中山大學
SUN YAT-SEN UNIVERSITY

第3章 线性模型

1. 基本形式
2. 回归问题
3. 分类问题

第3章 线性模型

1. 基本形式
2. 回归问题
3. 分类问题

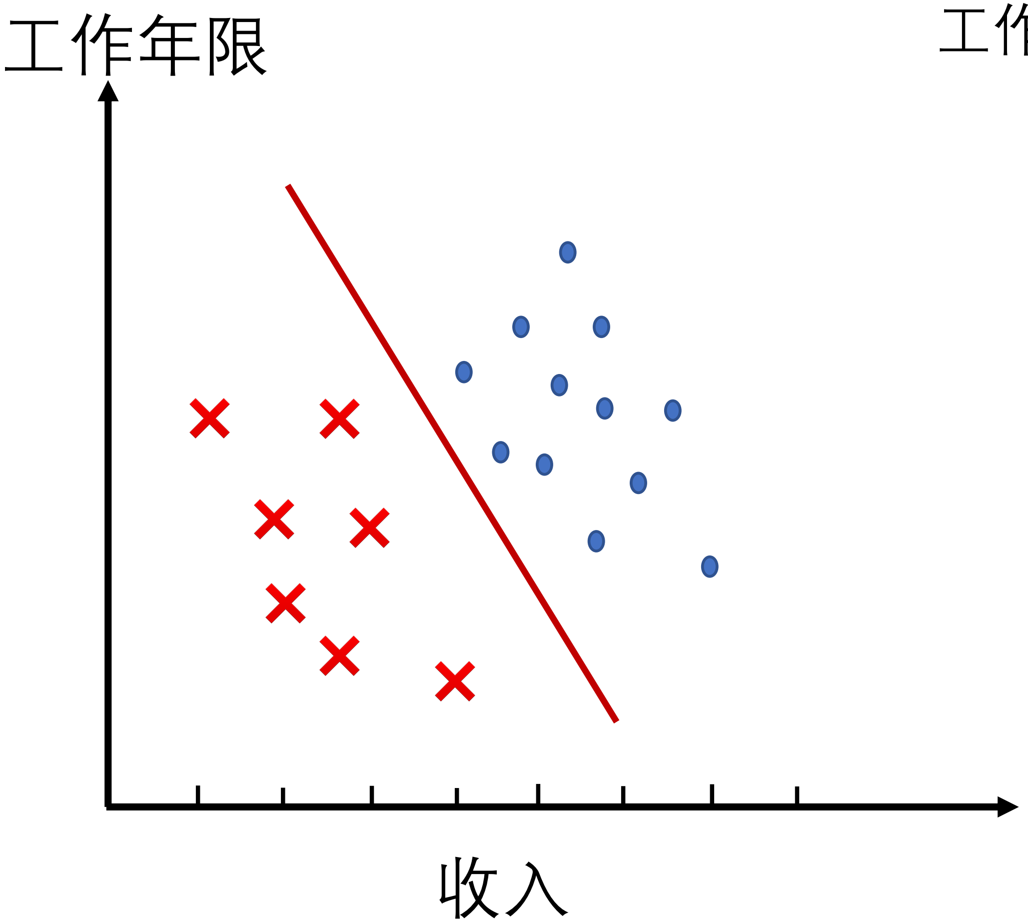
分类问题

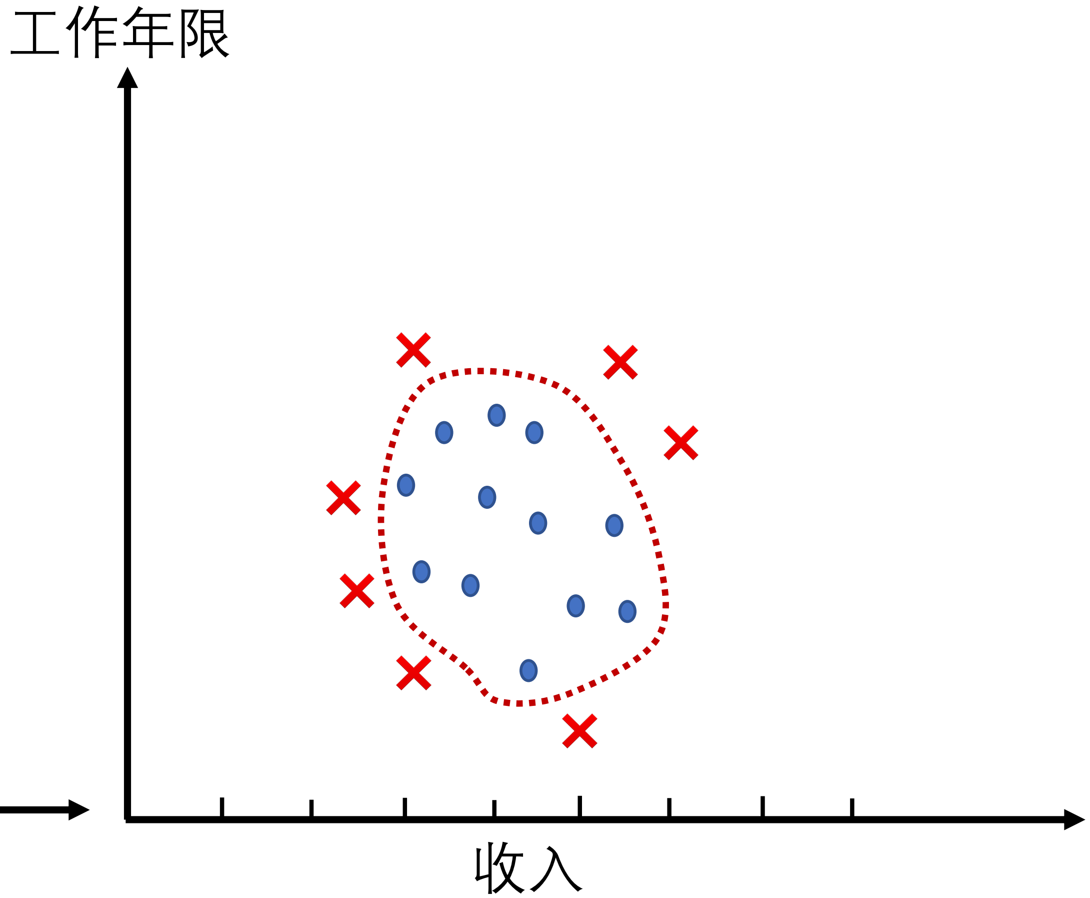
（1）什么是分类问题

（2）对数几率回归模型建立

（3）多分类学习

分类问题





分类问题

- **已知：**
  - 训练集 X， x ∈ X
  - 分类 $C = \{c_1, c_2, \dots, c_n\}$

- **决定：**
  - X 到 C 的一个映射关系
  - 二分类： $C = \{0, 1\}$ 或 $C = \{-1, +1\}$

  - 概率：

$$\mathrm{P}_{\boldsymbol{\omega}}(y|x) \quad \mathrm{P}(y|x; \boldsymbol{\omega})$$

$$\mathrm{P}(y = 1|x; \boldsymbol{\omega}) \qquad \mathrm{P}(y = 0|x; \boldsymbol{\omega}) = 1 - \mathrm{P}(y = 1|x; \boldsymbol{\omega})$$

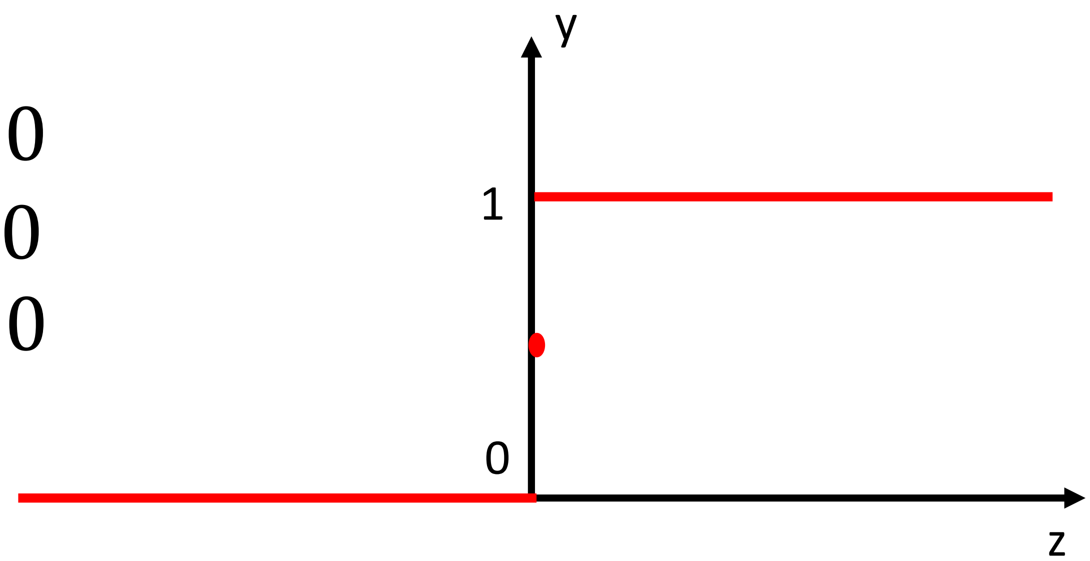
怎么转化成 {0,1}

- 线性回归输出

$$f(\boldsymbol{x}) = \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + b = z \qquad\qquad g(z) \in \{0,1\}$$

- 最理想的函数——单位阶跃函数

$$g(z) = \begin{cases} 0, & z < 0 \\ 0.5 & z = 0 \\ 1 & z > 0 \end{cases}$$



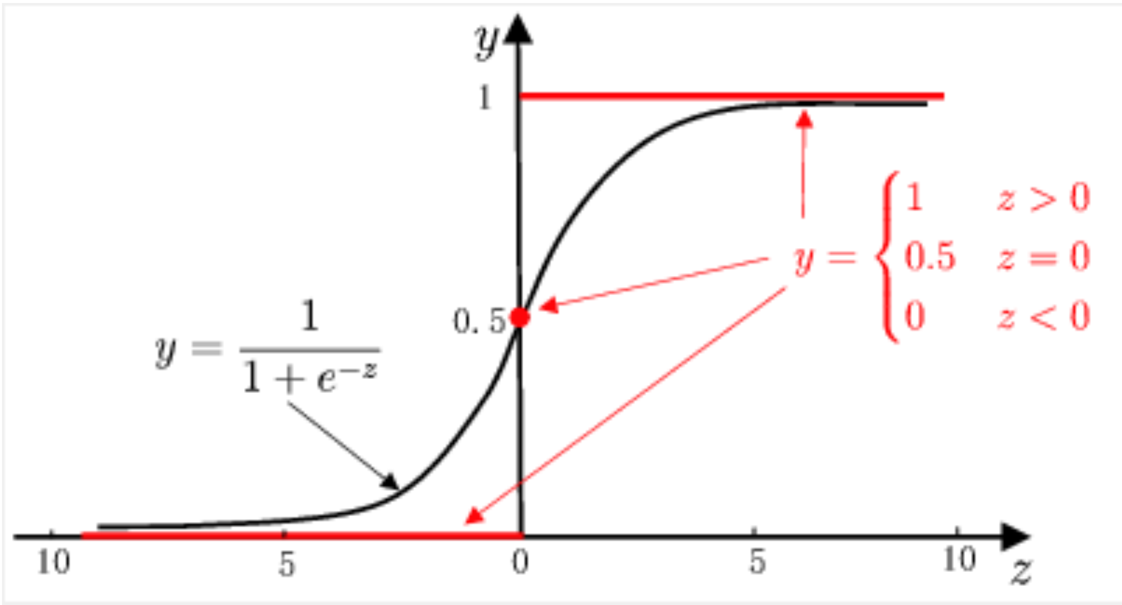
不连续

对数几率函数

- 替代函数——对数几率函数（logistic function）
  - ➢ 优点：单调可微、任意阶可导

单位阶跃函数与对数几率函数的比较

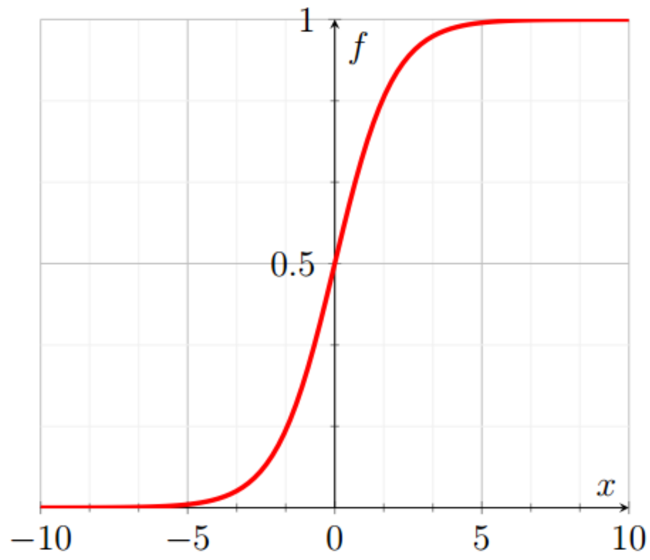
$$g(z) = \frac{1}{1+e^{-z}}$$



假设函数： $h_\omega(\boldsymbol{x}) = f(\boldsymbol{x}) = g(\boldsymbol{\omega}^T \boldsymbol{x} + b)$

对数几率函数性质

$$g(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}$$



$g(-\infty) = 0 \quad g(+\infty) = 1$

$g(0) = 0.5$

- 连续，单调可微
- Sigmoid function of z

$$g'(z) = g(z)(1 - g(z))$$

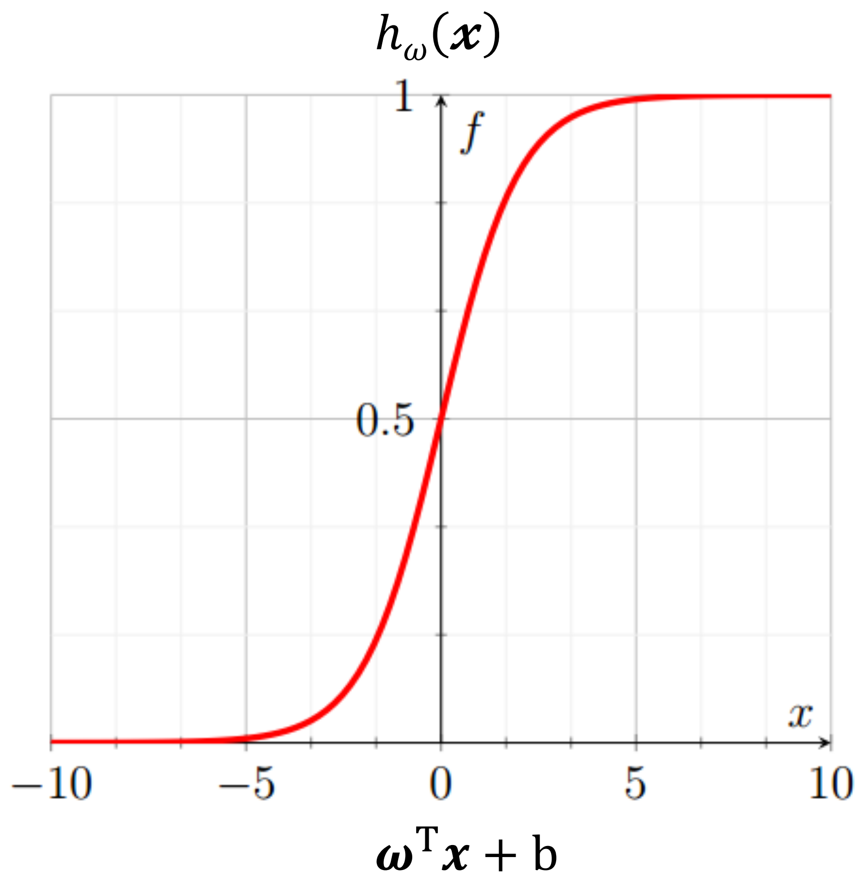
$$\begin{aligned} g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\ &= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\ &= \frac{1}{(1+e^{-z})} \cdot \left(1 - \frac{1}{(1+e^{-z})}\right) \\ &= g(z)(1-g(z)). \end{aligned}$$

判别规则和判别边界

- 判别规则： $h_w(\boldsymbol{x}) = f(\boldsymbol{x}) = g(\boldsymbol{\omega}^T\boldsymbol{x} + b)$
  - 预测 y=1 $h_\omega(\boldsymbol{x}) \geq 0.5$ $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + \mathrm{b} \geq 0$
  - 预测 y=0 $h_\omega(\boldsymbol{x}) < 0.5$ $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + \mathrm{b} < 0$

- 判别边界：

$$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x} + \mathrm{b} = 0$$



分类问题

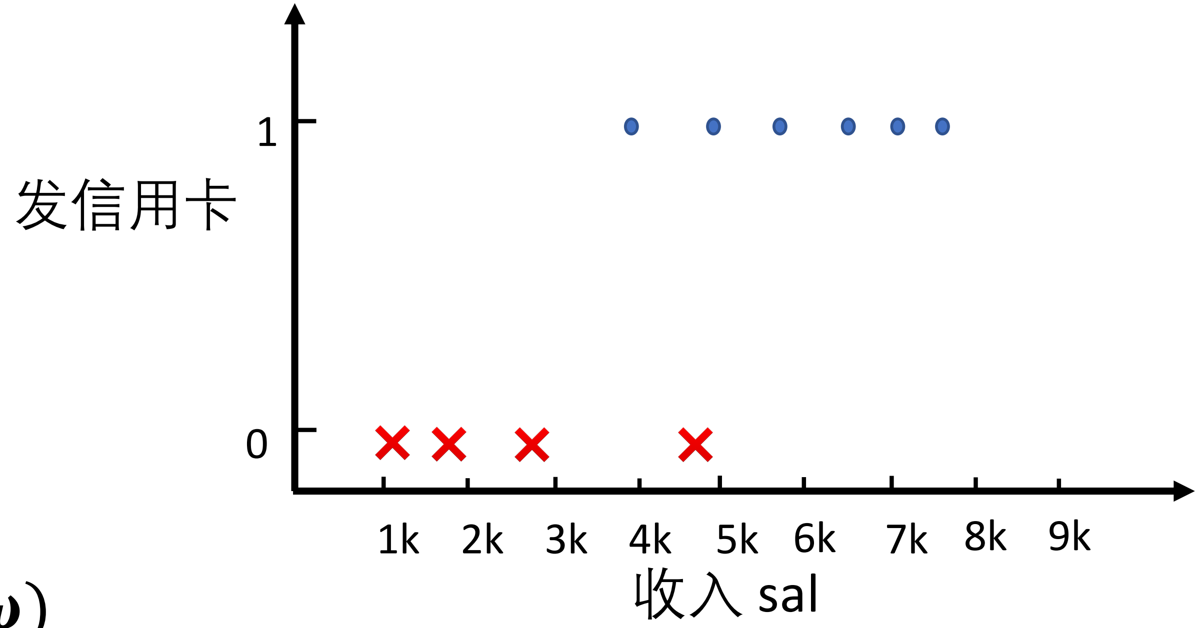
（1）什么是分类问题

（2）对数几率回归模型建立

~~（3）多分类学习~~

假设函数输出

收入：**sal**



$h_{\boldsymbol{\omega}}(\mathrm{sal})$代表什么？

$$h_{\boldsymbol{\omega}}(sal) = P(y = 1 \mid \boldsymbol{x};\ \boldsymbol{\omega})$$

$$P(y = 0 \mid \boldsymbol{x}\,; \boldsymbol{\omega}) = 1 - P(y = 1 \mid \boldsymbol{x}\,; \boldsymbol{\omega}) \quad = 1 - h_{\boldsymbol{\omega}}(\boldsymbol{x})$$

$$P(y = 1|\boldsymbol{x}; \boldsymbol{\omega}) = h_{\omega}(\boldsymbol{x})$$

$$P(y = 0|\boldsymbol{x}; \boldsymbol{\omega}) = 1 - h_{\omega}(\boldsymbol{x})$$

$$h_{\omega}(\boldsymbol{x}) = f(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{\omega}^T \boldsymbol{x})}$$

注意：此处引入 $x_0 = 1, \omega_0 = b$

Likelihood（似然）

$$h_\omega(\boldsymbol{x}) = f(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{\omega}^T\boldsymbol{x})}$$

$\mathrm{P}(y=1|x;\boldsymbol{\omega})=h_\omega(\boldsymbol{x})$ $\quad$ $\mathrm{P}(y=0|x;\boldsymbol{\omega})=1-h_\omega(\boldsymbol{x})$

- 假设我们有如下样本：

$$D=\{(x_1,+1),(x_2,0),\ldots,(x_m,0)\}$$

样本属于其真实标签的概率

$\mathrm{P}(y=1|x;\boldsymbol{\omega})$ $\quad$ $\mathrm{P}(y=0|x;\boldsymbol{\omega})$ $\quad$ … $\quad$ $\mathrm{P}(y=0|x;\boldsymbol{\omega})$

$\mathrm{P}(y=1|x;\boldsymbol{\omega})$ $\quad$ $1-\mathrm{P}(y=1|x;\boldsymbol{\omega})$ $\quad$ … $\quad$ $1-\mathrm{P}(y=1|x;\boldsymbol{\omega})$

$$P(y_i|\boldsymbol{x}_i;\omega)=y_i\mathrm{P}(y=1|x;\boldsymbol{\omega})+(1-y_i)\mathrm{P}(y=0|x;\boldsymbol{\omega})$$

- Likelihood（似然）

$$L(y|\boldsymbol{x};\boldsymbol{\omega})=\prod_{i=1}^{m}P(y_i|\boldsymbol{x}_i;\boldsymbol{\omega})$$

- 找到一个合适的 $h_\omega(\boldsymbol{x})$

$$\boxed{\max L(y|\boldsymbol{x};\boldsymbol{\omega})}$$

联乘？？

对数似然 $L(y|x;\omega)=\prod_{i=1}^{m} P(y_i|\boldsymbol{x}_i;\omega)$

- 对数似然：

$$\mathcal{L}(y|x;\omega)=\sum_{i=1}^{m} \ln P(y_i|\boldsymbol{x}_i;\omega)$$

- 最大化对数似然：

$$\max \mathcal{L}(y|x;\omega)$$

习惯解决最小化问题

对数几率回归（Logistic Regression）问题

$$\text{Min } \mathcal{L}(y|x;\omega)=-\sum_{i=1}^{m} \ln P(y_i|\boldsymbol{x}_i;\omega)$$

交叉熵损失 (Cross Entropy Loss)

Min $\mathcal{L}(y|x;\omega) = -\sum_{i=1}^{m} \ln P(y_i|\boldsymbol{x}_i;\omega)$

已知：

$$P(y_i|\boldsymbol{x}_i;\boldsymbol{\omega}) = y_i \mathrm{P}(y=1|\boldsymbol{x};\boldsymbol{\omega}) + (1-y_i)\mathrm{P}(y=0|\boldsymbol{x};\boldsymbol{\omega})$$

已知：

$$P(y=1|\boldsymbol{x};\boldsymbol{\omega}) = h_\omega(\boldsymbol{x}) \qquad P(y=0|\boldsymbol{x};\boldsymbol{\omega}) = 1-h_\omega(\boldsymbol{x})$$

已知：

$$h_\omega(\boldsymbol{x}) = \frac{\exp(\boldsymbol{\omega}^T\boldsymbol{x})}{1+\exp(\boldsymbol{\omega}^T\boldsymbol{x})} \qquad 1-h_\omega(\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{\omega}^T\boldsymbol{x})}$$

代入

$$\mathcal{L}(y|x;\omega) = -\sum_{i=1}^{m} \ln\left(y_i \frac{\exp(\boldsymbol{\omega}^T\boldsymbol{x})}{1+\exp(\boldsymbol{\omega}^T\boldsymbol{x})} + (1-y_i)\frac{1}{1+\exp(\boldsymbol{\omega}^T\boldsymbol{x})}\right)$$

进一步整理

$$\mathcal{L}(y|x;\omega) = -\sum_{i=1}^{m} \ln\left(y_i \frac{\exp(\boldsymbol{\omega}^T \boldsymbol{x}_i)}{1+\exp(\boldsymbol{\omega}^T \boldsymbol{x}_i)} + (1-y_i) \frac{1}{1+\exp(\boldsymbol{\omega}^T \boldsymbol{x}_i)} \right)$$

$$= -\sum_{i=1}^{m} \ln \frac{y_i \exp(\boldsymbol{\omega}^T \boldsymbol{x}_i) + 1 - y_i}{1+\exp(\boldsymbol{\omega}^T \boldsymbol{x}_i)}$$

$$= \sum_{i=1}^{m} \left(\ln(1+\exp(\boldsymbol{\omega}^T \boldsymbol{x}_i)) - \underbrace{\boxed{\ln(y_i \exp(\boldsymbol{\omega}^T \boldsymbol{x}_i) + 1 - y_i)}}_{xxx} \right)$$

阴影部分可统合为 $y_i \boldsymbol{\omega}^T \boldsymbol{x}_i$

y_i=0 时 $\ln 1 = 0$

y_i=1时 $\ln xxx = \ln \exp(\boldsymbol{\omega}^T \boldsymbol{x}_i) = \boldsymbol{\omega}^T \boldsymbol{x}_i$

Min $\mathcal{L}(\boldsymbol{\omega}) = \sum_{i=1}^{m} \left(\ln(1+\exp(\boldsymbol{\omega}^T \boldsymbol{x}_i)) - y_i \boldsymbol{\omega}^T \boldsymbol{x}_i \right)$

梯度下降

$$\mathcal{L}(\boldsymbol{\omega}) = \sum_{i=1}^{m}(\ln(1 + \exp(\boldsymbol{\omega}^T\boldsymbol{x}_i)) - y_i\boldsymbol{\omega}^T\boldsymbol{x}_i)$$

$$\frac{\partial\mathcal{L}(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}_j} = \sum_{i=1}^{m}(\frac{\partial\ln((1 + \exp(\boldsymbol{\omega}^T\boldsymbol{x}_i))}{\partial((1 + \exp(\boldsymbol{\omega}^T\boldsymbol{x}_i))}\frac{\partial((1 + \exp(\boldsymbol{\omega}^T\boldsymbol{x}_i))}{\partial(\boldsymbol{\omega}^T\boldsymbol{x}_i)}\frac{\partial(\boldsymbol{\omega}^T\boldsymbol{x}_i)}{\partial\omega_j} - \frac{\partial(y_i\boldsymbol{\omega}^T\boldsymbol{x}_i)}{\partial\omega_j})$$

$$= \sum_{i=1}^{m}(\frac{1}{((1 + \exp(\boldsymbol{\omega}^T\boldsymbol{x}_i))}\exp(\boldsymbol{\omega}^T\boldsymbol{x}_i)\,\boldsymbol{x}_{ij} - y_i\boldsymbol{x}_{ij})$$

$$= \sum_{i=1}^{m}(\mathrm{P}(y_i = 1|x_i;\boldsymbol{\omega})\,\boldsymbol{x}_{ij} - y_i\boldsymbol{x}_{ij})$$

For each feature *j*

$$\omega_j = \omega_j - \eta\frac{1}{m}\sum_{i=1}^{m}(\mathrm{P}(y_i = 1|x_i;\boldsymbol{\omega})\,\boldsymbol{x}_{ij} - y_i\boldsymbol{x}_{ij})$$

对数几率回归-预测

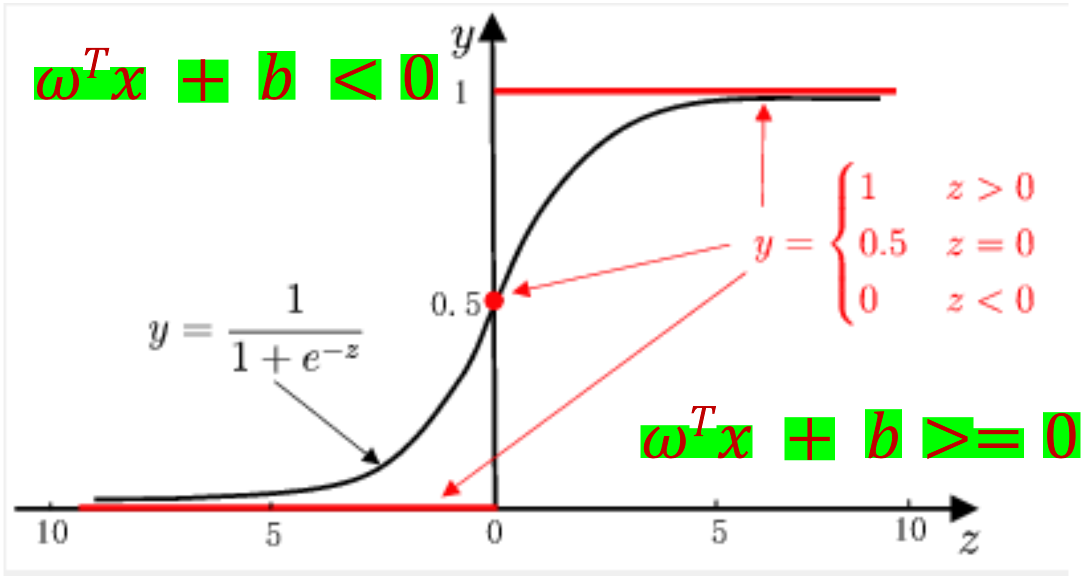
- $y=1$的概率

$$y = h_{\omega}(\boldsymbol{x}) = \frac{1}{1+e^{-(\boldsymbol{\omega}^T\boldsymbol{x}+\boldsymbol{b})}}$$

- 假设 预测$y = 1$ if

$\boldsymbol{\omega}^T\boldsymbol{x} + b >= 0$

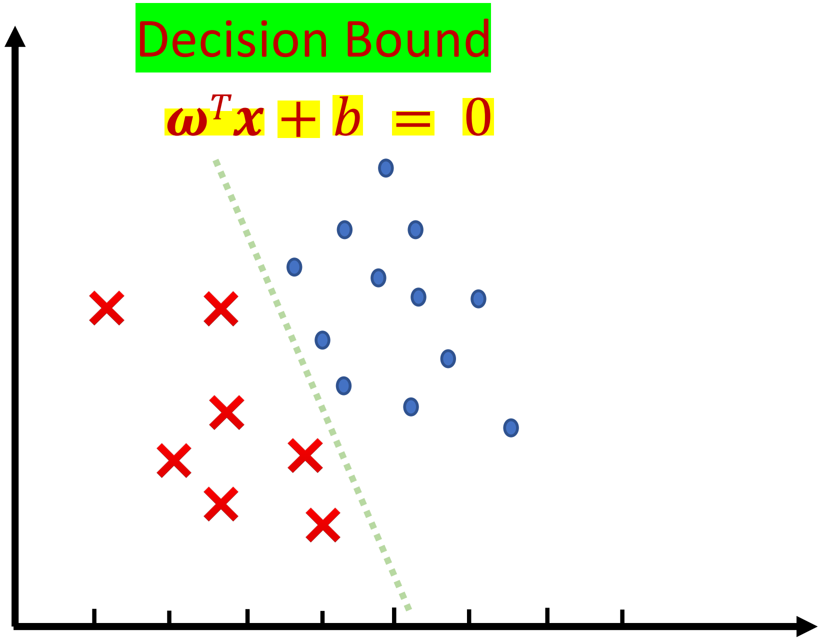
$h_{\omega}(\boldsymbol{x}) >= 0.5$



- 假设 预测$y = 0$ if

$h_{\omega}(\boldsymbol{x}) < 0.5$ $\boldsymbol{\omega}^T\boldsymbol{x} + b < 0$

$$\ln\frac{y}{1-y} = \boldsymbol{\omega}^T\boldsymbol{x} + b$$



对数几率：样本作为正例的相对可能性的对数

分类问题

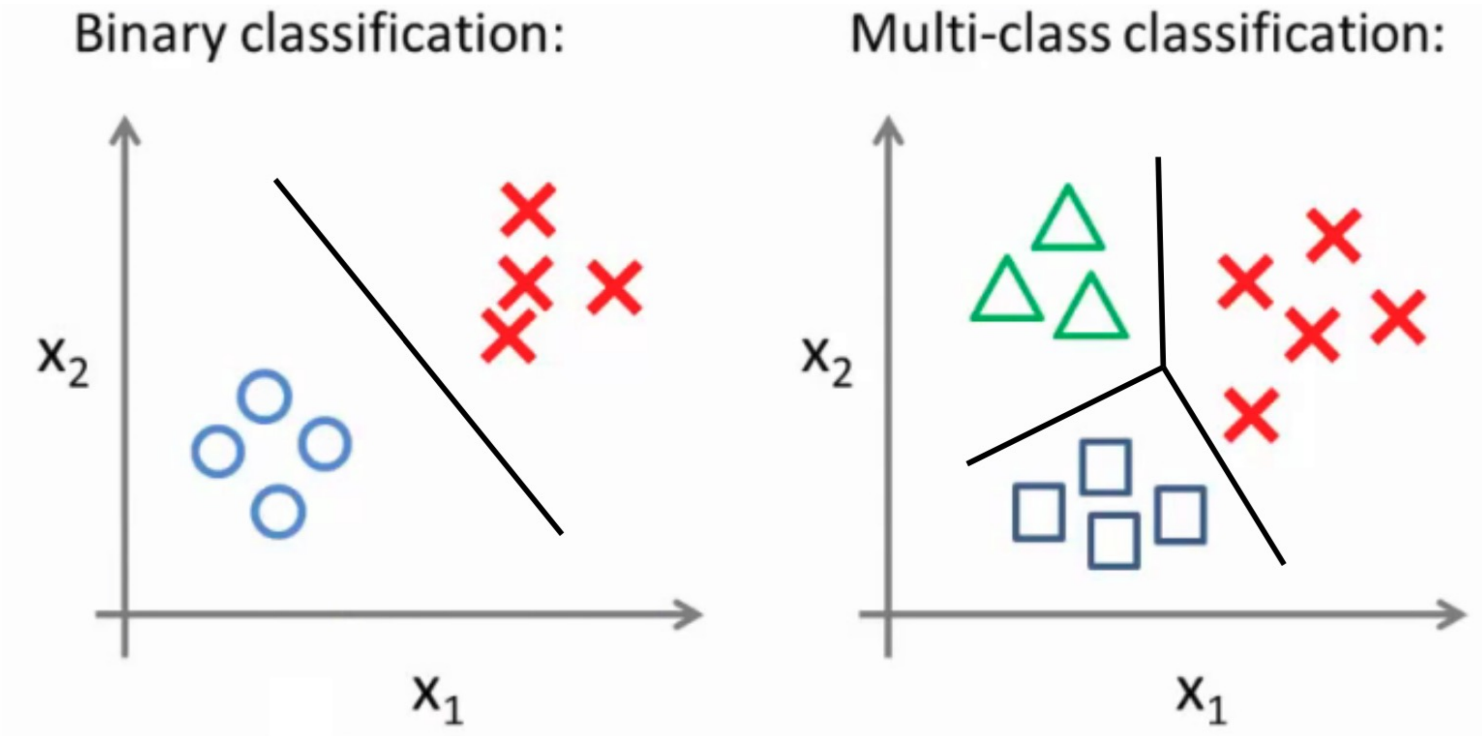
（1）什么是分类问题

（2）对数几率回归模型建立

（3）多分类学习

多分类学习

数据： $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$， $y_i\in\{C_i,\cdots,C_N\}$



多分类学习

数据：$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$，$y_i \in \{C_i, \cdots, C_N\}$

多分类学习方法

- 二分类学习方法推广到多类（如多项对数几率回归）
- 利用二分类学习器解决多分类问题（常用）
  - 对问题进行拆分，为拆出的每个二分类任务训练一个分类器
  - 对于每个分类器的预测结果进行集成以获得最终的多分类结果

拆分策略

- 一对一（One vs. One, OvO）
- 一对其余（One vs. Rest, OvR）
- 多对多（Many vs. Many, MvM）

多分类学习：一对一

数据：$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$，$y_i \in \{C_i, \cdots, C_N\}$

拆分阶段

- N个类别，两两配对（如 C_i和C_j ，将C_i作为正例C_j作为反例）
  - $N(N-1)/2$ 个二类任务
- 各个二类任务，学习分类器（注意数据集的切分和使用）
  - $N(N-1)/2$ 个二类分类器

测试阶段

- 新样本提交给所有分类器预测
  - $N(N-1)/2$个分类结果
- 投票产生最终分类结果
  - 被预测最多的类别为最终类别

多分类学习：一对其余

数据：$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$，$y_i \in \{C_i, \cdots, C_N\}$

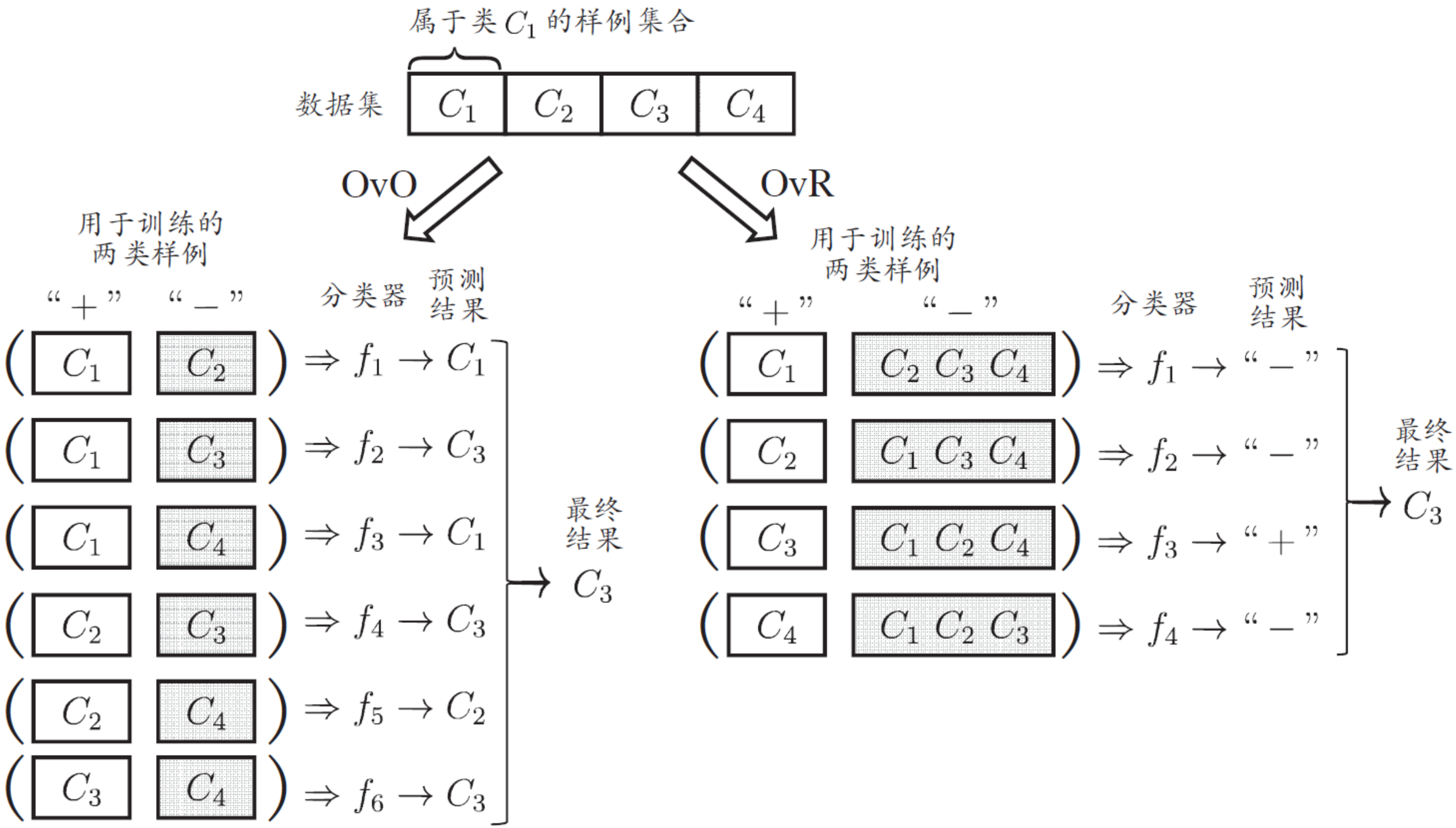
拆分阶段

- N个类别，某一类作为正例，所有其他类作为反例
  - N 个二类任务
- 各个二类任务，学习分类器（注意数据集的使用）
  - N 个二类分类器

测试阶段

- 新样本提交给所有分类器预测
  - N个分类结果
- 比较各分类器预测置信度
  - 置信度最大类别作为最终类别

多分类学习：两种策略比较



多分类学习：两种策略比较

一对一

- 训练 $N(N-1)/2$ 个分类器，存储开销大
- 训练只用两个类的样例，训练时间短
- 测试时间长

一对其余

- 训练N个分类器，存储开销小
- 训练用到全部训练样例，训练时间长
- 测试时间短

预测性能取决于具体数据分布，多数情况下两者差不多

谢谢！