北京交通大学硕士研究生《机器学习》课件

第11章 线性分类器

执其两端，用其中于民。

——《中庸·第六章》

北京交通大学《机器学习》课程组

目录

- 11.1 判别函数和判别模型
- 11.2 线性判别函数
- 11.3 线性感知机算法
  - 11.3.1 感知机数据表示
  - 11.3.2 感知机的归类判据
  - 11.3.3 感知机分类算法
- 11.4 支持向量机
  - 11.4.1 线性可分支持向量机
  - 11.4.2 近似线性可分支持向量机
  - 11.4.3 讨论

- **K近邻法：类认知表示不需要学习，输出类认知表示由输入直接决定。**
  - 输出类认知表示缺少凝练，没有给出输出类的整体描述或者内在本质描述。
- **单类回归分析：输出类的认知表示是一个函数**
  - 假设输入和输出对于对象的特征描述相同，如果特征位于欧氏空间R^p，可以假设输出类认知表示是$R^p \rightarrow R$的一个函数。
- **最简单的函数是线性函数，假设输出类认知表示是线性函数，根据归类公理可以导出线性分类模型。**

11.1 判别函数和判别模型

	输入端	输出端
四元组	$(X, U, \underline{X}, Sim_X)$	$(Y, V, \underline{Y}, Sim_Y)$
对象特性	$X = [x_{\tau k}]_{p \times N}$	$Y = [y_{\tau k}]_{d \times N}$
认知表示	$\underline{X} = \{\underline{X_1}, \underline{X_2}, ..., \underline{X_c}\}$ $\underline{X_i} = (x, f_i(x))$ $f_i(x)$是$R^p \to R$的函数	$\underline{Y} = \{\underline{Y_1}, \underline{Y_2}, ..., \underline{Y_c}\}$ $\underline{Y_i} = (y, F_i(y))$ $F_i(y)$是$R^d \to R$的函数
相似性映射	$Sim_X(x, \underline{X_i}) = \exp(f_i(x))$	$Sim_Y(y, \underline{Y_i}) = \exp(F_i(y))$

假设$Y = X$， 则有$y=x, \forall i, F_i(y) = F_i(x)$;　$\forall k, y_k = x_k$

分类问题可简化为$(X, U, \underline{Y}, Sim_Y)$

判别函数和判别模型

假设$Y = X$，则有$y=x, \forall i, F_i(y) = F_i(x); \quad \forall k, y_k = x_k$

分类问题可简化为$(X, U, \underline{Y}, Sim_Y)$

- (X, U)**为训练输入，** $(\underline{Y}, \mathrm{Sim}_Y)$**为待学习的分类器**

其中$Sim_Y\left(y, \underline{Y_i}\right) = Sim_Y(x, \underline{Y_i}) \quad \underline{Y_i} = (x, F_i(\mathrm{x}))$

根据归类公理：

$$x \in \underline{X_i} \Leftrightarrow \forall j \neq i, Sim_Y(x, \underline{Y_i}) > Sim_Y(x, \underline{Y_j})$$

$$\Leftrightarrow \forall j \neq i, \exp(F_i(x)) > \exp(F_j(x))$$

$$\Leftrightarrow \forall j \neq i, F_i(x) > F_j(x)$$

$F_i(x)$称为第 i 类的判别式函数

$\{F_1(x), F_2(x), ..., F_c(x)\}$称为判别式模型

判别函数和判别模型

$F_i(x)$ 称为**第i类的判别式函数**

$\{F_1(x), F_2(x), ..., F_c(x)\}$称为**判别式模型**

判别式函数值越大，表示该值对应的样本属于该类的概率越大，反之越小。

$\forall i \forall j, F_i(x) - F_j(x) = 0$

称为第i、j类的决策**超平面**。

$\forall i, F_i(x) > 0 \Rightarrow x \in X_i$且$x \notin X_i \Rightarrow F_i(x) \le 0$

称为**正则判别式模型**。

如果一个样本只能归为一个类，必然有

$\forall i \forall j, i \neq j \Rightarrow \{x \mid F_i(x) > 0\} \bigcap \{x \mid F_j(x) > 0\} = \varnothing$

正则判别式模型

目录

- 11.1 判别函数和判别模型
- **11.2 线性判别函数**
- 11.3 线性感知机算法
  - 11.3.1 感知机数据表示
  - 11.3.2 感知机的归类判据
  - 11.3.3 感知机分类算法
- 11.4 支持向量机
  - 11.4.1 线性可分支持向量机
  - 11.4.2 近似线性可分支持向量机
  - 11.4.3 讨论

11.2 线性判别函数

假设判别函数是线性函数：$F_i(x) = w_i^T x + w_{i0}$

此时 $Sim_Y(x, \underline{Y_i}) = \exp(w_i^T x + w_{i0})$

w_i 是一个$p \times 1$的向量，w_{i0} 是一个标量。

训练集要满足凸集分离定理，这样的类表示才有效。

对未知样本分类的判别函数

任意样本x属于类 X_i 的类判别函数为：

$$F_i(x) = \arg\max_{\underline{Y_j}} Sim_Y(x, \underline{Y_j}) = w_i^T x + w_{i0}$$

这里j表示变量，类判别函数被称为线性判别函数，线性分类器。

两分类线性判别分析

为简化类认知表示的线性判别函数，减少含有的参数

如果$F_1(x) = w_1{}^T x+w_{10}, F_2(x)=w_2{}^T x+w_{20}$

则其决策超平面为 $F_1(x) - F_2(x)= (w_1-w_2)^T x+w_{10}-w_{20}=0$
也是线性函数。

令$w' = w_1 - w_2, b' = w_{10} - w_{20}$，

则$(w')^T x+ b' = 0$

如果训练集线性可分，且样本为有限集，**必存在**$\gamma \neq 0 \in R$**使得**

$x \in X_1 \rightarrow (w')^T x+ b' - \gamma > 0$且

$x \in X_2 \rightarrow (w')^T x + b' + \gamma < 0$

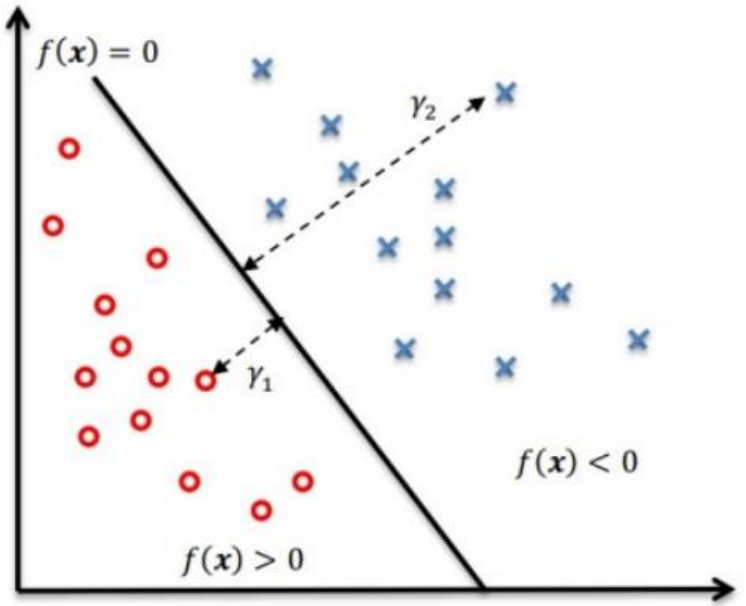


图 11.1　线性超平面二分类示例图

两分类线性判别分析

选择$F_1(x) = w^T x + b - 1, F_2(x) = -w^T x - b - 1$

此时分类器参数个数最少，是最简单的线性分类器

如果训练集线性可分，则:

$\forall x_k \in X_1$满足$w^T x + b - 1 \geq 0$　　$\forall x_k \in X_2$满足$-w^T x - b - 1 \geq 0$

$$2f(x) = F_1(x) - F_2(x)$$
$$= (w^T x + b - 1) - (-w^T x - b - 1)$$
$$= 2w^T x + 2b = 2(w^T x + b)$$

决策超平面为 $f(x) = w^T x + b = 0$

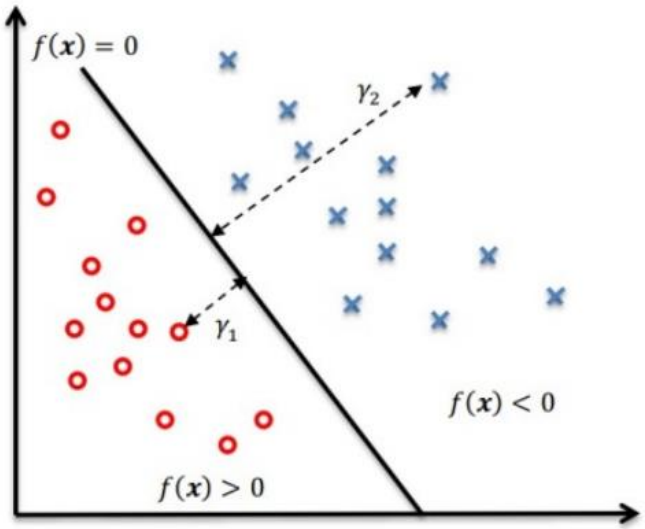


图 11.1　线性超平面二分类示例图

如果训练集线性可分，则:

$\forall x_k \in X_1$满足$w^T x + b > 0$

$\forall x_k \in X_2$满足$-w^T x - b > 0$

多类线性判别分析

假设共含有c个类，并且每个类都有正则线性判别式将该类与其他类别正确划分。

$$F_i(x) = w_i^T x + w_{i0}, \quad \forall i \in \{1, 2, \dots, c\}$$

$$\begin{cases} F_i(x) > 0 \Rightarrow x \in X_i \\ F_i(x) \leq 0 \Leftarrow x \notin X_i \end{cases}$$

理想情况：

对于每个类X_i都存在一个超平面H_i，使得所有

$x \in X_i$都在该超平面的正侧

$x \in X_j, i \neq j$都在该超平面的负侧

现实情况：

✓ 超平面正侧重叠

✓ **所有的判别式都小于零**

→ 样本指派到判别式值最大的类

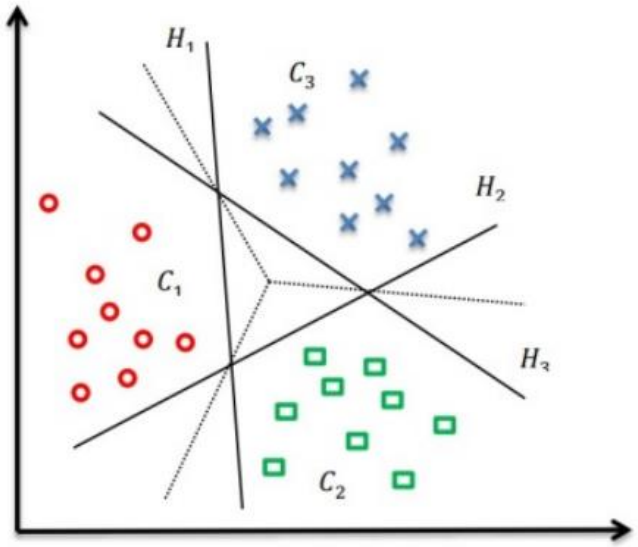


图 11.2　3类线性可分示意图

线性不可分情况

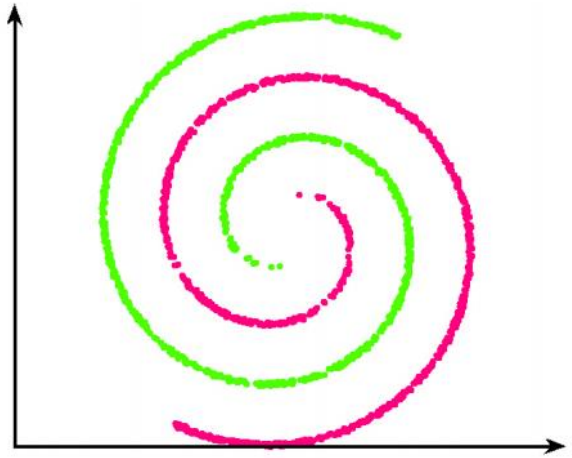


图 11.3　两类线性不可分情况

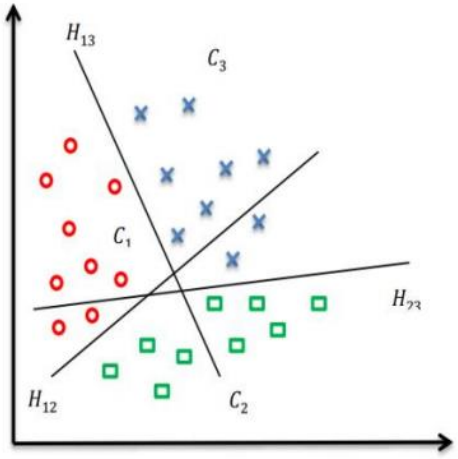


图 11.4　样例在不同分离平面下所给定的类标不同

解决线性不可分的一般方法是寻找新的特征空间，使该问题在新特征空间里转化为线性可分问题。

目录

- 11.1 判别函数和判别模型
- 11.2 线性判别函数
- **11.3 线性感知机算法**
  - 11.3.1 感知机数据表示
  - 11.3.2 感知机的归类判据
  - 11.3.3 感知机分类算法
- 11.4 支持向量机
  - 11.4.1 线性可分支持向量机
  - 11.4.2 近似线性可分支持向量机
  - 11.4.3 讨论

1.3 线性感知机算法

如何通过训练数据将每类的类表示(即线性函数)学习出来?

对于二分类问题，学习到决策超平面$w^T x+ b=0$ 就可以了。

此时两类的输出类表示可以设定为

$$\underline{Y_1} = (x, w^T x + b)$$
$$\underline{Y_2} = (x, -w^T x - b)$$

1957年Rosenblatt提出了线性感知机算法，就是基于这种思路。

对于二分类问题， 类标集合一般设定为 $I = \{-1，1\}$

问题描述

线性感知机是一个典型的二分类算法，

输入样本集$X = \{x_1, x_2, \cdots, x_N\}$

类标集合$U = \{u_1, u_2, \cdots, u_N\}$

$\forall k, x_k \in R^p, u_k \in \{-1,\ \ 1\}$。

旨在求出能够将训练数据进行线性划分的分类超平面
$w^T x + b = 0$,该超平面将实例划分为正负两类实例。

1.3.1 感知机数据表示

类相似性映射

$$Sim_Y(x,\underline{Y_1}) = \exp(w^T x + b)$$

$$Sim_Y(x,\underline{Y_2}) = \exp(-w^T x - b)$$

类预测函数

$$\mathrm{h}(x) = \mathrm{sign}(w^{\mathrm{T}} x + b)$$

w和b为参数，$w \in R^p$为权值向量，$b \in R$为偏置。

符号函数

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

假设前提是样本空间线性可分，学习的关键是根据样本学习得到分离超平面的参数w和b。

11.3.2 感知机算法的归类判据

类表示唯一公理对分类问题一般不再成立，只能尽可能成立。

类一致性准则要求误分类实例尽可能少。

类内紧致准则：

- 对同一个样本，其输入的类标与输出的类标如果一致，则类内相异度为零。
- 如果不一致，则错分样本到超平面的距离越近越好，离决策超平面越远表明该错误率越大。

感知机算法的归类判据

综合考虑**类一致性**和**类内紧致性，类内相异度定义为**

$$\blacksquare |\overrightarrow{x_k} - \widetilde{y_k}| = \left| \left(min(0, \frac{u_k(w^T x_k + b)}{\|w\|} \right) \right|$$

样本到超平面的距离记作

■点到直线的距离公式

$$\frac{(w^T x_k + b)}{\|w\|}$$

误分类的样本(x_k, u_k)满足
$-u_k(w^T x_k + b) > 0$,
而正确分类样本(x_k, u_k)满足
$u_k(w^T x_k + b) > 0$。

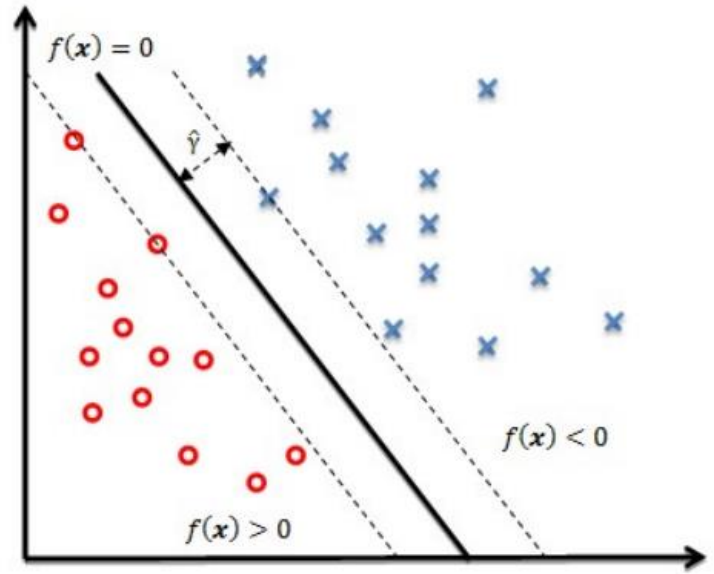


图 11.5　两类对象判决超平面与分离超平面

感知机算法的归类判据

错误分类样本到超平面的距离记作

$$-\frac{u_k(w^T x_k + b)}{\|w\|}$$

■点到直线的距离

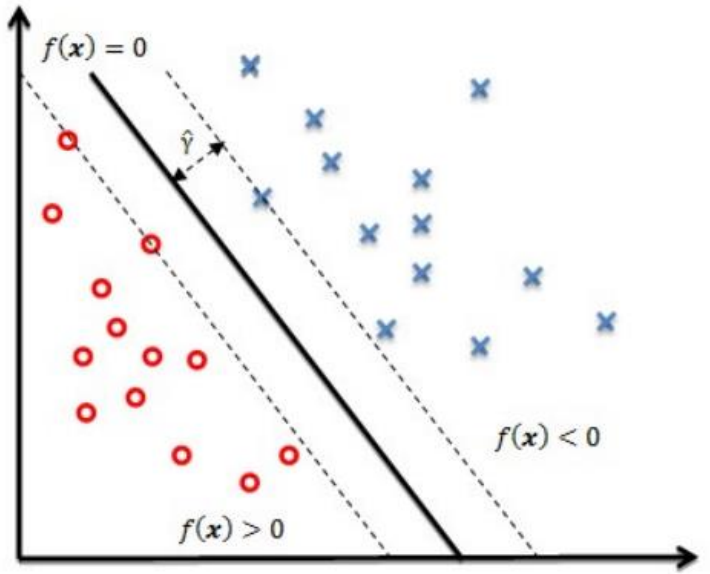


图 11.5　两类对象判决超平面与分离超平面

感知机算法的归类判据

■ $$\left|\vec{X}-\widetilde{Y}\right| = \sum_{k=1}^{N} |\overrightarrow{x_k} - \widetilde{y_k}| = \sum_{k=1}^{N} \left| \min\left(0, \frac{u_k(w^T x_k + b)}{\|w\|}\right) \right| \quad (11.7)$$

感知机算法的归类判据

任意正分类样本x_k使得$min(0, u_k(w^T x_k + b)) = 0$

公式11.7可以化简为错误分类样本到超平面的总距离

$$-\frac{1}{\|w\|}\sum_{x_k \in M} u_k(w^T x_k + b)$$

M为误分样本的集合。最小化该函数就可以学习决策超平面。

等价于最小化下面的目标函数：

$$L(w,b) = -\sum_{x_k \in M} u_k(w^T x_k + b) \qquad (11.9)$$

误分类样本越少，误分类样本离超平面越近，函数值越小。

11.3.3 感知机分类算法:原始形式

利用随机梯度下降法求解

目标函数 $$L(w, b) = -\sum_{x_k \in M} u_k (w^T x_k + b) \qquad (11.9)$$

$$\nabla_w L(w, b) = -\sum_{x_k \in M} u_k x_k$$

$$\nabla_b L(w, b) = -\sum_{x_k \in M} u_k$$

感知机分类算法:原始形式

随机选取一个误分类样本(x_k, u_k)，对w, b进行更新：

$$w \leftarrow w + \eta u_k x_k$$

$$b \leftarrow b + \eta u_k$$

η为步长，或学习率，是超参数。通过迭代使归类判据最小化。

$$L(w, b) = -\sum_{x_k \in M} u_k (w^T x_k + b)$$

算法11.1 感知机学习算法的原始形式

输入：训练数据集$X=(x_1,u_1),(x_2,u_2),\dots,(x_N,u_N)$，其中$x_k\in R^p,u_k\in\{-1,1\},k=1,2,\dots,N;0<\eta\leq 1$

输出：w,b

(1)选取初值w_0,b_0；

(2)在训练集中选取数据(x_k,u_k)；

(3)如果$u_k\left(w^Tx_k+b\right)\leq 0$，

$$w\leftarrow w+\eta x_k u_k$$

$$b\leftarrow b+\eta u_k$$

(4)转至(2),直至训练集中没有误分类样本。

当一个样本被误分类时，则调整w,b的值，使分离超平面向该误分类样本的一侧移动，以减少该误分类样本与超平面间的距离，直至超平面越过该误分类样本使其被正确分类。

Novikoff于1962年证明了感知机算法的收敛性。

感知机分类算法:对偶形式

误分类样本(x_k, u_k)，对w, b进行更新：

$$w \leftarrow w + \eta u_k x_k$$

$$b \leftarrow b + \eta u_k$$

逐步修改w, b，第k个样本修改n_k次，则w, b的增量分别是$\alpha_k u_k x_k$和$\alpha_k u_k$，$\alpha_k = n_k \eta$。最后学习到的w, b表示为：

$$w = \sum_{k=1}^{N} \alpha_k u_k x_k$$

$$b = \sum_{k=1}^{N} \alpha_k u_k$$

当学习率$\eta = 1$，$\alpha_k \geq 0$表示第k个样本由于误分更新的次数。

更新次数越多，离超平面越近，越难正确分类，对分类结果影响最大。只要学到样本组合系数α_k即可，得到对偶形式的算法。

算法11.2 感知机学习算法对偶形式

输入：训练数据集$X=(x_1,u_1),(x_2,u_2),\dots,(x_N,u_N)$，**其中**$x_k\in R^p, u_k\in\{-1,1\}, k=1,2,\dots,N; 0<\eta\leq 1$

输出：α, b，线性类预测函数

$$h(x)=sign\left(\sum_{k=1}^{N}\alpha_k u_k x_k{}^T x+b\right)$$；其中$\alpha\in R^N$

(1)$\alpha\leftarrow 0,\ b\leftarrow 0$

(2)在训练集中选取数据(x_k,u_k)

(3)如果$u_k\left(\sum_{l=1}^{N}\alpha_l u_l x_l{}^T x_k+b\right)\leq 0$，

$$\alpha_k\leftarrow\alpha_k+\eta$$

$$b\leftarrow b+\eta u_k$$

(4)转至(2),直至训练集中没有误分类样本。

对偶形式基于两两样本的相似性

感知机学习算法的对偶形式中的样本信息是以两两样本的内积形式出现的，其样本的原始特征已经消失。

两个样本的内积在一定意义上表示了两个样本之间的相似性。

样本相似性在模式识别中有重要作用，可以不知道样本的原始特征，基于样本的相似性设计算法。

目录

- 11.1 判别函数和判别模型
- 11.2 线性判别函数
- 11.3 线性感知机算法
  - 11.3.1 感知机数据表示
  - 11.3.2 感知机的归类判据
  - 11.3.3 感知机分类算法
- **11.4 支持向量机**
  - 11.4.1 线性可分支持向量机
  - 11.4.2 近似线性可分支持向量机
  - 11.4.3 讨论

11.4 支持向量机

当算法分类正确时，类分离准则要求最优的类表示应该具有最大间距。

两类线性分类器算法的输出类认知表示为

$$\underline{Y_1} = (x, w^T x + b - 1)$$
$$\underline{Y_2} = (x, -w^T x - b - 1)$$

这两个类表示为平行线，根据类分离准则，平行线间的距离最大，

可以引出支持向量机（Support vector machine,SVM）模型。

Vapnik于1995年提出，用来处理二分类问题。

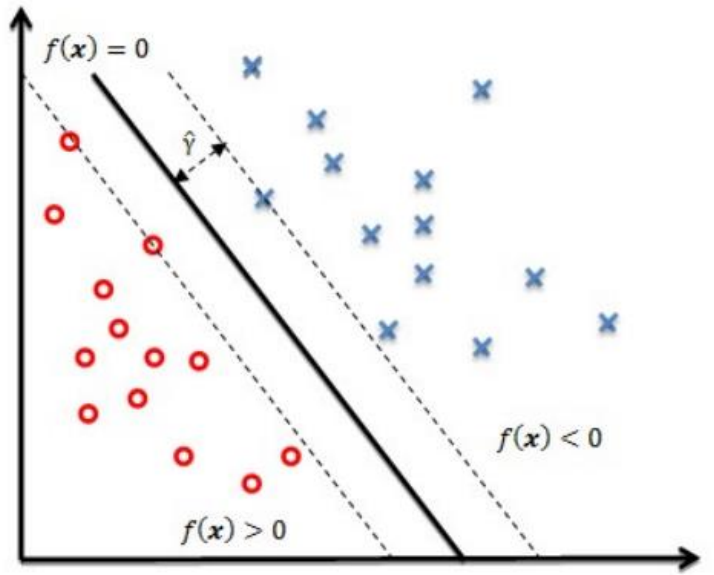


图 11.5　两类对象判决超平面与分离超平面

11.4.1 线性可分支持向量机：归类表示

输出类认知表示

$$\underline{Y_1} = (x, w^T x + b - 1)$$

$$\underline{Y_2} = (x, -w^T x - b - 1)$$

类判别函数为

$$F_1(x) = w^T x + b - 1,$$

$$F_2(x) = -w^T x - b - 1$$

如果训练集线性可分，则:

$x_k \in X_1$，则$w^T x + b - 1 \geq 0$

$x_k \in X_2$，则$-w^T x - b - 1 \geq 0$

类相似性映射

$$Sim_Y(x, \underline{Y_1}) = \exp(w^T x + b - 1)$$

$$Sim_Y(x, \underline{Y_2}) = \exp(-w^T x - b - 1)$$

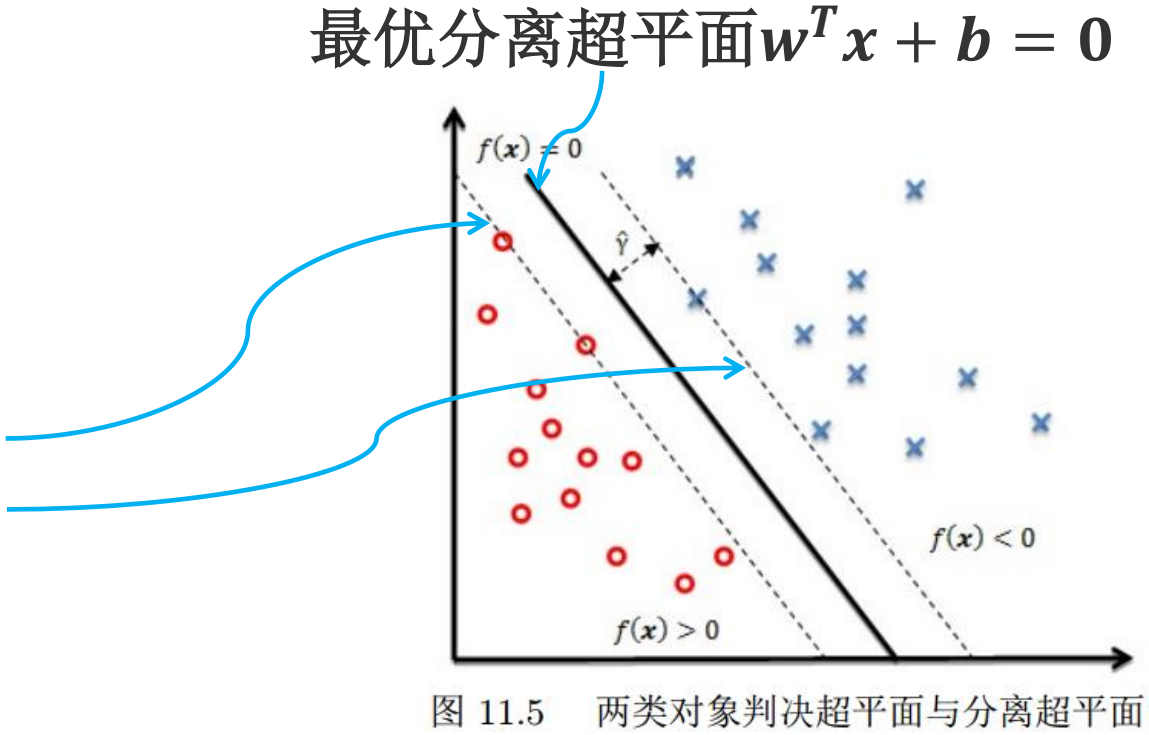


图 11.5　两类对象判决超平面与分离超平面

归类判据

如何得到最优的两个类判别超平面？

根据类分离准则，可以通过最大化两类函数对应的判别超平面的距离得到。

$w^T x + b - 1 = 0$

$-w^T x - b - 1 = 0$

最优分离超平面$w^T x + b = 0$

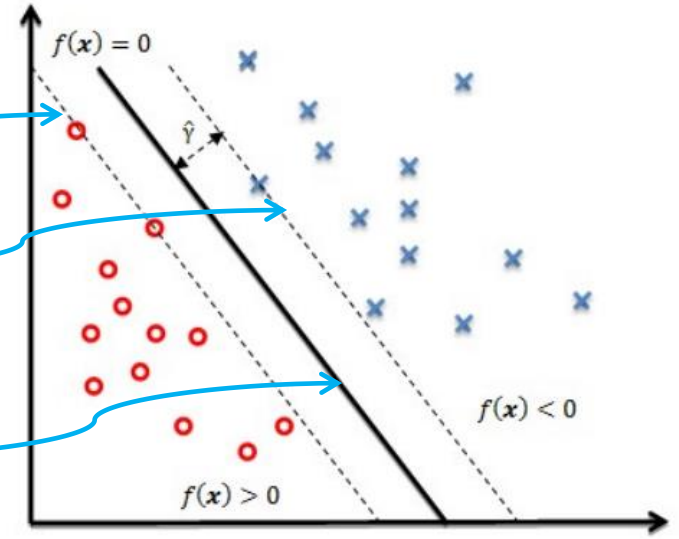


图 11.5　两类对象判决超平面与分离超平面

两个判别超平面平行，没有实例落在两个平面之间，分离超平面平行于这两个平面且位于两者中央。

归类判据

计算两类判别超平面之间的距离

$$\frac{w^T x+ b \ + \ 1 - (w^T x+ b \ - \ 1)}{\|w\|} = \frac{2}{\|w\|}$$

两个平面之间的距离成为间隔，依赖于分离超平面的法向量w, 等于$\frac{2}{\|w\|}$。

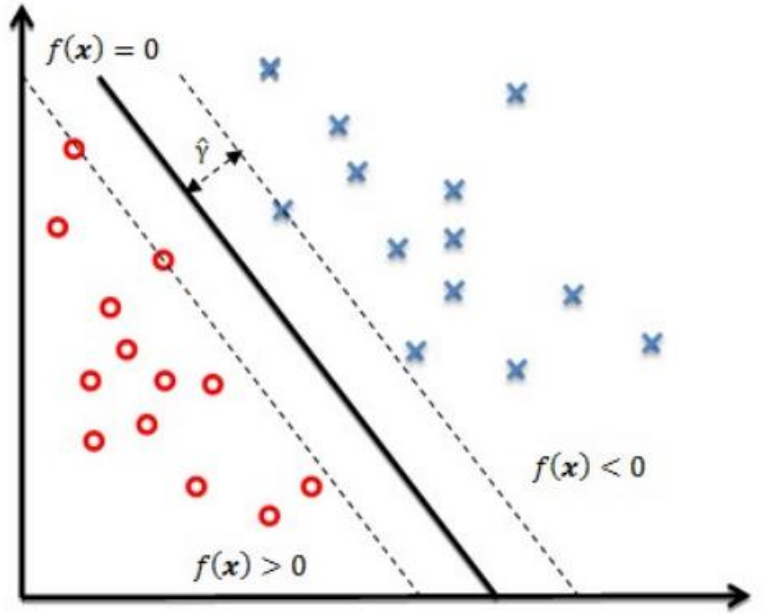


图 11.5　两类对象判决超平面与分离超平面

线性可分SVM的目标函数为

$$\max_{w,b} \frac{1}{\|w\|}$$
$$\text{s.t. } u_k(w^T x_k+ b) - 1 \geq 0, \ k = 1, 2, \dots, N \qquad (11.14)$$

线性可分支持向量机分类算法

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{等价} \quad \max_{w,b} \frac{1}{\|w\|}$$

$$\text{s.t. } u_k(w^T x_k + b) - 1 \geq 0,\ k = 1, 2, \ldots, N \quad (11.14)$$

凸二次规划问题，能利用现成软件包求解，还有更高效的办法，可以转化为对偶问题求解。

拉格朗日乘子法

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{k=1}^{N} \alpha_k \left[u_k\left(w^T x_k + b\right) - 1\right]$$

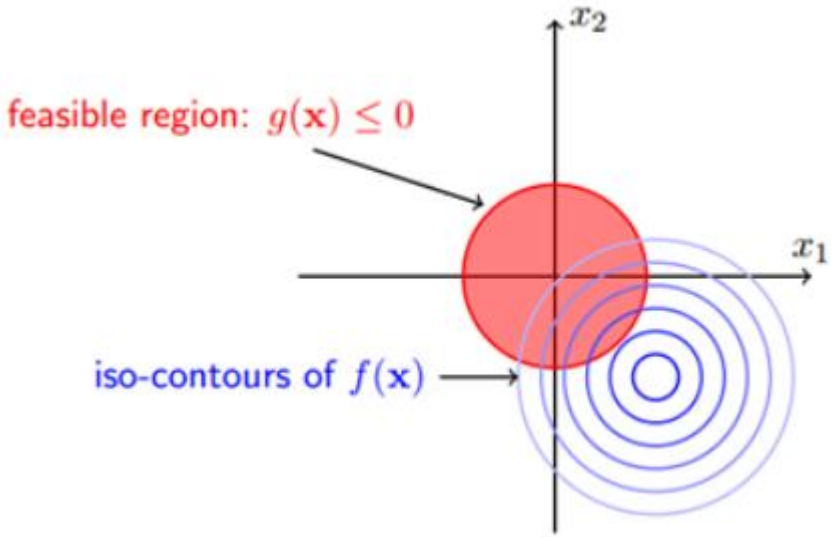
$$= \frac{1}{2}\|w\|^2 - \sum_{k=1}^{N} \alpha_k u_k\left(w^T x_k + b\right) + \sum_{k=1}^{N} \alpha_k$$

■ $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$ 为拉格朗日乘子向量 (11.16)

补充知识：约束优化方法

不等式约束优化

$$\min_{x} \ f(x)$$
$$s.t. \ \ g(x) \leq 0$$



由图可见可行解 x 只能在 $g(x) < 0$ 或者 $g(x) = 0$ 的区域里取得：

- 当可行解 x 落在 $g(x) < 0$ 的区域内，此时直接极小化 $f(x)$ 即可；
- 当可行解 x 落在 $g(x) = 0$ 即边界上，此时等价于等式约束优化问题.

补充知识：约束优化方法

对于不等式约束，只要满足一定的条件，依然可以使用拉格朗日乘子法解决，这里的条件便是 KKT 条件。

$$L(x, \alpha, \beta) = f(x) + \sum_{j=1}^{n} \beta_j g_j(x)$$

加上不等式约束后可行解 x 需要满足的就是以下的 KKT 条件：

$$\nabla_x L(x, \alpha, \beta) = 0$$ 拉格朗日乘子法取得可行解的必要条件

$$\beta_j g_j(x) = 0, \ j = 1, 2, \ldots, n$$ 松弛补充条件

$$g_j(x) \leq 0, \ j = 1, 2, \ldots, n$$ 初始约束条件

$$\beta_j \geq 0, \ j = 1, 2, \ldots, n$$ 拉格朗日乘子须满足的条件

线性可分支持向量机:分类算法

偏导数为0取得极值

$$w = \sum_{k=1}^{N} u_k \alpha_k x_k \qquad \sum_{k=1}^{N} u_k \alpha_k = 0$$

代入

对偶形式

- $\max_{\alpha} L(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} u_k u_l \alpha_k \alpha_l x_k^T x_l$

(11.18)

- s.t. $\sum_{k=1}^{N} u_k \alpha_k = 0, \alpha_k \geq 0; k = 1,2,\ldots,N$

算法11.3 线性可分SVM对偶问题学习算法

输入：线性可分训练数据集$X = \{(x_1, u_1), (x_2, u_2), \dots, (x_N, u_N)\}$, 其中$x_k \in R^p, u_k \in \{-1, +1\}$

输出：w, b和**类判别函数**

步骤：

(1) 基于类分离性准则构造如公式(11.18)的约束最优化问题，求得最优解$\alpha^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_N^*)^T$:

计算$w = \sum_{k=1}^{N} \alpha_k u_k x_k$,**并选择**$\alpha^*$**的一个正分量**$\alpha_l{}^* > 0$;

计算$b^* = u_l - \sum_{k=1}^{N} \alpha_k{}^* u_k (x_k \cdot x_l)$;

(2)求得两类的判别函数$F_1(x)$和$F_2(x)$

$F_1(x) = w^T x + b - 1 = \sum_{k=1}^{N} \alpha_k u_k x_k{}^T x + b - 1$

线性可分支持向量机:支持向量

KKT条件要求最优解满足：

$$\alpha_k^* \left[u_k \left((w^*)^T x_k + b^* \right) - 1 \right] = 0, k = 1, 2, \ldots, N$$

由此看出，如果样本位于类判别超平面，则α_k^*非零，否则α_k^*为零。

最终权重向量表达式只包含解为非零的位于类判别超平面的样本，这些样本被称作**支持向量**。

在决定分类超平面时，**只有支持向量起作用**，其他样本不起作用，所以称为**支持向量机分类模型**。

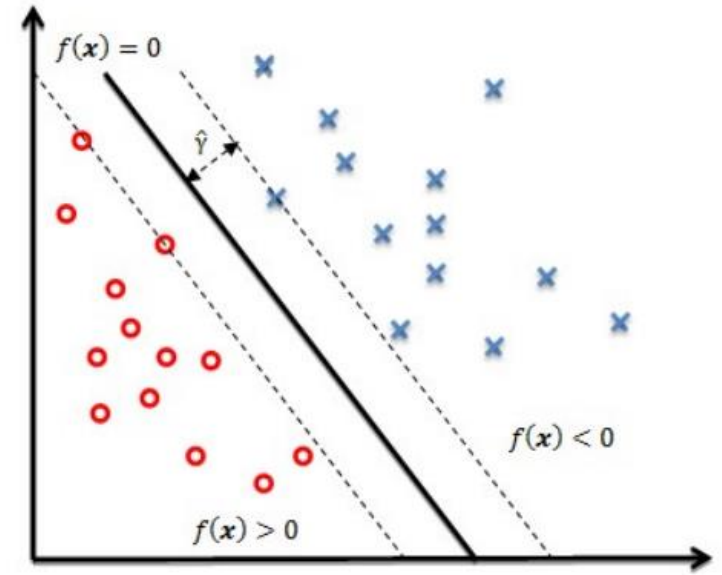


图 11.5　两类对象判决超平面与分离超平面

SVM改进训练算法

- SVM的算法特点
  - 不依赖于设计者的经验知识;
  - 能求全局最优值;
  - 有良好的泛化能力而不会出现过学习。
- 二值规划问题，可使用通用二次规划算法求解
- SVM算法复杂导致训练速度较慢，其中的主要原因是在算法寻优过程中涉及大量矩阵运算。目前提出的一些改进训练算法是基于循环迭代的思想，有3类改进算法：
  - Vapnik等提出的块算法
  - Qsuna等提出的分解算法
  - **Platt的SMO算法（应用最广）**

Platt J . Fast Training of Support Vector Machines using Sequential Minimal Optimization.[C]// MIT Press, 2000.

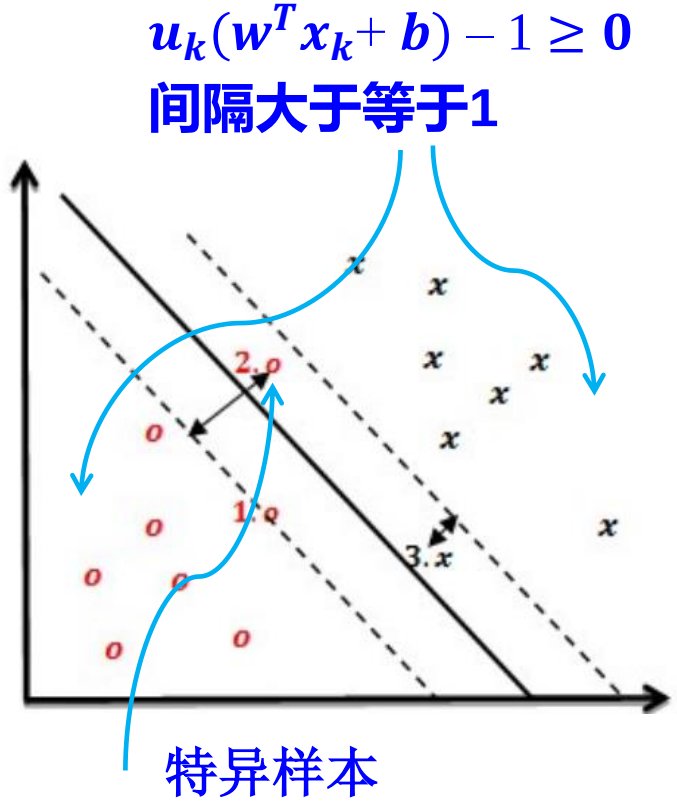
11.4.2 近似线性可分支持向量机:问题表示

存在特异性样本，不满足线性可分条件。近似线性可分是指特异性样本到判别超平面的函数间隔小于1，为此定义松弛变量$\xi_k \geq 0$表示间隔离差，使函数间隔加上ξ_k大于等于1。

定义松弛变量

$\xi_k \geq 0$表示间隔离差，又称软间隔

- (1) $\xi_k = 0$，样本不特异
- (2) $\xi_k \geq 1$，样本分类错误
- (3) $0 < \xi_k < 1$，样本分类正确



软间隔含义

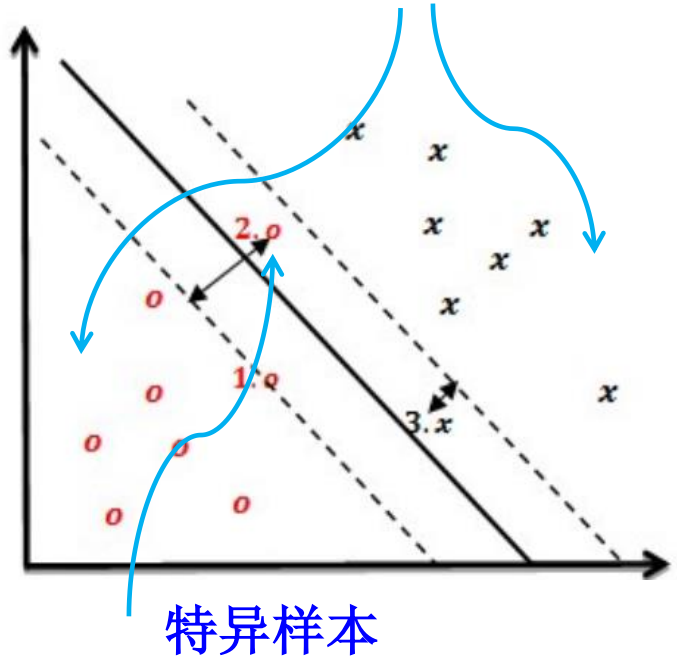
- (1) $\xi_k = 0$，分类正确，离超平面足够远；
- (2) $\xi_k = 1 + f(x)$，分类错误；
- (3) $\xi_k = 1 - f(x)$，样本在正确一侧，但在边缘内，离超平面不够远。

适当放宽约束条件：

- s.t. $u_k(w^T x_k + b) \geq 1 - \xi_k,\ k = 1,2,\ldots,N$
- $\xi_k \geq 0,\ k = 1,2,\ldots,N$

$u_k(w^T x_k + b) - 1 \geq 0$

间隔大于等于1



分类判据

软误差$\sum_{k=1}^{N}\xi_k$表示不能用规定边缘分开的程度，反映了类内紧致度。
$\sum_{k=1}^{N}\xi_k$越小越好，当$\sum_{k=1}^{N}\xi_k=0$，紧致性最好，完全线性可分。

■类间分离　■惩罚因子　■类内紧致

目标函数

$$\min_{\omega,b,\xi}\frac{1}{2}\|w\|^2+C\sum_{k=1}^{N}\xi_k$$

$$\text{s.t. } u_k(w^T x_k+b)\geq 1-\xi_k$$

$$\xi_k\geq 0,\ k=1,2,\ldots,N$$

惩罚因子C的大小代表对错误分类的惩罚力度，C值大则惩罚力度大，误分类的样本少些，反之则小。

对偶最优化问题

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2+C\sum_{k=1}^{N}\xi_k-\sum_{k=1}^{N}\alpha_k\left[u_k\left(w^T x_k+b\right)-1+\xi_k\right]$$

$$-\sum_{k=1}^{N}u_k\xi_k$$

$$\frac{\partial L(w,b,\alpha)}{\partial(w)}=w-\sum_{k=1}^{N}u_k\,\alpha_k x_k$$

$$\frac{\partial L(w,b,\alpha)}{\partial(b)}=-\sum_{k=1}^{N}u_k\,\alpha_k$$

$$\frac{\partial L(w,b,\alpha)}{\partial(\xi_k)}=C-\alpha_k-\mu_k$$

对偶最优化问题

$$w = \sum_{k=1}^{N} u_k \alpha_k x_k$$

$$\sum_{k=1}^{N} u_k \alpha_k = 0$$

$$C - \alpha_k - \mu_k = 0 \quad (11.26)$$

- $\max_{\alpha} W(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} u_k u_l \alpha_k \alpha_l x_k^T x_l$
- s.t. $\sum_{k=1}^{N} u_k \alpha_k = 0$, $0 \leq \alpha_k \leq C; k = 1,2,\ldots,N$

对偶最优化问题

■$\max_{\alpha} W(\alpha) = \sum_{k=1}^{N} \alpha_k - \frac{1}{2} \sum_{k=1}^{N} \sum_{l=1}^{N} u_k u_l \alpha_k \alpha_l x_k^T x_l$

■s.t. $\sum_{k=1}^{N} u_k \alpha_k = 0,\ 0 \le \alpha_k \le C;\ k = 1,2,\ldots,N$ (11.28)

$$w^* = \sum_{k=1}^{N} u_k \alpha_k^* x_k$$

■$b^* = u_l - \sum_{k=1}^{N} \alpha_k^* u_k (x_k \cdot x_l)$

对应于$\alpha_k^* > 0$的样例为软边缘支持向量，距离类判别超平面(间隔边界)有间隔离差，$\xi_k/\|w\|$表示样例x_k到判别超平面的距离。

有3类软边缘支持向量：

1）在判别平面上

2）在判别超平面和分离超平面之间

3）在分离超平面误分一侧

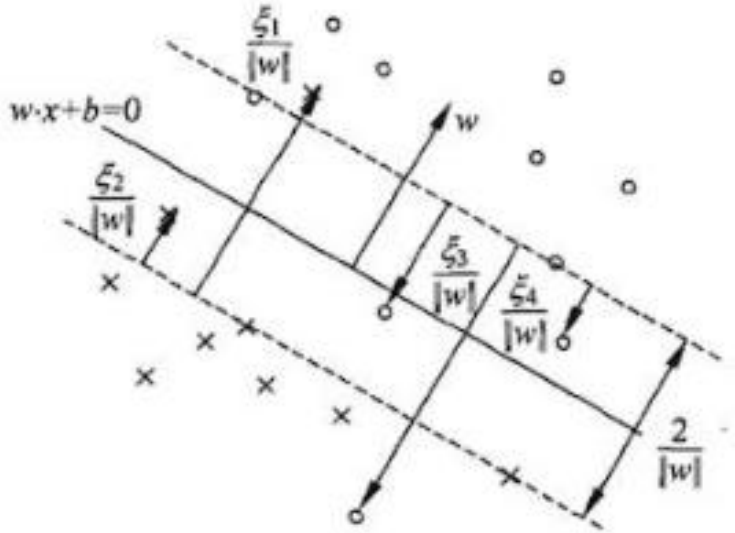
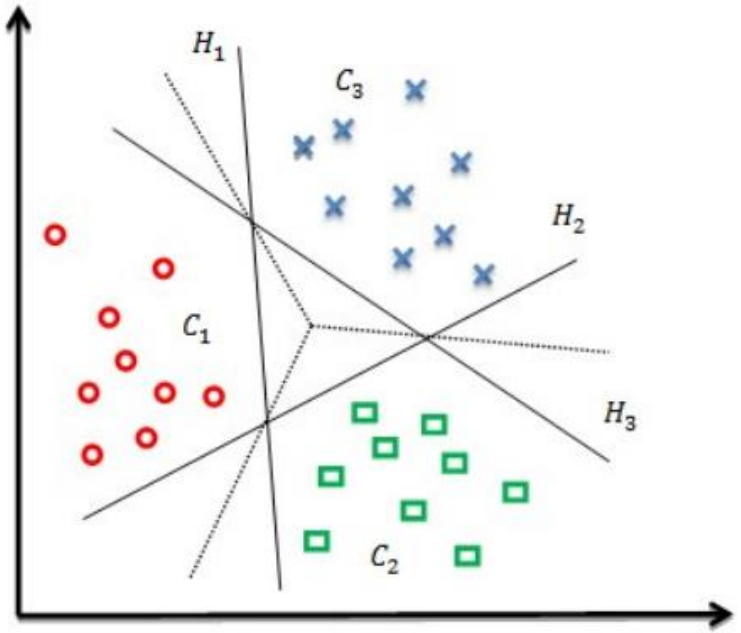


图 7.5 软间隔的支持向量

11.4.3 多类分类问题：一对多分类

每个类都有判别式可以将该类与其它类正确划分

$$
\begin{aligned}
F_1(x|w_1,b_1) &= w_1^T x + b_1 \\
F_2(x|w_2,b_2) &= w_2^T x + b_2 \\
&\vdots \\
F_c(x|w_c,b_c) &= w_c^T x + b_c
\end{aligned}
$$



对于每个类C_i都存在一个超平面h_i，使得
所有$x \in X_i$都在该超平面的正侧
所有$x \in X_j, i \neq j$都在其负侧。

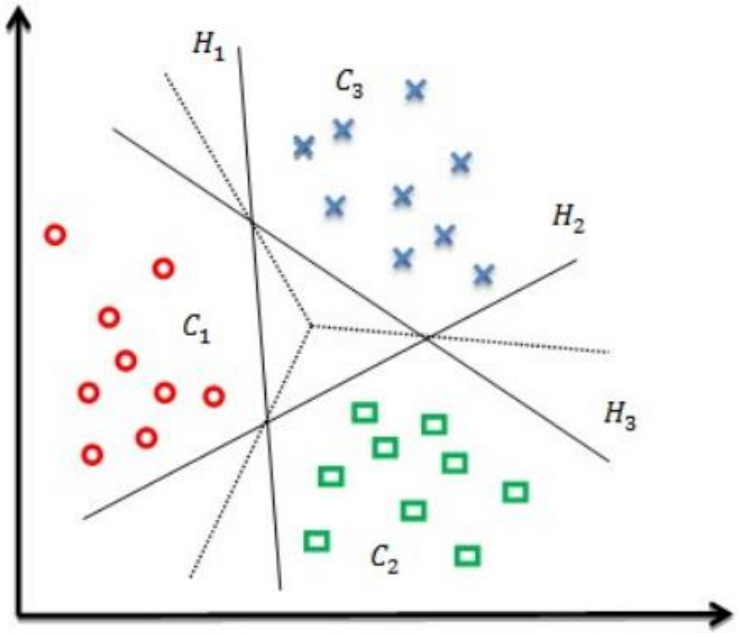
$$
F_i(x|w_i,b_i) = \begin{cases} > 0 \Rightarrow x \in X_i \\ \leq 0 \Rightarrow x \notin X_i \end{cases}
$$

一对多分类

- 现实很难达到理想情况，超平面正侧出现重叠
- 对于某个输入x，所有的判别式结果都小于0，这种输入样例会被拒绝
- 样本指派到判别式值最大的类或距离类超平面最远的类

$$若F_i(x) = max_j F_j(x), 则x \in X_i$$

- 假设有N个输入样例，总共需要计算$c \times N$个判别式结果



多类分类问题：一对一分类

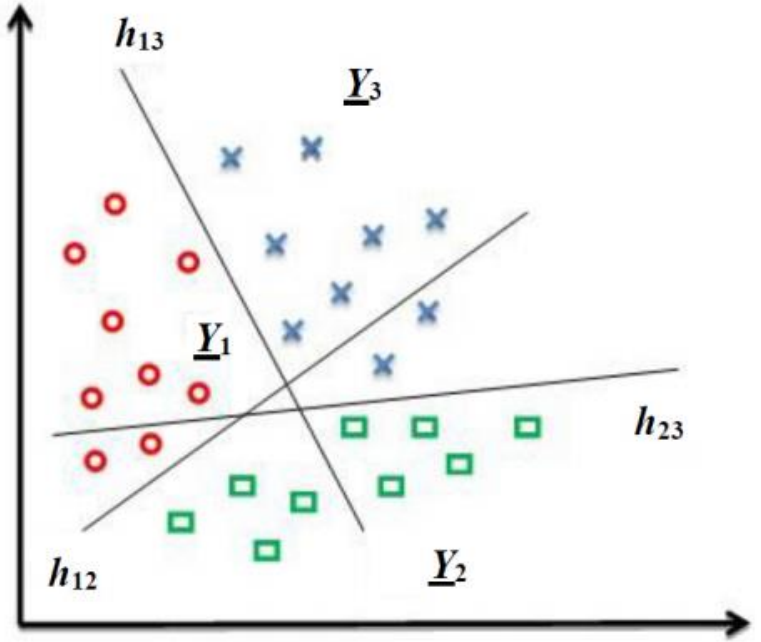
对c个类别进行**两两判别**，即一对一分类

对c个类别中的每一类X_i都能找到能与$X_j(j \neq i)$分开的超平面，共有个 $c(c-1)/2$ 个线性判别式$f_{ij}(x)$，每对不同的类都由一个超平面来划分：

$$f_i(x|w_{ij}, b_{ij}) = w_{ij}x^T + b_{ij}$$

$$f_{ij}(x) = \begin{cases} > 0, x \in X_i \\ \leq 0, x \in X_j \end{cases}$$

$i, j = 1, 2, \cdots, c$且$i \neq j$.



多类分类问题：一对一分类

$$f_i(x|w_{ij}, b_{ij}) = w_{ij}x^T + b_{ij}$$

$$f_{ij}(x) = \begin{cases} > 0, x \in X_i \\ \leq 0, x \in X_j \end{cases}$$

$i, j = 1, 2, \cdots, c$且$i \neq j$.

- 如果存在样例 $x_t \in X_k, k \neq i, k \neq j,$

则在训练时不使用该样例，认为该样例对于划分X_i、X_j**没有贡献。**

- 在检验测试集样本时

  $\forall j \neq i$**都有**$f_{ij}(x) > 0$，则样例划分到X_i

- **不满足上述条件，则取判别式** $f_i(x) = \sum_{j \neq i} f_{ij}(x)$ **取得最大值的类;**
- $f_i(x)$**结果依赖所有对**X_i**进行划分的判别式结果**

11.4.4 支持向量回归

支持向量机也可以用来解决回归问题。

假设有一组训练数据集$X=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$，其中$x_k\in R^p$是独立变量的测量集，$y_k\in R$是相应变量的测量值。

SVM约束条件为

- $u_k(w^Tx_k+b)\geq 1-\xi_k$,
- $\xi_k\geq 0,\quad k=1,2,\ldots,N$

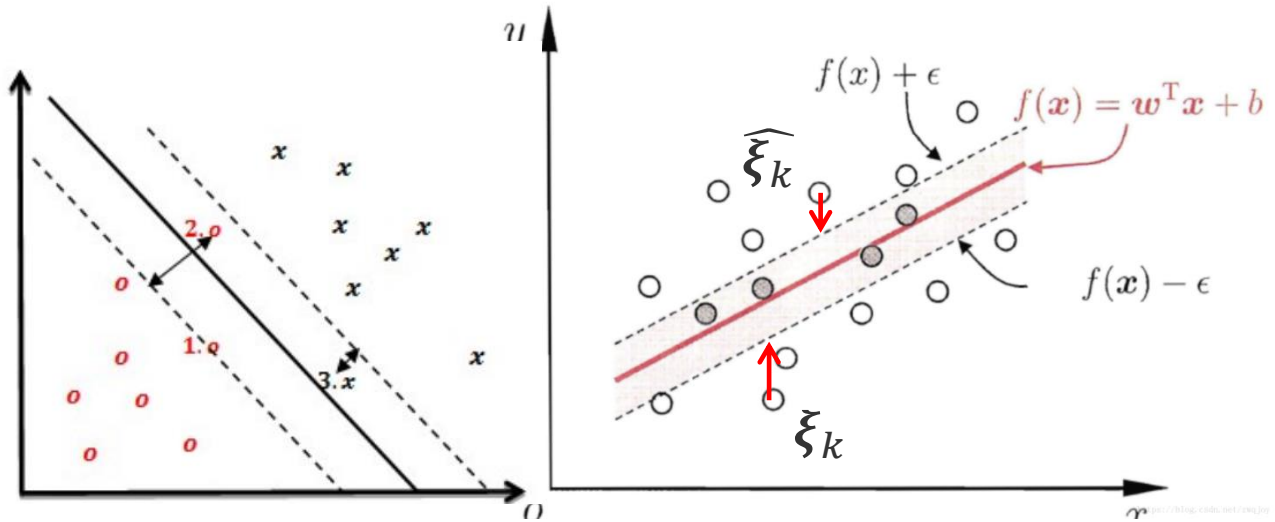
$$(w^Tx_k+b)-(-1)\leq \xi_k,$$

$$1-(w^Tx_k+b)\leq \xi_k,$$

- $\xi_k\geq 0,\quad k=1,2,\ldots,N$

SVR约束条件为

- $(w^Tx_k+b)-y_k\leq \varepsilon+\xi_k$,
- $y_k-(w^Tx_k+b)\leq \varepsilon+\hat{\xi}_k$,
- $\xi_k,\hat{\xi}_k\geq 0,\quad k=1,2,\ldots,N$



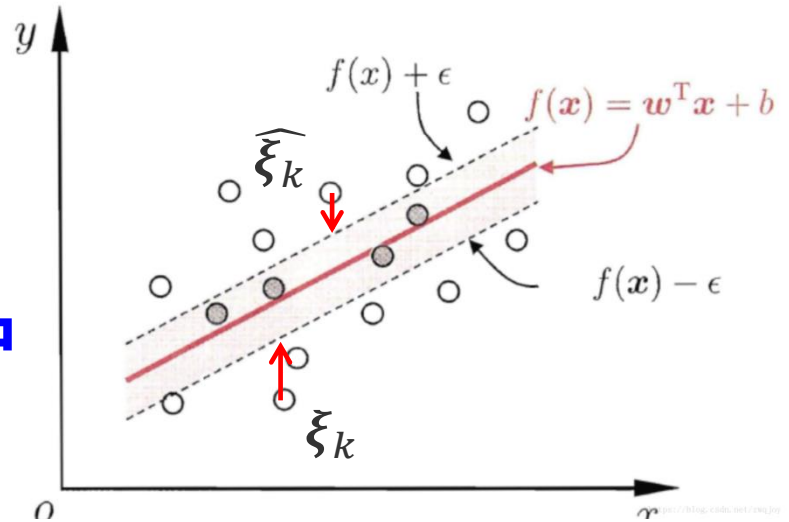
$\xi_k,\hat{\xi}_k$为函数值超出和低于目标值的值大于ε所设

- $-\varepsilon-\xi_k\leq y_k-(w^Tx_k+b)\leq \varepsilon+\hat{\xi}_k$,
- $\xi_k,\hat{\xi}_k\geq 0,\quad k=1,2,\ldots,N$

目标函数

约束条件

- $-\varepsilon - \xi_k \le y_k - \left(w^T x_k + b\right) \le \varepsilon + \widehat{\xi_k}$,
- $\xi_k, \widehat{\xi_k} \ge 0, \quad k = 1,2,\ldots,N$



允许目标值y_k和函数$f(x_k)$之间存在偏差，其中$f(x_k) = w^T x_k + b$.引入两个松弛变量ξ_k和$\widehat{\xi_k}$。

SVR的目标函数为

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{N}(\xi_k + \widehat{\xi_k})$$

$$\text{s.t. } -\varepsilon - \xi_k \le y_k - \left(w^T x_k + b\right) \le \varepsilon + \widehat{\xi_k},$$

$$\xi_k, \widehat{\xi_k} \ge 0, \quad k = 1,2,\ldots,N$$

拉格朗日函数

约束条件

- $-\varepsilon-\xi_k \le y_k-\left(w^T x_k+b\right) \le \varepsilon+\widehat{\xi_k}$,
- $\xi_k, \widehat{\xi_k} \ge 0, \quad k=1,2,\ldots,N$

拉格朗日函数为

$$L(w,b,\alpha,\hat{\alpha},\xi,\hat{\xi},u,\hat{u}) = \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{N}(\xi_k+\widehat{\xi_k}) - \sum_{k=1}^{N}\alpha_k\left[\xi_k+\varepsilon-\left(w^T x_k+b-y_k\right)\right] - \sum_{k=1}^{N}u_k\xi_k - \sum_{k=1}^{N}\widehat{\alpha_k}\left[\widehat{\xi_k}+\varepsilon-\left(y_k-w^T x_k-b\right)\right] - \sum_{k=1}^{N}\widehat{u_k}\,\widehat{\xi_k}$$

其中拉格朗日乘子$\alpha_k, \widehat{\alpha_k} \ge 0$和$u_k, \widehat{u_k} \ge 0$

拉格朗日函数的对偶形式

分别对w, b, ξ_k及$\widehat{\xi_k}$求偏导，并令偏导数为0，可得

$$w + \sum_{k=1}^{N}(\alpha_k - \widehat{\alpha_k})x_k = 0 \qquad \sum_{k=1}^{N}(\alpha_k - \widehat{\alpha_k}) = 0$$

$$C - \alpha_k - u_k = 0 \qquad C - \widehat{\alpha_k} - \widehat{u_k} = 0$$

把上边的式子代入原来的拉格朗日函数，得其对偶形式为

$$\max_{\alpha,\hat{\alpha}} W(\alpha,\hat{\alpha}) = \sum_{k=1}^{N}(\widehat{\alpha_k} - \alpha_k)y_k - \varepsilon\sum_{k=1}^{N}(\widehat{\alpha_k} + \alpha_k)$$

$$-\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\widehat{\alpha_k} - \alpha_k)(\widehat{\alpha_l} - \alpha_l)\, x_k^T x_l$$

$$\text{s.t.} \sum_{k=1}^{N}(\widehat{\alpha_k} - \alpha_k) = 0,\ 0 \le \alpha_k, \widehat{\alpha_k} \le C;\ k = 1,2,\ldots,N$$

求解

通过二次规划数值求解，可得优化参数$\alpha_k, \widehat{\alpha_k}$，进一步求得

$$w^* = \sum_{k=1}^{N} (\widehat{\alpha_k} - \alpha_k) x_k$$

$$b^* = y_l + \varepsilon - \sum_{k=1}^{N} (\widehat{\alpha_k} - \alpha_k)\, x_k^T x_l$$

回归函数为

$$f(x) = \sum_{k=1}^{N} (\widehat{\alpha_k} - \alpha_k) x_k x + b$$

小结

数据特性		SVM
线性可分	→	hard margin
线性近似可分	→	soft margin
线性不可分	→	kernel function

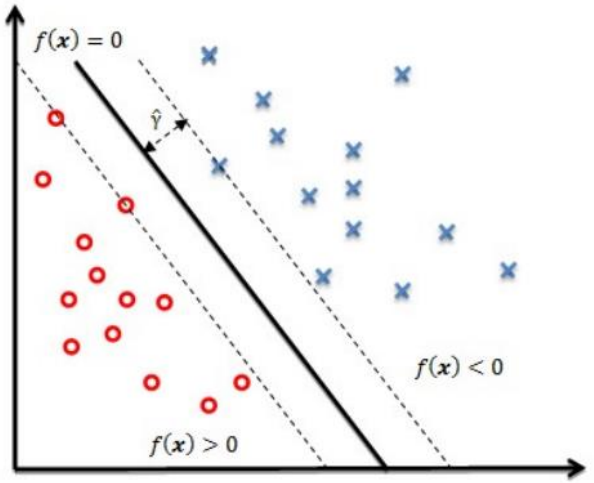
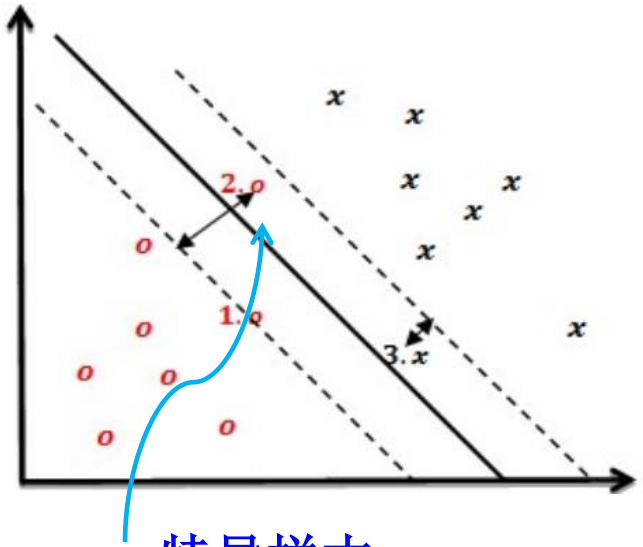
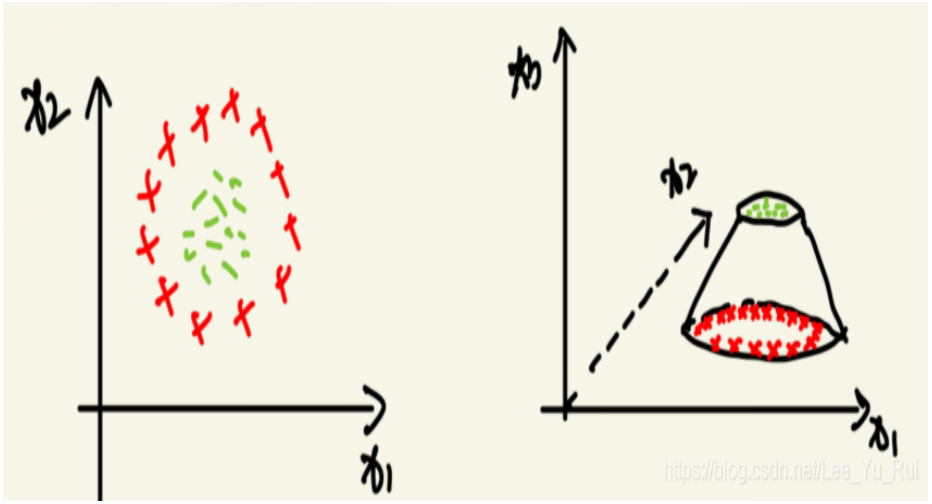


图 11.5　两类对象判决超平面与分离超平面





SVM算法的主要优点

- 解释性很强；
- 解决高维特征的分类问题和回归问题很有效,在特征维度大于样本数时依然有很好的效果；
- 仅仅使用一部分支持向量来做超平面的决策，无需依赖全部数据；
- 有大量的核函数可以使用，从而可以很灵活地解决各种非线性的分类回归问题；
- 样本量不是海量数据的时候，分类准确率高，泛化能力强。

SVM算法的主要缺点

- 如果特征维度远远大于样本数，则SVM表现一般。
- SVM在样本量非常大，核函数映射维度非常高时，计算量过大，不太适合使用。
- 解释能力虽强，但其类表示能力有限，即使引入了核函数，类表示能力还是提高有限；
- 非线性问题的核函数的选择没有通用标准，至今没有好的解决方案；
- SVM对缺失特征数据敏感。

应用实例

Prediction of Individual Brain Maturity Using fMRI

Nico U. F. Dosenbach,[1*] Binyam Nardos,[1] Alexander L. Cohen,[1] Damien A. Fair,[2] Jonathan D. Power,[1] Jessica A. Church,[1] Steven M. Nelson,[1,3] Gagan S. Wig,[1,4,5] Alecia C. Vogel,[1] Christina N. Lessov-Schlaggar,[6] Kelly Anne Barnes,[1] Joseph W. Dubis,[1] Eric Feczko,[6] Rebecca S. Coalson,[1,7] John R. Pruett Jr.,[6] Deanna M. Barch,[3,6,7] Steven E. Petersen,[1,3,7,8] Bradley L. Schlaggar[1,7,8,9*]

Group functional connectivity magnetic resonance imaging (fcMRI) studies have documented reliable changes in human functional brain maturity over development. Here we show that support vector machine-based multivariate pattern analysis extracts sufficient information from fcMRI data to make accurate predictions about individuals' brain maturity across development. The use of only 5 minutes of resting-state fcMRI data from 238 scans of typically developing volunteers (ages 7 to 30 years) allowed prediction of individual brain maturity as a functional connectivity maturation index. The resultant functional maturation curve accounted for 55% of the sample variance and followed a nonlinear asymptotic growth curve shape. The greatest relative contribution to predicting individual brain maturity was made by the weakening of short-range functional connections between the adult brain's major functional networks.

■Dosenbach et al., Science, 2010

应用实例

特征提取

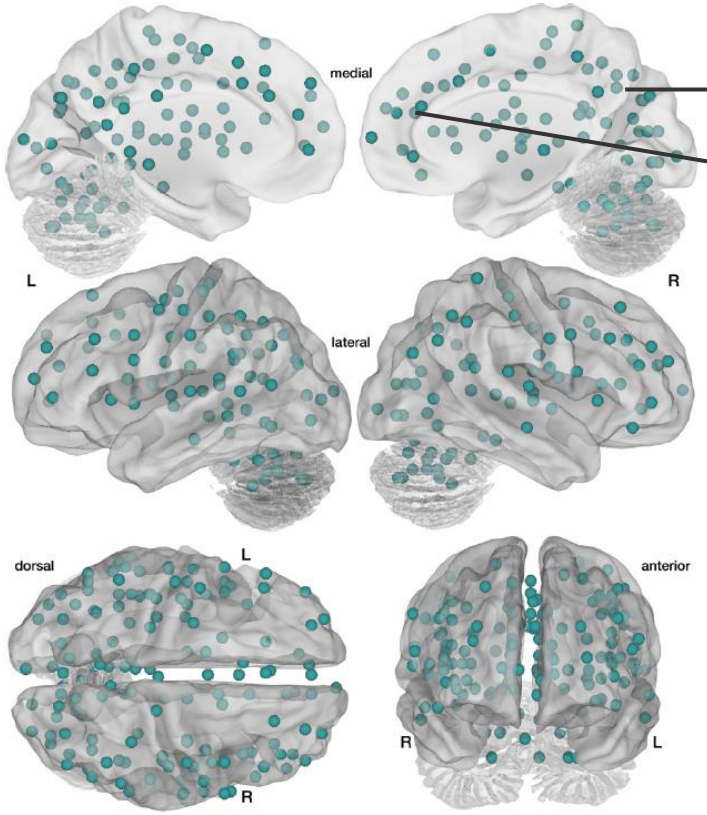
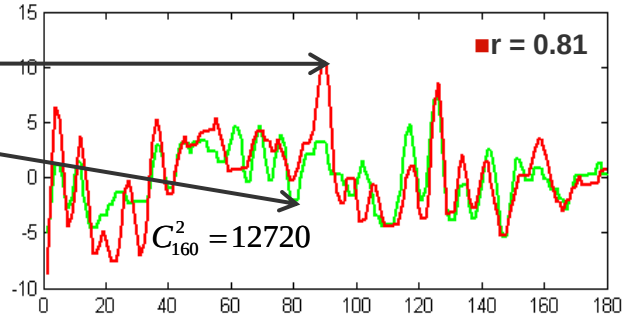


Fig. S1. Regions of interest (ROIs). All 160 ROIs utilized in the analyses are displayed on a surface rendering of the brain (CARET 5.614).



$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$C^2_{160} = 12720$$

$$Y_{238\times 1}$$

$$X_{238\times 12720}$$

特征选择

学习目标1：分类

children (61 scans of 7- to 11-year-olds; mean = 9.4) or adults (61 scans of 24- to 30-year-olds; mean = 26.2)

$$t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}-\frac{1}{n_2}\right)}}$$

学习目标2：回归

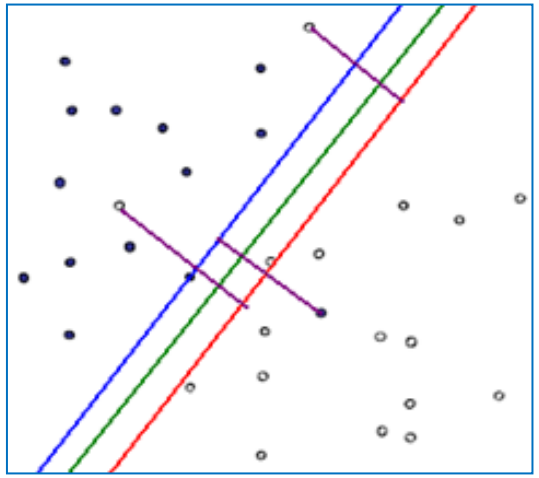
$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

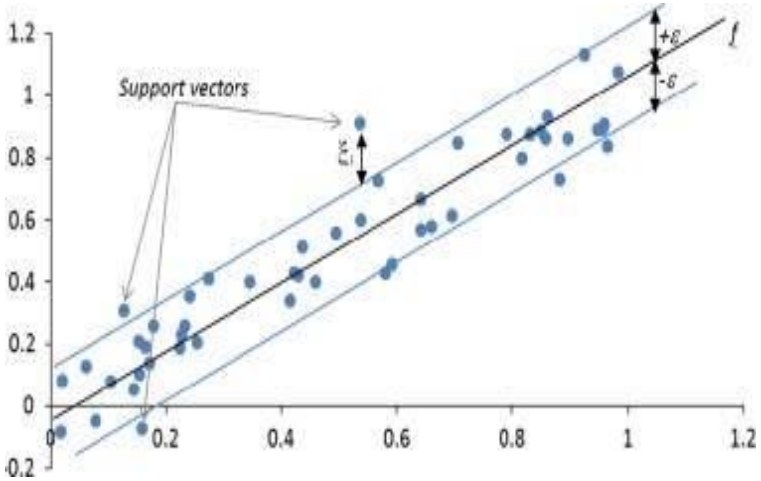
$Y_{238\times1}$

$X_{238\times200}$

应用实例

■模型构建





$$f(x) = w^T x + b$$

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

$$\min \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{R} \varepsilon_i, s.t., y_i (w^T x_i + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0$$

应用实例

交叉验证

$$input: y\_Orig_{238\times1}, X_{238\times12720}, \ 200$$

$$output: y\_Predicted_{238\times1}, weight_{238\times200}$$

$$for\ i = 1:n$$

$$TrainNo_{237\times1} = find(i \neq [1:n])$$

$$TestNo_{1\times1} = i;$$

$$for\ \ j = 1:d$$

$$R(j) = corrcoef(y(TrainNo), X(TrainNo, j));$$

$$end$$

$$R\_Sorted = sort(abs(R), 'descend');$$

$$FeatNo_{1\times200} = find(R \geq R\_Sorted(200));$$

$$\mathrm{Model\{i\}} = svmtrain(y(TrainNo), X(TrainNo, FeatNo));$$

$$y\_Predicted(i) = svmpredict(X(TestNo, FeatNo), Model\{\mathrm{i}\});$$

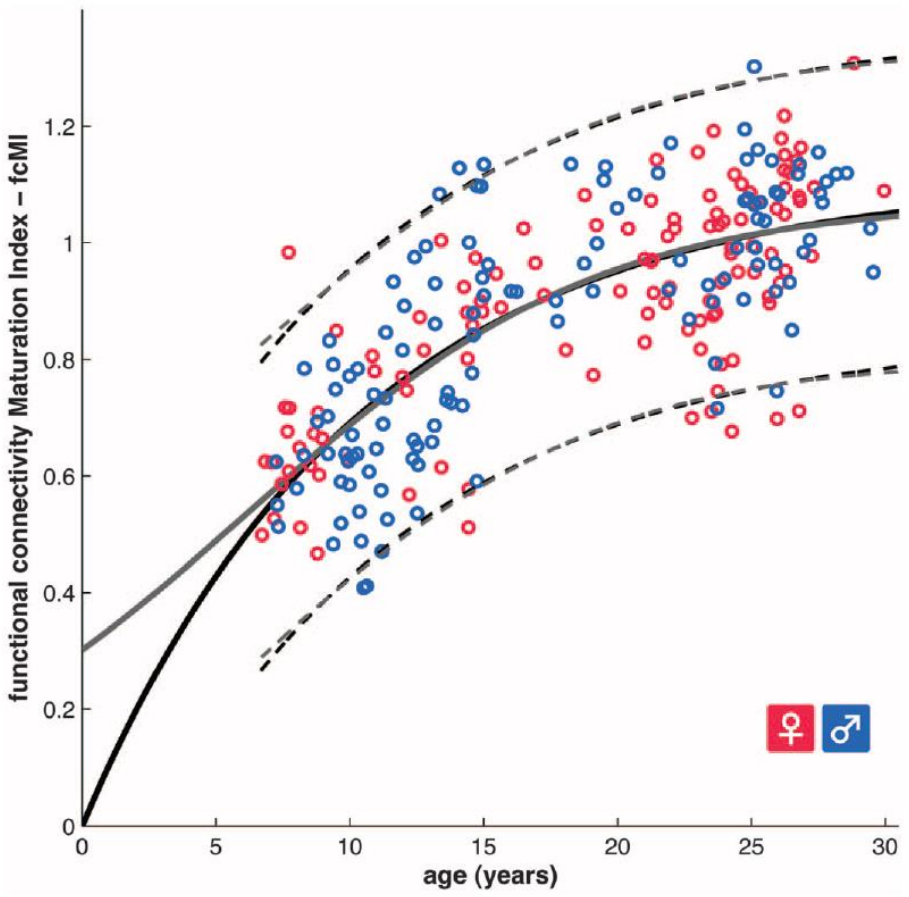
$$end$$

应用实例

■结果：

- 分类：正确率91%
- 回归：

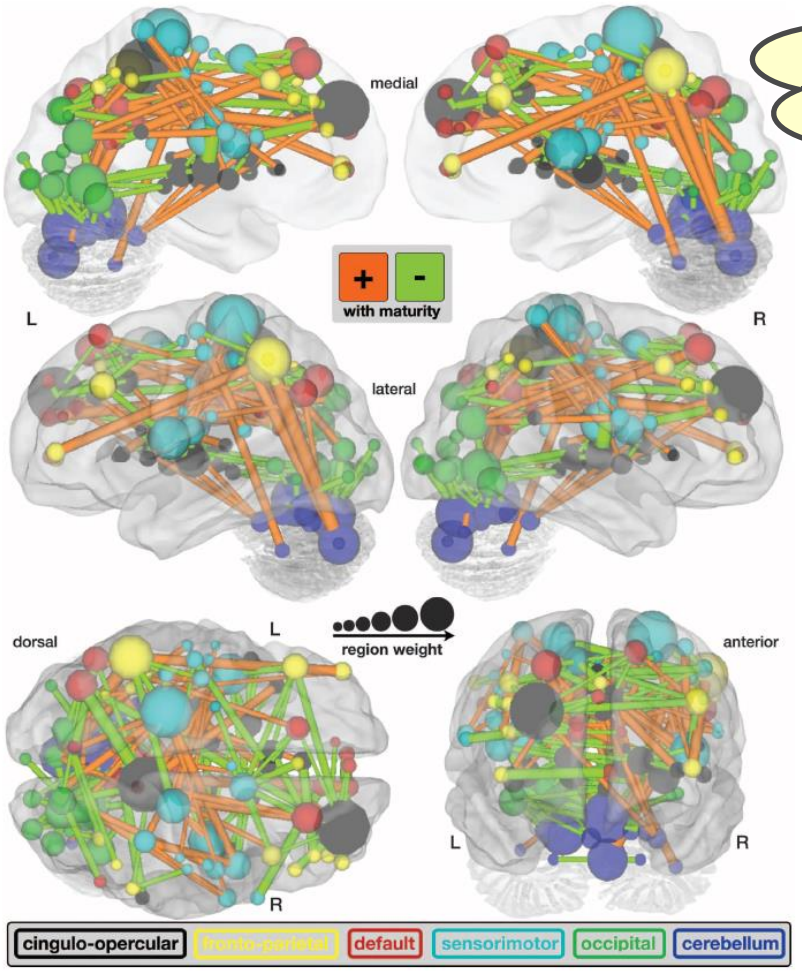
Fig. 1. Functional brain maturation curve. Individual functional brain maturity levels of 238 rs-fcMRI scans (115 females) between the ages of 7 to 30 years. Chronological age is shown on the *x* axis and the fcMI on the *y* axis (females pink, males blue). The fit for the Von Bertalanffy's equation [$a \bullet (1 - e^{-bx})$, $r^2 = 0.553$, permutation test, $P < 0.001$, AIC weight = 0.3] is shown with a solid black line. The fit for the Pearl-Reed equation [$a/(1 + b \bullet e^{-cx})$, $r^2 = 0.555$, AIC weight = 0.23] is shown with a solid gray line. The 95% prediction limits are shown with dashed lines.



■Dosenbach et al., Science, 2010

应用实例

■结果



$$f(x) = w^T x + b$$

$$f(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

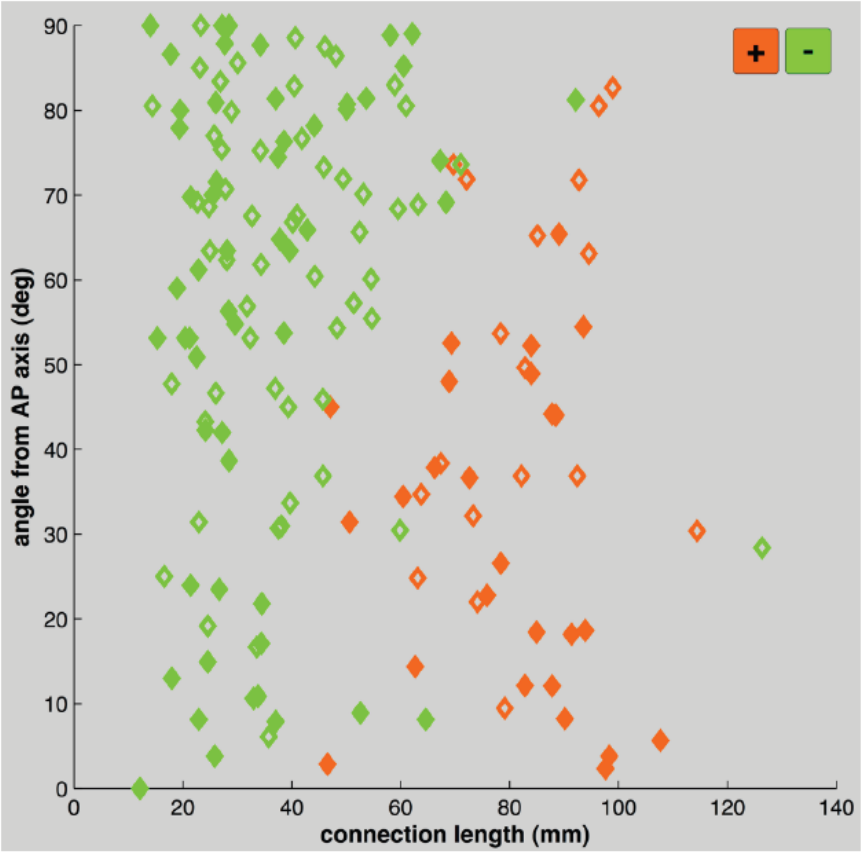
$$X_{238\times12720}$$

$$weight_{238\times200}$$

$$weight_{238\times156}$$

应用实例

■结果



■Lengths and angles of consensus functional connections. Consensus functional connections (diamonds) are displayed in **orange** if they are positively correlated with chronological age (stronger with age), they are displayed in **light green** if they are negatively correlated with chronological age (weaker with age). Displayed on the **x-axis** is the length of functional connections in mm. Displayed on the **y-axis** is the absolute angle (in degrees) functional connections make with the anterior-posterior (AP) axis in the horizontal plane. **Filled diamonds** represent cerebral and cerebellar interhemispheric connections, as well as ipsilateral connections between the cerebellum and cortex. **Empty diamonds** indicate connections within cerebral and cerebellar hemispheres as well as contralateral connections between the cerebellum and cerebrum.

■Dosenbach et al., Science, 2010

应用实例

■结果

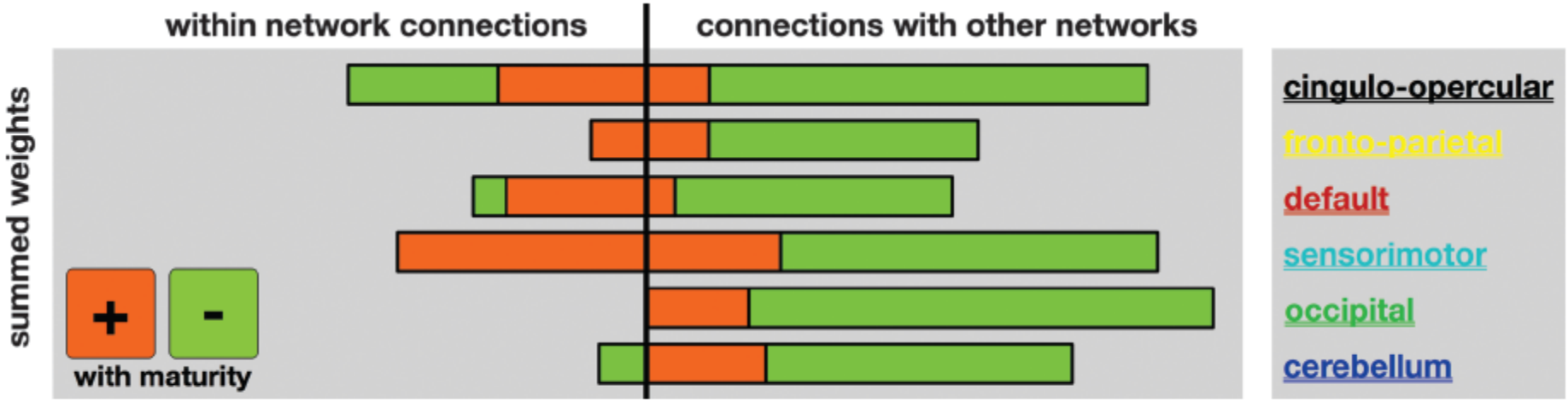
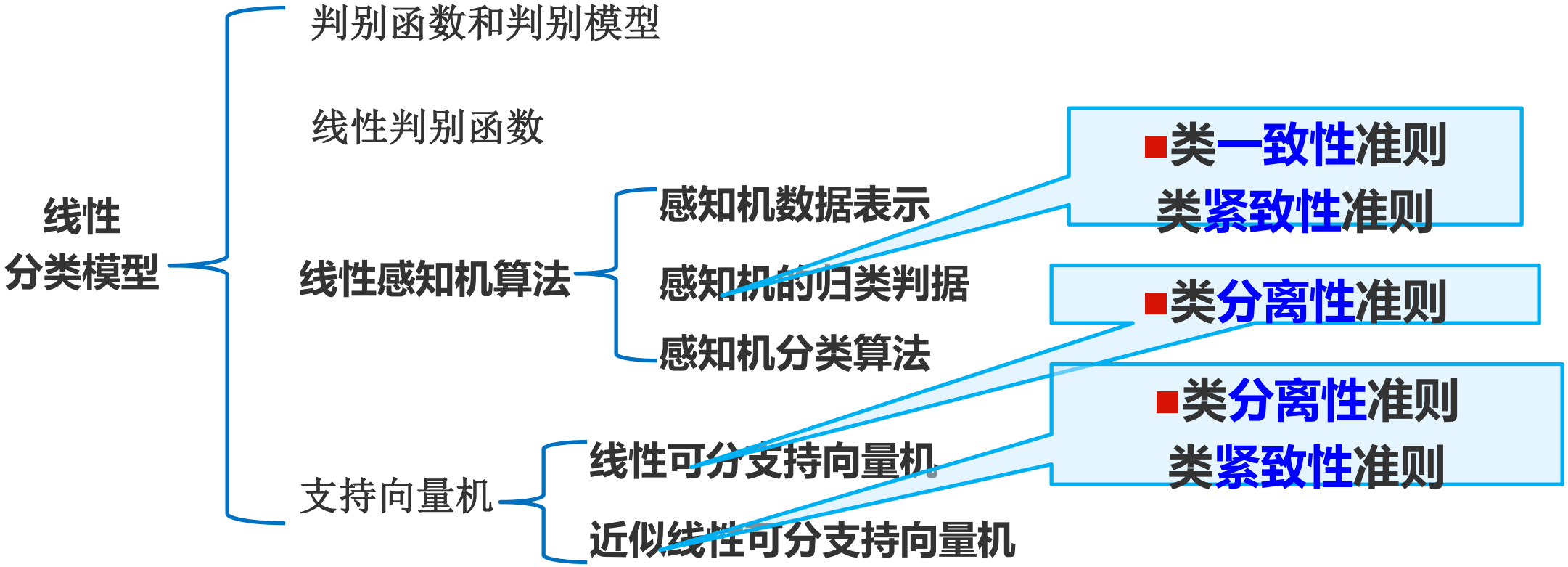


Fig. 3. SVR brain maturity weights by adult rs-fcMRI networks. The sums of all the functional connection weights within each network are shown to the left of the vertical black line. The sums of all the functional connection weights between networks are shown to the right.

■Dosenbach et al., Science, 2010

本章总结


线性
分类模型
判别函数和判别模型
线性判别函数
线性感知机算法
感知机数据表示
感知机的归类判据
感知机分类算法
支持向量机
线性可分支持向量机
近似线性可分支持向量机
■类一致性准则
类紧致性准则
■类分离性准则
■类分离性准则
类紧致性准则

作业

■理论部分

教材165页第2、6、9题

■实践部分

1. 阅读论文：Minimal Optimization: A Fast Algorithm for Training Support Vector Machines，理解SMO算法；
2. 基于给定的样本特征和标签数据，采用支持向量机对样本进行分类，观察惩罚因子C对分类结果的影响。

- 假设有两类样例点如下：
- 第1类$\begin{pmatrix}1\\1\end{pmatrix}\begin{pmatrix}1\\2\end{pmatrix}\begin{pmatrix}2\\1\end{pmatrix}$，第2类$\begin{pmatrix}0\\0\end{pmatrix}\begin{pmatrix}0\\1\end{pmatrix}\begin{pmatrix}1\\0\end{pmatrix}$
- 在坐标系中画出这些点并且寻找最佳分离超平面，计算支持向量和间隔

BJTU “Machine Learning” Group

于　剑：jianyu@bjtu.edu.cn；

景丽萍：lpjing@bjtu.edu.cn；

田丽霞：lxtian@bjtu.edu.cn；

黄惠芳：hfhuang@bjtu.edu.cn；

李晓龙：hlli@bjtu.edu.cn；

吴　丹：wudan@bjtu.edu.cn；

万怀宇：hywan@bjtu.edu.cn；

王　晶：wj@bjtu.edu.cn.