



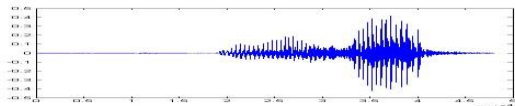
華東師範大學
EAST CHINA NORMAL UNIVERSITY

机器学习

机器学习 \approx 构建一个映射函数

□ 语音识别

$f($



$) = \text{“你好”}$

□ 图像识别

$f($



$) = \text{“猫”}$

□ 围棋

$f($



$) = \text{“5-5”}$ (落子位置)

□ 对话系统

$f($

“你好”

用户输入

$) = \text{“今天天气真不错”}$

机器输出

机器学习基本概念（以监督学习为例）

□ 输入空间（input space）

- 所有可能输入取值的集合，输入变量用 X 表示

□ 输出空间（output space）

- 所有可能输出取值的集合，输入变量用 Y 表示
- 输入空间和输出空间都可以是有限，也可以是无限
- 一般输出空间远远小于输入空间

□ 特征空间（feature space）

- 每个具体的输入是一个实例，通常由特征向量表示
- 所有特征向量存在的空间称为特征空间
- 将实例从输入空间映射到特征空间

机器学习基本概念（以监督学习为例）

- 输入变量的取值用 x 表示

$$x^{(i)} = (x_1, x_2, \dots, x_n)^T$$

- 输出变量的取值用 y 表示

- 训练集（training set）和测试集（test set）

$$T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

- 样本点 $(x^{(i)}, y^{(i)})$

- 回归问题、分类问题和标注问题

机器学习基本概念（以监督学习为例）

□ 联合概率分布

- 假设输入与输出随机变量遵循联合概率分布 $P(X, Y)$
- 联合概率分布的具体定义是未知的。
- 训练数据和测试数据依联合概率分布独立同分布

□ 假设空间 (hypothesis space)

- 输入空间到输出空间的映射的集合
- 决策函数 $Y = f(X)$
- 条件概率分布 $P(Y|X)$

什么是机器学习？

机器学习：从数据中获得决策（预测）函数使得机器可以根据数据进行自动学习，通过算法使得机器能从大量历史数据中学习规律从而对新的样本做决策。

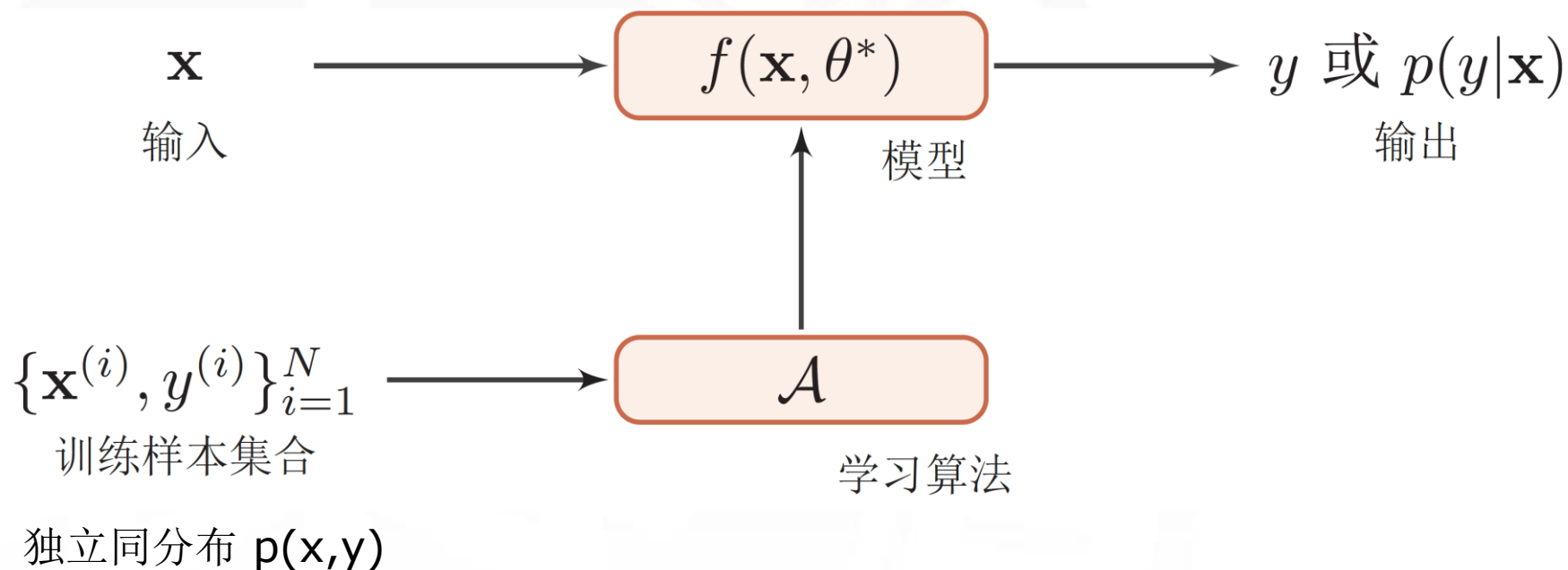
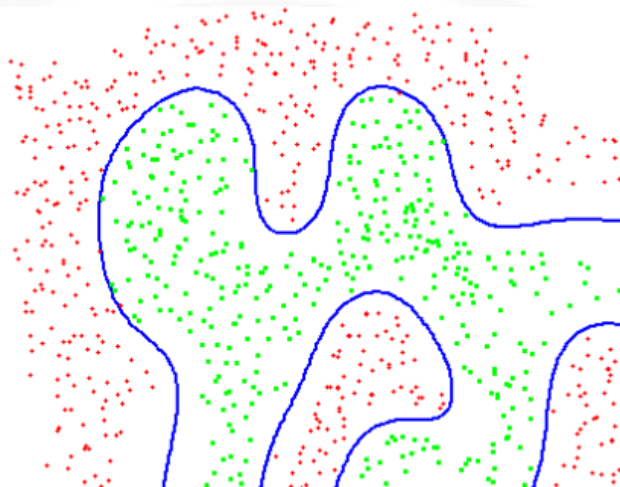
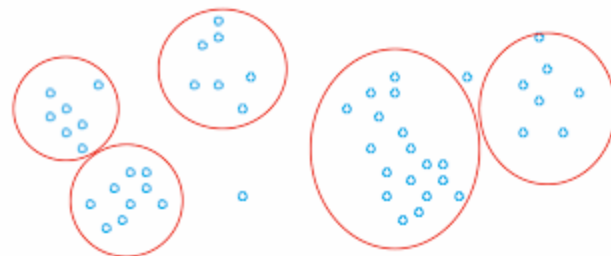


图 监督学习问题

常见的机器学习问题



分类



聚类

常见的机器学习类型

	监督学习	无监督学习	强化学习
训练样本	训练集 $\{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$	训练集 $\{\mathbf{x}^n\}_{n=1}^N$	智能体和环境交互的 轨迹 τ 和累积奖励 G_τ
优化目标	$y = f(\mathbf{x})$ 或 $p(y \mathbf{x})$	$p(\mathbf{x})$ 或带隐变量 \mathbf{z} 的 $p(\mathbf{x} \mathbf{z})$	期望总回报 $\mathbb{E}_\tau[G_\tau]$
学习准则	期望风险最小化 最大似然估计	最大似然估计 最小重构错误	策略评估 策略改进

机器学习三要素

□ 学习三要素：模型+策略+算法

□ 模型 (What)

■ $\mathcal{F} = \{f | Y = f_{\theta}(X), \theta \in R^n\}$ 或

$\mathcal{F} = \{P | P_{\theta}(Y|X), \theta \in R^n\}$

■ 线性方法: $f(x, \theta) = w^T x + b$

■ 非线性方法: $f(x, \theta) = w^T \phi(x) + b$

➤ 如果 $\phi(x)$ 为可学习的非线性基函数, $f(x, \theta)$ 就等价于神经网络

■ 神经网络

■ 概率图模型

策略

□ 策略 (Why)

■ 损失函数: $L(Y, f(X))$

■ 0-1 损失函数: $L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$

■ 平方损失函数:

$$L(Y, f(X)) = (Y - f(X))^2$$

■ 绝对损失函数:

$$L(Y, f(X)) = |Y - f(X)|$$

■ 对数损失函数:

$$L(Y, P(Y|X)) = -\log P(Y|X)$$

策略

- 风险函数或期望函数:

$$R_{\text{exp}}(f) = \int_{X \times Y} L(y, f(x)) P(x, y) dx dy$$

- 训练集上的经验风险:

$$R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^m L(y_i, f(x_i))$$

- 经验风险最小化 (ERM) : (极大似然估计)

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^m L(y_i, f(x_i))$$

- 结构风险最小化 (SRM) : (最大后验概率估计)

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^m L(y_i, f(x_i)) + \lambda J(f)$$

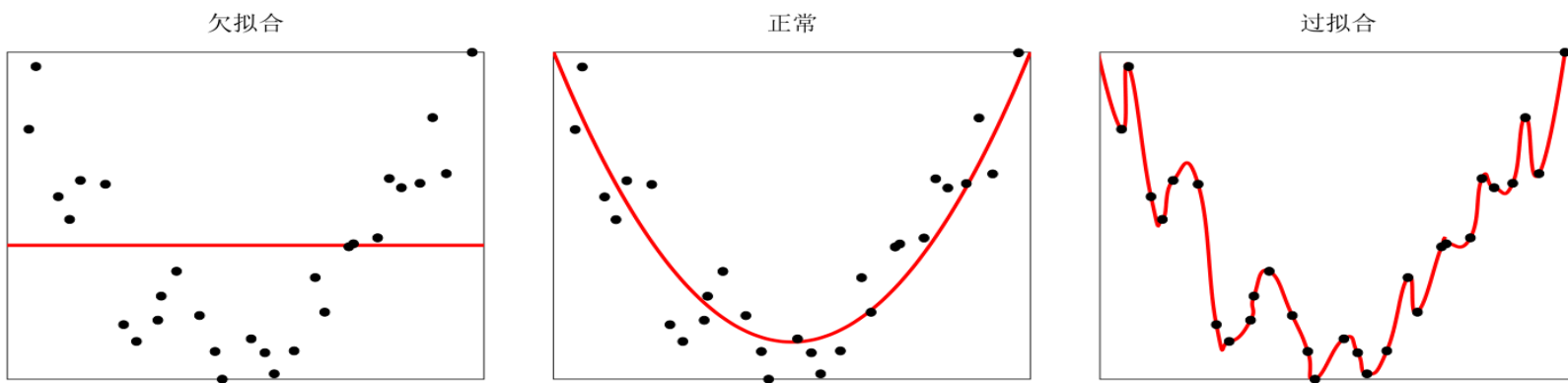
➤ $J(f)$ 是模型的复杂度

- 测试误差: $e_{\text{test}}(f) = \frac{1}{N'} \sum_{i=1}^{N'} L(y_i, f(x_i))$

算法

□ 算法 (How)

- 机器学习基于训练数据集，根据学习策略，通过计算方法从假设空间中选择最优模型。
- 最优化问题：解析解，数值解
- 梯度下降法（见课件）



过拟合：**经验风险最小化原则**很容易导致模型在训练集上错误率很低，但是在未知数据上错误率很高。
过拟合问题往往是由于训练数据少和噪声等原因造成。

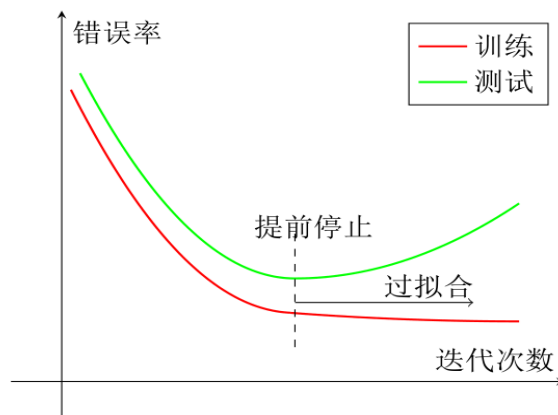
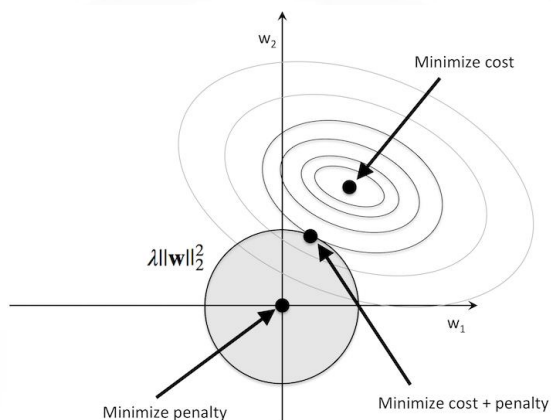
模型选择：正则化

□ 正则化
$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^m L(y_i, f(x_i)) + \lambda J(f)$$

- 第1项是经验风险，第2项是正则化项
- $\lambda \geq 0$ 调整两者之间的关系

□ 正则化方法

- 增加优化约束： L_1 约束 ($\frac{\lambda}{2} \|w\|^2$)、 L_2 约束 $\lambda \|w\|_1$ 数据增强
- 干扰优化过程：权重衰减、随机梯度下降、提前停止



模型选择：交叉验证

- 交叉验证(cross validation)
 - 验证集 (Validation Set):用于模型选择
 - 训练集:用于训练模型
 - 测试集:用于最终对学习方法的评估
- 交叉验证方法
 - 简单交叉验证
 - S折交叉验证
 - 留一交叉验证

泛化误差

□ 泛化误差: $R_{\text{exp}}(f) = \int_{X \times Y} L(y, f(x)) P(x, y) dx dy$

□ 训练误差: $R_{\text{emp}}(f) = \frac{1}{N} \sum_{i=1}^m L(y_i, f(x_i))$

□ 定理（泛化误差上界）对二类分类问题，当假设空间是有限个函数的集合 $\mathcal{F} = \{f_1, f_2, \dots, f_d\}$ ，对任意一个函数 $f \in \mathcal{F}$ ，至少以概率 $1 - \delta$ ，以下不等式成立：

$$R_{\text{exp}}(f) \leq R_{\text{emp}}(f) + \varepsilon(d, N, \delta)$$

$$\text{其中 } \varepsilon(d, N, \delta) = \sqrt{\frac{1}{2N} \left(\log d + \log \frac{1}{\delta} \right)}$$

分类问题的评价指标

- TP: true positive
- FN: false negative
- FP: false positive
- TN: true negative
- 精准率: $P = \frac{TP}{TP+FP}$
- 召回率: $R = \frac{TP}{TP+FN}$
- F_1 值: $\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R}$

生成模型与判别模型

- 生成方法由数据学习概率分布 $P(X, Y)$ ，进而利用

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

- 生成模型：朴素贝叶斯法和隐马尔科夫模型
- 判别方法由数据直接学习决策函数 $f(X)$ 或条件概率分布 $P(Y|X)$ 作为预测的模型。
- 判别模型：K近邻法、感知机、决策树、Logistic回归模型、最大熵模型、支持向量机和条件随机场

生成模型与判别模型

- 生成方法：可还原出联合概率分布 $P(X, Y)$ ，而判别方法不能，生成方法的收敛速度更快，当样本容量增加的时候，学到的模型可以更快地收敛于真实模型；当存在隐变量时，仍然可以使用生成方法，而判别方法则不能用。
- 判别方法：直接学习到条件概率或决策函数，直接进行预测，往往学习的准确率更高；由于直接学习 $f(X)$ 或 $P(Y|X)$ ，可对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习过程。

线性模型

- 线性模型是通过样本特征的线性组合来进行预测的模型。

$$\begin{aligned} f(\mathbf{x}, \mathbf{w}) &= w_1x_1 + w_2x_2 + \cdots + w_dx_d + b \\ &= \mathbf{w}^T \mathbf{x} + b \end{aligned}$$

- 在回归问题中，最小二乘法：采用均方误差最小化
- 在分类问题中，引入非线性的决策函数(decision function) $g(\cdot)$ 来预测输出目标

$$y = g(f(\mathbf{x}, \mathbf{w}))$$

线性模型：线性回归

- 分析不同变量之间存在关系的研究叫回归分析
- 刻画不同变量之间关系的模型被称为回归模型。
如果这个模型是线性的，则称为线性回归模型
- 一旦确定了回归模型，就可以进行预测等分析工作，如从碳排放量预测气候变化程度、从广告投入量预测商品销售等

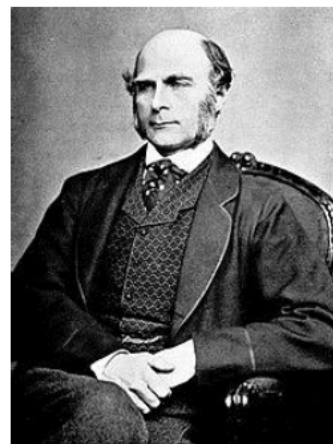
线性模型: 线性回归

$$y = 3.78 + 0.516x$$

y : 子女平均身高

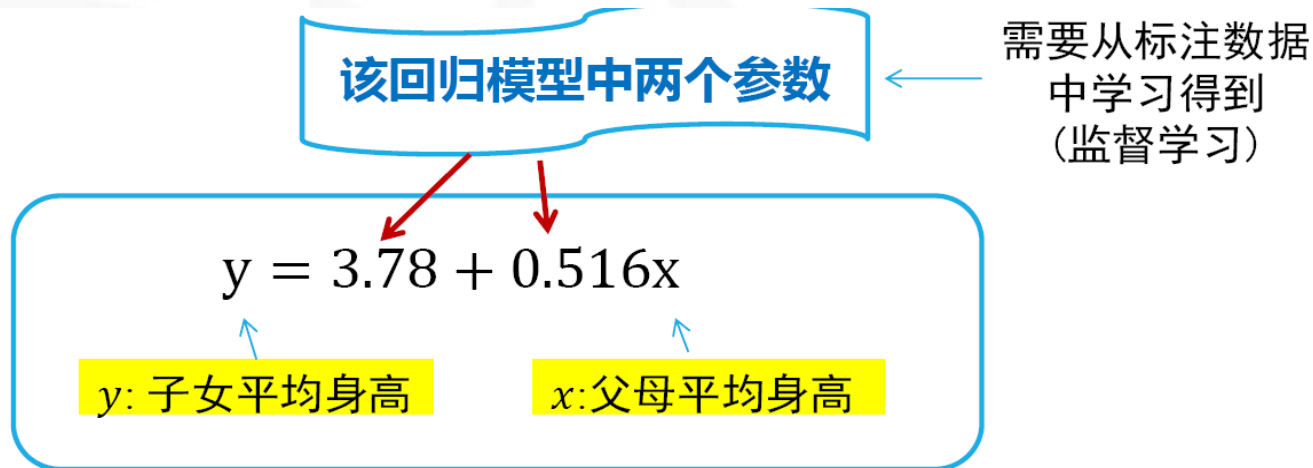
x : 父母平均身高

- 父母平均身高每增加一个单位, 其成年子女平均身高只增加0.516个单位, 它反映了这种“衰退 (regression)”效应 (“回归”到正常人平均身高)。
- 虽然 x 和 y 之间并不总是具有“衰退” (回归) 关系, 但是“线性回归”这一名称就保留下来了。



英国著名生物学家兼
统计学家高尔顿
Sir Francis Galton
(1822-1911)

线性模型：线性回归



- 给出任意一对父母平均身高，则可根据上述方程，计算得到其子女平均身高
- 从父母平均身高来预测其子女平均身高
- 如何求取上述线性方程（预测方程）的参数？

线性模型：线性回归

线性回归模型例子

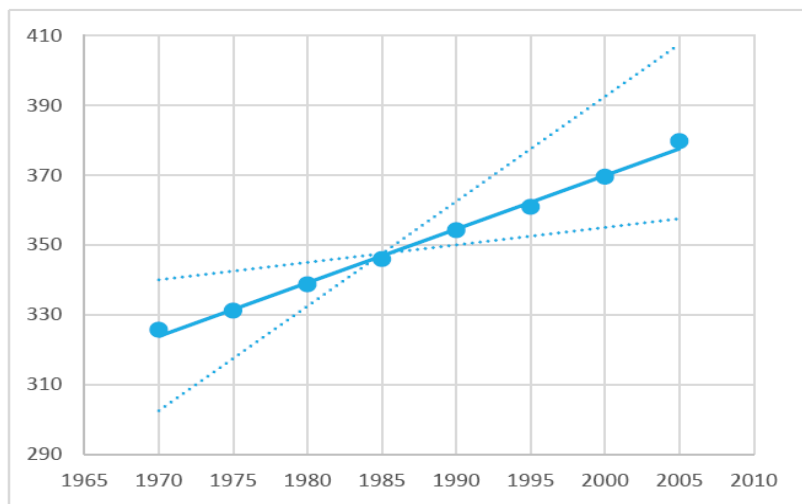
下表给出了莫纳罗亚山（夏威夷岛的活火山）从1970年到2005年每5年的二氧化碳浓度，单位是百万分比浓度（Parts Per Million, ppm）。

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2 (y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

问题：1) 给出1984年二氧化碳浓度值；2) 预测2010年二氧化碳浓度值

线性模型：线性回归

线性回归模型例子



莫纳罗亚山地区时间年份与二氧化碳浓度之间的一元线性回归模型（实线为最佳回归模型）

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2(y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67



代入

回归模型： $y = ax + b$

求取：最佳回归模型是最小化残差平方和的均值，即要求8组 (x, y) 数据得到的残差平均值 $\frac{1}{N} \sum (y - \hat{y})^2$ 最小。残差平均值最小只与参数 a 和 b 有关，最优解即是使得残差最小所对应的 a 和 b 的值。

线性模型：线性回归

回归模型参数求取： $y_i = ax_i + b \ (1 \leq i \leq n)$

- 记在当前参数下第 i 个训练样本 x_i 的预测值为 \hat{y}_i
- x_i 的标注值（实际值） y_i 与预测值 \hat{y}_i 之差记为 $(y_i - \hat{y}_i)^2$
- 训练集中 n 个样本所产生误差总和为： $L(a, b) = \sum_{i=1}^n (y_i - a \times x_i - b)^2$

目标：寻找一组 a 和 b ，使得误差总和 $L(a, b)$ 值最小。在线性回归中，解决如此目标的方法叫最小二乘法。

一般而言，要使函数具有最小值，可对 $L(a, b)$ 参数 a 和 b 分别求导，令其导数值为零，再求取参数 a 和 b 的取值。

线性模型：线性回归

回归模型参数求取： $y_i = ax_i + b$ ($1 \leq i \leq n$) $\min_{a,b} L(a,b) = \sum_{i=1}^n (y_i - a \times x_i - b)^2$

$$\frac{\partial L(a,b)}{\partial a} = \sum_{i=1}^n 2(y_i - ax_i - b)(-x_i) = 0$$

将 $b = \bar{y} - a\bar{x}$ ($\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$, $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$)

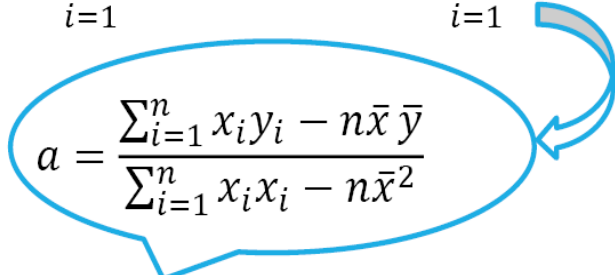
代入上式

$$\rightarrow \sum_{i=1}^n (y_i - ax_i - \bar{y} + a\bar{x})(x_i) = 0$$

$$\rightarrow \sum_{i=1}^n (y_i x_i - ax_i x_i - \bar{y} x_i + a\bar{x} x_i) = 0$$

$$\rightarrow \sum_{i=1}^n (y_i x_i - \bar{y} x_i) - a \sum_{i=1}^n (x_i x_i - \bar{x} x_i) = 0$$

$$\rightarrow (\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}) - a(\sum_{i=1}^n x_i x_i - n\bar{x}^2) = 0$$


$$a = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i x_i - n\bar{x}^2}$$

线性模型：线性回归

回归模型参数求取： $y_i = ax_i + b$ ($1 \leq i \leq n$) $\min_{a,b} L(a,b) = \sum_{i=1}^n (y_i - a \times x_i - b)^2$

$$\frac{\partial L(a,b)}{\partial b} = \sum_{i=1}^n 2(y_i - ax_i - b)(-1) = 0$$

$$\rightarrow \sum_{i=1}^n (y_i - ax_i - b) = 0$$

$$\rightarrow \sum_{i=1}^n (y_i) - a \sum_{i=1}^n x_i - \sum_{i=1}^n b = 0$$

$$\rightarrow n\bar{y} - an\bar{x} - nb = 0 \Rightarrow b = \bar{y} - a\bar{x}$$

可以看出：只要给出了训练样本 (x_i, y_i) ($i = 1, \dots, n$)，我们就可以从训练样本出发，建立一个线性回归方程，使得对训练样本数据而言，该线性回归方程预测的结果与样本标注结果之间的差值和最小。

$$a = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

线性模型：线性回归

回归模型参数求取： $y_i = ax_i + b$ ($1 \leq i \leq n$) $\min_{a,b} L(a,b) = \sum_{i=1}^n (y_i - a \times x_i - b)^2$

$$b = \bar{y} - a\bar{x}$$

$$a = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

年份(x)	1970	1975	1980	1985	1990	1995	2000	2005
CO2(y)	325.68	331.15	338.69	345.90	354.19	360.88	369.48	379.67

训练样本数据



$$a = \frac{x_1 y_1 + x_2 y_2 + \dots + x_8 y_8 - 8\bar{x}\bar{y}}{x_1^2 + x_2^2 + \dots + x_8^2 - 8\bar{x}^2} = 1.5344$$

$$b = \bar{y} - a\bar{x} = -2698.9$$

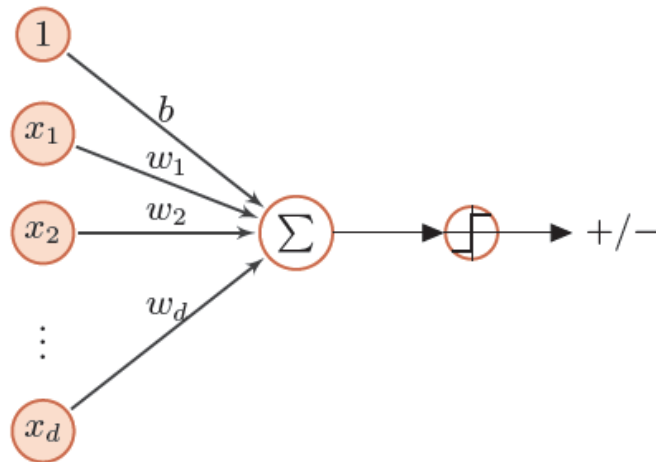
预测莫纳罗亚山地区二氧化碳浓度的一元线性回归模型为“二氧化碳浓度 = $1.5344 \times$ 时间年份 - 2698.9”，即 $y = 1.5344x - 2698.9$ 。

线性模型: 分类问题

- 对于两分类问题, $g(\cdot)$ 可以是符号函数(sign function)

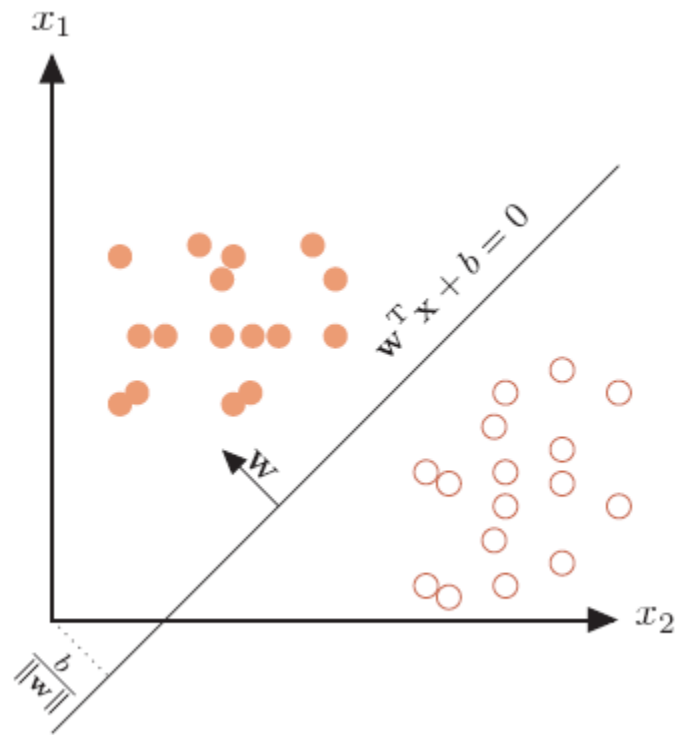
$$g(f(\mathbf{x}, \mathbf{w})) = \text{sgn}(f(\mathbf{x}, \mathbf{w}))$$
$$\triangleq \begin{cases} +1 & \text{if } f(\mathbf{x}, \mathbf{w}) > 0, \\ -1 & \text{if } f(\mathbf{x}, \mathbf{w}) < 0. \end{cases}$$

- 其结构如下图所示:



线性模型: 分类问题

- 在两分类中，对于线性判别函数 $f(\mathbf{x}, \mathbf{w}) = \mathbf{w}^T \mathbf{x} + b$ 特征空间中所有满足 $f(\mathbf{x}, \mathbf{w}) = 0$ 的点组成一个分割超平面，称为决策边界或决策平面。
- 决策边界将特征空间一分为二，划分为两个区域，每个区域对应一个类别。
- 决策边界是线性超平面属于线性分类模型。
- 样本点到决策平面的距离为： $\gamma = \frac{f(\mathbf{x}, \mathbf{w})}{\|\mathbf{w}\|}$



线性模型: 分类问题

- 给定 N 个样本的训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$, 其中 $y^{(n)} \in \{+1, -1\}$
- 线性模型试图学习到参数 \mathbf{w}^* , 使得对于每个样本尽量满足:

$$\begin{aligned} f(\mathbf{x}^{(n)}, \mathbf{w}^*) &> 0 & \text{if } y^{(n)} &= 1 \\ f(\mathbf{x}^{(n)}, \mathbf{w}^*) &< 0 & \text{if } y^{(n)} &= -1 \end{aligned}$$

- 也就是: $y^{(n)} f(\mathbf{x}, \mathbf{w}^*) > 0, \quad \forall n \in [1, N]$

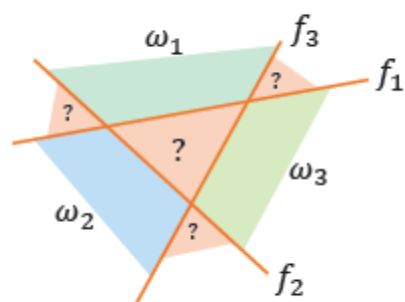
定义 3.1 – 两类线性可分: 对于训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$, 如果存在权重向量 \mathbf{w}^* , 对所有样本都满足 $yf(\mathbf{x}, \mathbf{w}^*) > 0$, 那么训练集 \mathcal{D} 是线性可分的。

- 损失函数: $\mathcal{L}(y, f(\mathbf{x}, \mathbf{w})) = I(yf(\mathbf{x}, \mathbf{w}) > 0)$

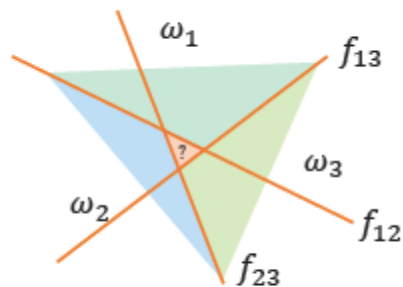
线性模型：多分类问题

- 多分类问题：分类的类别数 $C > 2$
- 一对其余方式：把多分类问题转化 C 个一对其余的两分类问题。
- 一对一方式：把多分类问题转换为 $C(C-1)/2$ 个一对一的两分类问题。
- argmax方式： $y = \arg \max_c f_c(\mathbf{x}, \mathbf{w}_c)$

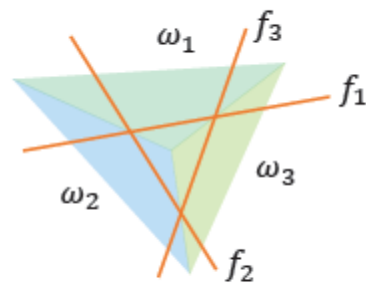
线性模型：多分类问题



(a) “一对其余”方式



(b) “一对一”方式



(c) “argmax”方式

定义 3.2—多类线性可分：对于训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$ ，如果存在 C 个权重向量 $\mathbf{w}_c^*, 1 \leq c \leq C$ ，对所有第 c 类的样本都满足 $f_c(\mathbf{x}, \mathbf{w}_c) > f_{\bar{c}}(\mathbf{x}, \mathbf{w}_{\bar{c}}), \forall \bar{c} \neq c$ ，那么训练集 \mathcal{D} 是线性可分的。

线性模型

- ☐ Logistic 回归
- ☐ Softmax 回归
- ☐ 感知机
- ☐ 支持向量机

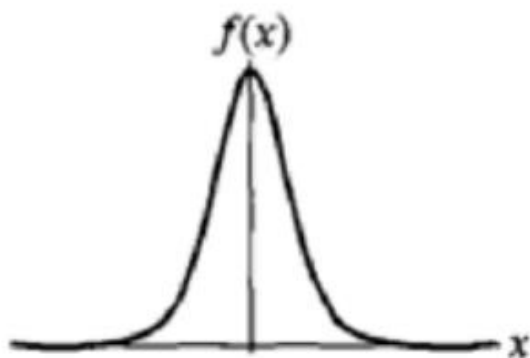
线性模型：Logistic回归

- Logistic分布：设 X 是连续随机变量， X 服从逻辑斯谛分布是指其具有下列分布函数和密度函数：

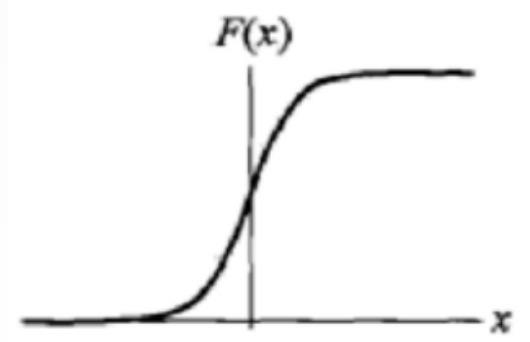
$$F(x) = P\{X \leq x\} = \frac{1}{1+e^{-(x-u)/\gamma}}$$

$$f(x) = F'(x) = \frac{e^{-(x-u)/\gamma}}{\gamma(1+e^{-(x-u)/\gamma})^2}$$

其中， μ 是位置参数， γ 是形状参数



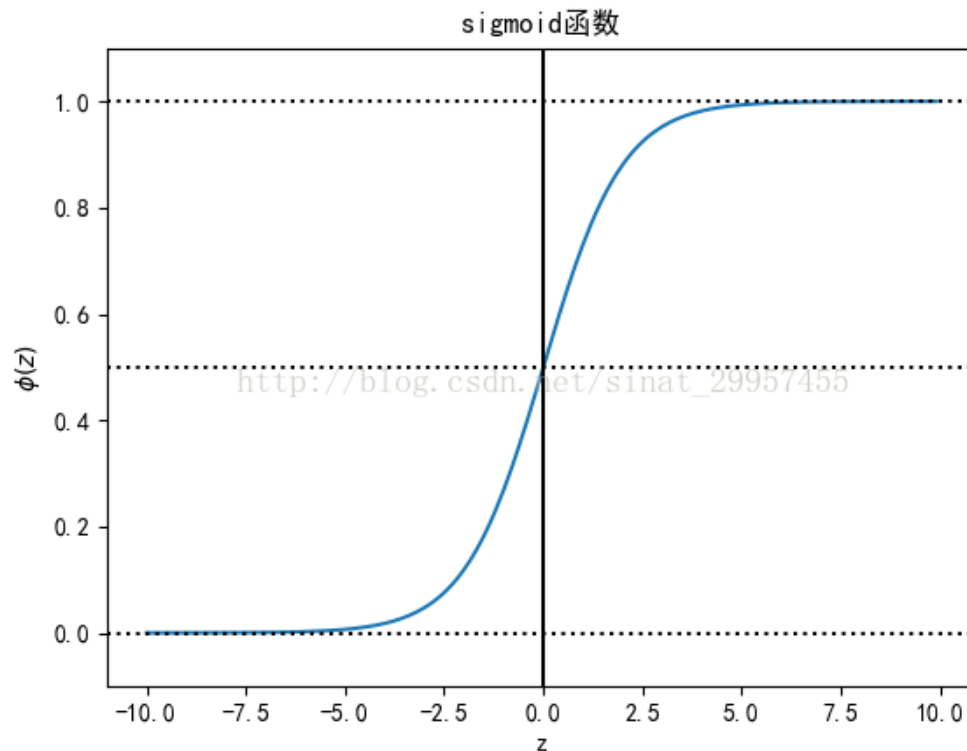
密度函数



分布函数

线性模型：Logistic回归

■ Sigmoid函数: $\phi(z) = \frac{1}{1+e^{-z}}$

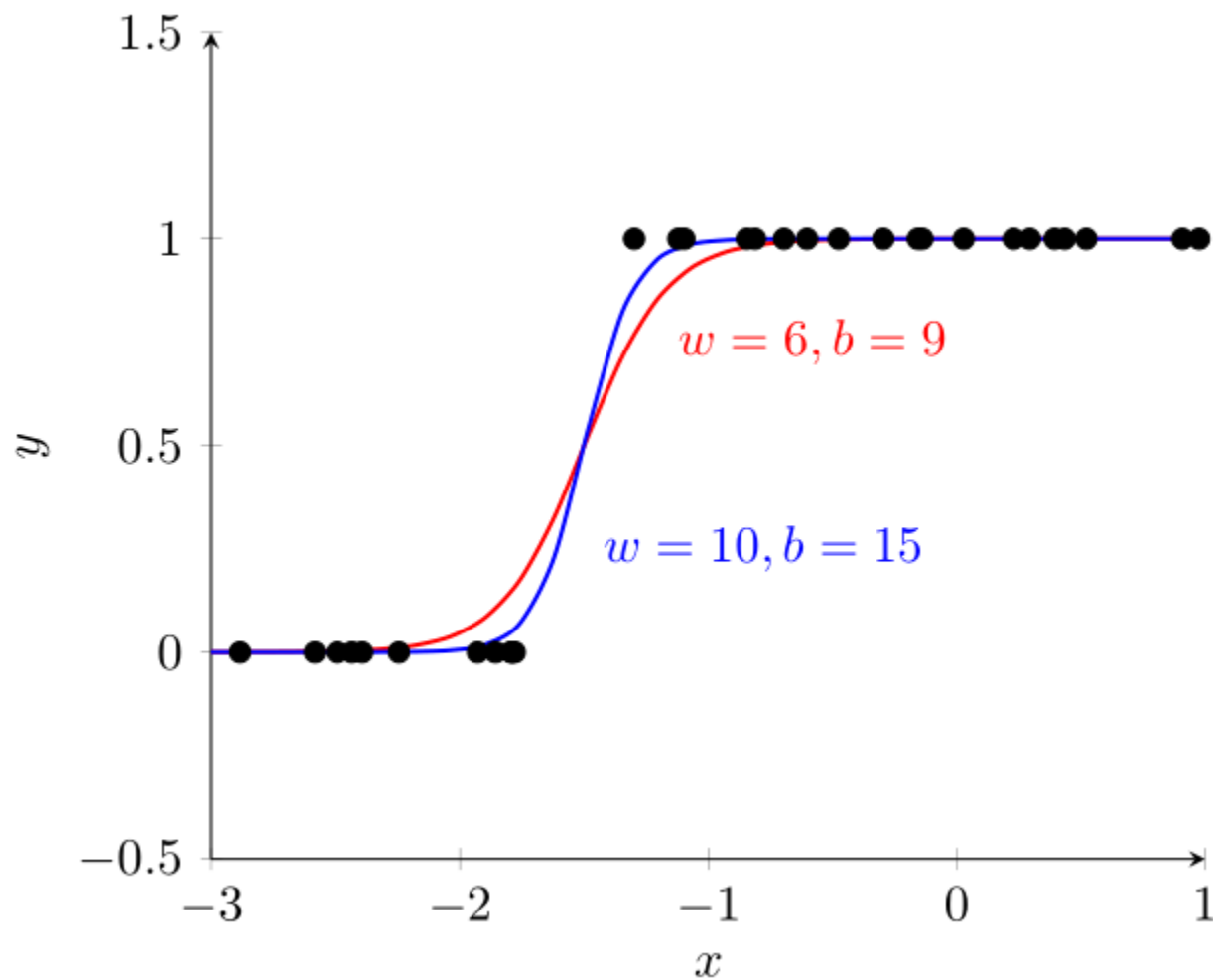


$$\phi'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \phi(z)(1 - \phi(z))$$

线性模型：Logistic回归

- Logistic回归（LR）是一种常见的处理两分类问题的线性模型
- 引入非线性函数 $g: \mathbb{R}^d \rightarrow (0, 1)$ 来预测类别标签的后验概率 $P(y = 1|\mathbf{x})$ ，即：
$$\begin{aligned} P(y = 1|\mathbf{x}) &= g(f(\mathbf{x}, \mathbf{w})) \\ &= \phi(\mathbf{w}^T \mathbf{x}) \\ &= \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})} \end{aligned}$$
- 标签的对数几率： $\mathbf{w}^T \mathbf{x} = \log \frac{P(y=1|\mathbf{x})}{P(y=0|\mathbf{x})}$

线性模型：Logistic回归



线性模型：Logistic回归

■ 设 $P(y = 1|\mathbf{x}) = \phi(\mathbf{w}^T \mathbf{x}) = \phi(z)$, $P(y = 0|\mathbf{x}) = 1 - \phi(z)$

似然函数为: $\prod_{i=1}^N [\phi(z^{(i)})]^{y^{(i)}} [1 - \phi(z^{(i)})]^{1-y^{(i)}}$

对数似然函数为:

$$L(\mathbf{w}) = \sum_{i=1}^N [y^{(i)} \log \phi(z^{(i)}) + (1-y^{(i)}) \log (1-\phi(z^{(i)}))]$$

$$\begin{aligned} \frac{\partial L(\mathbf{w})}{\partial w_j} &= \sum_{i=1}^N \frac{\partial L(\mathbf{w})}{\partial \phi(z^{(i)})} \cdot \frac{\partial \phi(z^{(i)})}{\partial z^{(i)}} \cdot \frac{\partial z^{(i)}}{\partial w_j} \\ &= \sum_{i=1}^N (y^{(i)} \frac{1}{\phi(z^{(i)})} - (1-y^{(i)}) \frac{1}{1-\phi(z^{(i)})}) \phi(z^{(i)}) (1-\phi(z^{(i)})) x_j^{(i)} \\ &= \sum_{i=1}^N (y^{(i)} - \phi(z^{(i)})) x_j^{(i)} \\ &= \sum_{i=1}^N x_j^{(i)} (y^{(i)} - \phi(\mathbf{w}^T \mathbf{x}^{(i)})) \end{aligned}$$

线性模型：Logistic回归

□ 交叉熵损失函数，模型在训练集的风险函数为

$$\mathcal{R}(\mathbf{w}) = -\frac{1}{N} \sum_{i=1}^N \left(y^{(i)} \log \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) + (1 - y^{(i)}) \log \left(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) \right) \right).$$

□ 梯度为

$$\frac{\partial \mathcal{R}(\mathbf{w})}{\partial \mathbf{w}} = \frac{1}{N} \sum_{i=1}^N \left(\mathbf{x}^{(i)} \cdot \left(\sigma(\mathbf{w}^T \mathbf{x}^{(i)}) - y^{(i)} \right) \right)$$

线性模型：Softmax回归

Softmax回归

线性模型：Softmax回归

- Softmax回归是logistic回归的多类推广。

$$\hat{y} = \arg \max_{c=1}^C \mathbf{w}_c^T \mathbf{x}$$

- 利用softmax函数，我们定义目标类别 $y = c$ 的后验概率为：

$$\begin{aligned} P(y = c | \mathbf{x}) &= \text{softmax}(\mathbf{w}_c^T \mathbf{x}) \\ &= \frac{\exp(\mathbf{w}_c^T \mathbf{x})}{\sum_{i=1}^C \exp(\mathbf{w}_i^T \mathbf{x})}. \end{aligned}$$

线性模型：Softmax回归

□ 交叉熵损失函数，模型在训练集的风险函数为

$$\mathcal{R}(W) = -\frac{1}{N} \sum_{i=1}^N (\mathbf{y}^{(i)})^T \log \left(\text{softmax}(W^T \mathbf{x}^{(i)}) \right).$$

□ 梯度为

$$\frac{\partial \mathcal{R}(W)}{\partial W} = -\frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)} \left(\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)} \right)^T.$$

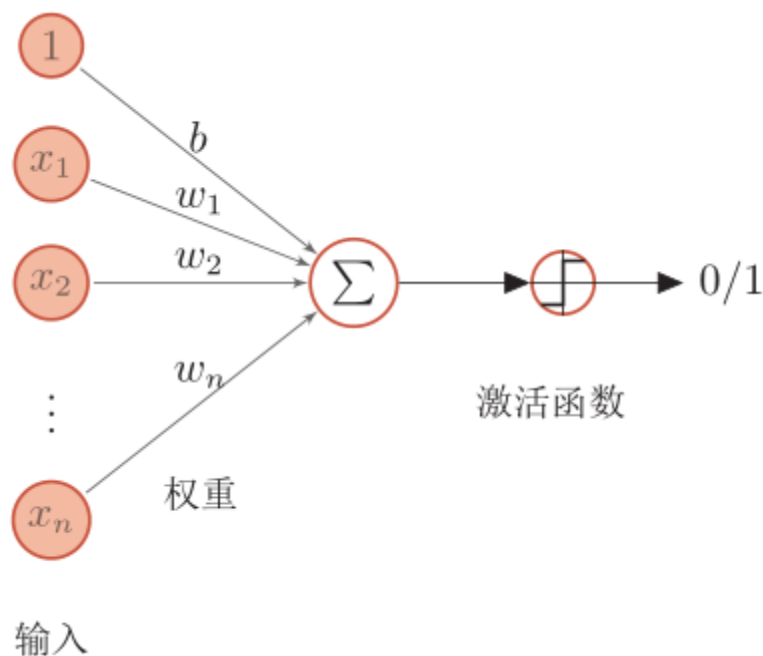
线性模型：感知机

感知机

感知机

- 模拟生物神经元行为的机器，有与生物神经元相对应的部件，如权重（突触）、偏置（阈值）及激活函数（细胞体），输出为0或1。

$$\hat{y} = \begin{cases} +1 & \text{当 } \mathbf{w}^T \mathbf{x} > 0 \\ -1 & \text{当 } \mathbf{w}^T \mathbf{x} \leq 0 \end{cases},$$



感知器的学习过程

输入: 训练集: $(\mathbf{x}_i, y_i), i = 1, \dots, N$, 迭代次数: T

```
1 初始化:  $\mathbf{w}_0 = 0$  ;  
2  $k = 0$  ;  
3 for  $t = 1 \dots T$  do  
4   for  $i = 1 \dots N$  do  
5     选取一个样本  $(\mathbf{x}_i, y_i)$ , if  $\mathbf{w}^T(y_i \mathbf{x}_i) < 0$  then  
6        $\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$ , ;  
7        $k = k + 1$ ;  
8     end  
9   end  
10 end  
    输出:  $\mathbf{w}_k$ 
```

表示分错

对比Logistic回归的更新方式:

$$\frac{\partial \mathcal{R}(W)}{\partial W} = -\frac{1}{N} \sum_{i=1}^N \mathbf{x}^{(i)} \left(\mathbf{y}^{(i)} - \hat{\mathbf{y}}^{(i)} \right)^T.$$

感知器的学习过程

- 在两分类问题中，当数据集是线性可分时，存在一个正的常数 $\gamma > 0$ 和权重向量 \mathbf{w}^* ，且 $\|\mathbf{w}^*\| = 1$ ，则对所有点都满足 $(\mathbf{w}^*)^T(y\mathbf{x}) \geq \gamma$ 。

定理 3.1 – 感知器收敛性：给定一个训练集 $\mathcal{D} = \{(\mathbf{x}^{(n)}, y^{(n)})\}_{n=1}^N$ ，假设 R 是训练集中最大的特征向量的模，

$$R = \max_n \|\mathbf{x}^{(n)}\|.$$

如果训练集 \mathcal{D} 线性可分，感知器学习算法3.1的权重更新次数不超过 $\frac{R^2}{\gamma^2}$ 。

感知机

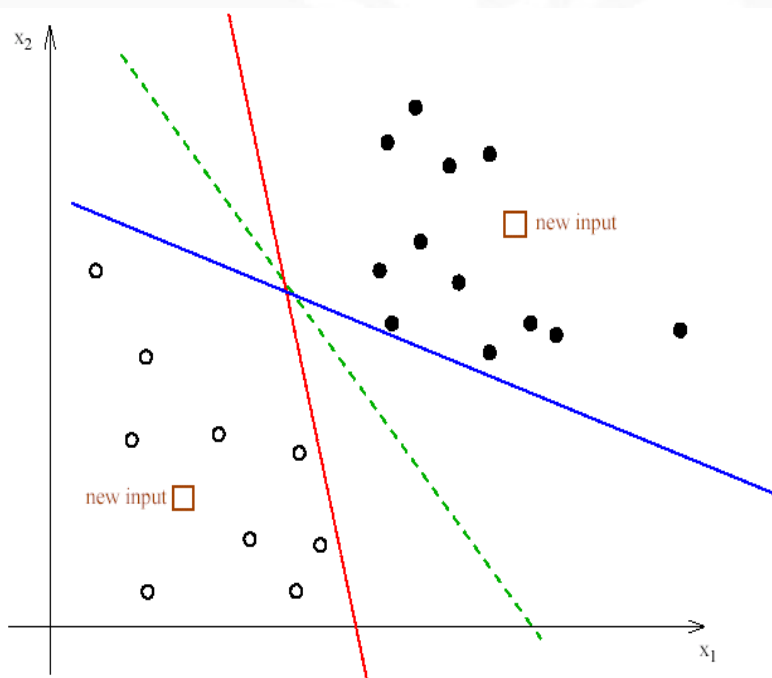
- ❑ 在数据集线性可分时，感知机虽然可以找到一个超平面把两类数据分开，但并不能保证其泛化能力。
- ❑ 感知器对样本顺序比较敏感。每次迭代的顺序不一致时，找到的分割超平面也往往不一致。
- ❑ 如果训练集不是线性可分的，就永远不会收敛。

线性模型：支持向量机

支持向量机

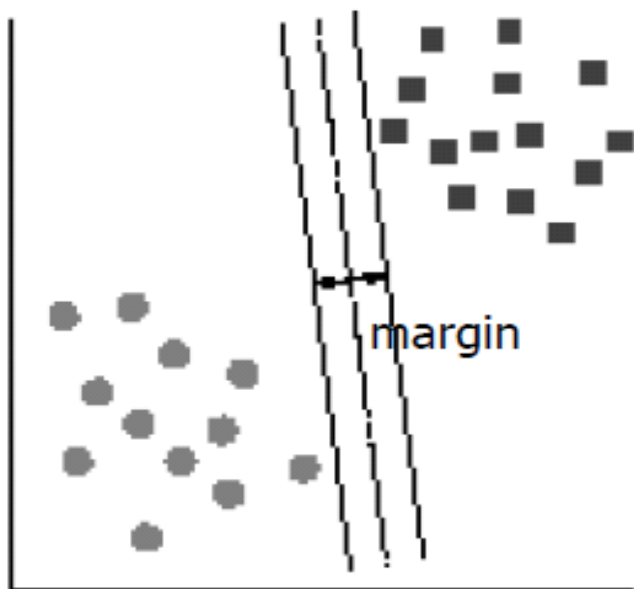
支持向量机 (SVM)

- **目标：** 找到一个超平面，使得它能够尽可能多的将两类数据点正确的分开，同时使分开的两类数据点距离分类面最远。

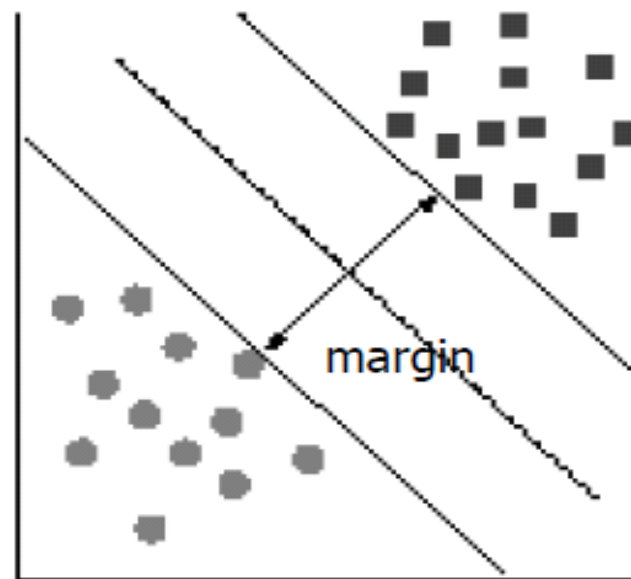


线性可分情形

支持向量机 (SVM)



(a)



(b)

最大边缘超平面(MMH)

边缘：从超平面到其边缘的侧面的最短距离等于到其边缘的另一个侧面的最短距离，边缘侧面平行于超平面

支持向量机 (SVM)

- 数据集中每个样本到分割超平面的距离

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

- 我们定义整个数据集中所有样本到分割超平面的最短距离为间隔: $\gamma = \min_n \gamma^{(n)}$

- 支持向量机的目标: $\max_{\mathbf{w}, b} \gamma$

$$s.t. \quad \frac{y^{(n)}(\mathbf{w}^T \mathbf{x}^{(n)} + b)}{\|\mathbf{w}\|} \geq \gamma$$

- 令 $\|\mathbf{w}\| \cdot \gamma = 1$, 该目标等价于: $\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|^2}$

$$s.t. \quad y^{(n)}(\mathbf{w}^T \mathbf{x}^{(n)} + b) \geq 1$$

支持向量机 (SVM)

□ 数据集中所有满足 $y^{(n)}(\mathbf{w}^T \mathbf{x}^{(n)} + b) = 1$ 的样本点，都称为支持向量

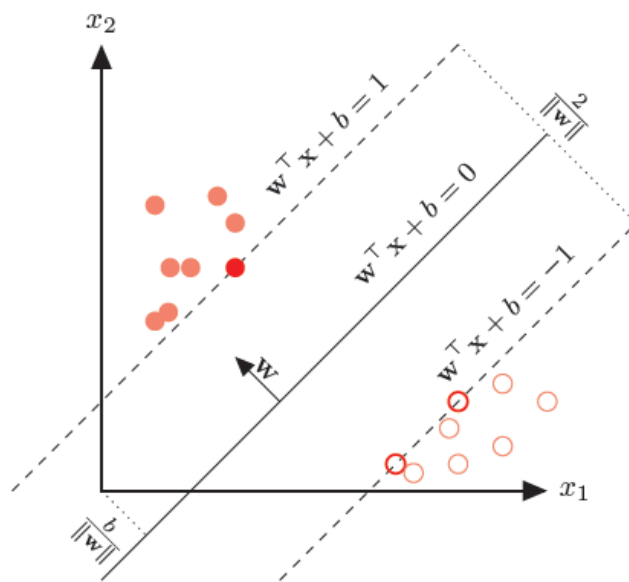
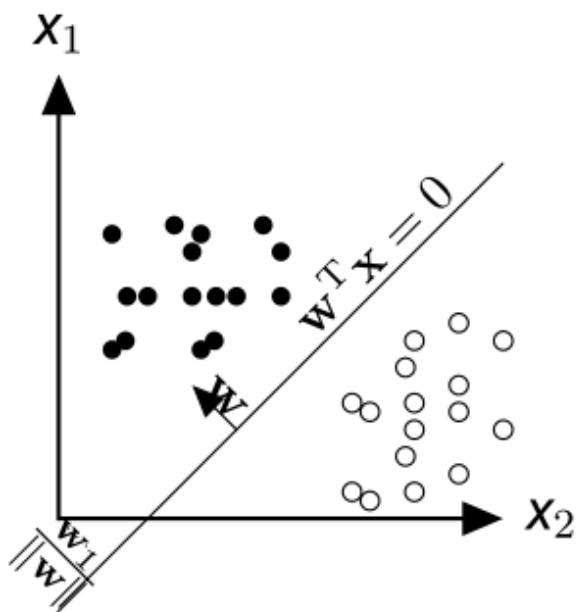


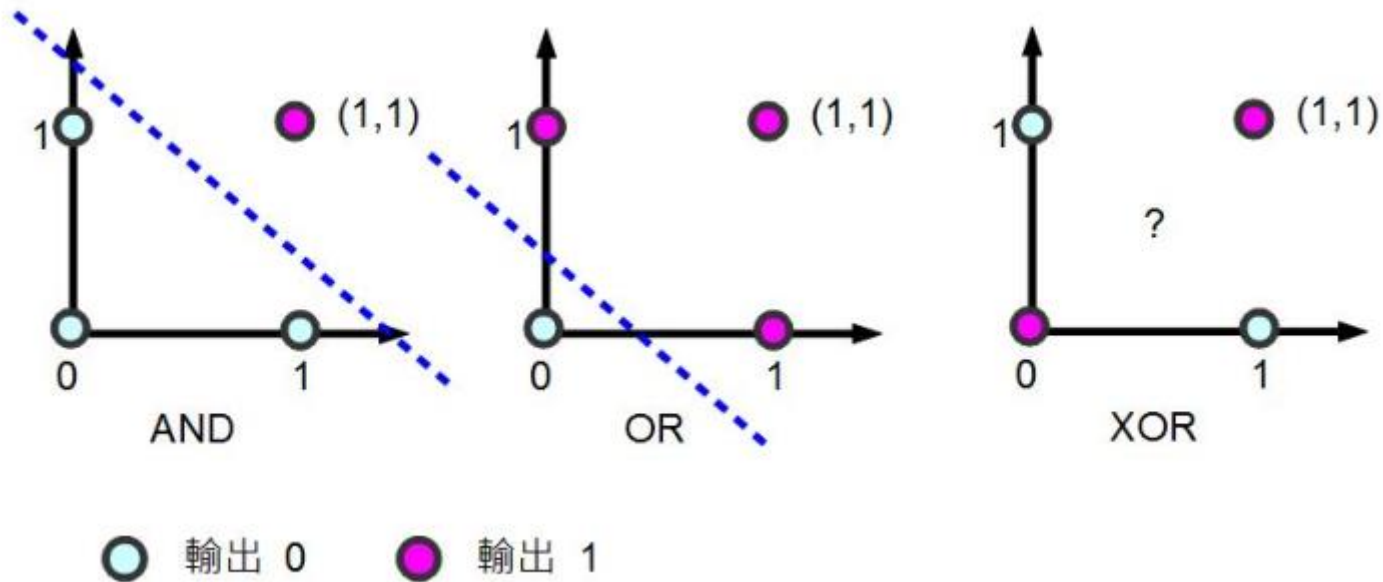
图 3.6 支持向量机示例

线性分类器小结

	损失函数	优化方法
线性回归	平方误差	最小二乘、梯度下降
Logistic 回归	交叉熵	梯度下降
感知器	0-1 损失	$\mathbf{w}_{k+1} = \mathbf{w}_k + y_i \mathbf{x}_i$
支持向量机	Hinge 损失	SMO 等



XOR问题





THE END