

# 经典模型

## 第三章 逻辑斯蒂回归

1. 二项逻辑斯蒂回归
2. 多项逻辑斯蒂回归
3. 小练习

# 1. 二项逻辑斯蒂回归

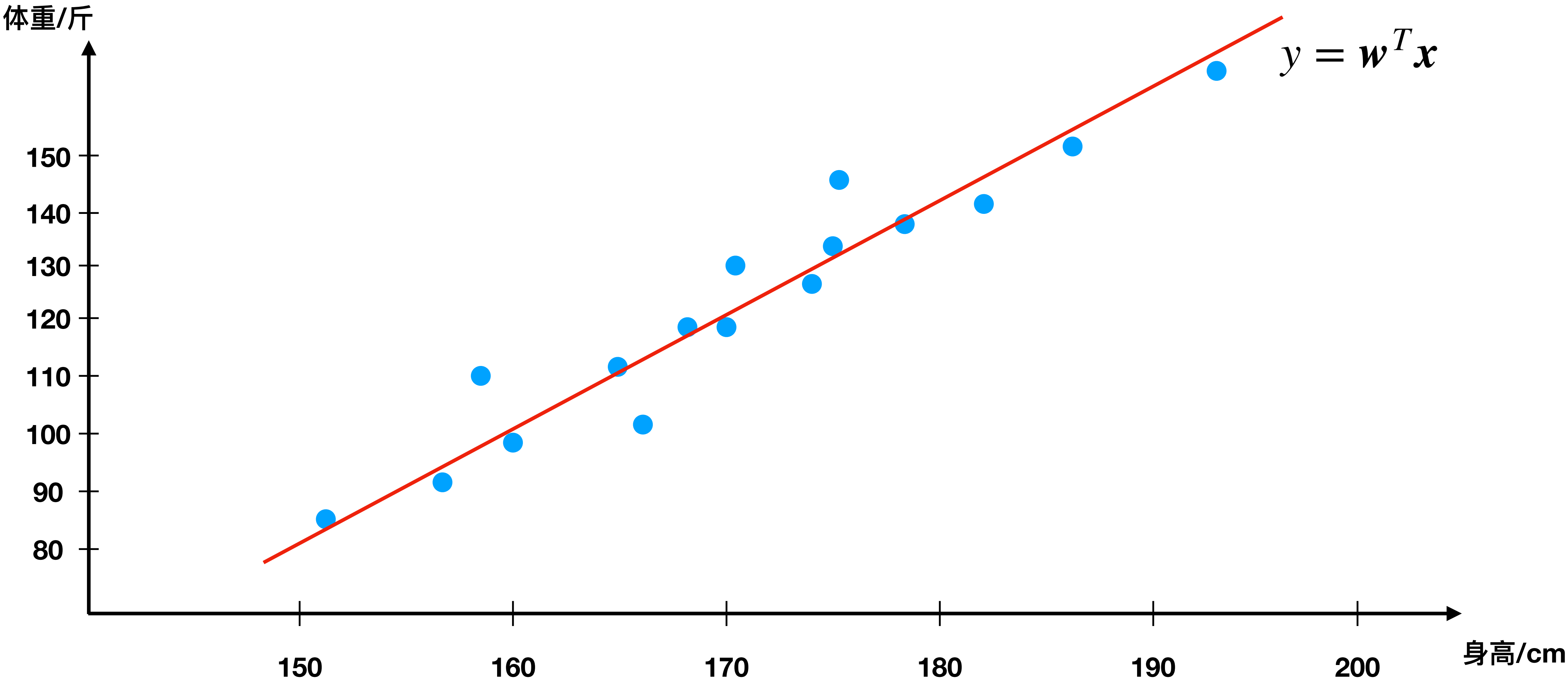
# 1.1 二项逻辑斯蒂回归简介

# 二项逻辑斯蒂回归

- **二项逻辑斯蒂回归**（binomial logistic regression），简称**逻辑回归**也被称为**对数几率回归**，是一种**二分类模型**。
- 从概率的角度，研究分类变量与样本特征之间的关系。属于概率型、非线性回归方法。

考虑：如何使用线性模型做分类任务？

找一个单调可微函数将分类任务的真实标记  $y$  与线性回归模型的预测值联系起来。



# 二项逻辑斯蒂回归

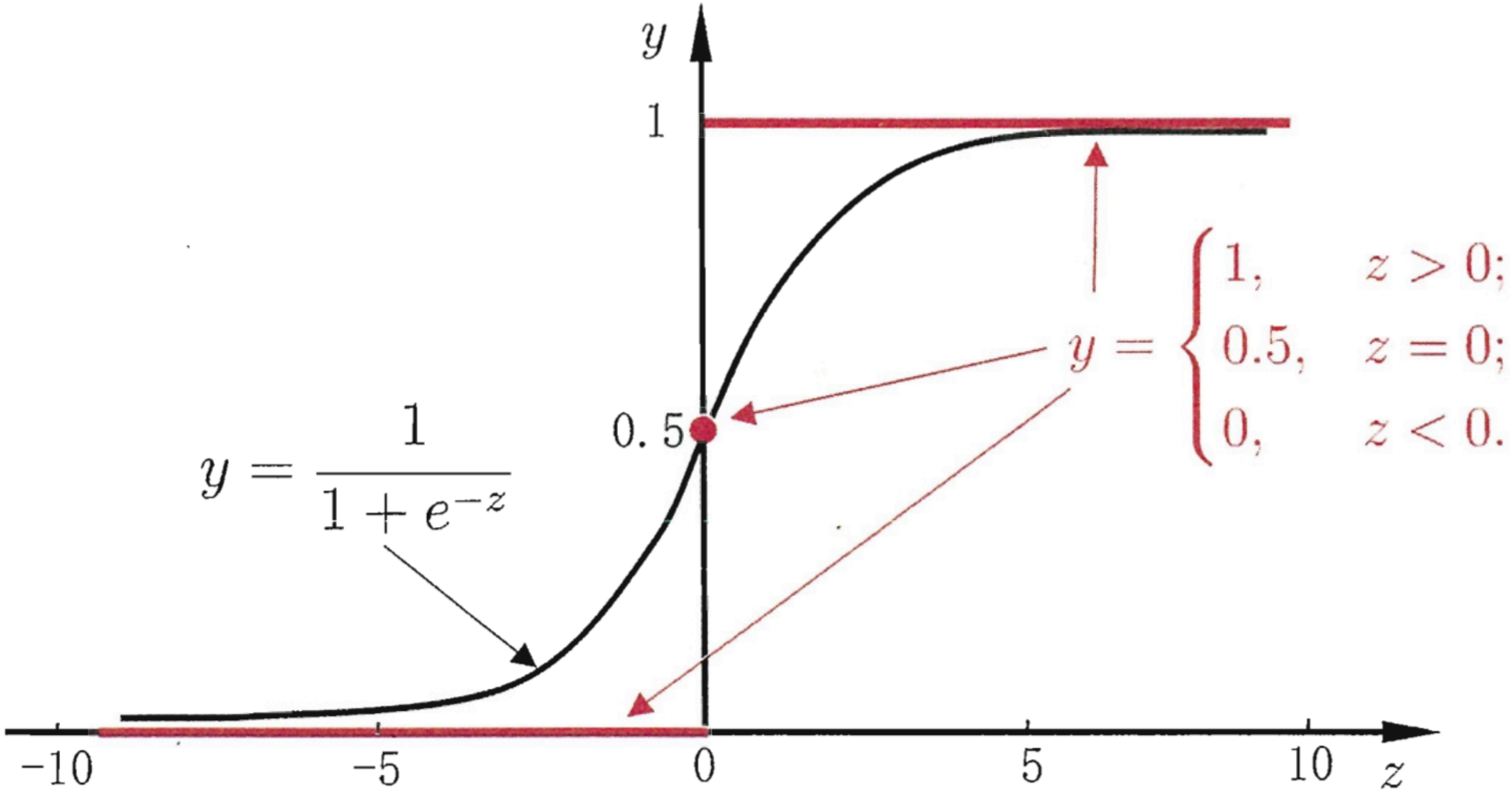
- 考虑二分类任务，其输出标记  $y \in \{0,1\}$ ，而线性回归模型产生的预测值  $z = \mathbf{w}^T \mathbf{x} + b$  是实值，于是，我们将实值  $z$  转换为 0/1 值，最理想的是**单**

**位阶跃函数 (unit-step function)**  $y = \begin{cases} 0, & z < 0 \\ 0.5, & z = 0 \\ 1 & z > 0 \end{cases}$

- 单位阶跃函数不连续，我们希望找到在一定程度上接近单位阶跃函数的替代函数，并希望它单调可微，对数几率函数正是这样一个常用的替代函数，**对**

**数几率函数**  $y = \frac{1}{1 + e^{-z}}$

若预测值  $z$  大于零就判为正例，小于零就判为反例，预测值为临界值则可任意判别。





## 1.2 对数几率函数

# 对数几率函数的意义

- 若将  $y$  视为样本  $x$  作为正例的可能性，则  $1 - y$  是其反例的可能性，两者的比值  $\frac{y}{1 - y}$  称为**几率 (odds)**，反映了  $x$  作为正例的相对可能性，对几

率取对数则得到**对数几率 (log odds, 即logit)**  $\ln \frac{y}{1 - y}$ 。令

$$z = \ln \frac{y}{1 - y}, \text{ 可得 } y = \frac{1}{1 + e^{-z}}, \text{ 即如果输入 } z \text{ 表示对数几率时, 对数}$$

几率函数将  $z$  转化为正例的概率。

# 对数几率函数

- 对数几率函数  $y = \frac{1}{1 + e^{-z}}$  是一种**Sigmoid函数**（即形似S的函数），它的值域是  $(0,1)$ ，它将  $z$  值转化为一个接近 0 或 1 的  $y$  值，并且其输出值在  $z = 0$  附近变化很陡。
- 将  $z = \mathbf{w}^T \mathbf{x} + b$  带入对数几率函数得到  $h = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$ ，即得到逻辑回归模型形式。
- 可以看出，实际上是在用线性回归模型的预测结果去逼近真实标记的对数几率，因此也将逻辑回归叫做对数几率回归。

# 1.3 模型参数估计

训练样本集  $\{(\mathbf{x}_i, y_i), i = 1, 2, \dots, n\}$  , 样本彼此独立, 类别标号  $y_i \in \{0, 1\}$

则样本  $(\mathbf{x}_i, y_i)$  出现的概率:

$$P(\mathbf{x}_i, y_i) = [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i} P(\mathbf{x}_i)$$

$n$  个独立样本出现的似然函数为  $l = \prod_{i=1}^n P(\mathbf{x}_i, y_i)$

似然函数为  $l = \prod_{i=1}^n P(\mathbf{x}_i, y_i)$  , 带入  $P(\mathbf{x}_i, y_i) = [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i} P(\mathbf{x}_i)$  ,

可得  $l = \prod_{i=1}^n [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i} \prod_{i=1}^n P(\mathbf{x}_i)$

其中  $\prod_{i=1}^n P(\mathbf{x}_i)$  与模型参数无关,

所以似然函数可简化为  $l = \prod_{i=1}^n [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i}$ 。

对似然函数取对数  $l = \prod_{i=1}^n [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i}$  ,

得到对数似然函数：

$$l' = \ln l = \ln \left\{ \prod_{i=1}^n [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i} \right\}$$

$$= \sum_{i=1}^n \ln \{ [P(y_i | \mathbf{x}_i)]^{y_i} [1 - P(y_i | \mathbf{x}_i)]^{1-y_i} \}$$

$$= \sum_{i=1}^n y_i \ln[P(y_i | \mathbf{x}_i)] + (1 - y_i) \ln[1 - P(y_i | \mathbf{x}_i)]$$

平均对数似然函数为：

$$\hat{l} = \frac{1}{n} \sum_{i=1}^n y_i \ln[P(y_i | \mathbf{x}_i)] + (1 - y_i) \ln[1 - P(y_i | \mathbf{x}_i)]$$

**目标** 最大化平均对数似然求解参数：  $\arg \max_{w,b} \hat{l}$

**等价于** 最小化负的平均对数似然：  $\arg \min_{w,b} (-\hat{l})$



$P(y_i | \mathbf{x}_i)$  表示类后验概率，即模型预测结果，所以将  $h = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$  带

入平均负平均对数似然函数得到：

$$J = -\frac{1}{n} \sum_{i=1}^n y_i \ln h(\mathbf{x}_i) + (1 - y_i) \ln[1 - h(\mathbf{x}_i)] , \text{ 即为代价函数。}$$

最小化代价函数求得参数  $(\mathbf{w}^*, b^*)$  即可。

## 1.4 使用GD方法求解参数

# 使用小批量梯度下降法求解参数

训练集  $\{(\mathbf{x}_i, y_i), i = 1, 2, \dots, n\}$ ,  $\mathbf{x} \in R^d, y \in \{0, 1\}$  , 随机打乱样本顺序。设小批量大小为  $m$  , 使用固定学习率  $\alpha$  , 设定一种停止迭代规则。

① 从打乱训练集中顺序选择  $m$  个样本, 组成一个小批次样本

② 使用抽取出的小批次样本计算模型代价  $J(\mathbf{w}, b) = -\frac{1}{m} \sum_{i=1}^m y_i \ln h(\mathbf{x}_i) + (1 - y_i) \ln[1 - h(\mathbf{x}_i)]$

③ 计算  $m$  个样本的平均梯度  $\bar{\mathbf{g}}_{\mathbf{w}}, \bar{g}_b$

④ 更新模型参数  $\mathbf{w} = \mathbf{w} - \alpha \bar{\mathbf{g}}$  ,  $b = b - \alpha \bar{g}_b$ , 判断是否满足停止迭代规则, 如满足, 则停止迭代, 将此时参数作为最优参数, 如不满足, 则回到 ① 继续迭代。

# 1.5 逻辑回归模型分类决策

- 若  $\text{logit}(\boldsymbol{x}) = \ln \frac{h(\boldsymbol{x})}{1 - h(\boldsymbol{x})} = \boldsymbol{w}^{*T} \boldsymbol{x} + b^* > 0$  , 即  $h(\boldsymbol{x}) > 0.5$  判别为正类

- 若  $\text{logit}(\boldsymbol{x}) = \ln \frac{h(\boldsymbol{x})}{1 - h(\boldsymbol{x})} = \boldsymbol{w}^{*T} \boldsymbol{x} + b^* < 0$  , 即  $h(\boldsymbol{x}) < 0.5$  判别为负类

## 2. 多项逻辑斯蒂回归

## 2.1 多项逻辑斯蒂回归简介

# 多项逻辑斯蒂回归

- 多项逻辑斯蒂回归 (multi-nominal logistic regression) 也被称为 **softmax回归**，是二项逻辑斯蒂回归的推广，用于多类别分类。
- 从概率的角度，研究分类变量与样本特征之间的关系。属于概率型、非线性回归方法。



## 2.2 模型

对于多类问题，设训练样本集  $\{(\mathbf{x}_i, y_i) \mid i = 1, 2, \dots, n\}$ ，样本彼此独立，类别标号  $y_i \in \{1, 2, \dots, C\}$ 。

几率定义为  $odds_j(\mathbf{w}) = \ln \frac{P(y = j \mid \mathbf{x})}{P(y = C \mid \mathbf{x})} \quad j = 1, 2, \dots, C - 1$ ，

可得  $P(y = j \mid \mathbf{x}) = P(y = C \mid \mathbf{x})e^{odds_j(\mathbf{x})}$

对一个样本  $\mathbf{x}$  预测各类概率之和为  $P(y = C | \mathbf{x}) + \sum_{j=1}^{C-1} P(y = j | \mathbf{x}) = 1$ , 带入

$P(y = j | \mathbf{x}) = P(y = C | \mathbf{x})e^{odds_j(\mathbf{x})}$  得:

$$P(y = C | \mathbf{x}) + \sum_{j=1}^{C-1} P(y = C | \mathbf{x})e^{odds_j(\mathbf{x})} = 1,$$

$$P(y = C | \mathbf{x}) \left[ 1 + \sum_{j=1}^{C-1} e^{odds_j(\mathbf{x})} \right] = 1,$$

根据  $odds_j(\mathbf{x}) = \mathbf{w}_j^T \mathbf{x} + b_j$ ,  $P(y = C | \mathbf{x}) \left[ 1 + \sum_{j=1}^{C-1} e^{odds_j(\mathbf{x})} \right] = 1$ , 得到

$P(y = C | \mathbf{x}) \left[ 1 + \sum_{j=1}^{C-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j} \right] = 1$ , 所以模型预测类别概率为:

$$\begin{cases} P(y = j | \mathbf{x}) = \frac{e^{\mathbf{w}_j^T \mathbf{x} + b_j}}{1 + \sum_{j=1}^{C-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} & j = 1, 2, \dots, C-1 \\ P(y = C | \mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{C-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \end{cases}$$

## 2.3 模型参数估计

训练样本集  $\{(\mathbf{x}_i, y_i), i = 1, 2, \dots, n\}$  , 样本彼此独立, 类别标号  $y_i \in \{1, 2, \dots, C\}$

则样本  $(\mathbf{x}_i, y_i)$  出现的概率:

$$P(\mathbf{x}_i, y_i) = \left[ \prod_{k=1}^C \left[ P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right] \right] P(\mathbf{x}_i)$$

$n$  个独立样本出现的似然函数为  $l = \prod_{i=1}^n P(\mathbf{x}_i, y_i)$

将  $P(\mathbf{x}_i, y_i) = \left[ \prod_{k=1}^C \left[ P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right] \right] P(\mathbf{x}_i)$  带入似然函数得：

$$l = \prod_{i=1}^n P(\mathbf{x}_i, y_i) = \prod_{i=1}^n \prod_{k=1}^C \left[ P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right] \prod_{i=1}^n P(\mathbf{x}_i)$$

其中  $\prod_{i=1}^n P(\mathbf{x}_i)$  与模型参数无关，

故似然函数简化为  $l = \prod_{i=1}^n \prod_{k=1}^C \left[ P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right]$

取对数得对数似然代价函数： $\hat{J} = \sum_{i=1}^n \sum_{k=1}^C \left[ \ln P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right]$

带入  $\begin{cases} P(y = j | \mathbf{x}) = \frac{e^{\mathbf{w}_j^T \mathbf{x} + b_j}}{1 + \sum_{j=1}^{C-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} & j = 1, 2, \dots, C-1 \\ P(y = C | \mathbf{x}) = \frac{1}{1 + \sum_{j=1}^{C-1} e^{\mathbf{w}_j^T \mathbf{x} + b_j}} \end{cases}$  可得到完整形式

负平均对数似然函数即代价函数为：

$$J = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^C \left[ \ln P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right], \text{ 最终可使用梯度下降等方法求得模型参数。}$$



## 2.4 分类决策

- 若  $\mathbf{w}_j^T \mathbf{x} + b_j = \max_{k \in \{1, 2, \dots, C-1\}} (\mathbf{w}_k^T \mathbf{x} + b_k) > 0$  则将  $\mathbf{x}$  判断为第  $j$  类，否

则，将  $\mathbf{x}$  判断为第  $C$  类。

## 2.5 softmax回归的其它形式

softmax 函数  $y_j = \frac{e^{z_j}}{\sum_{i=1}^C e^{z_i}}$  , 将  $z_j = \mathbf{w}_j^T \mathbf{x} + b_j$  带入, 得到

$$P(y = j | \mathbf{x}) = \frac{e^{\mathbf{w}_j^T \mathbf{x} + b_j}}{\sum_{i=1}^C e^{\mathbf{w}_i^T \mathbf{x} + b_i}} \quad j = 1, 2, \dots, C$$

同样使用最大似然估计可得代价函数  $J = -\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^C \left[ \ln P(y_i | \mathbf{x}_i)^{I(y_i=k)} \right]$

相比于前面所述的多项逻辑斯蒂回归它的形式更加简单, 但冗余了一组参数。

# 3. 小练习

1. 求对数几率函数的导函数，观察导函数值与原函数值有什么关系？

2. 画出  $y_i$  分别为 0,1 时的代价函数  $J = -\frac{1}{n} \sum_{i=1}^n y_i \ln h(\mathbf{x}_i) + (1 - y_i) \ln[1 - h(\mathbf{x}_i)]$  图像（设  $n = 1$ ），观察代价函数随  $h$  如何变化？求对数似然代价函数对  $h$  的偏导数  $\frac{\partial J}{\partial h_i}$ ，观察  $y_i$  分别为 0,1 时， $\frac{\partial J}{\partial h_i}$  是什么？

3. 梯度下降法求解模型逻辑回归模型参数时， $\bar{\mathbf{g}}_{\mathbf{w}}, \bar{g}_b$  具体的形式是什么？

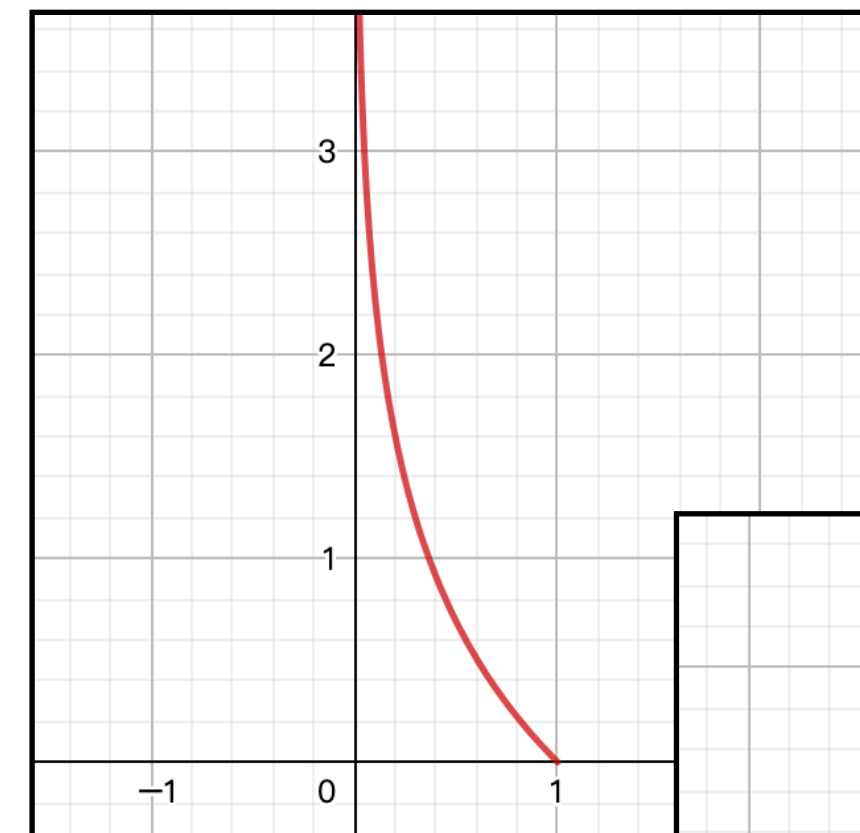
4. 思考：为什么不使用均方误差代价函数作为逻辑回归模型的代价函数？尝试使用均方误差作为代价函数，观察求得的  $\bar{\mathbf{g}}_{\mathbf{w}}, \bar{g}_b$  有什么不同？

# 答案

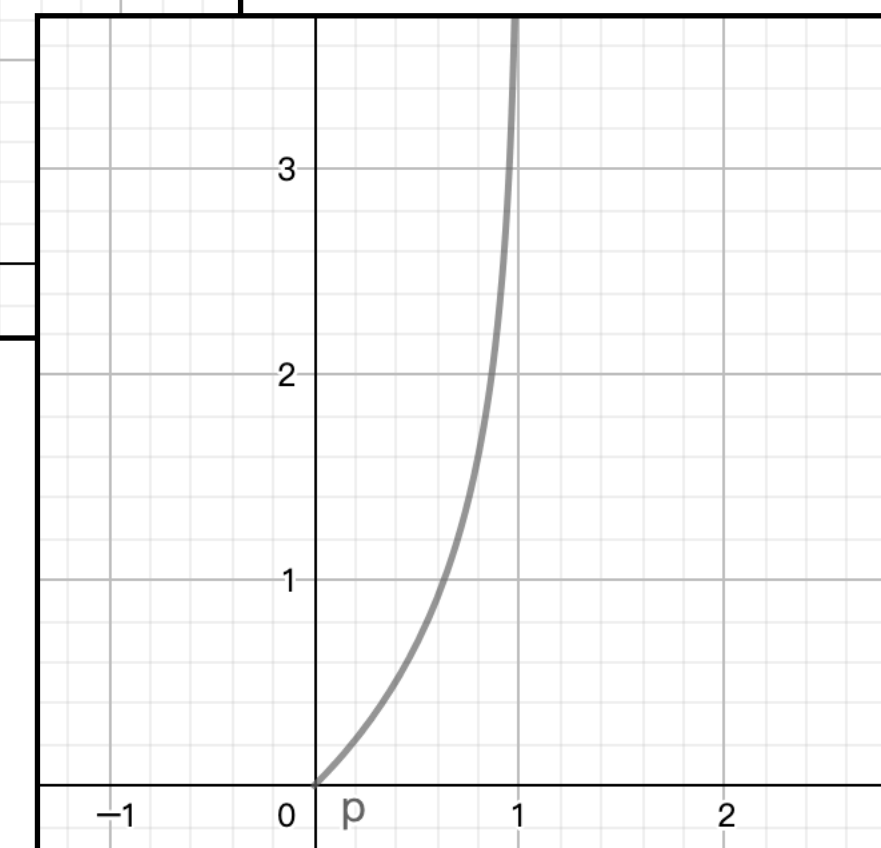
1.  $y = \frac{1}{1 + e^{-z}}$  的导函数为  $y' = \frac{e^{-z}}{(1 + e^{-z})^2} = y(1 - y)$ , 关系:  $y' = y(1 - y)$

2. 如下:

1. 若  $y_i = 1$ , 则  $J = -\ln h_i$ ,  $\frac{\partial J}{\partial h_i} = -\frac{1}{h_i}$ , 代价函数图像:



2. 若  $y_i = 0$ , 则  $J = \ln(1 - h_i)$ ,  $\frac{\partial J}{\partial h_i} = \frac{1}{1 - h_i}$ , 代价函数图像:



3.  $\bar{g}_{w_i} = \frac{1}{n} \sum_{i=1}^n (h_i - y_i)x_i, \quad \bar{b} = \frac{1}{n} \sum_{i=1}^n (h_i - y_i)$

4. 使用均方误差代价函数后，代价函数为非凸函数（[证明请参考此处](#)），求解容易陷入局部极值点，且使用均方误差代价函数后，logistic 函数值在接近0,1时，梯度接近于 0，使得模型参数更新速度变慢。

**THANKS**