon+y_i))$$

$$\text{s.t.} \sum_{i=1}^{m}(\alpha_i-\hat{\alpha}_i)=0,$$

$$0\leq\alpha_i\leq C,\ 0\leq\hat{\alpha}_i\leq C.$$

预测 $$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\phi(\boldsymbol{x})+b=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)y_i\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b$$

本页课件来源：周志华《机器学习》及其课件，致谢：张腾

72

表示定理

支持向量机 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} \alpha_i y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b$

支持向量回归 $f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + b = \sum_{i=1}^{m} (\hat{\alpha}_i - \alpha_i) y_i \kappa(\boldsymbol{x}_i, \boldsymbol{x}) + b$

结论: 无论是支持向量机还是支持向量回归, 学得的模型总可以表示成核函数的线性组合.

更一般的结论(表示定理): 对于任意单调增函数Ω和任意非负损失函数l, 优化问题

$$\min_{h \in \mathbb{H}} \quad F(h) = \Omega(\|h\|_{\mathbb{H}}) + l(h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_m))$$

的解总可以写为 $h^* = \sum_{i=1}^{m} \alpha_i \kappa(\cdot, \boldsymbol{x}_i)$

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
- 将支持向量的思想应用到回归问题上得到支持向量回归
- 将核方法推广到其他学习模型

成熟的SVM软件包

- LIBSVM
  http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- LIBLINEAR
  http://www.csie.ntu.edu.tw/~cjlin/liblinear/
- SVM^{light}、SVM^{perf}、SVM^{struct}
  http://svmlight.joachims.org/svm_struct.html
- Pegasos
  http://www.cs.huji.ac.il/~shais/code/index.html