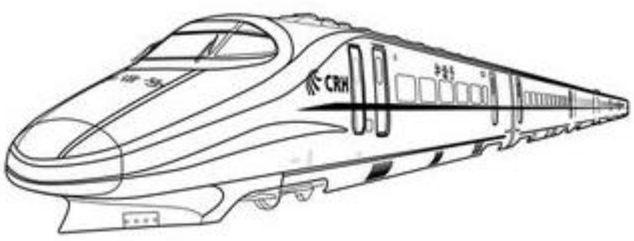
北京交通大学
BEIJING JIAOTONG UNIVERSITY

智能信息技术教育中心
The Education Center of Intelligence Information Technologies

人工智能基础

机器学习

耿阳李敖
2022年3月



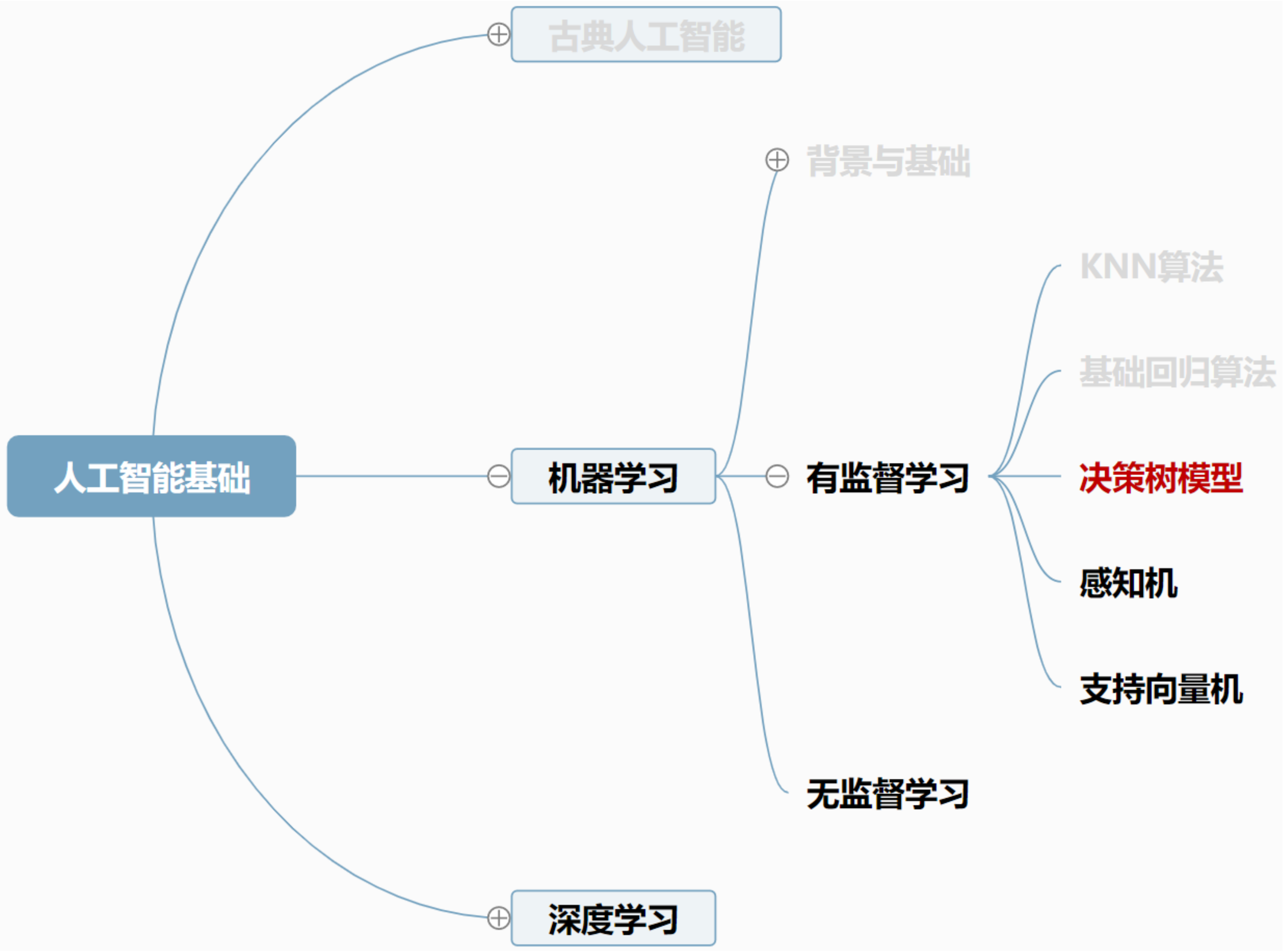
课程反馈

- 讲授内容过难，数学内容多
- 课堂缺乏互动
- 作业
- 节课方式

内容概要

- **内容回顾**
- 感知机
- 支持向量机
- 总结

课程总览

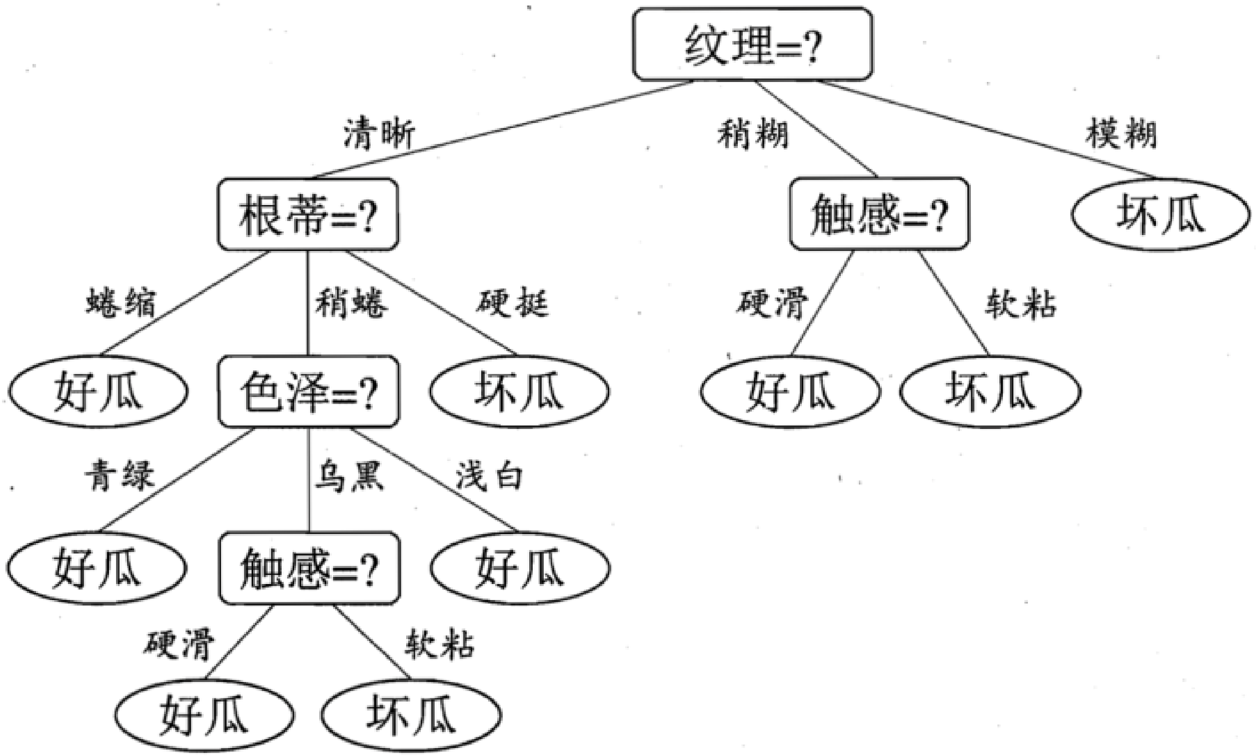


人工智能基础
古典人工智能
机器学习
深度学习
背景与基础
有监督学习
无监督学习
KNN算法
基础回归算法
决策树模型
感知机
支持向量机

决策树算法

主要思想

设定判定规则，根据判定结果不断分枝，形成决策树结构



决策树算法

算法流程

输入：训练集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_m,y_m)\}$;
属性集 $A=\{a_1,a_2,\ldots,a_d\}$.

过程：函数 TreeGenerate(D, A)

生成结点 node;
if D 中样本全属于同一类别 C **then**
  将 node 标记为 C 类叶结点; **return**
end if
if $A=\varnothing$ **OR** D 中样本在 A 上取值相同 **then**
  将 node 标记为叶结点，其类别标记为 D 中样本数最多的类; **return**
end if
从 A 中选择最优划分属性 a_*;
for a_* 的每一个值 a_*^v **do**
  为 node 生成一个分支; 令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集;
  if D_v 为空 **then**
    将分支结点标记为叶结点，其类别标记为 D 中样本最多的类; **return**
  else
    以 TreeGenerate(D_v, $A\setminus\{a_*\}$)为分支结点
  end if
end for

输出：以 node 为根结点的一棵决策树

例

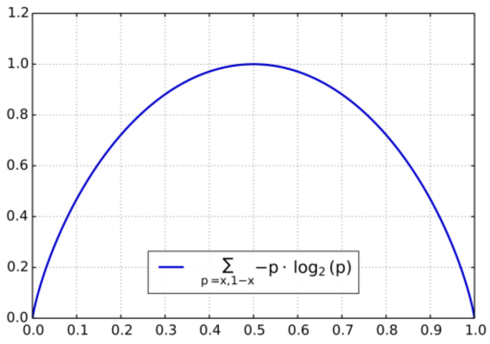
x：清晰、硬挺、乌黑、硬滑
y：好瓜/坏瓜
A：纹理、根蒂、色泽、触感

决策树的分支结点所包含的样本尽可能属于同一类别，即结点的“纯度”尽量高。

决策树算法

信息熵

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$



- 信息熵越小，纯度越高
- 平均分布取得最大值为 $log_2|y|$

划分准则

- 信息增益

$$\mathrm{Gain}(D,a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

对取值数目较多的类别有所偏好

- 信息增益率

$$\mathrm{Gain\_ratio}(D,a) = \frac{\mathrm{Gain}(D,a)}{\mathrm{IV}(a)}$$

$$\mathrm{IV}(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|}\log_2 \frac{|D^v|}{|D|}$$

- 基尼指数

$$\mathrm{Gini\_index}(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|}\mathrm{Gini}(D^v)$$

$$\mathrm{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|}\sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2$$

决策树算法

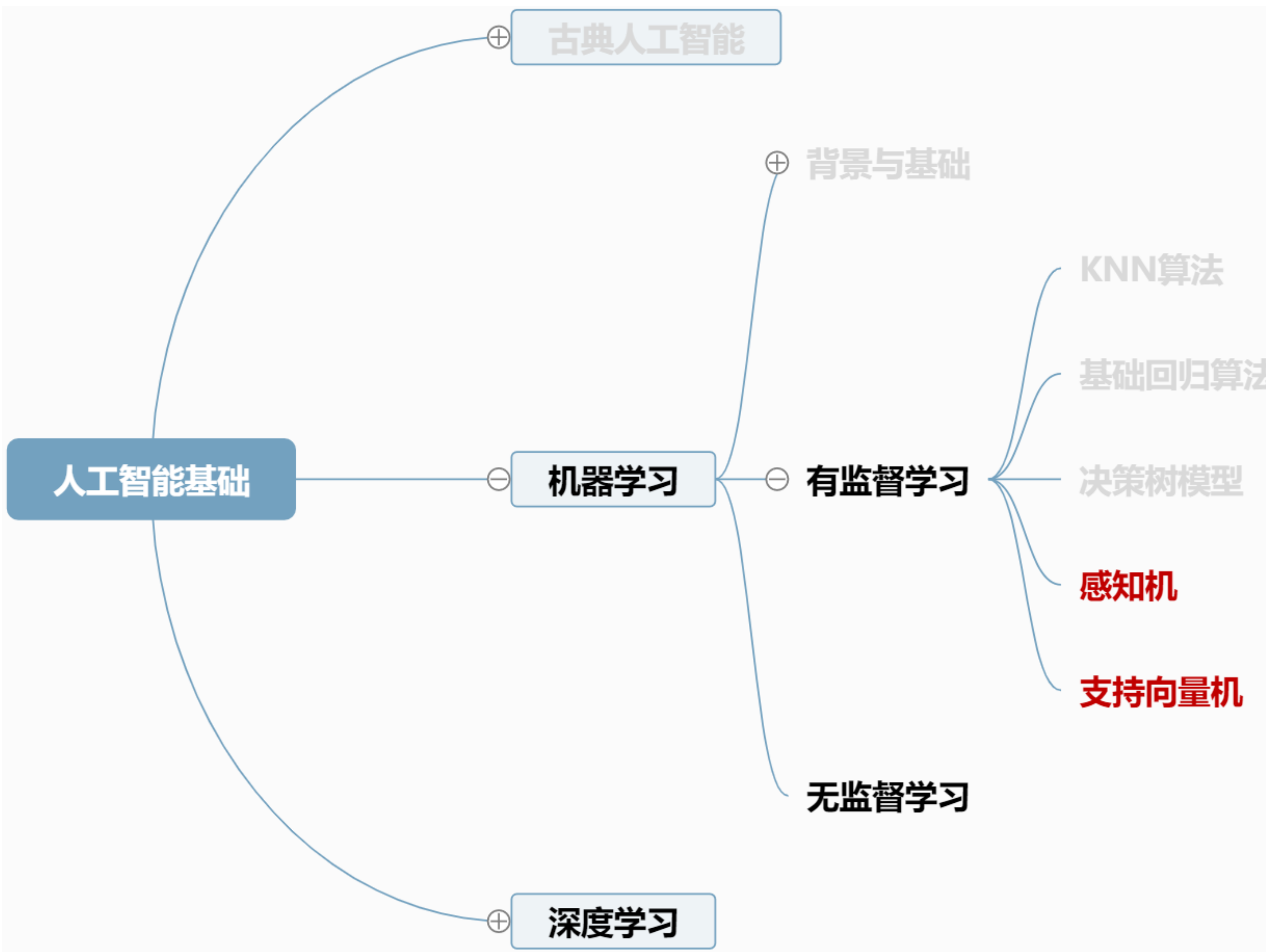
决策树剪枝

- 预剪枝：在决策树生成过程中进行剪枝
- 后剪枝：先生成一颗完整的决策树，再进行剪枝

适用性增强

- 连续属性值：二分法进行最优离散划分
- 缺失值处理：样本加权+概率划分
- 多变量决策树：一次性利用多个属性进行划分

本节概况



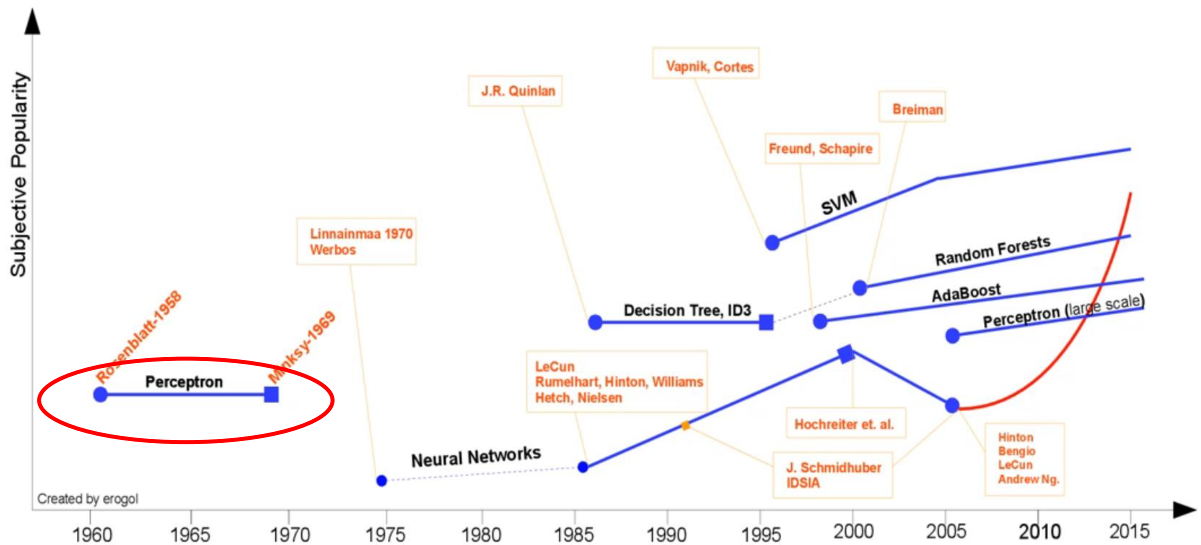
人工智能基础
古典人工智能
机器学习
背景与基础
有监督学习
KNN算法
基础回归算法
决策树模型
感知机
支持向量机
无监督学习
深度学习

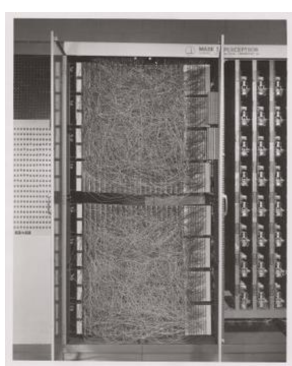
内容概要

- 内容回顾
- **感知机**
- 支持向量机
- 总结

感知机

诞生

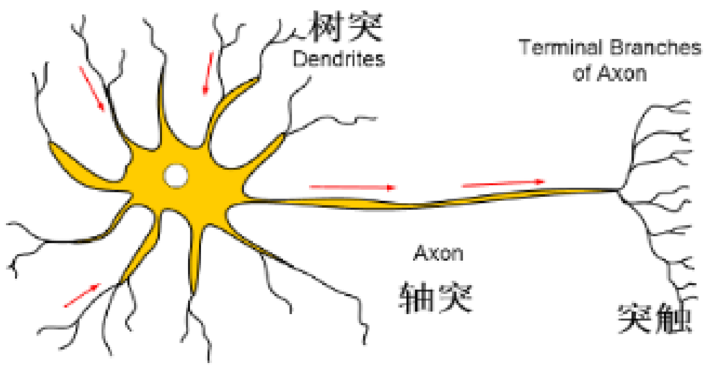


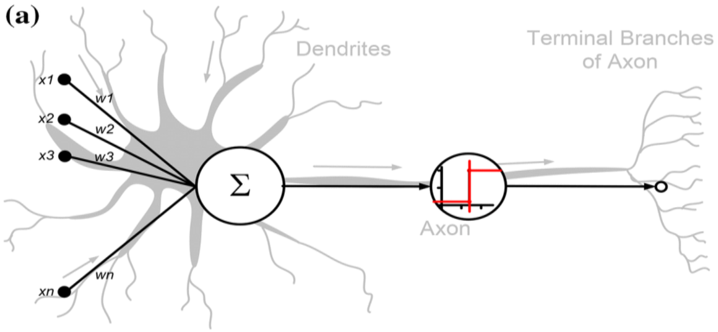


The **perceptron** algorithm was invented in 1958 at the Cornell Aeronautical Laboratory by **Frank Rosenblatt**. The perceptron was intended to **be a machine, rather than a program**, and while its first implementation was in software for the IBM 704.

感知机

设计原理





- 接收神经信号：

$$x_{(1)}, x_{(2)}, x_{(3)}, \dots, x_{(D)}$$

- 内部汇总处理：

$$w_1 x_{(1)} + w_2 x_{(2)} + \cdots + w_D x_{(D)} + b = \sum_{i=1}^{D} w_i x_{(i)} + b$$

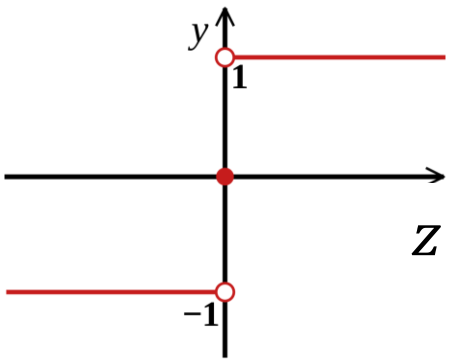
- 激活输出：

$$y = \sigma\left(\sum_{i=1}^{D} w_i x_{(i)} + b\right)$$

感知机

激活函数

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



感知机向量形式

$$y = \sigma\left(\sum_{i=1}^{d} w_i x_{(i)} + b\right) = \sigma(\mathbf{w}^T \mathbf{x} + b)$$

$$\mathbf{w} = (w_1, w_2, \dots, w_D)^T \in \mathbb{R}^D \qquad \mathbf{x} = \left(x_{(1)}, x_{(2)}, \dots, x_{(D)}\right)^T \in \mathbb{R}^D$$

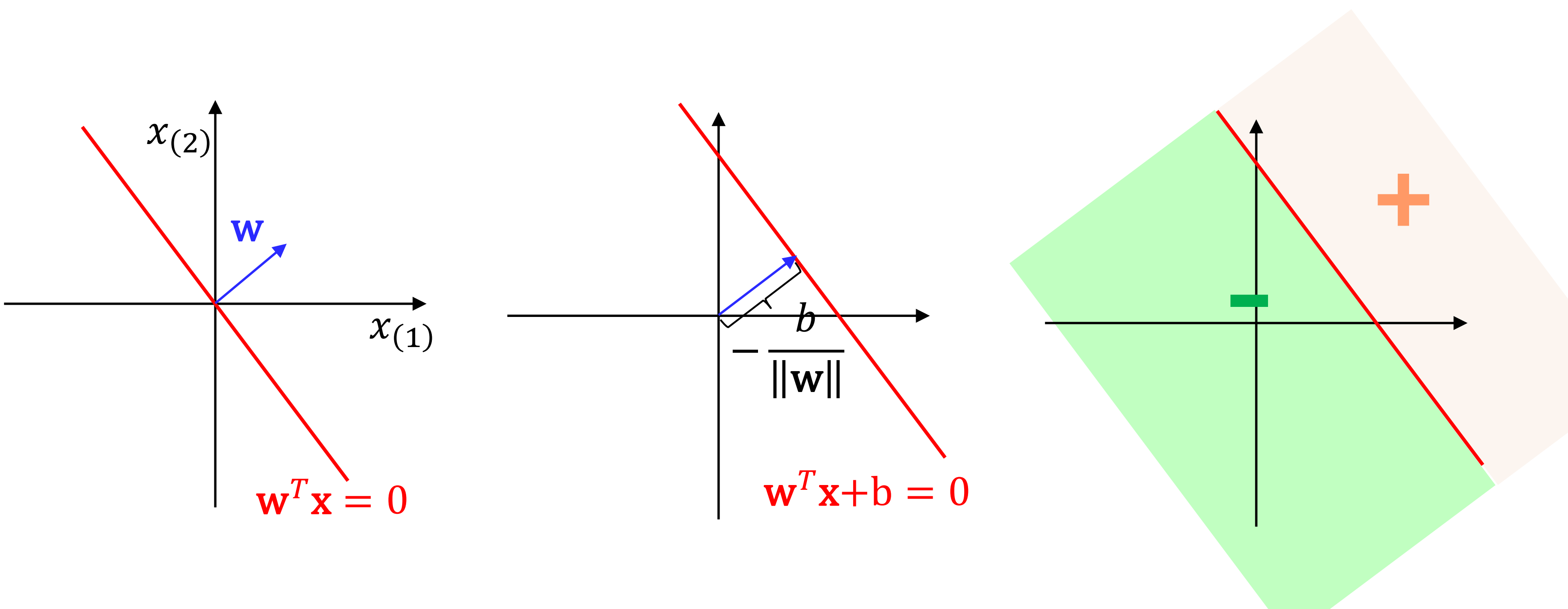
感知机

几何意义

$$\sigma(\mathbf{w}^T\mathbf{x}+b) \qquad \sigma(z)=\begin{cases}1, & z>0\\ 0, & z=0\\ -1, & z<0\end{cases}$$

最终输出只与 $\mathbf{w}^T\mathbf{x}+b$ 的符号有关

$$\mathbf{w},\mathbf{x}\in\mathbb{R}^2$$



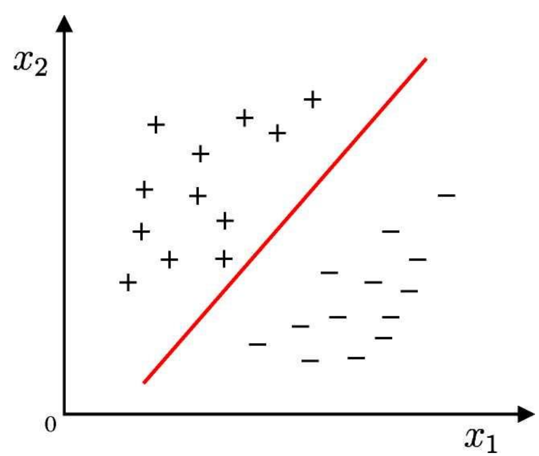
感知机

二分类任务

- 训练数据：$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$　$\mathbf{x}_i \in \mathbb{R}^D$　$y_i \in \{-1, 1\}$
- 找到一组 $(\mathbf{w}, b)$ 使得：

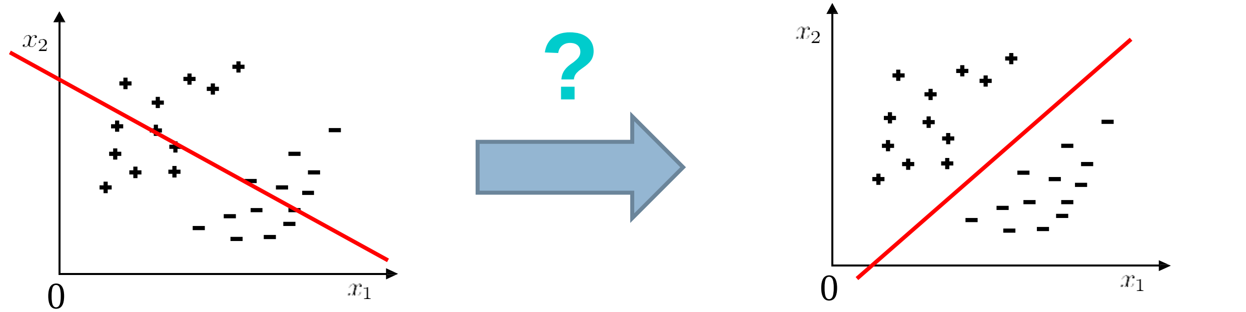
$$y_i \sigma(\mathbf{w}^T \mathbf{x}_i + b) = 1 \quad \Longleftrightarrow \quad y_i(\mathbf{w}^T \mathbf{x}_i + b) > 0$$

- 找到一个 D 维超平面能够将 N 个训练数据完美分开



感知机

感知机学习



- 目标函数导出：

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) > 0$$

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

$$\min_{\mathbf{w},b} f(\mathbf{w}, b)$$

$$f(\mathbf{w}, b) = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

感知机

感知机学习

- 目标函数

$$\min_{\mathbf{w},b} f(\mathbf{w},b) \qquad f(\mathbf{w},b) = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

- 梯度下降

$$\nabla_{\mathrm{b}} f = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} \nabla_{\mathrm{b}} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

$$= -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\nabla_{\mathrm{b}}\mathbf{w}^T\mathbf{x}_i + \nabla_{\mathrm{b}} b)$$

$$= -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(0+1) = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i$$

感知机

感知机学习

- 目标函数

$$\min_{\mathbf{w},b} f(\mathbf{w},b) \qquad f(\mathbf{w},b) = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

- 梯度下降

$$\nabla_{\mathbf{w}} f = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} (\ ?\ )$$

A	B	C	D
$y_i(\mathbf{x}_i + b)$	$y_i\mathbf{x}_i$	$\mathbf{x}_i + b$	$y_i b$

感知机

感知机学习

- **梯度**

$$f(\mathbf{w}, b) = -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\mathbf{w}^T\mathbf{x}_i + b)$$

$$\begin{aligned}
\nabla_{\mathbf{w}} f &= -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} \nabla_{\mathbf{w}} y_i(\mathbf{w}^T\mathbf{x}_i + b) \\
&= -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\nabla_{\mathbf{w}}\mathbf{w}^T\mathbf{x}_i + \nabla_{\mathbf{w}} b) \\
&= -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i(\mathbf{x}_i + 0) \\
&= -\sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i\mathbf{x}_i
\end{aligned}$$

$\mathbf{w}^T\mathbf{x}_i = \sum_{j=1}^{d} w_j x_{i(j)}$

感知机

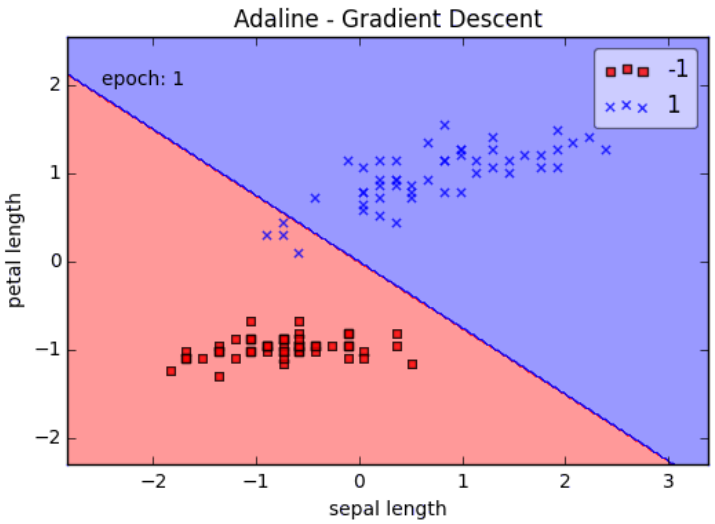
感知机学习

◆ **梯度**

$$\nabla_{\mathbf{w}} f = - \sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i \mathbf{x}_i \qquad \nabla_{\mathrm{b}} f = - \sum_{y_i(\mathbf{w}^T\mathbf{x}_i+b)<0} y_i$$

$$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} f$$

$$b \leftarrow b - \eta \nabla_{b} f$$



感知机

例子

- 假设给定训练集为

$$\text{正类}(y=+1):\quad \{(1,2),(1,3),(2,3),(2,4)\}$$

$$\text{负类}(y=-1):\quad \{(3,2),(4,2),(2,1),(3,1)\}$$

- 下面哪些参数对应的感知机可以对训练集正确分类（多选）

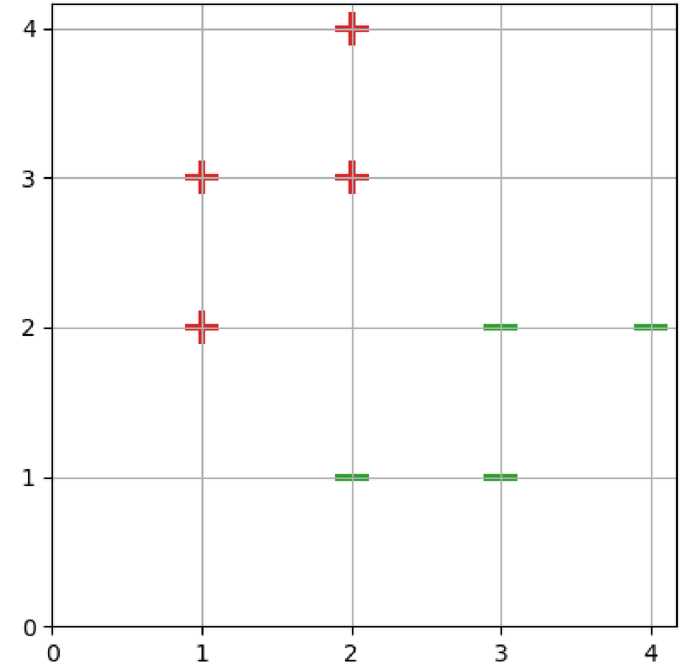
A	B	C	D
$\mathbf{w}=[-2,1]$ $b=2$	$\mathbf{w}=[1,1]$ $b=-3$	$\mathbf{w}=[-1,1]$ $b=0$	$\mathbf{w}=[-1,2]$ $b=-2$

感知机

例子

正类$(y=+1)$：　$\{(1,2),(1,3),(2,3),(2,4)\}$

负类$(y=-1)$：　$\{(3,2),(4,2),(2,1),(3,1)\}$



感知机

例子

正类$(y=+1)$： $\{(1,2),(1,3),(2,3),(2,4)\}$

负类$(y=-1)$： $\{(3,2),(4,2),(2,1),(3,1)\}$

A

$$\mathbf{w}=[-2,1]$$
$$b=2$$

B

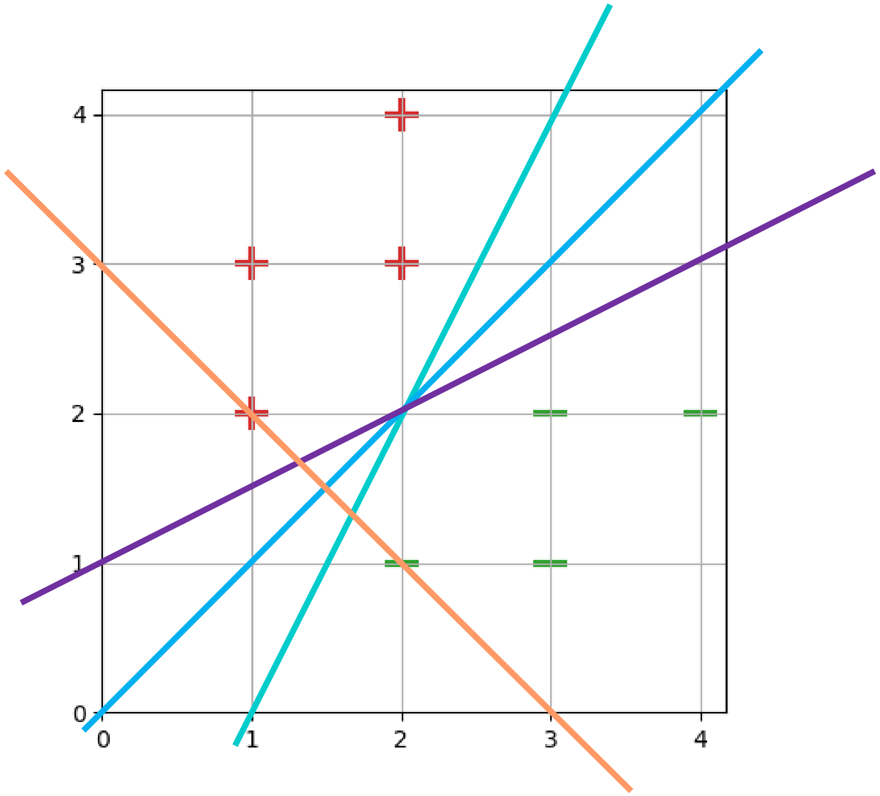
$$\mathbf{w}=[1,1]$$
$$b=-3$$

C

$$\mathbf{w}=[-1,1]$$
$$b=0$$

D

$$\mathbf{w}=[-1,2]$$
$$b=-2$$

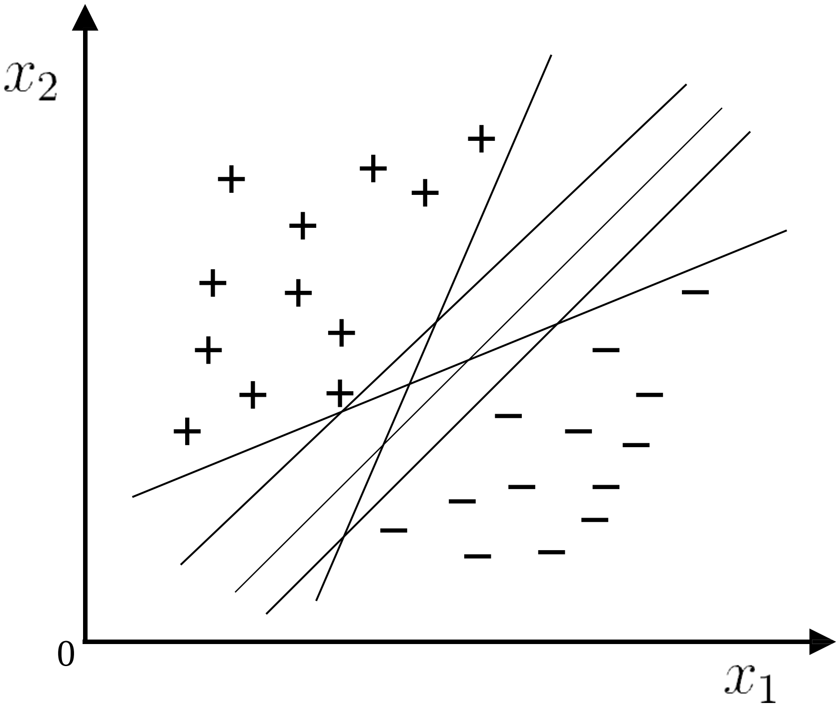


内容概要

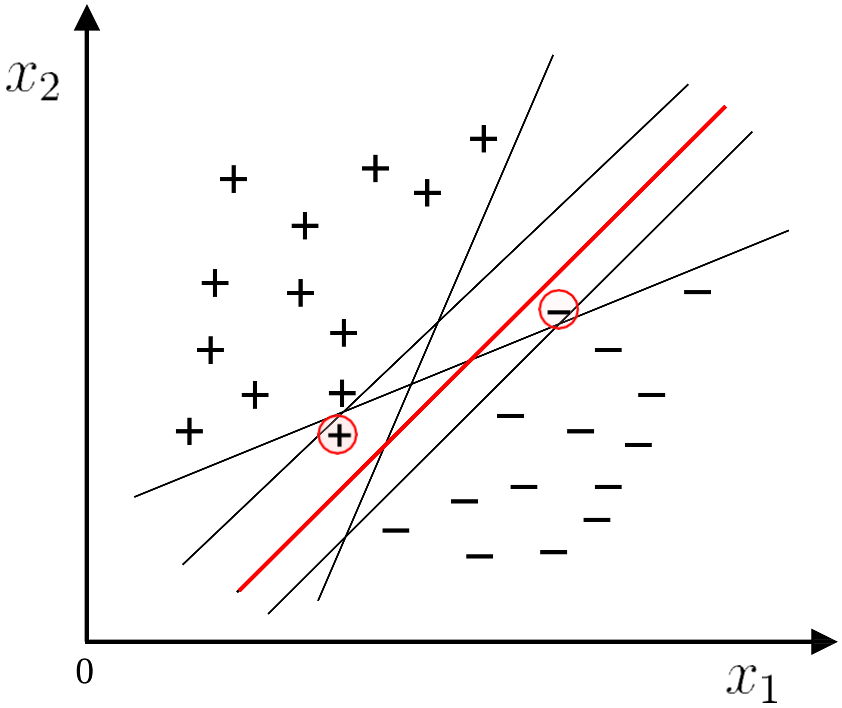
- 内容回顾
- 感知机
- **支持向量机**
- 总结

支持向量机

感知机的问题



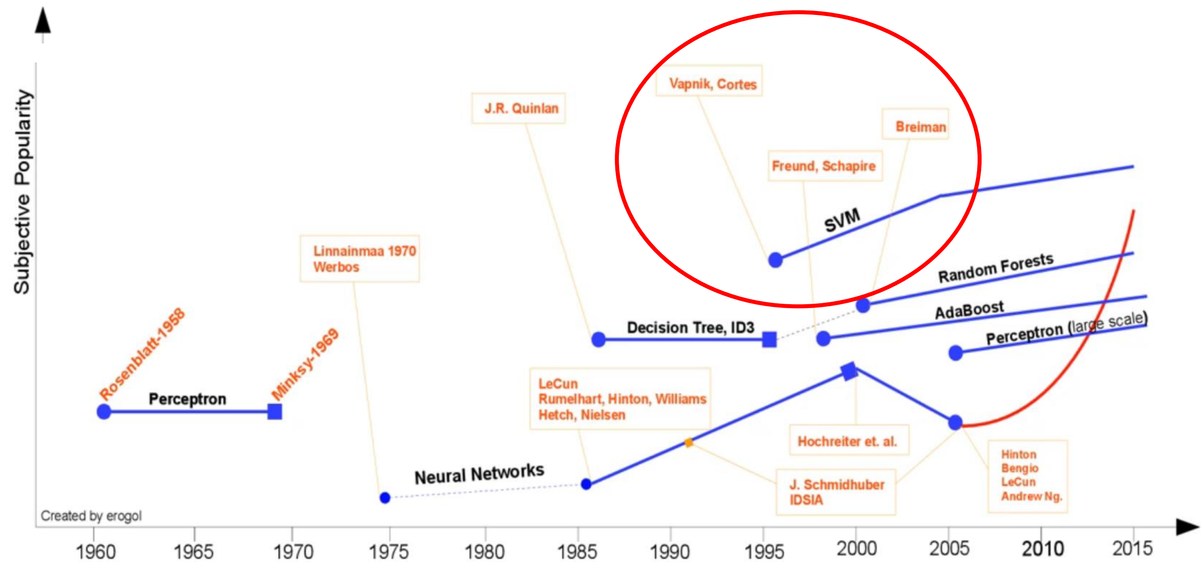
哪个分类超平面效果最好?

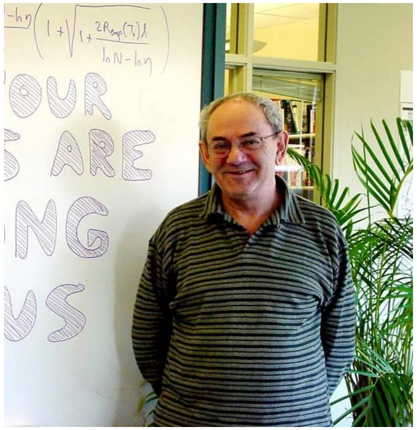


应选择“正中间”，容忍性好,鲁棒性高

支持向量机

支持向量机





In machine learning, **Support-Vector Machines (SVMs)** are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Developed at AT&T Bell Laboratories by **Vladimir Vapnik** with colleagues in 1995. SVMs are one of the most robust prediction methods.

支持向量机

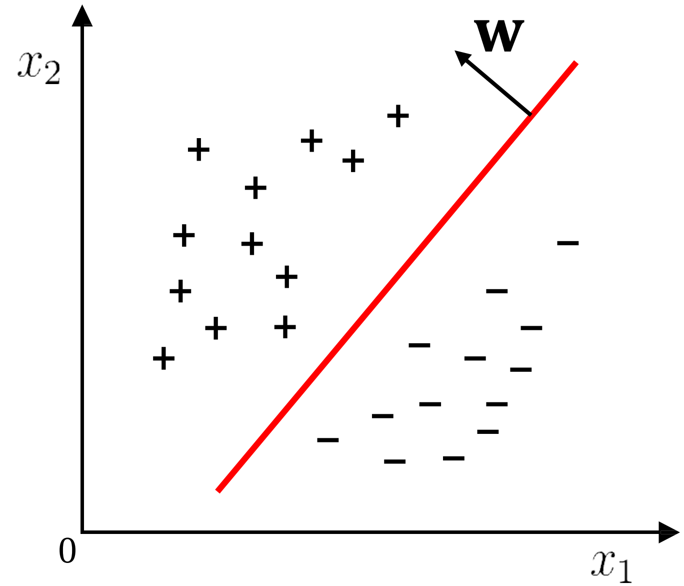
分离间隔最大化

分类超平面方程： $\mathbf{w}^T\mathbf{x} + b = 0$

点到超平面有向距离： $\dfrac{\mathbf{w}^T\mathbf{x} + b}{\|\mathbf{w}\|}$

训练数据： $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\}$

$\mathbf{x}_i \in \mathbb{R}^D \qquad y_i \in \{-1, 1\}$



- 原始目标函数

$$\max_{\mathbf{w},b} \min_{i} \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|}$$

支持向量机

问题转化

$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_{i} \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i + \tilde{b}\right)}{\|\tilde{\mathbf{w}}\|}$$

对于任意超平面

$$\forall \gamma \neq 0$$

$$\tilde{\mathbf{w}}^T\mathbf{x} + \tilde{b} = 0 \iff \frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x} + \frac{\tilde{b}}{\gamma} = 0$$

支持向量机

问题转化

$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_i \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i + \tilde{b}\right)}{\|\tilde{\mathbf{w}}\|}$$

对于任意超平面

$\forall\gamma \neq 0$

$$\tilde{\mathbf{w}}^T\mathbf{x} + \tilde{b} = 0 \quad \Longleftrightarrow \quad \frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x} + \frac{\tilde{b}}{\gamma} = 0$$

$$\exists\gamma \text{ 使得 } 1 = \min_i \left|\frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}_i + \frac{\tilde{b}}{\gamma}\right|$$

此时间隔恰为 $\frac{2}{\|\mathbf{w}\|}$

支持向量机

问题转化

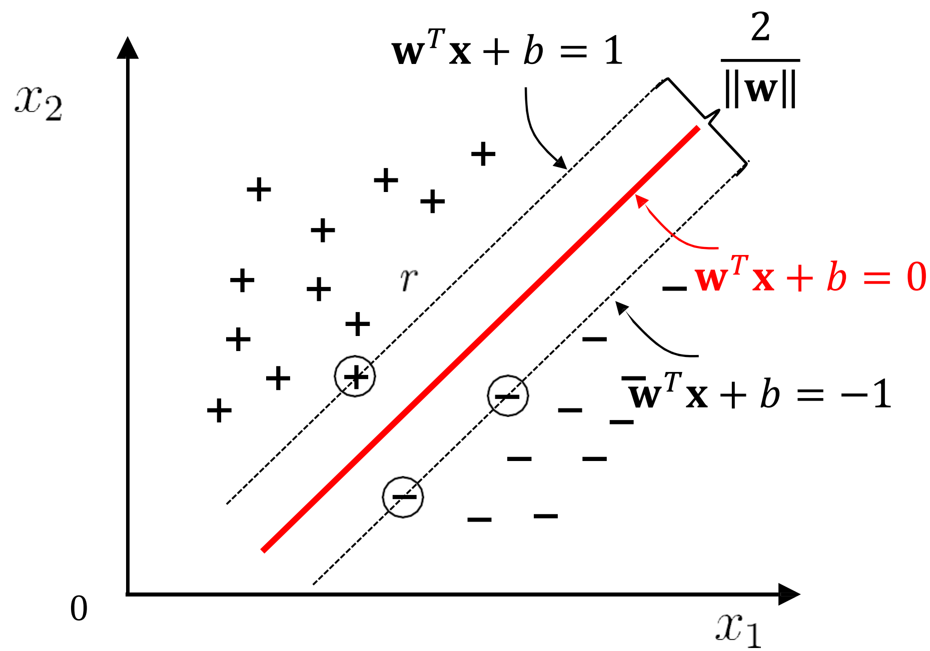
$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_i \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i+\tilde{b}\right)}{\|\tilde{\mathbf{w}}\|}$$

对于任意超平面 $\forall\gamma \neq 0$

$$\tilde{\mathbf{w}}^T\mathbf{x}+\tilde{b}=0 \iff \frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}+\frac{\tilde{b}}{\gamma}=0$$

$\exists\gamma$ 使得 $1=\min_i\left|\frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}_i+\frac{\tilde{b}}{\gamma}\right|$

此时间隔恰为 $\frac{2}{\|\mathbf{w}\|}$



支持向量机

问题转化

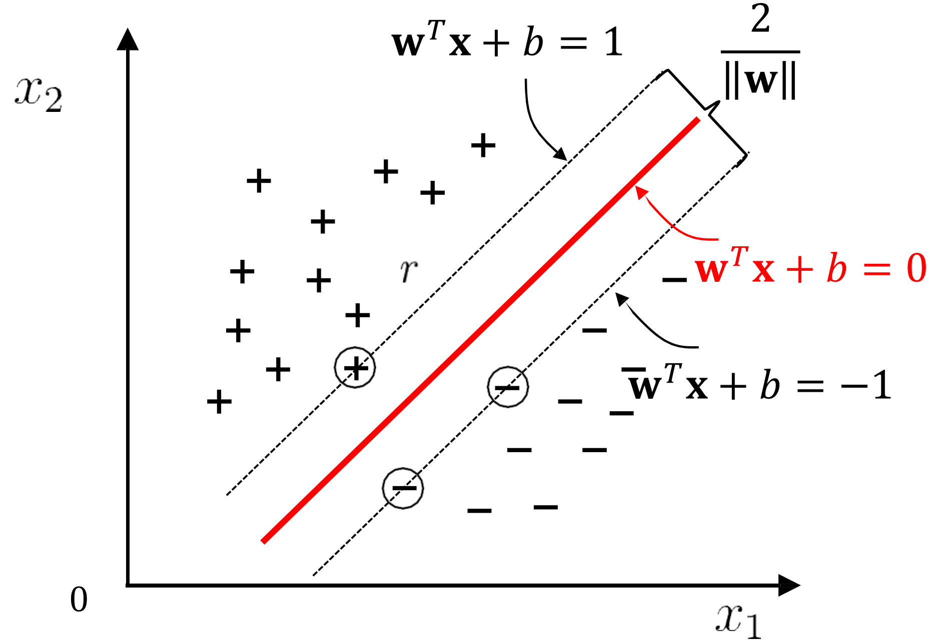
$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_{i} \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i+\tilde{b}\right)}{\|\tilde{\mathbf{w}}\|}$$

对于任意超平面

$$\tilde{\mathbf{w}}^T\mathbf{x}+\tilde{b}=0 \quad \xleftrightarrow{\forall\gamma\neq 0} \quad \frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}+\frac{\tilde{b}}{\gamma}=0$$

$\exists\gamma$ 使得 $1=\min_{i}\left|\frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}_i+\frac{\tilde{b}}{\gamma}\right|$

此时间隔恰为 $\frac{2}{\|\mathbf{w}\|}$



令得$\mathrm{w}=\frac{\tilde{\mathbf{w}}}{\gamma}$，$b=\frac{\tilde{b}}{\gamma}$，则原问题转化为：

$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_{i} \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i+\tilde{b}\right)}{\|\tilde{\mathbf{w}}\|} \quad \Longrightarrow \quad \max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$\text{s.t.} \quad y_i\left(\mathbf{w}^T\mathbf{x}_i+b\right)\geq 1 \ \ (i=1,2,\ldots,N)$$

支持向量机

问题转化

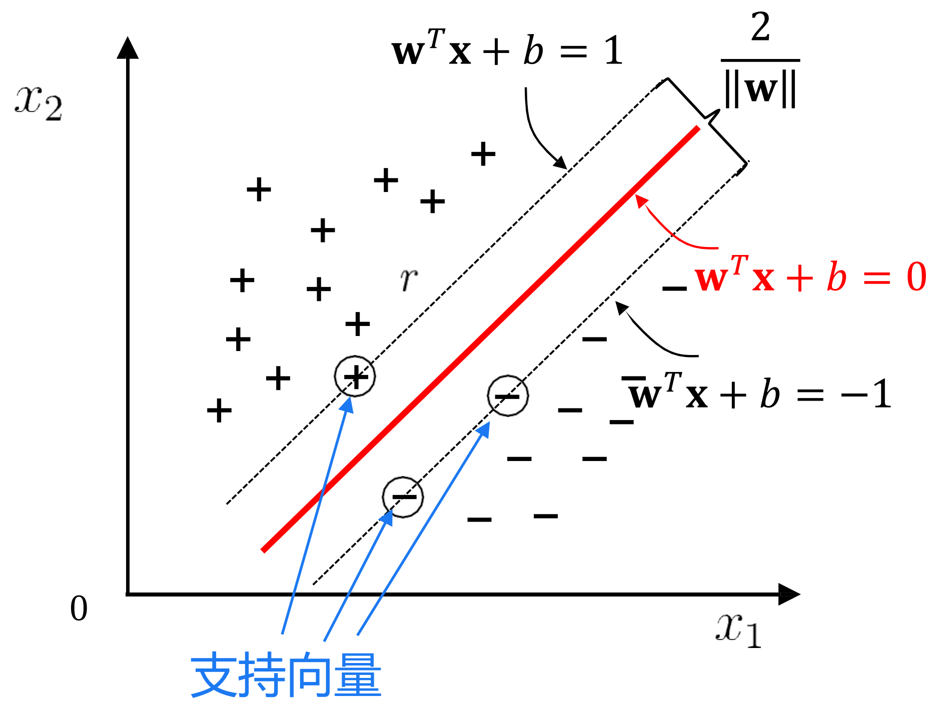
$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_i \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i+\tilde{b}\right)}{\|\tilde{\mathbf{w}}\|}$$

对于任意超平面 $\forall \gamma \neq 0$

$$\tilde{\mathbf{w}}^T\mathbf{x}+\tilde{b}=0 \iff \frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}+\frac{\tilde{b}}{\gamma}=0$$

$\exists\gamma$ 使得 $1=\min_i\left|\frac{\tilde{\mathbf{w}}^T}{\gamma}\mathbf{x}_i+\frac{\tilde{b}}{\gamma}\right|$

此时间隔恰为 $\frac{2}{\|\mathbf{w}\|}$



令得$\mathrm{w}=\frac{\tilde{\mathbf{w}}}{\gamma}$，$b=\frac{\tilde{b}}{\gamma}$，则原问题转化为：

$$\max_{\tilde{\mathbf{w}},\tilde{b}} \min_i \frac{y_i\left(\tilde{\mathbf{w}}^T\mathbf{x}_i+\tilde{b}\right)}{\|\tilde{\mathbf{w}}\|} \Rightarrow \max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$\text{s.t. } y_i(\mathbf{w}^T\mathbf{x}_i+b)\geq 1 \ \ (i=1,2,\dots,N)$$

支持向量机

对偶问题

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|} \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \Longleftrightarrow \quad \min_{\mathbf{w},b} \frac{\|\mathbf{w}\|^2}{2} \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

- 拉格朗日乘子法
  - 第一步：引入拉格朗日乘子$\alpha_i \geq 0 \ (i = 1, \dots, N)$得到拉格朗日函数

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{N} \alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1)$$

  - 第二步：构建对偶目标函数

$$\max_{\boldsymbol{\alpha} \geqslant \mathbf{0}} \min_{\mathbf{w},b} L(\mathbf{w}, b, \boldsymbol{\alpha}) \Rightarrow \begin{cases} \dfrac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \mathbf{0} \\ \dfrac{\partial L}{\partial b} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases} \Rightarrow \begin{cases} \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

(若原优化问题为凸，上方优化问题与原问题等价)

支持向量机

对偶问题

- 拉格朗日乘子法
  - 第三步：回代得出对偶目标函数

第二步所得

$$\begin{cases} \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \end{cases}$$

代入

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

$$= \frac{1}{2}\left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i\right)^T \left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i\right) - \sum_{i=1}^{N} \alpha_i \left(y_i \left(\left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i\right)^T \mathbf{x}_i + b \right) - 1 \right)$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

对偶问题

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \alpha_i \geq 0 \ (i = 1, \dots N)$$

支持向量机

对偶问题

- 求解对偶问题

$$\max_{\boldsymbol{\alpha}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i y_i \alpha_i y_i \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{s.t.} \quad \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \alpha_i y_i = 0, \alpha_i \geq 0 \; (i = 1, \dots N)$$

- 第一步：按照规则选取一对需要更新的变量α_i, α_j
- 第二步：固定其他变量，只针对α_i, α_j 进行优化

$$\max_{\alpha_i, \alpha_j} \; \alpha_i + \alpha_j + \frac{1}{2}\left(\alpha_i^2 \mathbf{x}_i^T \mathbf{x}_i + \alpha_j^2 \mathbf{x}_j^T \mathbf{x}_j + 2\alpha_i \alpha_j \mathbf{x}_i^T \mathbf{x}_j\right) + Const$$

$$\text{s.t.} \quad \alpha_j = \frac{1}{y_j}\left(\alpha_i y_i - \sum_{k \neq i,j} \alpha_k y_k\right), \qquad \alpha_i, \alpha_j \geq 0$$

单变量凸二次优化
（课下推导练习）

- 第三步：重复上述操作，直到算法收敛

支持向量机

对偶问题

- 求解对偶问题
  - 第四步：变量回代，得到原问题的最优解

$$\mathbf{w}^\star = \sum_{i=1}^{N} \alpha_i^\star y_i \mathbf{x}_i$$

利用互补松弛条件，所有$\alpha_i^\star > 0$对应的 $\mathbf{x}_i$ 为支持向量

$$b^\star = -\frac{1}{2}\left(\underset{\alpha_i^\star>0, y_i=1}{\text{mean}}\left\{\mathbf{x}_i^T\mathbf{w}^\star - 1\right\} + \underset{\alpha_i^\star>0, y_i=-1}{\text{mean}}\left\{\mathbf{x}_i^T\mathbf{w}^\star + 1\right\}\right)$$

- 测试
  - 对于测试样本$\mathbf{z}$

$$\begin{cases} \text{正类}, & \mathbf{z}^T\mathbf{w}^\star + b^\star > 0 \\ \text{负类}, & \mathbf{z}^T\mathbf{w}^\star + b^\star < 0 \end{cases}$$

支持向量机

例子

- 假设给定训练集为

$$\text{正类}(y=+1)\text{：}\quad \{(1,2),(1,3),(2,3),(2,4)\}$$

$$\text{负类}(y=-1)\text{：}\quad \{(3,2),(4,2),(2,1),(3,1)\}$$

- 下面哪个参数是SVM学习得到的结果（单选）

A	B	C	D
$\mathbf{w}=[-2,1]$ $b=2$	$\mathbf{w}=[1,1]$ $b=-3$	$\mathbf{w}=[-1,1]$ $b=0$	$\mathbf{w}=[-1,2]$ $b=-2$

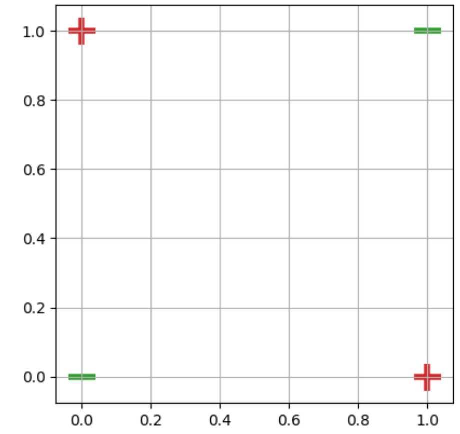
支持向量机

线性不可分挑战

- 异或分类问题

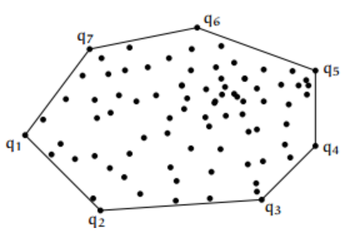
$$正类(y=+1):\quad \{(0,1),(1,0)\}$$

$$负类(y=-1):\quad \{(0,0),(1,1)\}$$



- 线性可分适用范围

线性可分当且仅当正类样本的凸包与负类样本的凸包不存在交集



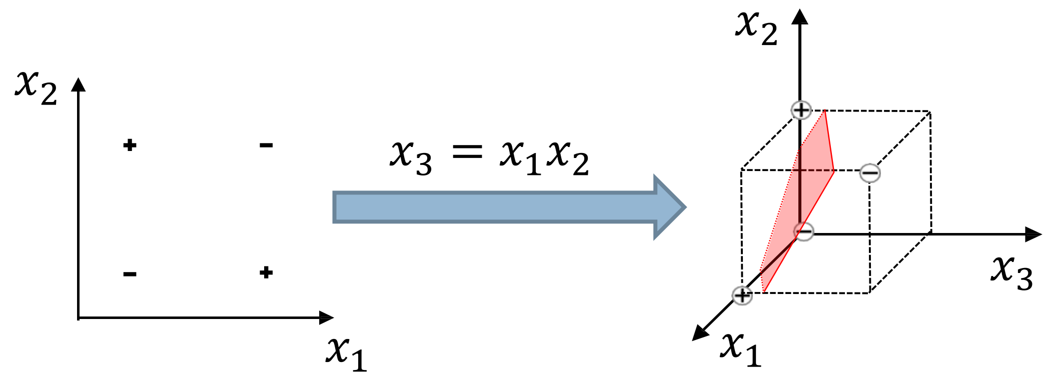
- 如何使SVM适用于非线性可分情况?

支持向量机

核方法

- 基本思想

  通过构造特征变到高维空间从而线性可分



- 引入特征映射$\phi(\mathbf{x})$升维

  一般维度越高，线性可分性越好，希望能够映射至无穷维

$$\phi(\mathbf{x}) \in \mathbb{R}^\infty$$

$$\mathbf{w}^T \phi(\mathbf{x}) + b = 0$$

如何求得一个无穷维的线性超平面？

支持向量机

核方法

- 对偶表示法

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

$$\text{s.t.} \quad y_i\left(\mathbf{w}^T\phi(\mathbf{x}_i)+b\right) \geq 1$$

原问题

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i y_i \alpha_i y_i \boxed{\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)}$$

$$\text{s.t.} \quad \sum_{i=1}^{N} \alpha_i y_i = 0, \alpha_i \geq 0 \ (i=1,\ldots N)$$

对偶问题

- 核函数

$$\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) = \kappa(\mathbf{x}_i,\mathbf{x}_j)$$

$$\mathbf{K} = \begin{bmatrix} \kappa(\mathbf{x}_1,\mathbf{x}_1) & \kappa(\mathbf{x}_1,\mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_N,\mathbf{x}_1) \\ \kappa(\mathbf{x}_2,\mathbf{x}_1) & \kappa(\mathbf{x}_2,\mathbf{x}_2) & & \kappa(\mathbf{x}_N,\mathbf{x}_2) \\ \vdots & & \ddots & \vdots \\ \kappa(\mathbf{x}_N,\mathbf{x}_1) & \kappa(\mathbf{x}_N,\mathbf{x}_2) & \cdots & \kappa(\mathbf{x}_N,\mathbf{x}_N) \end{bmatrix} \in \mathbb{R}^{N\times N}$$

κ被称为核函数，寻找合适的核函数往往比直接寻找ϕ更加方便

支持向量机

核方法

- 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	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lVert\boldsymbol{x}_i-\boldsymbol{x}_j\rVert^2}{2\delta^2}\right)$	$\delta > 0$为高斯核的带宽(width)
无穷维	拉普拉斯核	$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\lVert\boldsymbol{x}_i-\boldsymbol{x}_j\rVert}{\delta}\right)$	$\delta > 0$
	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, $\beta > 0$, $\theta < 0$

- 选择适合的核函数取决于具体任务，需要用后验方式进行验证

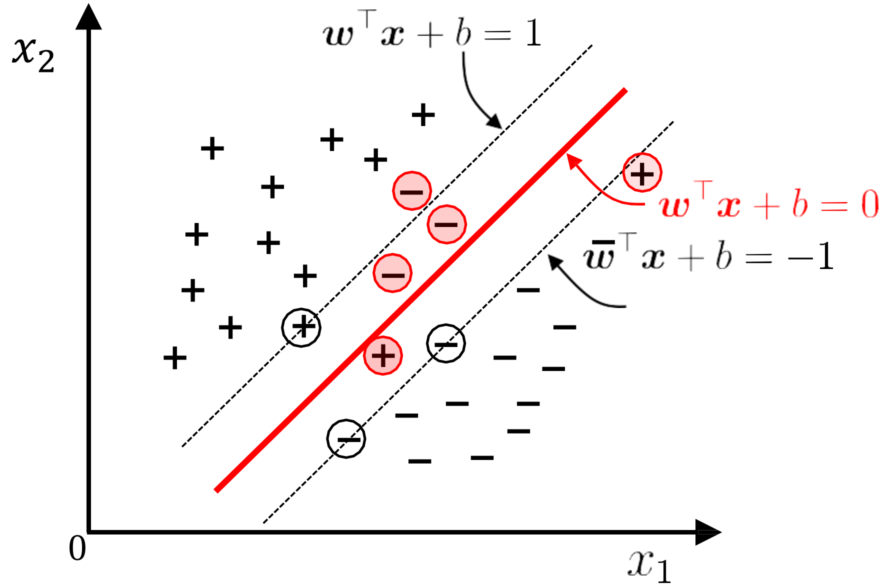
支持向量机

容错策略

◆ 软间隔法

- 现实中很难保证数据线性可分; 即便找到某个核函数使其线性可分，也很难断定是否是有过拟合造成的。

红色：不满足约束的样本



- 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。

支持向量机

容错策略

- 软间隔法

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

SVM原始优化问题

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{N} \xi_i$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad (i = 1, \dots, N)$$

ξ_i为第i个训练样本的容错间隔

多分类问题

- 1 vs Others

对于每个类别训练一个SVM，将当前类别当做正类，其他类别当做负类

支持向量机

支持向量回归

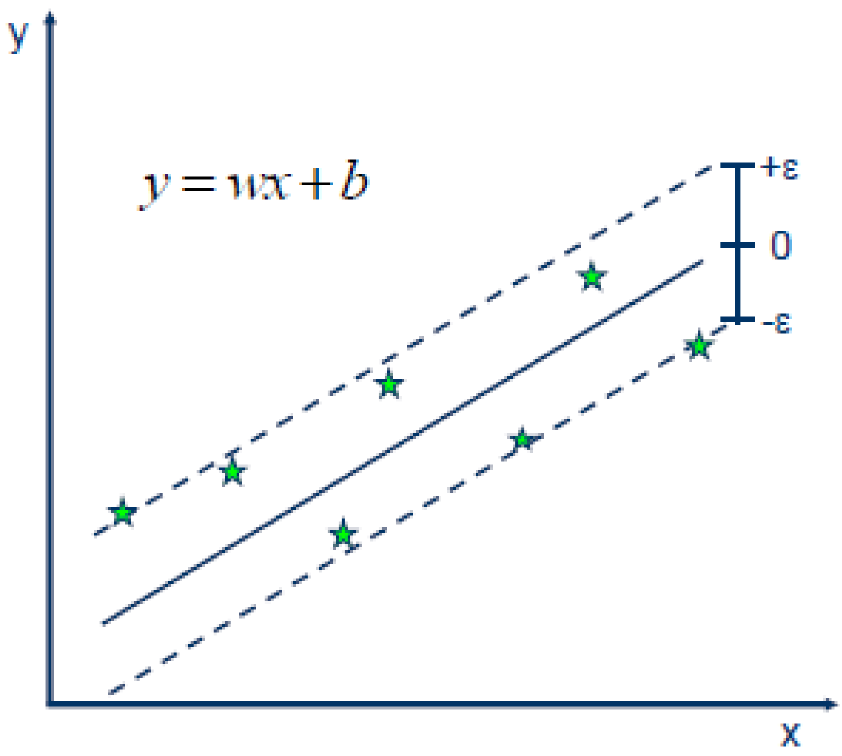
- 软间隔法

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

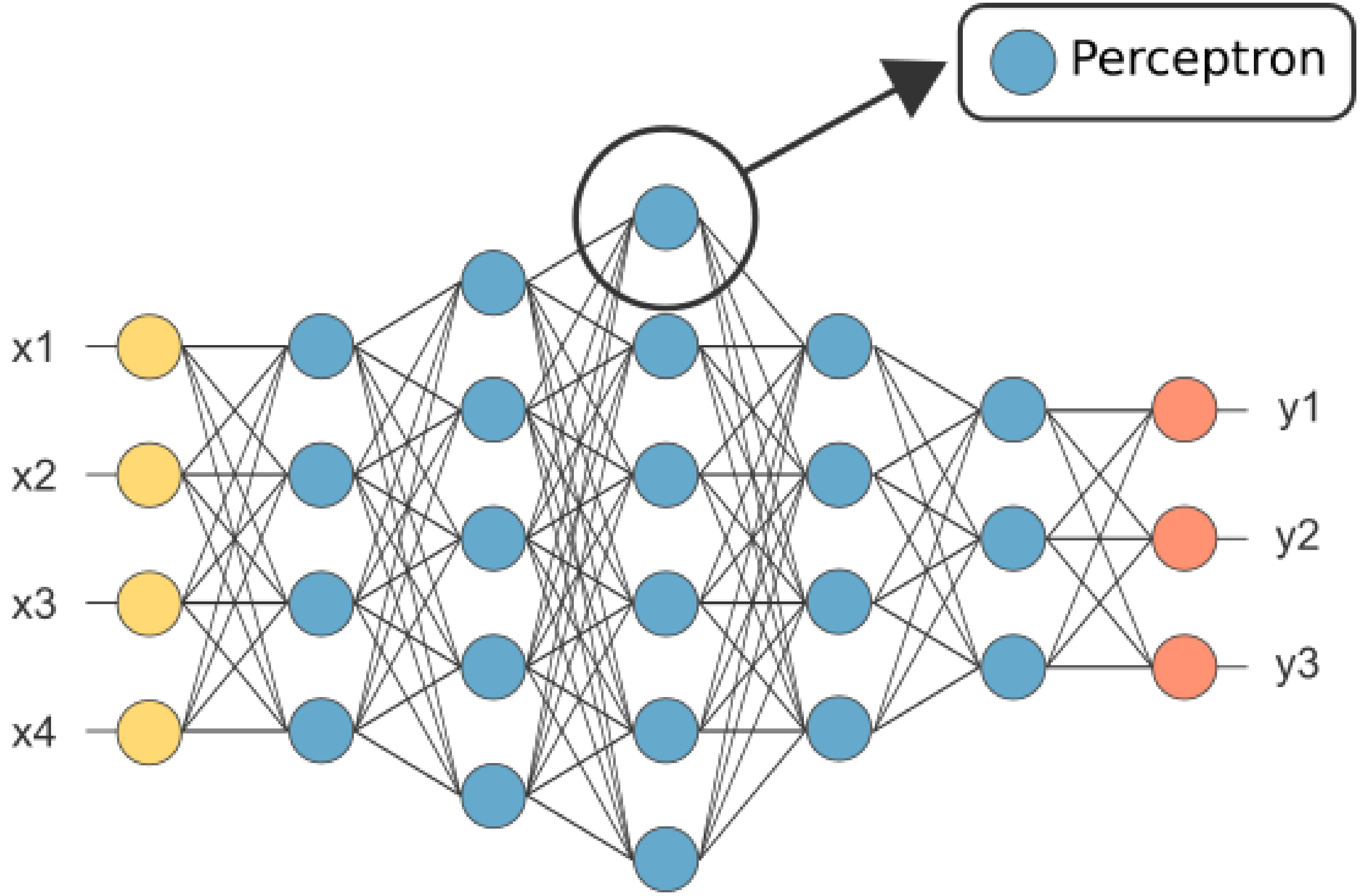
$$\text{s.t.} \ \ |\mathbf{w}^T\mathbf{x}_i + b - y_i| \geq \epsilon$$

- 本质上是使用了特殊损失函数的线性回归

- 核方法与软间隔法同样适用



感知机与神经网络



内容概要

- 内容回顾
- 感知机
- 支持向量机
- **总结**

总结

- 感知机
  - 感知机的生物学原理
  - 感知机的几何意义
  - 感知机的求解
- 支持向量机
  - 最大分类间隔
  - 拉格朗日函数与对偶优化
  - 核方法解决线性不可分问题
  - 容错机制、多分类问题、支持向量回归
  - 感知机与神经网络的关系

北京交通大学
BEIJING JIAOTONG UNIVERSITY