

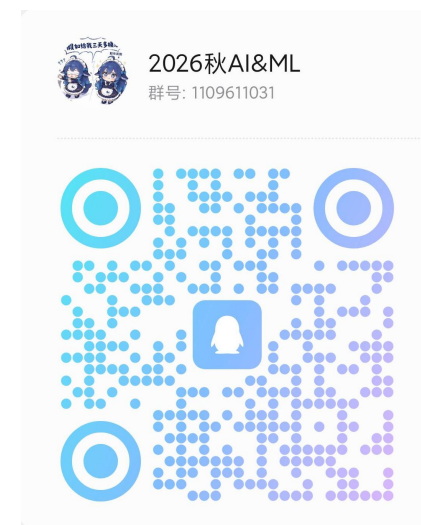
# 人工智能与机器学习基础

## Lecture 05: Logistic Regression

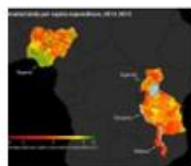
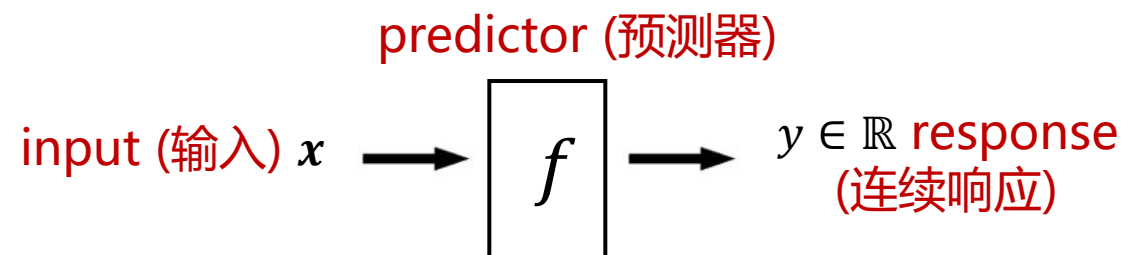
---

冯文杰

人工智能与数据科学学院  
中国科学技术大学



# Recap: 回归问题



贫困地区预测: 卫星图像  $\rightarrow$  资产财富指数

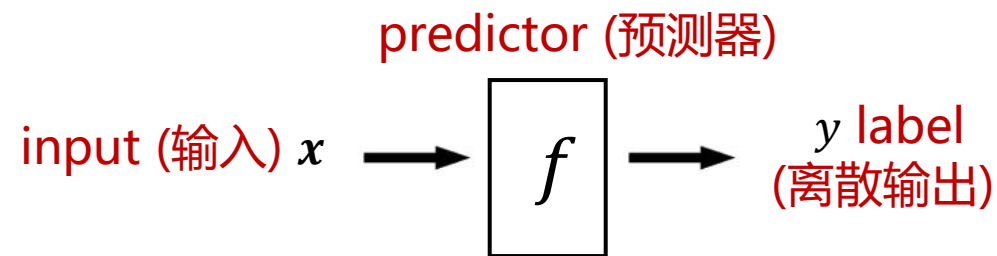


住房价格预测: 房屋信息  $\rightarrow$  价格



到达时间预测: 目的地、天气、出发时间  $\rightarrow$  预计到达时间

# Recap: 分类问题



二分类任务



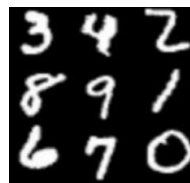
欺诈检测: 信用卡交易  $\rightarrow$  欺诈或非欺诈



有毒评论检测: 在线评论  $\rightarrow$  有毒或无毒

$$y \in \{+1, -1\}$$

多分类任务



手写数字识别: 数字图像  $\rightarrow$  0, 1, 2, ...

$$y \in \{0, 1, \dots, 9\}$$

$$y \in \{1, \dots, K\}$$

# Recap: 线性回归

- 模型向量表示:  $f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x)$ 
  - 例如: 参数向量/模型参数  $\mathbf{w} = [\omega_1, \omega_2]$ , 特征提取器  $\phi(x) = [1, x]$
- 假设类:  $\mathcal{F} = \{f_{\mathbf{w}}: \mathbf{w} \in \mathbb{R}^2\}$
- 损失函数:  $Loss(x, y, \mathbf{w}) = (f_{\mathbf{w}}(x) - y)^2$

$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y) \in \mathcal{D}_{\text{train}}} (f_{\mathbf{w}}(x) - y)^2$$
$$\operatorname{argmax}_{\mathbf{w}} \ln p(y|x, \mathbf{w}) = \operatorname{argmin}_{\mathbf{w}} \sum_{(x,y) \in \mathcal{D}} (y - \mathbf{w} \cdot \phi(x))^2$$

高斯假设下的最大似然估计 = 最小化平方误差

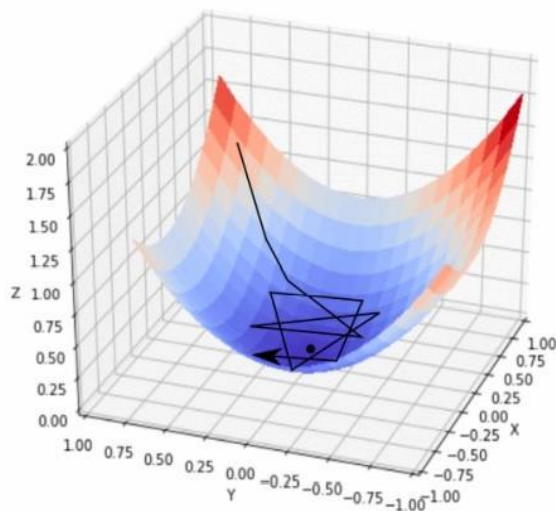
# Recap: 线性回归

- 优化算法:  $\min_{\mathbf{w}} \text{TrainLoss}(\mathbf{w})$



## Definition: gradient

The gradient  $\nabla_{\mathbf{w}} \text{TrainLoss}(\mathbf{w})$  is the direction that increases the training loss the most.



## Algorithm: gradient descent

Initialize  $\mathbf{w} = [0, \dots, 0]$

For  $t = 1, \dots, T$ : epochs

$$\mathbf{w} \leftarrow \mathbf{w} - \underbrace{\eta}_{\text{step size}} \underbrace{\nabla_{\mathbf{w}} \text{TrainLoss}(\mathbf{w})}_{\text{gradient}}$$

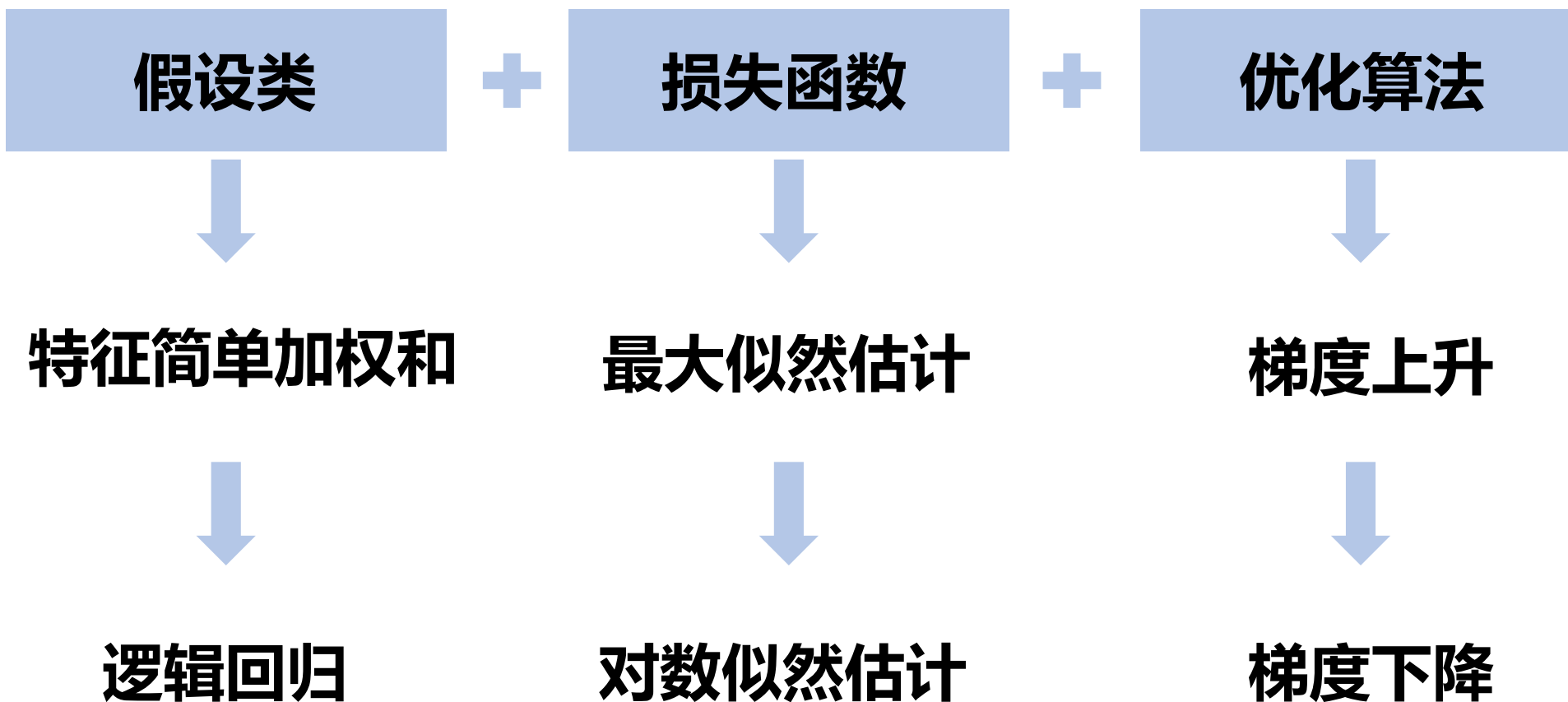
# 提纲

---

- Logistic Regression 框架
- 过拟合问题
- 正则化

# 逻辑回归 Logistic Regression: 框架

---



# 逻辑回归 Logistic Regression

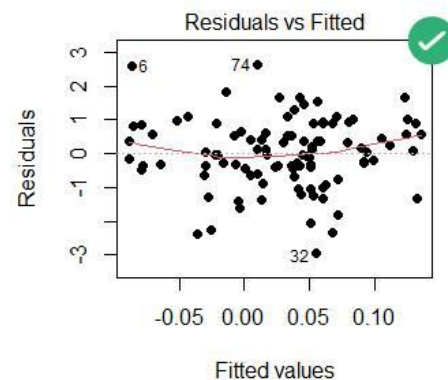
- 基本假设

- 线性 Linearity
- 独立性 Independence

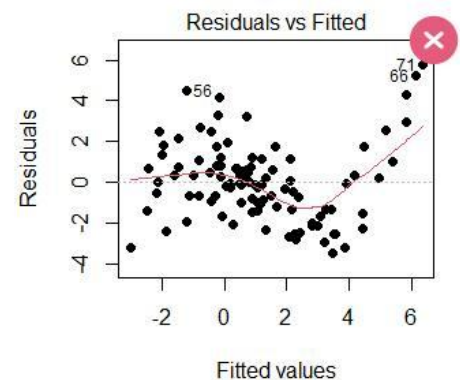
- 确保因变量和自变量之间关系的一致性

- 影响模型**准确性**和统计分析的**可靠性**
- 违反假设会引入偏差和混杂因素

A: Linearity assumption satisfied:



B: Linearity assumption violated:



# 提纲

---

- **Logistic Regression 框架**

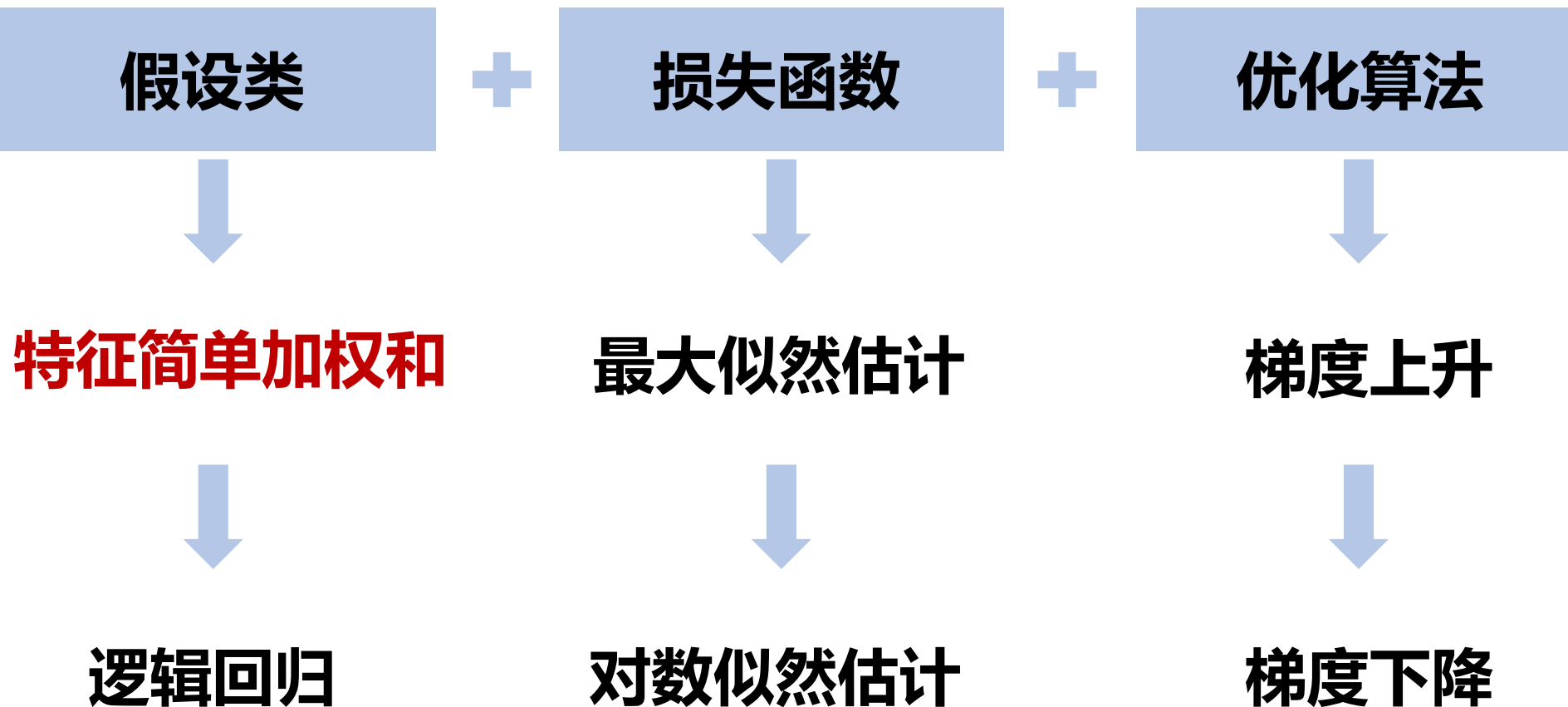
- 假设类
- 损失函数
- 优化算法

- 过拟合问题

- 正则化

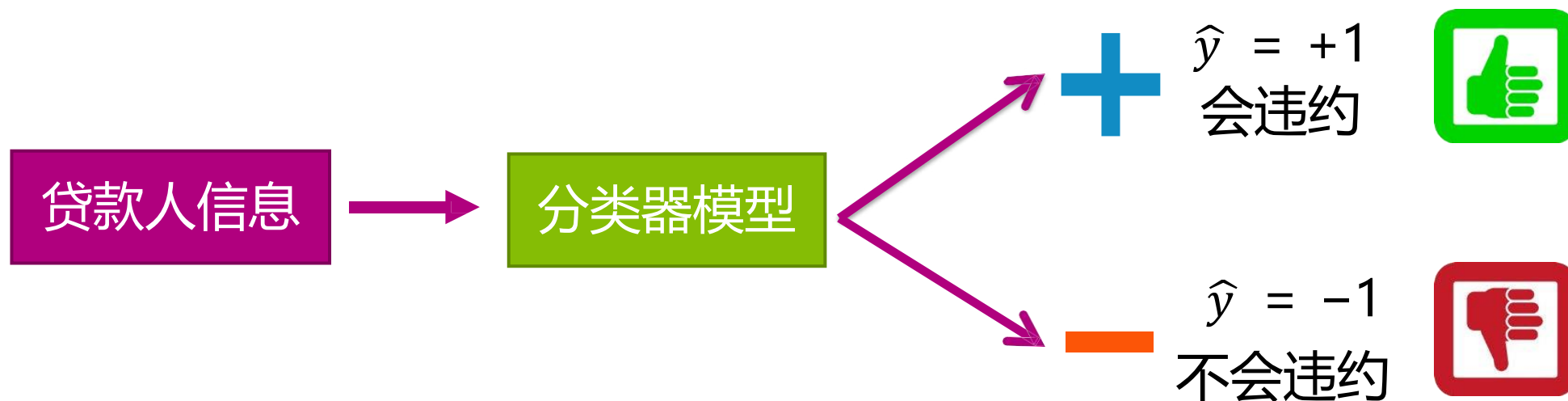
# 逻辑回归 Logistic Regression: 框架

---



# 二分类问题：贷款违约检测

- 输入input:  $x$ 
  - 如: “年收入4.5万 + 贷款4万, 利率13.67%”。
- 输出output:  $y$ 
  - 预测类别



# 分类器：简单线性分类器

| 特征  | 系数  |
|-----|-----|
| ... | ... |



贷款人信息



输入input:  $x$

简单线性分类器：

$\text{Score}(x)$  = 各特征的加权和

If  $\text{Score}(x) > 0$ :

$\hat{y} = +1$

Else:

$\hat{y} = -1$

# 一个例子：简单线性分类器

- 假设每个特征的系数已知

| 特征   | 系数   |
|------|------|
| 年收入  | 1.0  |
| 贷款金额 | -0.8 |
| 贷款利率 | -3.3 |
| ...  | ...  |

输入  $x$ :

- 贷款人年收入4.5万元
- 贷款金额4万元
- 贷款利率13.76%

- $Score(x) = 1.0 \times 4.5 - 0.8 \times 4 - 3.3 \times 0.1376 > 0$
- $\hat{y} = +1$

# 一个例子：简单线性分类器

---

- $\hat{y} = \text{sign}(\text{Score}(x))$
- $\text{Score}(x) = w_0 \cdot \phi(x)_0 + w_1 \cdot \phi(x)_1 + \dots + w_d \cdot \phi(x)_d$   
 $= \sum_i w_i \cdot \phi(x)_i = w^T \phi(x)$
- 特征提取器  $\phi$ :
  - 特征0 =  $\phi(x)_0$  (e.g., 1)
  - 特征1 =  $\phi(x)_1$  (e.g.,  $x[1]$  = 年收入)
  - 特征2 =  $\phi(x)_2$  (e.g.,  $x[2]$  = 贷款金额 or  $\log x[2]$  or  $\log x[2]/x[1]$ )
  - ...
  - 特征d =  $\phi(x)_d$  其他关于  $x$  的函数)

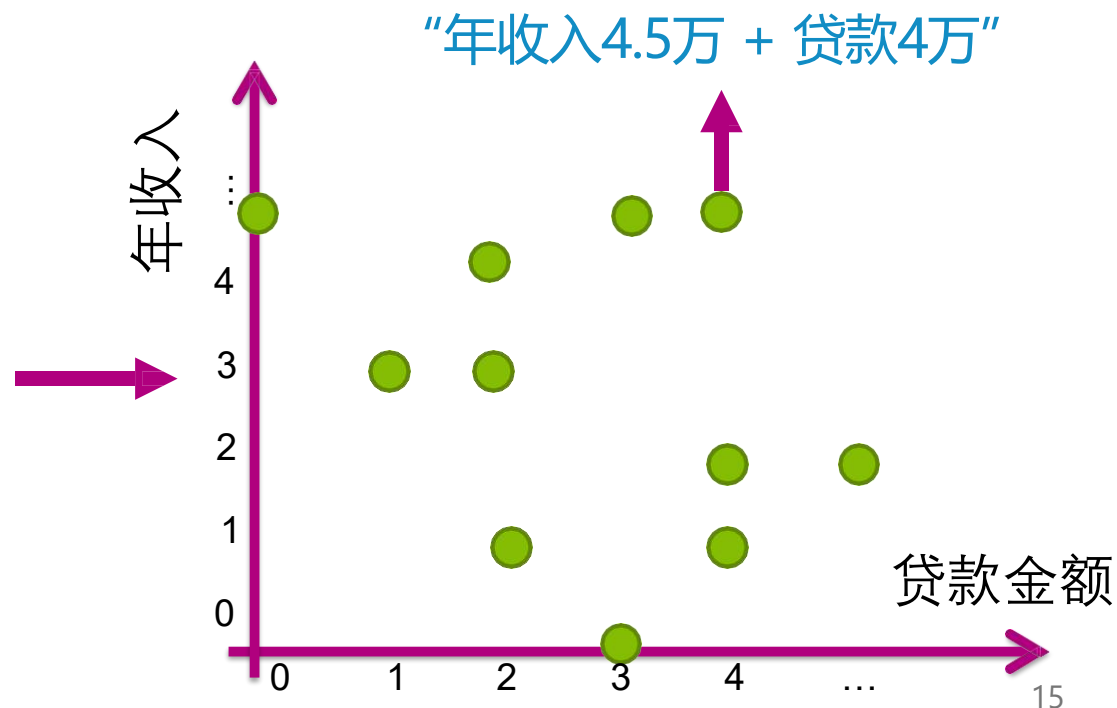
# 决策边界：Decision Boundary

- 假设只有两个特征具有非零系数

| 特征   | 系数    | 值    |
|------|-------|------|
|      | $w_0$ | 0.0  |
| 贷款金额 | $w_1$ | 1.0  |
| 年收入  | $w_2$ | -1.5 |

| 贷款人ID | 贷款金额 | 年收入 |
|-------|------|-----|
| 1     | 4.0  | 4.5 |
| 2     | 0.0  | 3.0 |
| 3     | 2.0  | 1.0 |
| 4     | 3.0  | 5.3 |
| ...   | ...  | ... |

$$Score(x) = 1.0 \times \text{贷款金额} - 1.5 \times \text{年收入}$$

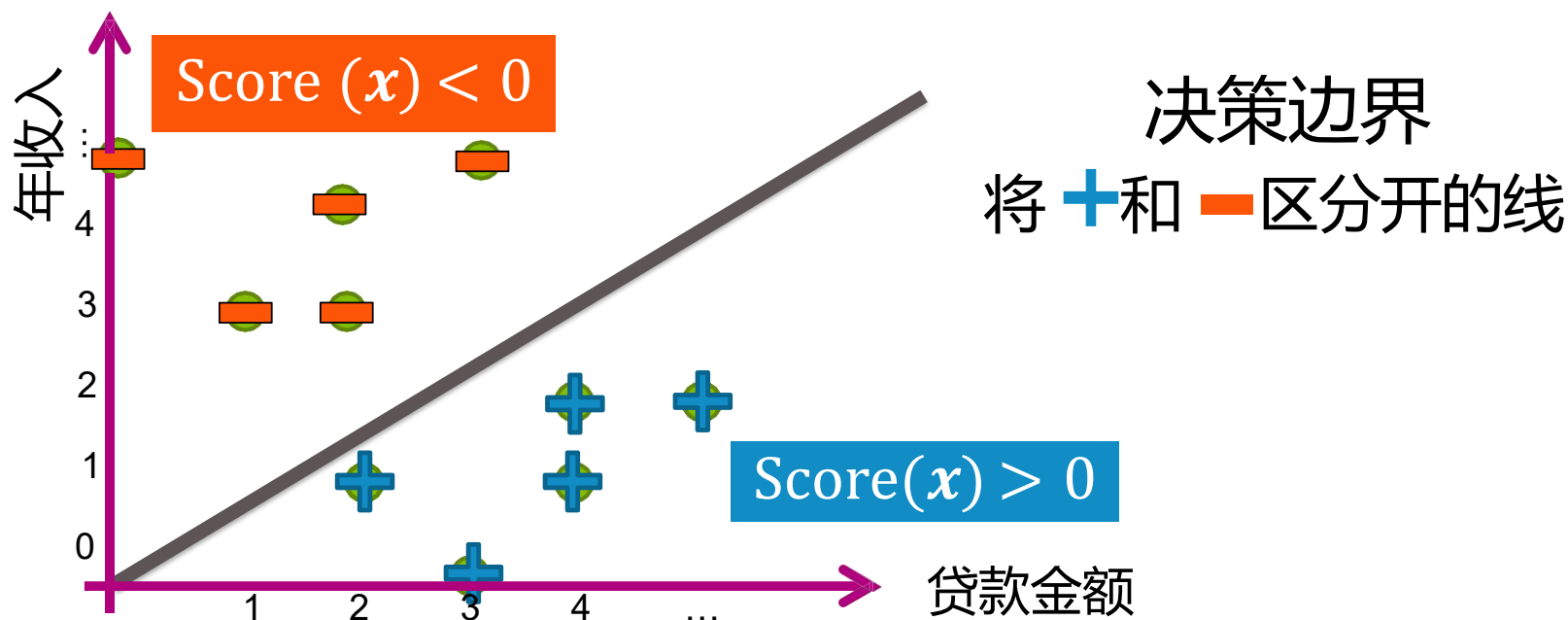


# 决策边界：Decision Boundary

- 假设只有两个特征具有非零系数

| 特征   | 系数    | 值    |
|------|-------|------|
|      | $w_0$ | 0.0  |
| 贷款金额 | $w_1$ | 1.0  |
| 年收入  | $w_2$ | -1.5 |

$$Score(x) = 1.0 \times \text{贷款金额} - 1.5 \times \text{年收入}$$

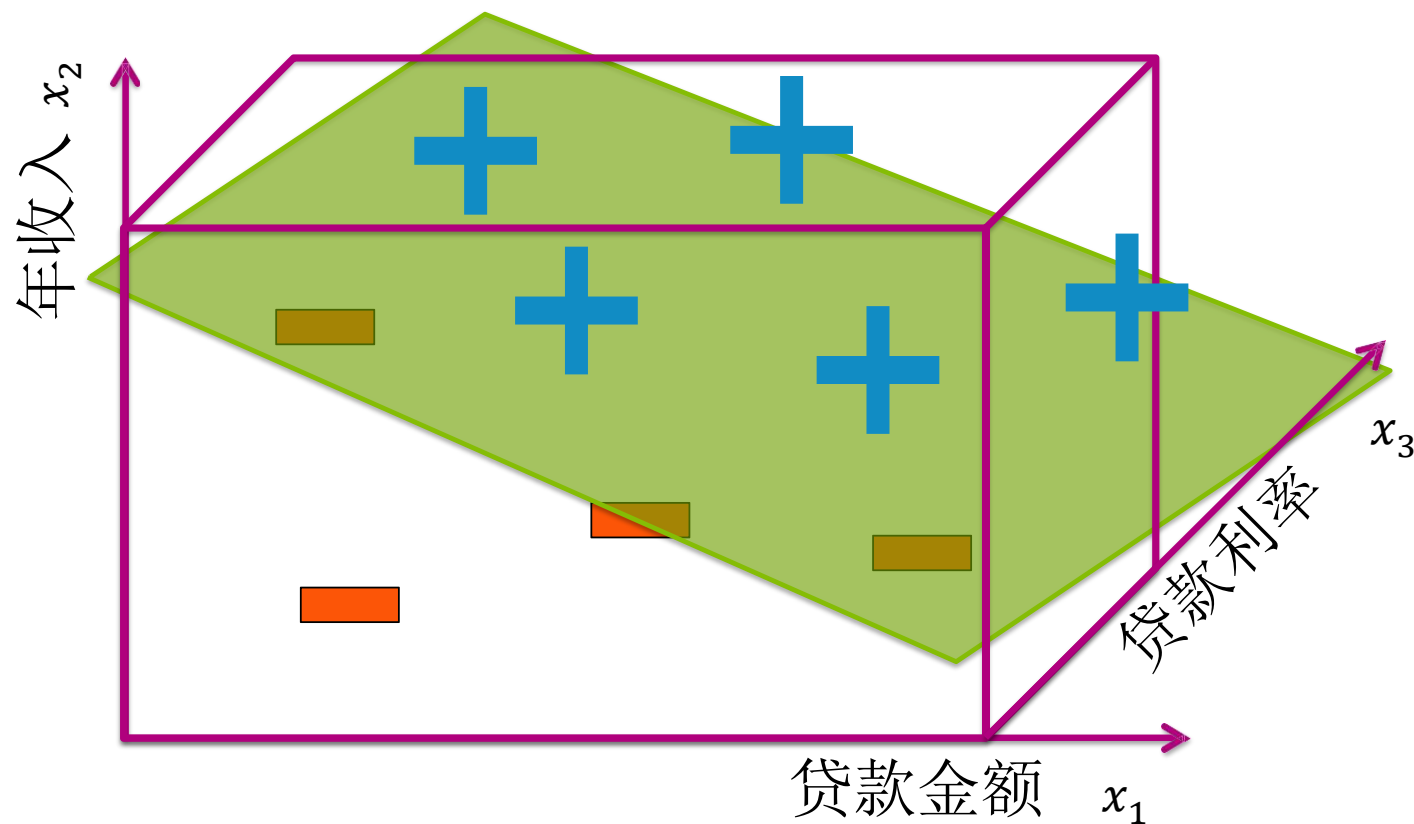


# 决策边界：Decision Boundary

- 对于更多线性特征

- $\text{Score}(x) = w_0 + w_1 * \text{贷款金额} + w_2 * \text{年收入} + w_3 * \text{贷款利率}$

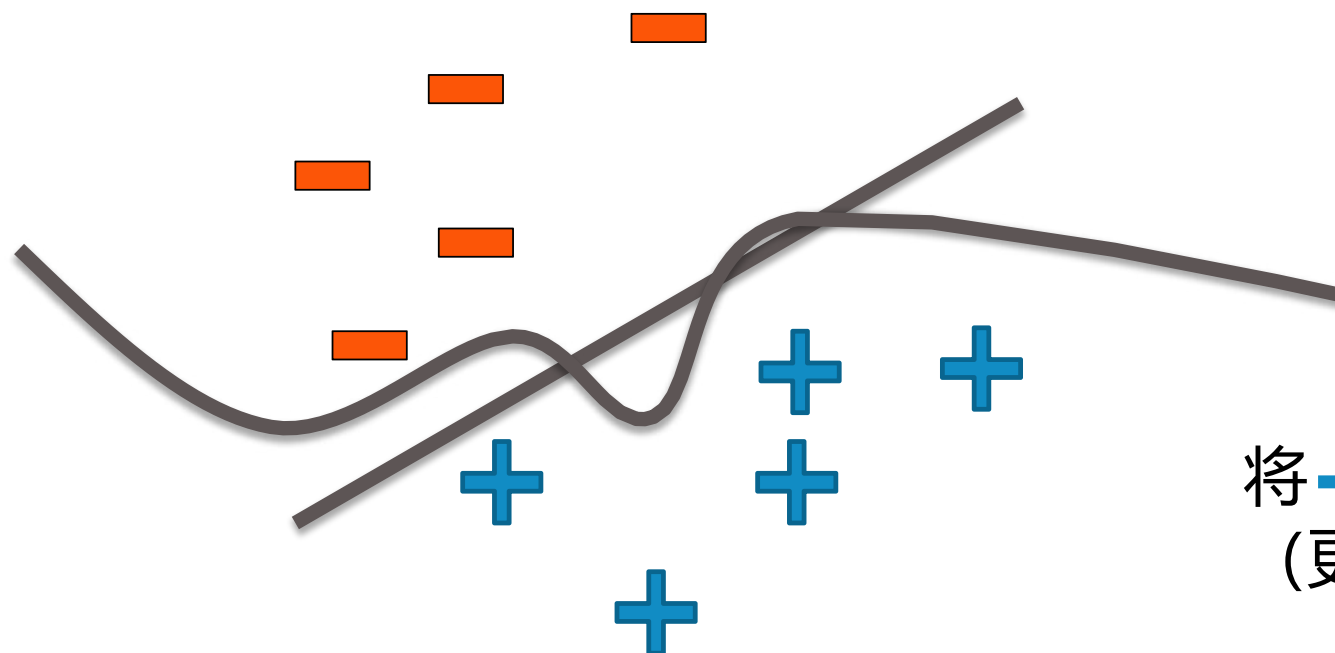
决策边界：  
将 **+** 和 **-** 区分开的面



# 决策边界：Decision Boundary

---

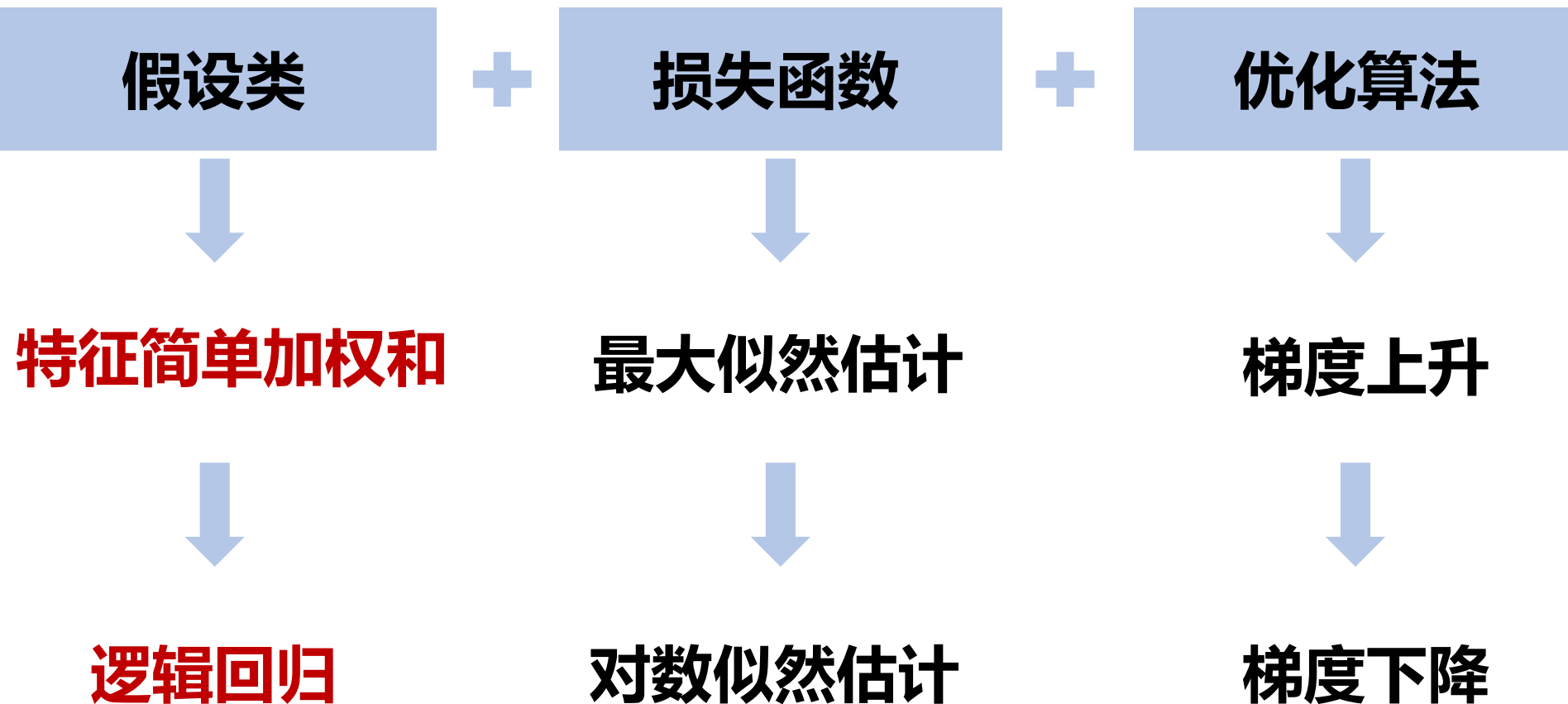
- 对于更一般的特征



决策边界  
将+和-区分开的曲面  
(更为复杂的形状)

# 逻辑回归 Logistic Regression: 框架

---



# 你的预测有多大把握？

---

- 一个简单分类器可以获得+1或-1的预测
- 但对于这个预测到底有多确定呢？

“年收入4.5万 + 贷款4万，  
利率13.67%”



很确定：+1

“年收入7.5万 + 贷款2.8万，  
利率20%”



不确定

# 在分类中使用概率

"年收入4.5万 + 贷款4万,  
利率13.67%"

很确定: **+1**

$$p(y = +1|x) = 0.99$$

"年收入7.5万 + 贷款2.8万,  
利率20%"

不确定

$$p(y = +1|x) = 0.5$$

许多分类器提供了一定程度的确定性:

输出类别

输入特征

$$P(y|x)$$

# 目标：从数据中学习条件概率

- 训练数据：N个观测值  $\{(x, y)_i\}_{i=1}^N$

| 贷款金额    | 年收入  | y = 是否违约 |
|---------|------|----------|
| 317.96k | 635k | -1       |
| 866.1k  | 305k | +1       |
| 136.08k | 45k  | +1       |
| 95.21k  | 100k | -1       |
| ...     | ...  | ...      |

在训练数据上学习  
(拟合训练数据分布)

通过寻找最优参数 $\hat{M}$   
确定最佳模型 $\hat{P}$   
可用来预测 $\hat{y}$

# 使用概率估计做好预测

- 使用概率估计做预测可以提升可解释性：
  - 预测  $\hat{y}$  并提供有多确定

贷款人信息

输入input:  $x$



使用概率估计做预测:

$\hat{p}(y|x)$  = 对类别概率的估计

If  $\hat{p}(y = +1|x) > 0.5$

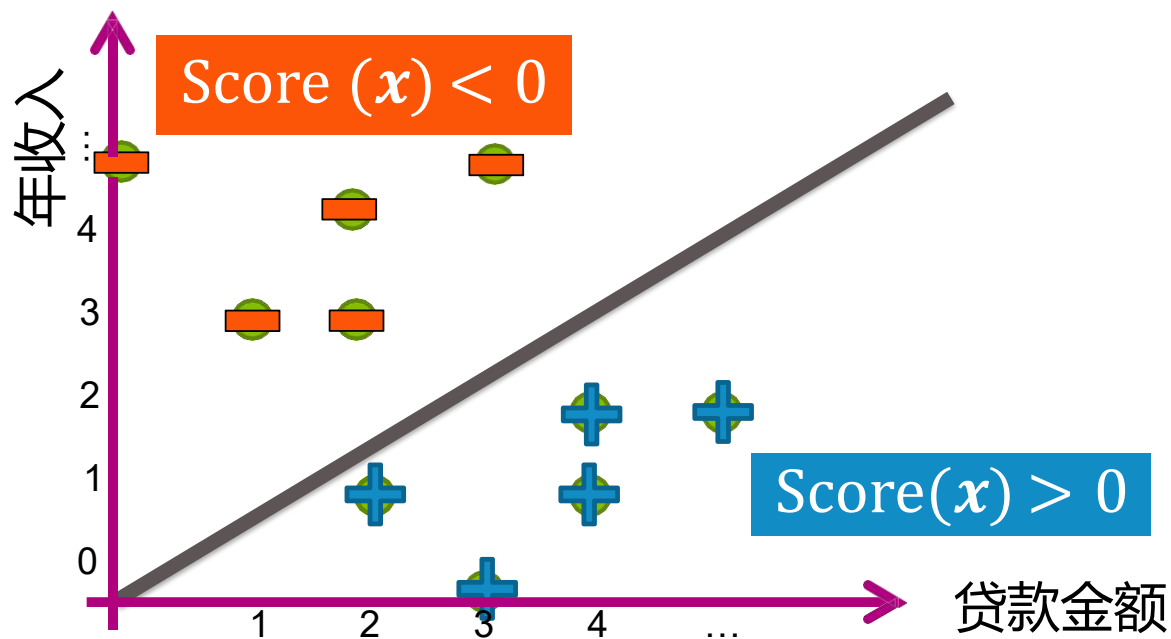
$\hat{y} = +1$

Else:

$\hat{y} = -1$

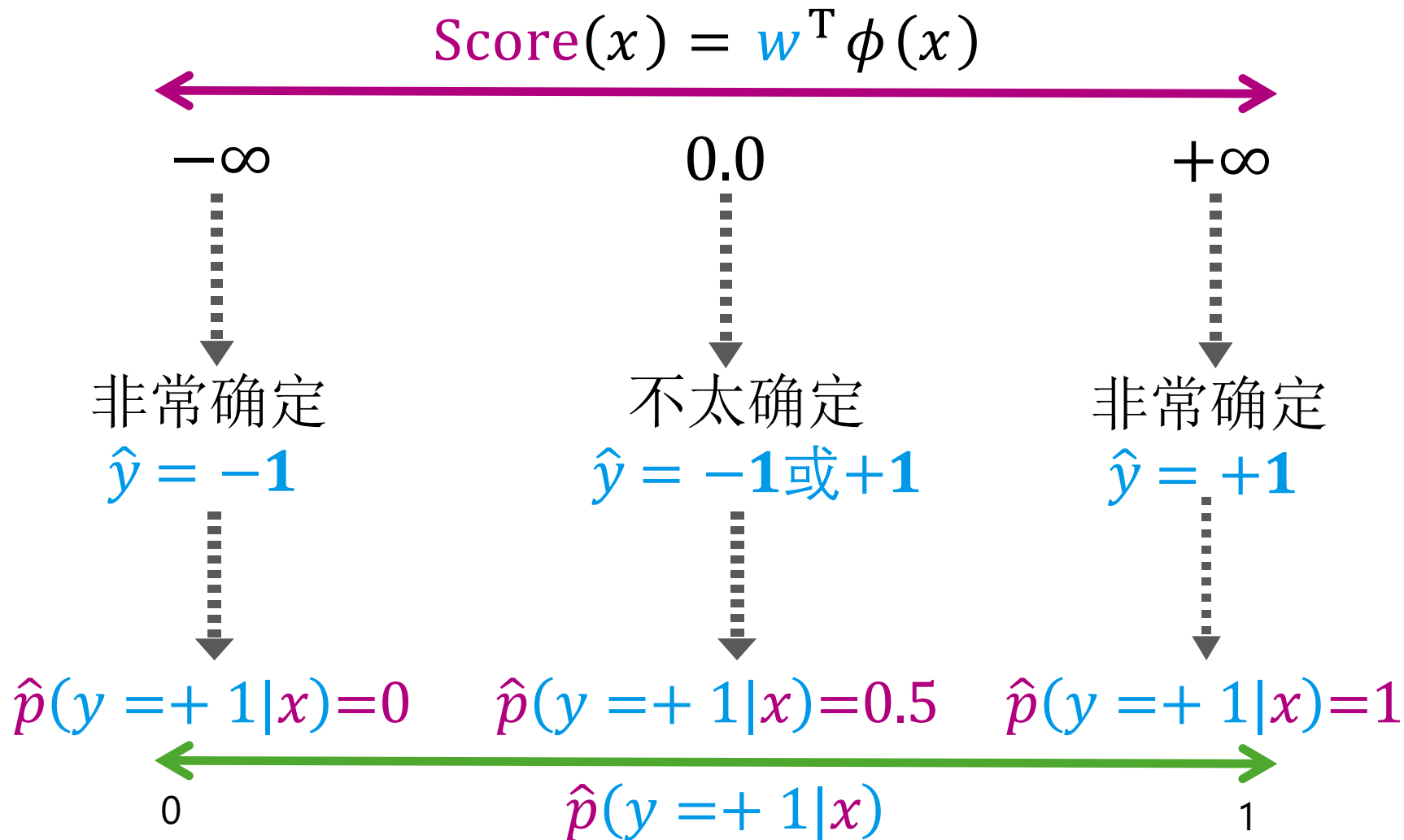
# 决策边界：Decision Boundary

- $\text{Score}(x) = w_0 \cdot \phi(x)_0 + w_1 \cdot \phi(x)_1 + \dots + w_d \cdot \phi(x)_d$   
 $= \sum_i w_i \cdot \phi(x)_i = w^T \phi(x)$



如何将  $\text{Score}(x)$  与  $\hat{p}(y|x)$  关联起来呢？

# 如何理解 $\text{Score}(x)$

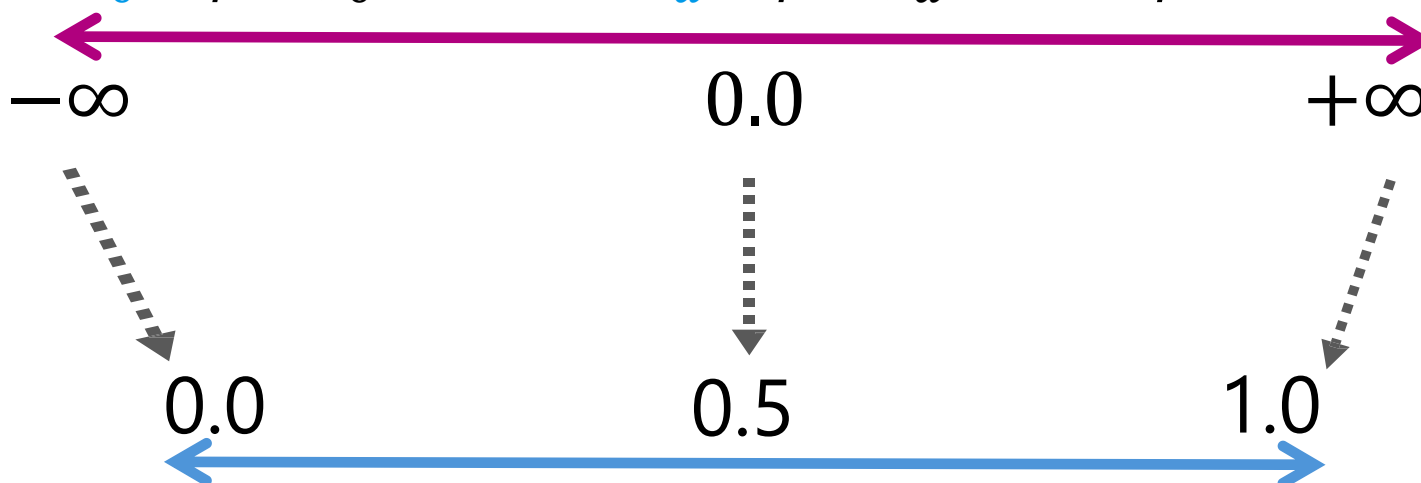


# 为什么不直接用回归来构建分类器?

$$-\infty < \text{Score}(x) < +\infty$$

如何将  $-\infty, +\infty$  映射到  $[0, 1]$

$$\text{Score}(x) = w_0 \cdot \phi(x)_0 + \dots + w_d \cdot \phi(x)_d = w^T \phi(x)$$



$$\hat{p}(y = +1|x)$$

但概率处于 0 到 1 之间

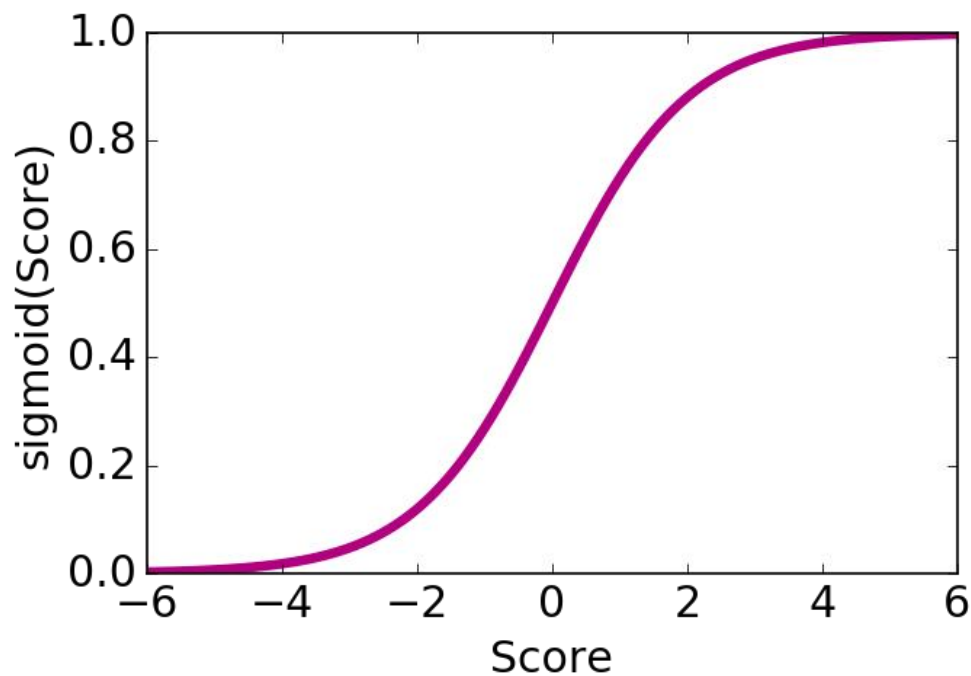
(x)

# 逻辑函数: Logistic function

- 逻辑函数 (Logistic, 也称sigmoid)

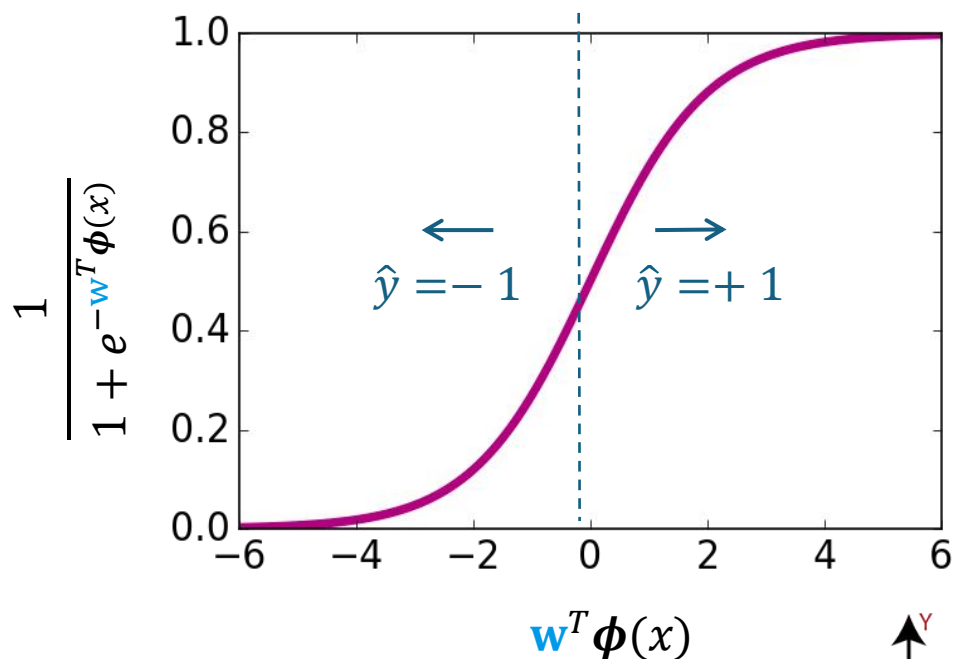
- $\text{sigmoid}(\text{Score}) = \frac{1}{1+e^{-\text{Score}}}$

| <i>Score</i> | <i>sigmoid(Score)</i> |
|--------------|-----------------------|
| $-\infty$    | 0                     |
| -2           | 0.12                  |
| 0            | 0.5                   |
| 2            | 0.88                  |
| $\infty$     | 1                     |

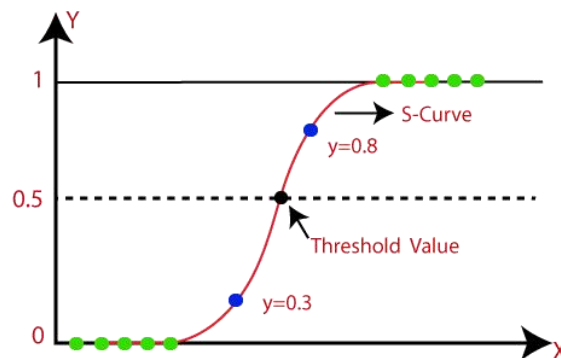


# 逻辑回归: Logistic Regression

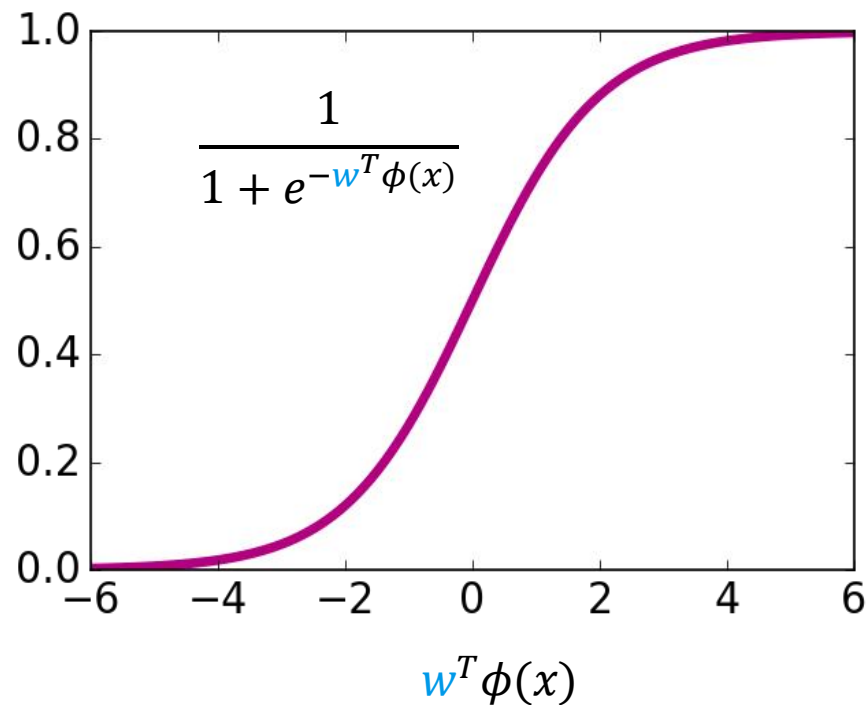
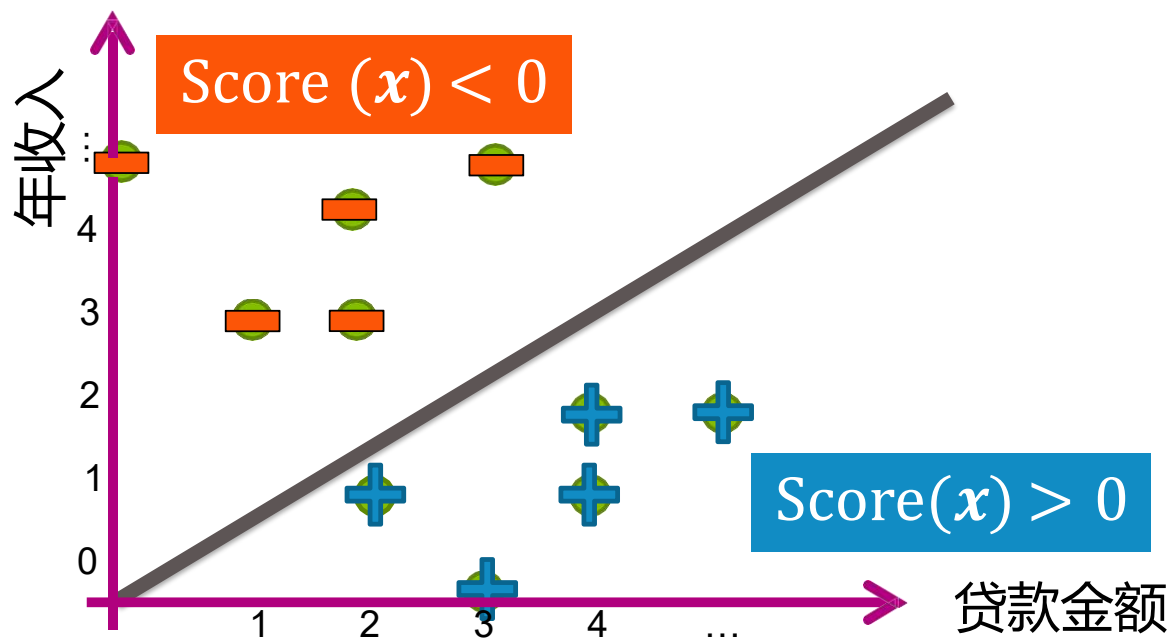
- $\hat{P}(y = +1 | x, w) = \text{sigmoid}(\text{Score}(x)) = \frac{1}{1 + e^{-w^T \phi(x)}}$



| $\text{Score}(x)$ | $\text{sigmoid}(\text{Score}(x))$ |
|-------------------|-----------------------------------|
| 0                 | 0.5                               |
| -2                | 0.12                              |
| 2                 | 0.88                              |
| 4                 | 0.98                              |



# 逻辑回归 → 线性决策边界

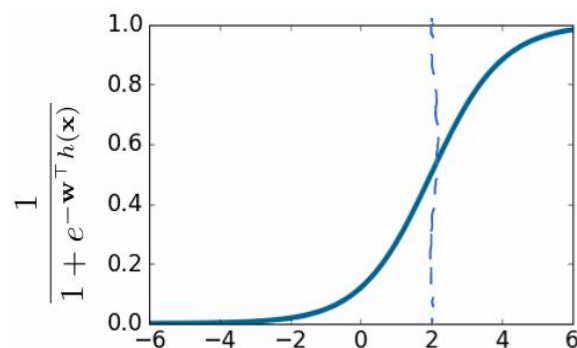


逻辑回归的决策边界由方程  $w^T \phi(x) = 0$  定义，是一个线性超平面

# 逻辑回归：模型参数的影响

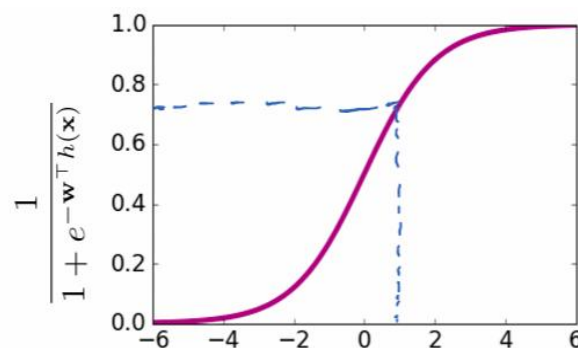
- Score( $x$ ) =  $w_0 + w_1 * \text{贷款金额} + w_2 * \text{年收入} = w^T \phi(x)$

|       |    |
|-------|----|
| $w_0$ | -2 |
| $w_1$ | +1 |
| $w_2$ | -1 |



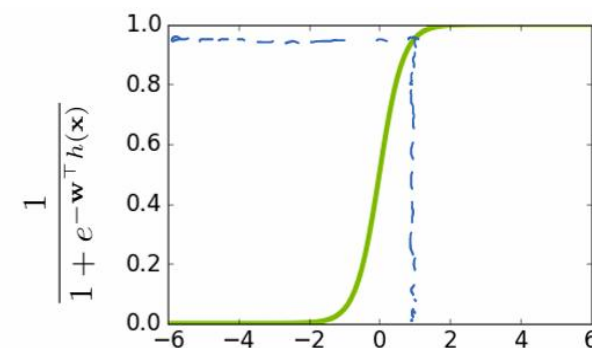
#贷款金额-#年收入

|       |    |
|-------|----|
| $w_0$ | 0  |
| $w_1$ | +1 |
| $w_2$ | -1 |



#贷款金额-#年收入

|       |    |
|-------|----|
| $w_0$ | 0  |
| $w_1$ | +3 |
| $w_2$ | -3 |

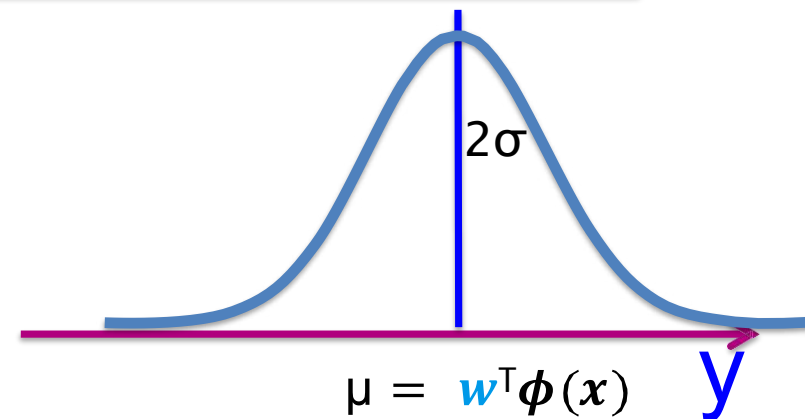


#贷款金额-#年收入

**$|w_j|$ 越大，对应特征 $x_j$ 对预测结果的影响越大；  
系数大表示该特征能显著改变预测概率**

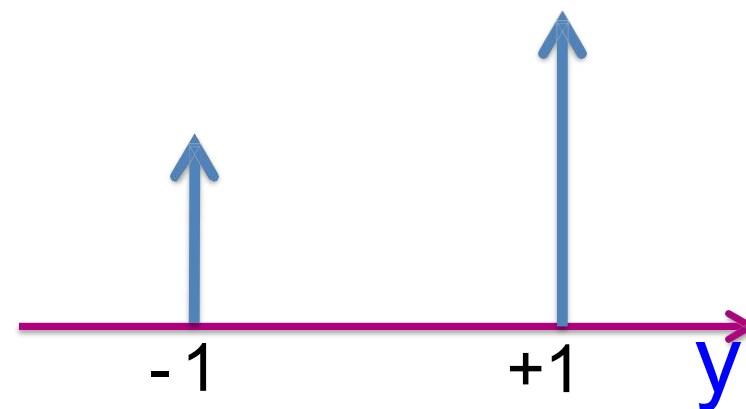
# 回归模型比较：线性回归 vs 逻辑回归

- 具有高斯误差的线性回归
- $y = \mathbf{w}^T \boldsymbol{\phi}(x) + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2)$   
 $\Rightarrow p(y|x, \mathbf{w}) = \mathcal{N}(y; \mathbf{w}^T \boldsymbol{\phi}(x), \sigma^2)$



- 逻辑回归

$$P(y|x) = \begin{cases} \frac{1}{1+e^{-\mathbf{w}^T \boldsymbol{\phi}(x)}} , & y = +1 \\ \frac{e^{-\mathbf{w}^T \boldsymbol{\phi}(x)}}{1+e^{-\mathbf{w}^T \boldsymbol{\phi}(x)}} , & y = -1 \end{cases}$$



# 提纲

---

- **Logistic Regression 框架**

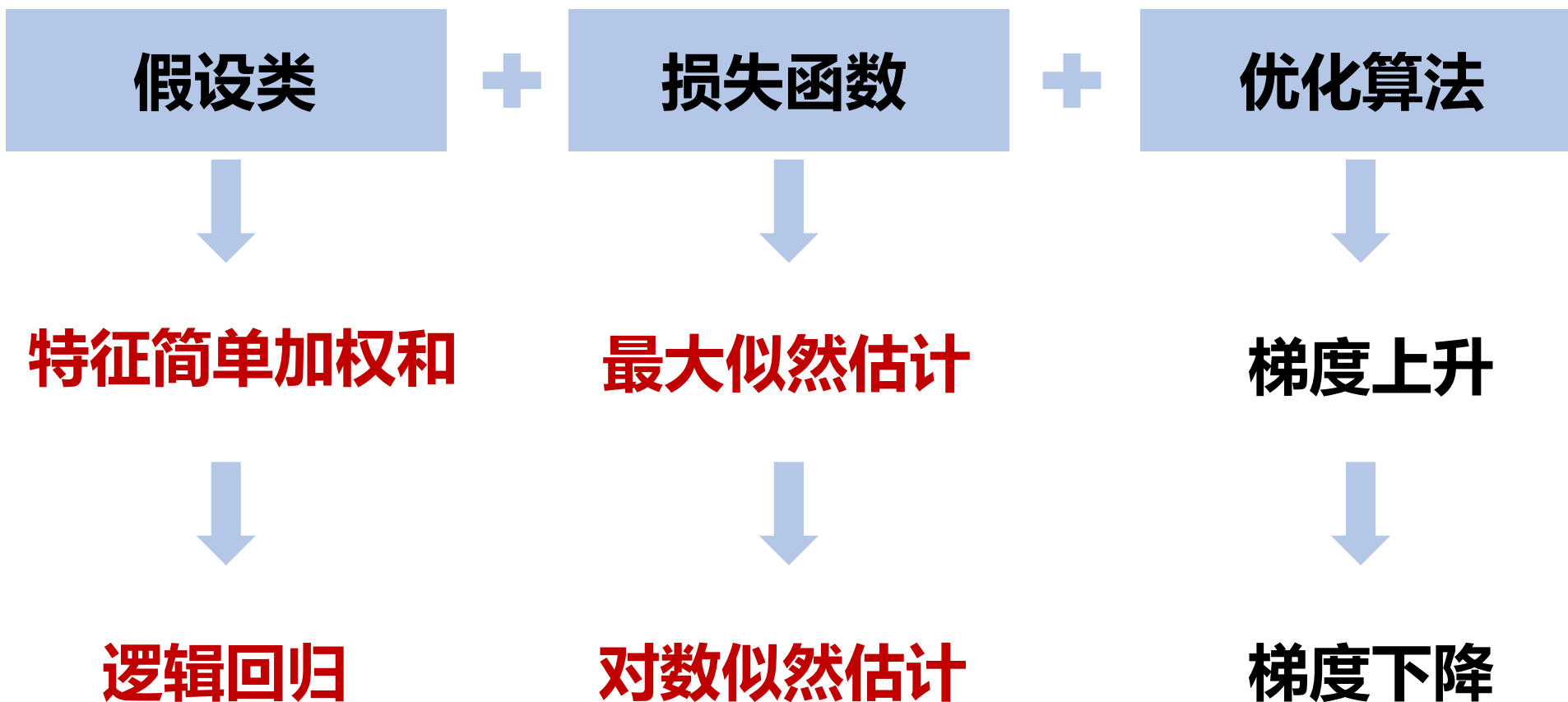
- 假设类
- 损失函数
- 优化算法

- 过拟合问题

- 正则化

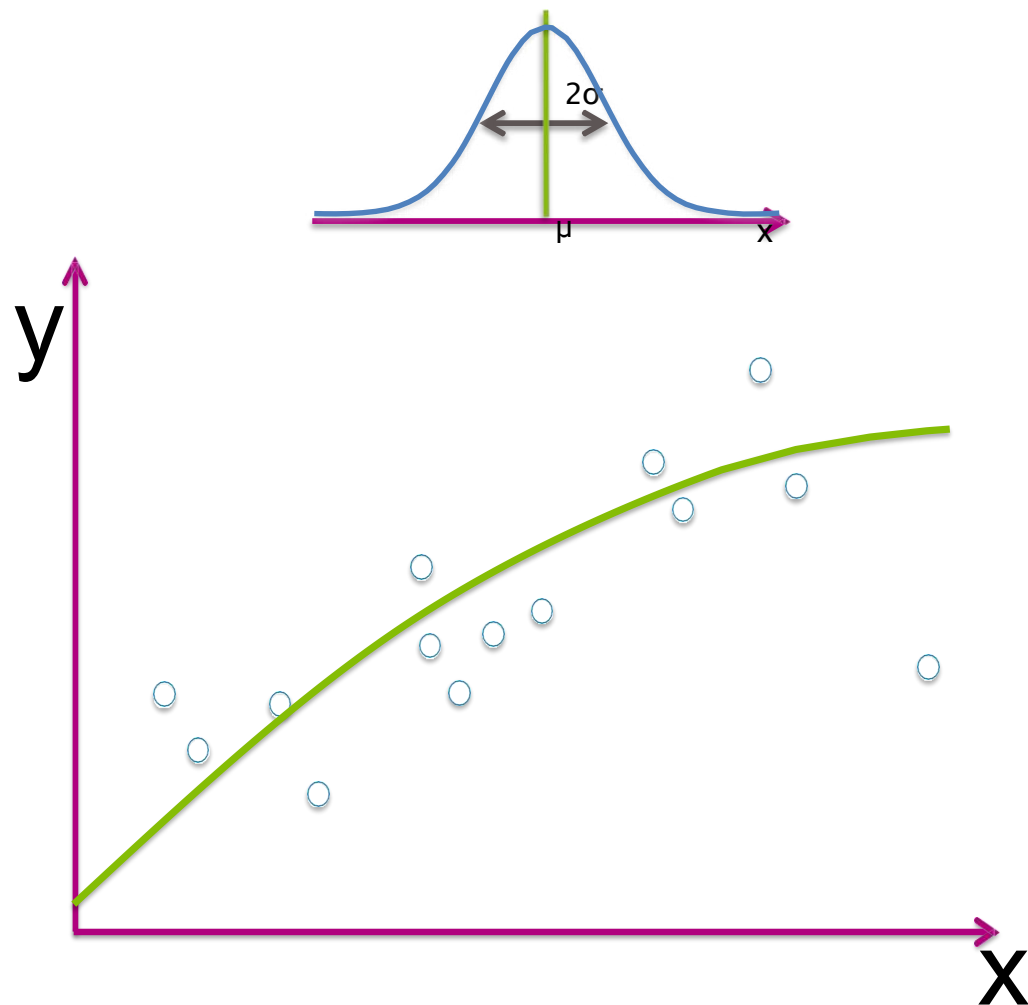
# 逻辑回归 Logistic Regression: 框架

---



# Recap: 线性回归模型 (高斯)

- $\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $\epsilon$ 
  - 基础假设:  $\mathbb{E}(\epsilon) = 0$
  - 进一步假设:  $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- $y$ 
  - $y = w^T \phi(x) + \epsilon$
  - $p(y|x; w) \sim \mathcal{N}(w^T \phi(x), \sigma^2)$



# Recap: 最大似然估计 MLE

- 最大似然函数

$$\operatorname{argmax}_{\mathbf{w}} p(y|x, \mathbf{w}) = \operatorname{argmax}_{\mathbf{w}} \prod_{(x,y) \in \mathcal{D}} p(y|x, \mathbf{w})$$

- 对数似然函数

$$\operatorname{argmax}_{\mathbf{w}} \ln p(y|x, \mathbf{w}) = \operatorname{argmax}_{\mathbf{w}} \ln \prod_{(x,y) \in \mathcal{D}} p(y|x, \mathbf{w})$$

概率密度函数

$$\begin{aligned} &= \operatorname{argmax}_{\mathbf{w}} \sum_{(x,y) \in \mathcal{D}} \ln \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - \mathbf{w}^T \cdot \phi(x))^2}{2\sigma^2}} \\ &= \operatorname{argmin}_{\mathbf{w}} \sum_{(x,y) \in \mathcal{D}} (y - \mathbf{w}^T \cdot \phi(x))^2 \end{aligned}$$

高斯模型下的最大似然估计 = 最小化平方损失

# 逻辑回归：需要寻找最佳参数

---

| 贷款金额    | 年收入  | y = 是否违约 |
|---------|------|----------|
| 317.96k | 635k | -1       |
| 866.10k | 305k | +1       |
| 136.08k | 45k  | +1       |
| 95.21k  | 100k | -1       |
| ...     | ...  | ...      |

# 逻辑回归：需要寻找最佳参数

| 贷款金额    | 年收入  | y = 是否违约 |
|---------|------|----------|
| 866.1k  | 305k | +1       |
| 136.08k | 45k  | +1       |
| ...     | ...  | ...      |

$$P(y = +1 | x, \mathbf{w}) = 1.0$$

| 贷款金额    | 年收入  | y = 是否违约 |
|---------|------|----------|
| 317.96k | 635k | -1       |
| 95.21k  | 100k | -1       |
| ...     | ...  | ...      |

$$P(y = -1 | x, \mathbf{w}) = 1.0$$

我们希望 $\hat{\mathbf{w}}$ 满足

# 逻辑回归：利用最大似然估计学习逻辑回归

- 数据点
  - $(x_i, y_i)$

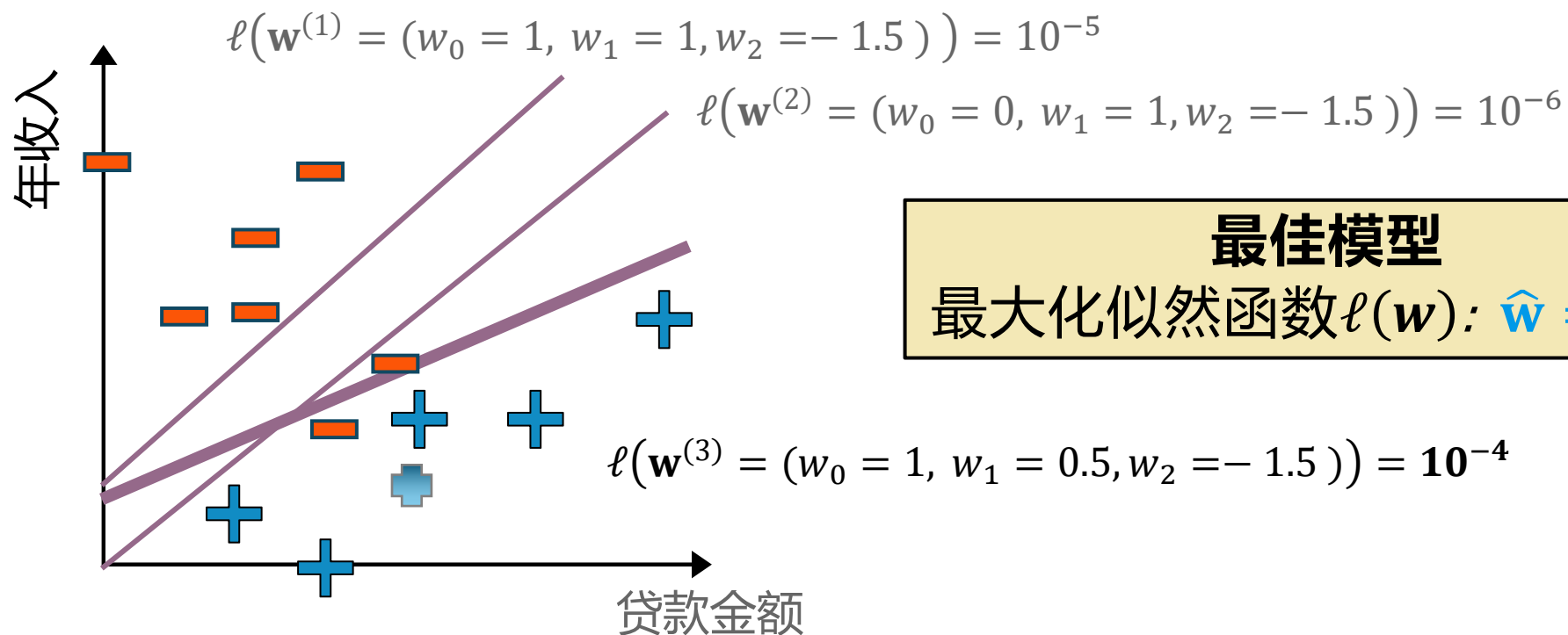
| 贷款金额    | 年收入  | y = 是否违约 |
|---------|------|----------|
| 317.96k | 635k | -1       |
| 866.1k  | 305k | +1       |
| 136.08k | 45k  | +1       |
| 95.21k  | 100k | -1       |
| ...     | ...  | ...      |

- $P(\mathbf{w}) = P(y_1|x_1, \mathbf{w}) \ P(y_2|x_2, \mathbf{w}) \ P(y_3|x_3, \mathbf{w}) \ P(y_4|x_4, \mathbf{w})$   
$$\prod_{i=1}^N P(y_i|x_i, \mathbf{w})$$

# 逻辑回归：确定“最佳”分类器

- 对于所有可能的  $w_0, w_1, w_2$ , 选择似然性最大的

- $\ell(w) = \prod_{i=1}^N P(y_i | x_i, w)$



**最佳模型**  
最大化似然函数  $\ell(w)$ :  $\hat{w} = w^{(3)}$

# 提纲

---

- **Logistic Regression 框架**

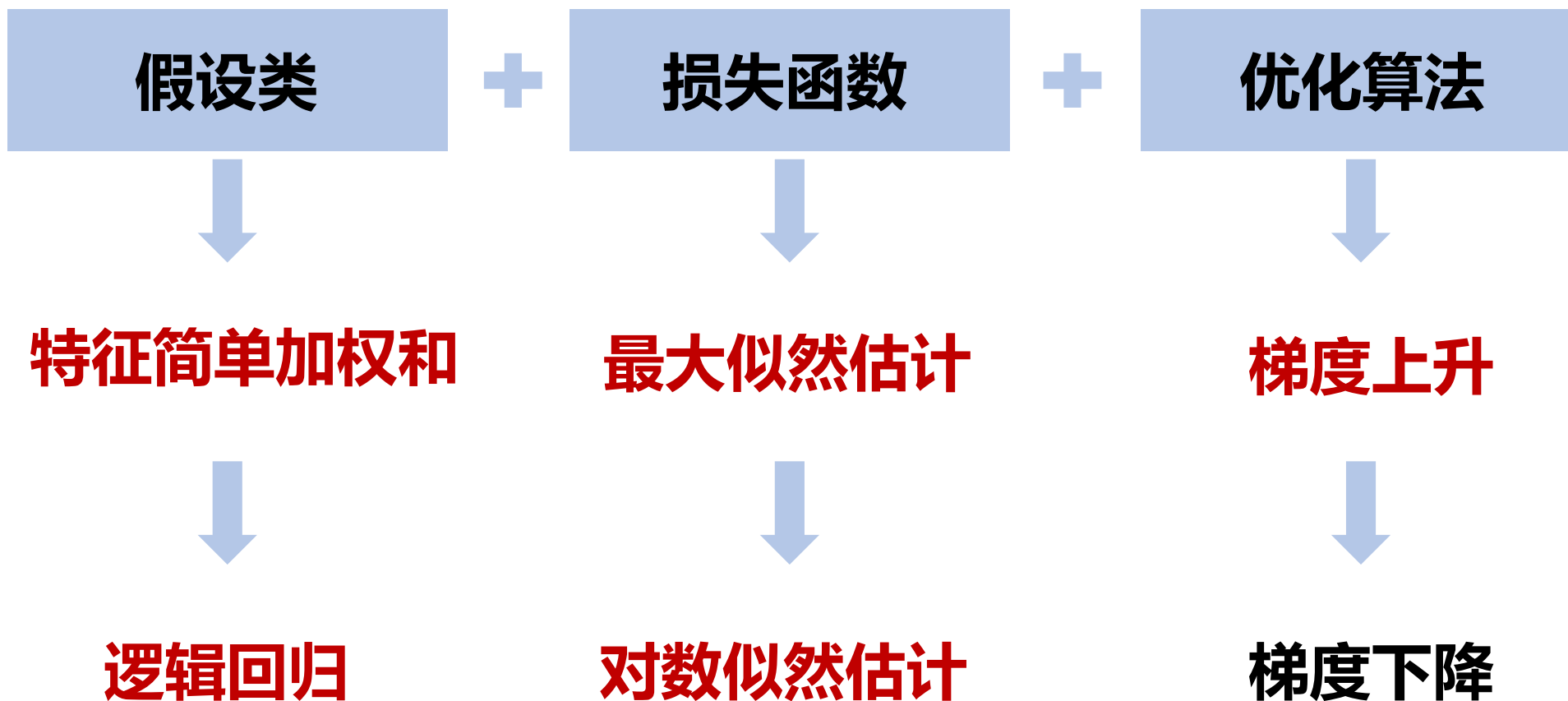
- 假设类
- 损失函数
- 优化算法

- 过拟合问题

- 正则化

# 逻辑回归 Logistic Regression: 框架

---

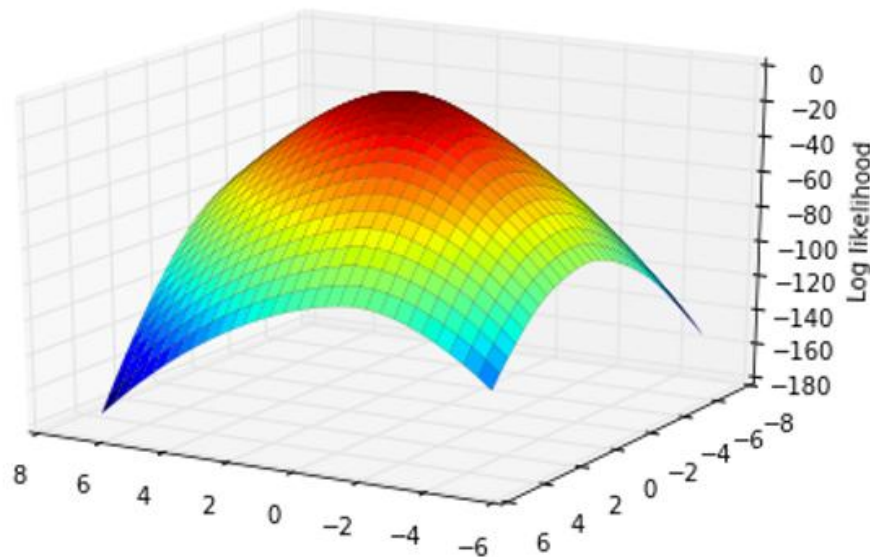


# 梯度上升优化MLE

- 最大化似然估计
  - 对于所有可能的 $w_0, w_1, w_2$ , 最大化函数:

$$\prod_{i=1}^N P(y_i|x_i, \mathbf{w})$$

$\ell(w_0, w_1, w_2)$  为三变量函数



二变量函数可视化

# 梯度上升优化MLE

- $P(y = +1|x, w) = \text{sigmoid}(\text{Score}(x)) = \sigma(w^T \phi(x)) = \frac{1}{1 + e^{-w^T \phi(x)}}$
- 对于N个样本的数据集  $(\{(x_i, y_i)\}_{i=1}^N)$  , 似然 (likelihood) 估计为:

$$\max_w l(w) = \prod_{i=1}^N P(y_i|x_i, w)$$

- 重写为对数似然 (log-likelihood) 估计:

$$\max_w ll(w) = \sum_{i=1}^N \log(P(y_i|x_i, w))$$

# 逻辑回归对数似然的梯度

---

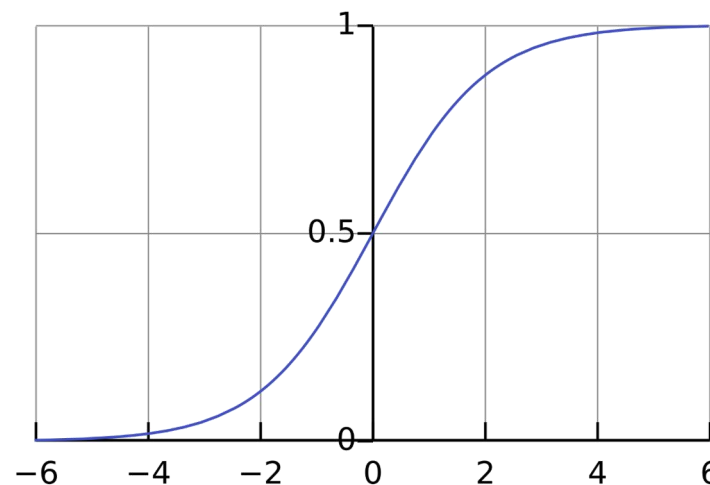
- Sigmoid函数求导

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

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

# 逻辑回归对数似然的梯度

- 逻辑回归假设正类概率为：
  - $P(y_i|x_i, \mathbf{w}) = \sigma(\mathbf{w}^T \phi(x_i))$
- 负类概率为：
  - $P(y_i|x_i, \mathbf{w}) = 1 - \sigma(\mathbf{w}^T \phi(x))$
- Sigmoid的性质:  $\sigma(-z) = 1 - \sigma(z)$
- 当  $y_i \in \{-1, +1\}$ ,  $P(y_i|x_i, \mathbf{w}) = \sigma(y_i \cdot \mathbf{w}^T \phi(x_i))$



# 逻辑回归对数似然的梯度

- $ll(\mathbf{w})$ 对 $w_j$ 求偏导

$$\begin{aligned}\frac{\partial ll(\mathbf{w})}{\partial w_j} &= \sum_{i=1}^N \frac{\partial}{\partial w_j} \log \left( \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) \right) \\ &= \sum_{i=1}^N \frac{1}{\sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i))} \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) \left( 1 - \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) \right) \frac{\partial (y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i))}{\partial w_j} \\ &= \sum_{i=1}^N \left( 1 - \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) \right) \frac{\partial (y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i))}{\partial w_j} \\ &= \sum_{i=1}^N \left( 1 - \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) \right) y_i \boldsymbol{\phi}_j(x_i)\end{aligned}$$

# 逻辑回归对数似然的梯度

- $\frac{\partial ll(\mathbf{w})}{\partial w_j} = \sum_{i=1}^N (1 - \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i))) y_i \boldsymbol{\phi}_j(x_i)$

- 当  $y_i = +1$  时

$$\begin{aligned}\sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) &= \sigma(\mathbf{w}^T \boldsymbol{\phi}(x_i)) \\ (1 - \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i))) y_i \boldsymbol{\phi}_j(x_i) &= (1 - \sigma(\mathbf{w}^T \boldsymbol{\phi}(x_i))) \boldsymbol{\phi}_j(x_i) \\ &= (1 - P(y_i = +1 | x_i, \mathbf{w})) \boldsymbol{\phi}_j(x_i) \\ &= \boldsymbol{\phi}_j(x_i) (\mathbf{1}[y_i = +1] - P(y_i = +1 | x_i, \mathbf{w}))\end{aligned}$$

- 当  $y_i = -1$  时

$$\begin{aligned}\sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i)) &= \sigma(-\mathbf{w}^T \boldsymbol{\phi}(x_i)) \\ &= 1 - \sigma(\mathbf{w}^T \boldsymbol{\phi}(x_i)) \\ (1 - \sigma(y_i \cdot \mathbf{w}^T \boldsymbol{\phi}(x_i))) y_i \boldsymbol{\phi}_j(x_i) &= - \left[ 1 - (1 - \sigma(\mathbf{w}^T \boldsymbol{\phi}(x_i))) \right] \boldsymbol{\phi}_j(x_i) \\ &= \boldsymbol{\phi}_j(x_i) (\mathbf{1}[y_i = +1] - P(y_i = +1 | x_i, \mathbf{w}))\end{aligned}$$

# 逻辑回归对数似然的梯度

- $\frac{\partial ll(\mathbf{w})}{\partial w_j} = \sum_{i=1}^N \boldsymbol{\phi}_j(x_i) (\mathbf{1}[y_i = +1] - P(y_i = +1 | x_i, \mathbf{w}))$

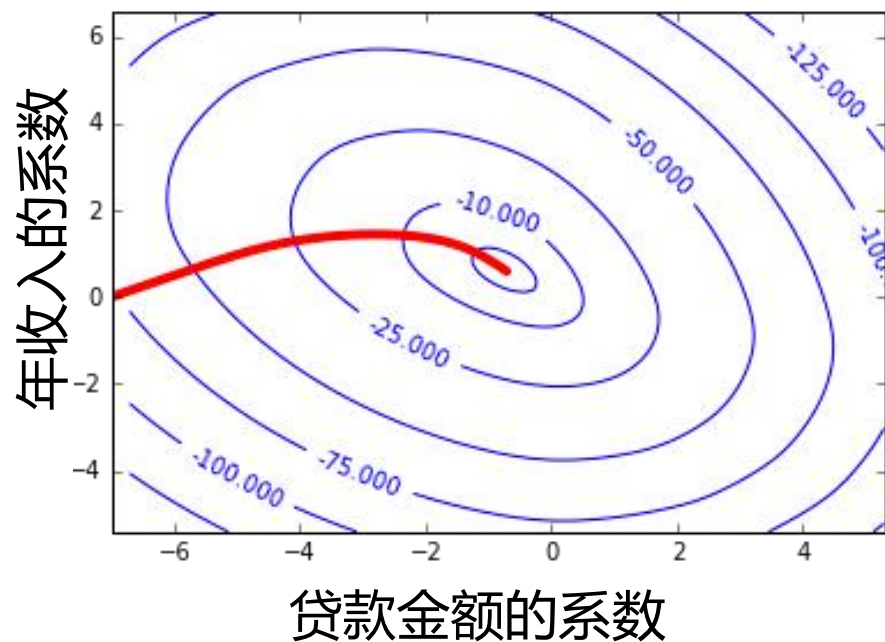
- $\boldsymbol{\phi}_j(x_i)$ : 特征向量对应值

- $\Delta_i = \mathbf{1}[y_i = +1] - P(y_i = +1 | x_i, \mathbf{w})$ : 真实值与预测值的差

|            | $P(y_i = +1   x_i, \mathbf{w}) \approx 1$                                             | $P(y_i = +1   x_i, \mathbf{w}) \approx 0$                                            |
|------------|---------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|
| $y_i = +1$ | $\Delta_i \approx 0$<br>$P(y_i = +1   x_i, \mathbf{w})$ 不变                            | $\Delta_i \approx 1 \rightarrow \mathbf{w}$ 增加<br>$P(y_i = +1   x_i, \mathbf{w})$ 增加 |
| $y_i = -1$ | $\Delta_i \approx -1 \rightarrow \mathbf{w}$ 下降<br>$P(y_i = +1   x_i, \mathbf{w})$ 下降 | $\Delta_i \approx 0$<br>$P(y_i = +1   x_i, \mathbf{w})$ 不变                           |

# 梯度上升优化MLE

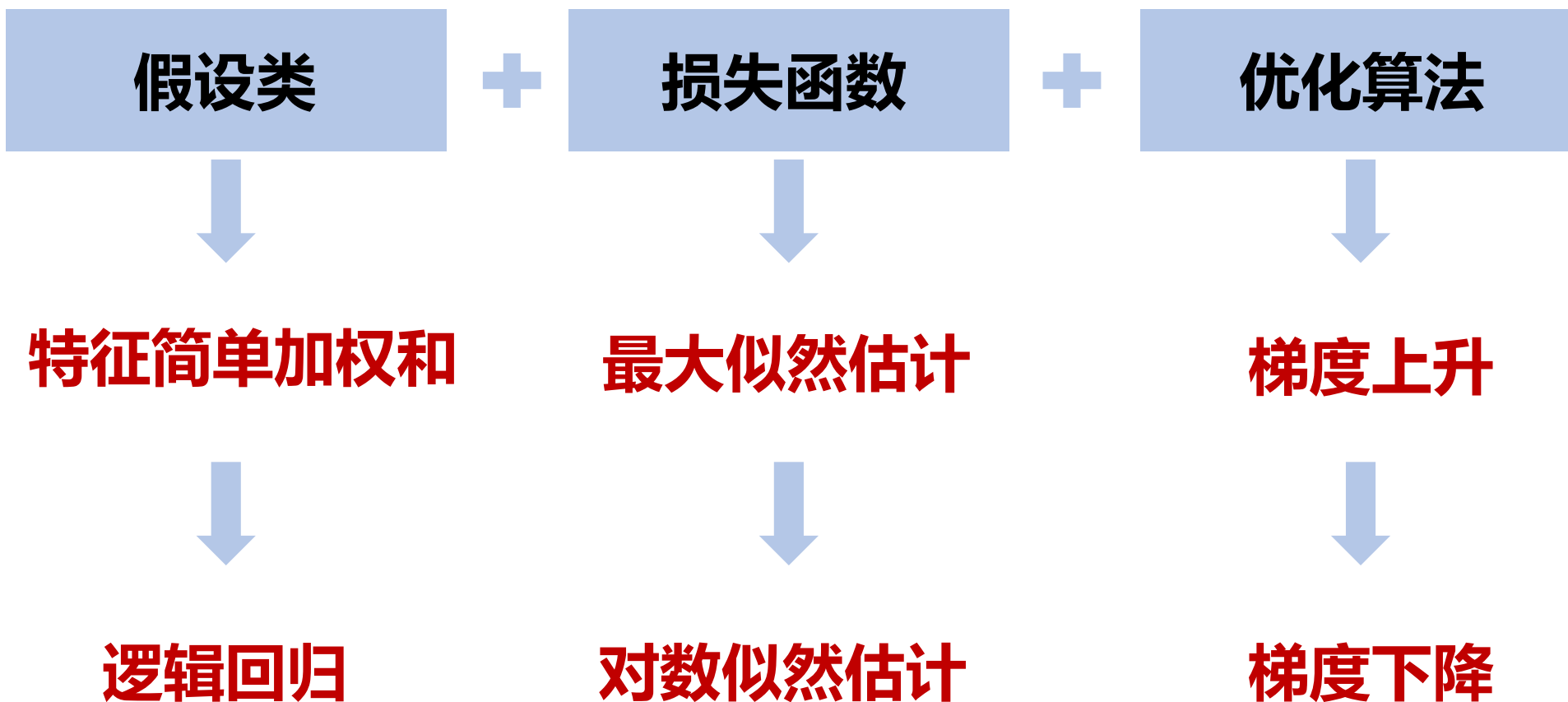
- 逻辑回归的梯度上升



```
init  $\mathbf{w}^{(1)} = 0, t = 1$   
while  $\|\nabla \ell(\mathbf{w}^{(t)})\| > \varepsilon$   
  for  $j = 0, \dots, D$   
     $partial[j] = \sum_{i=1}^N \phi_j(x_i) (\mathbf{1}[y_i = +1] - P(y_i = +1 | x_i, \mathbf{w}^{(t)}))$   
     $\mathbf{w}_j^{(t+1)} \leftarrow \mathbf{w}_j^{(t)} + \eta partial[j]$   
   $t \leftarrow t + 1$ 
```

# 逻辑回归 Logistic Regression: 框架

---



# 随机梯度下降

---

- 梯度上升
  - 最大化对数似然
- 梯度下降
  - 最小化负对数似然

二者等价

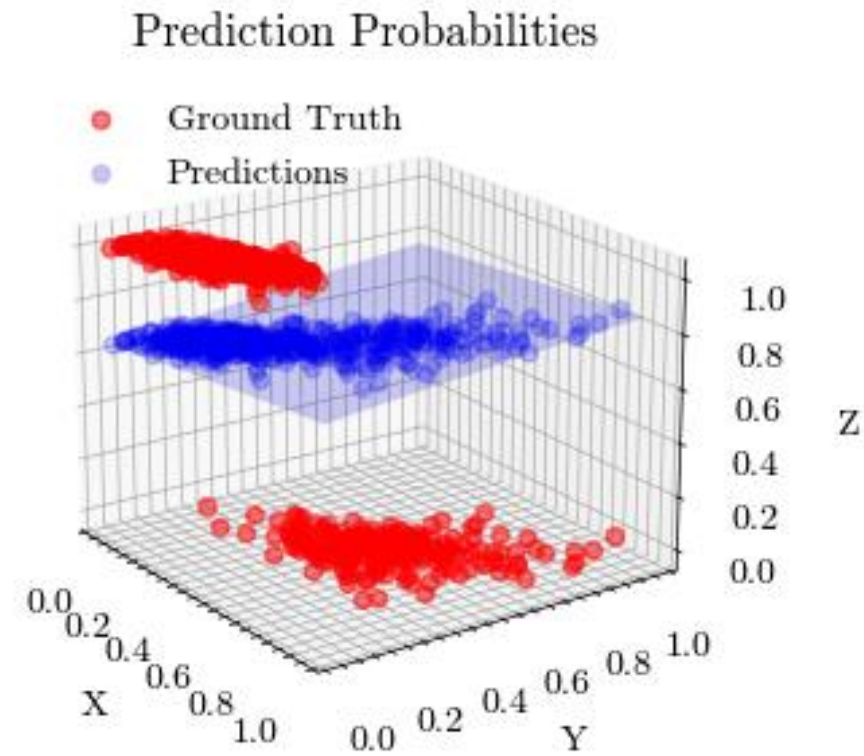
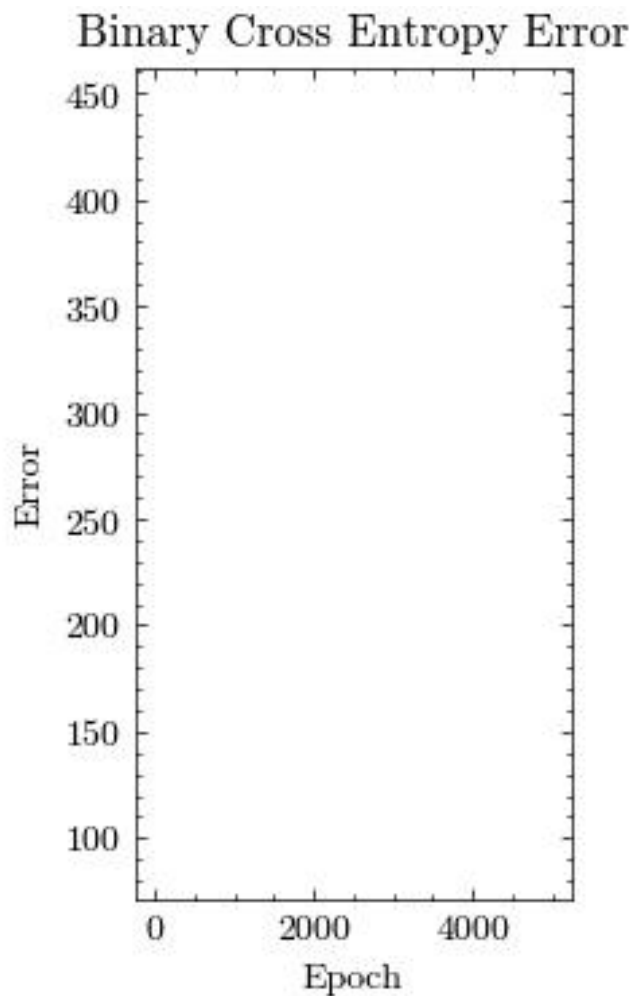
- 梯度下降
  - Batch Gradient Descent (BGD)
  - Stochastic Gradient Descent (SGD)
  - Mini-batch Gradient Descent

# 总结

---

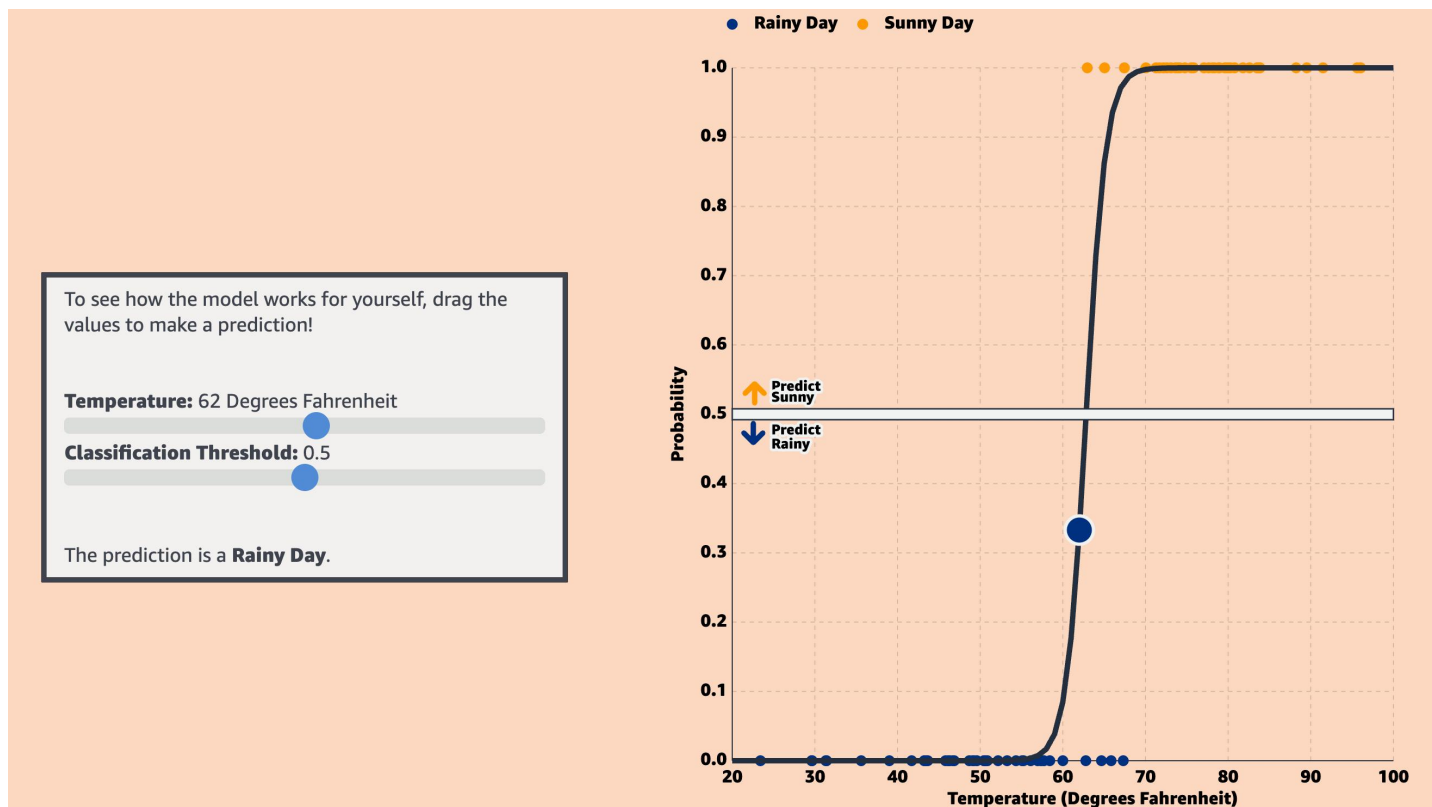
- 描述决策边界与线性分类器
- 利用类别概率表达预测的置信程度
- 定义逻辑回归模型
- 将逻辑回归的输出解释为类别概率
- 说明回归系数数值对逻辑回归输出的影响
- 使用似然函数来度量分类器的质量
- 通过最小化负对数似然损失，用梯度下降学习逻辑回归模型

# 逻辑回归示例

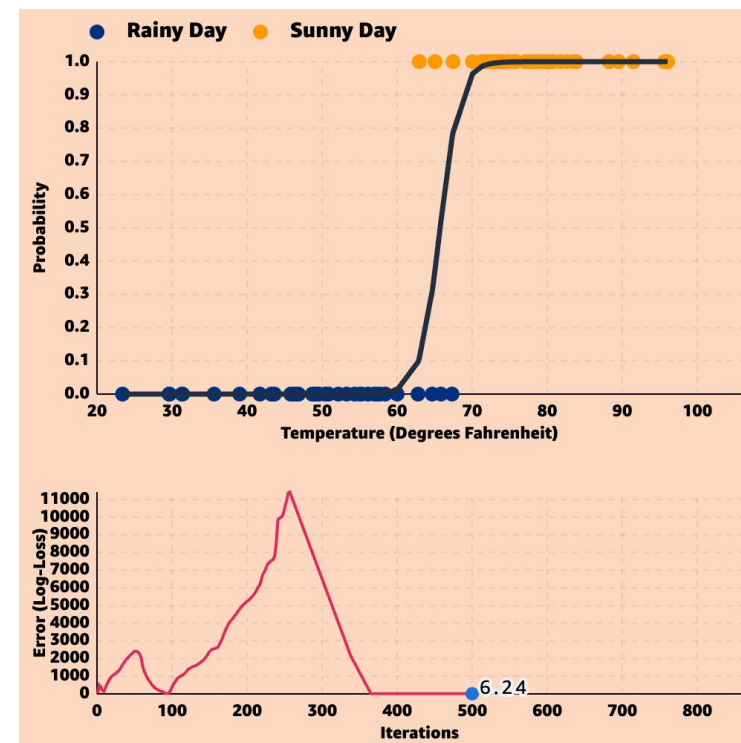
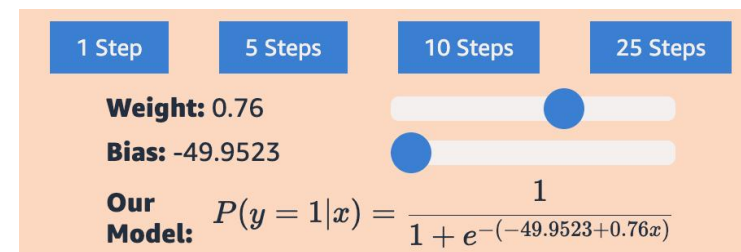


[Interactive Example](#)

# 逻辑回归示例



## Interactive Example



# 提纲

---

- Logistic Regression 框架
- **过拟合问题**
- 正则化

# 更复杂的模型往往具有更小的偏差

---

使用一个特征（如年收入）判断是否违约，可能达到预期效果，但...



对于年收入尚可、贷款金额巨大的人而言，这个人是否会违约？



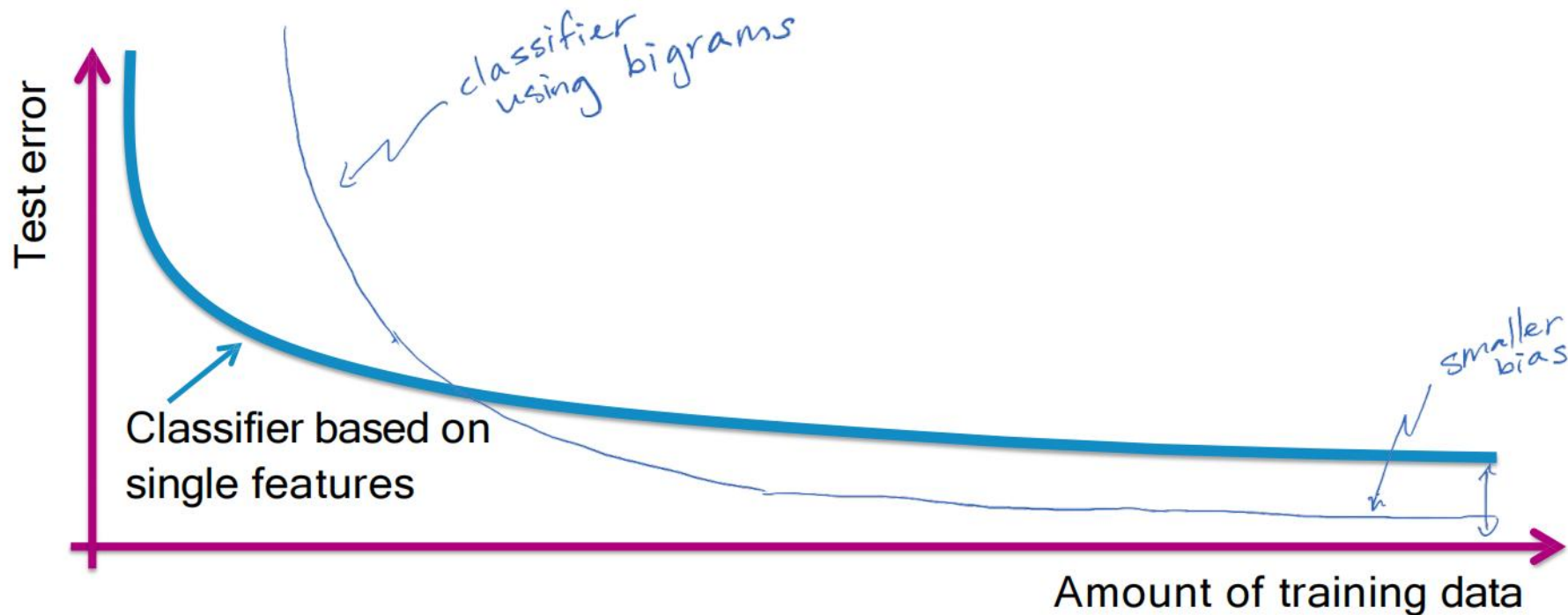
所以，我们需要更复杂的模型，来考虑更多的特征？



更少的偏差 → 可能更准确，需要更多数据来学习

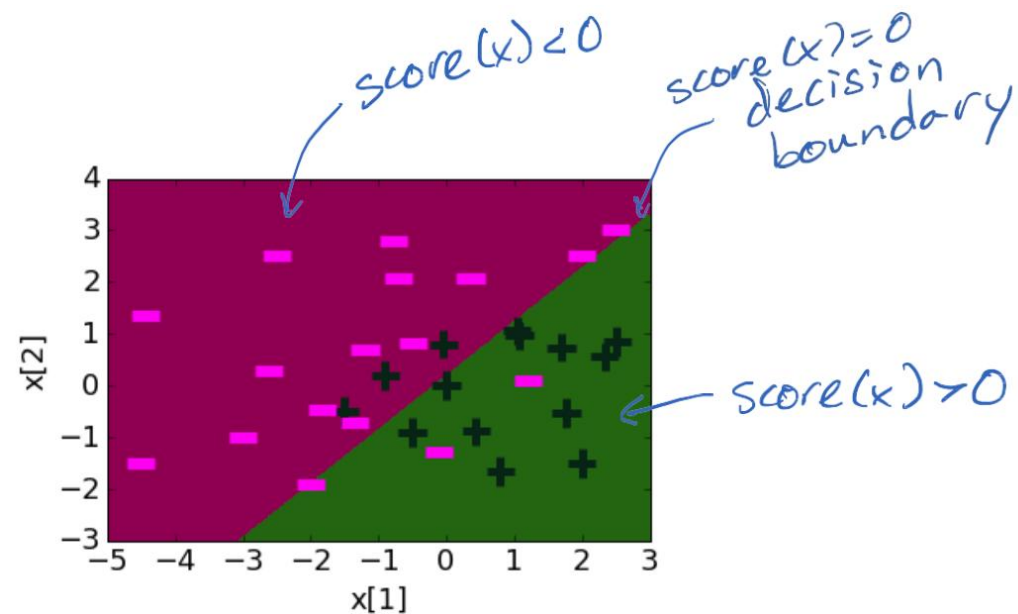
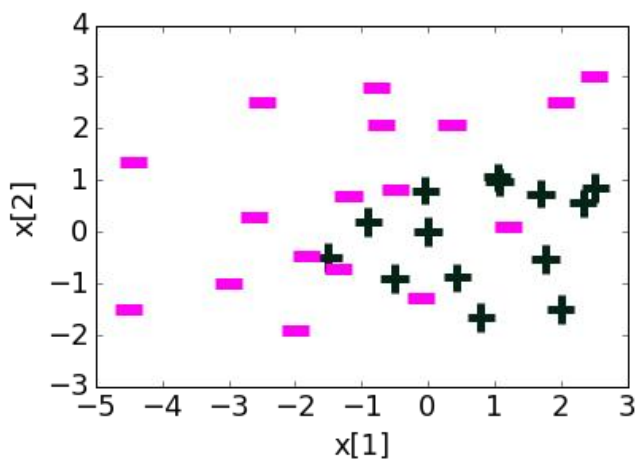
# 偏差较小的模型

- 往往需要更多数据才能学得好
- 数据少时→简单模型（高偏差）更稳定
- 数据多时→复杂模型（低偏差）能胜出



# 基于一阶特征的决策边界

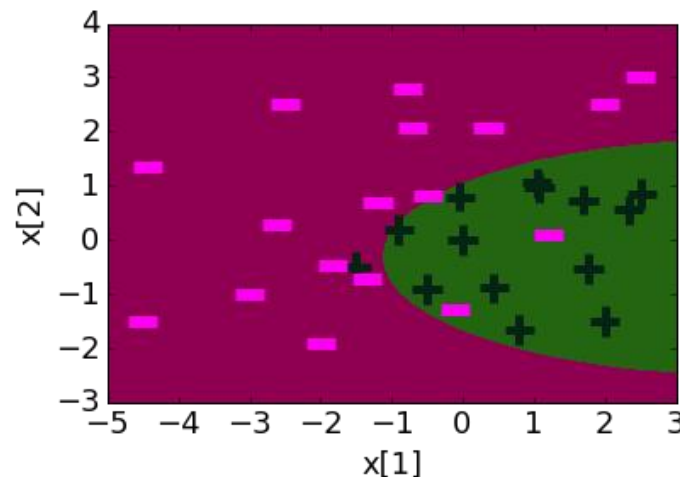
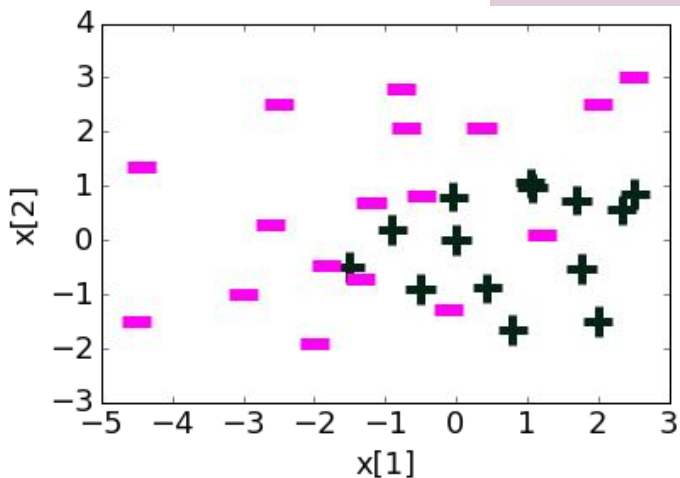
| 特征          | 值      | 学到的系数 |
|-------------|--------|-------|
| $\phi_0(x)$ | 1      | 0.23  |
| $\phi_1(x)$ | $x[1]$ | 1.12  |
| $\phi_2(x)$ | $x[2]$ | -1.07 |



# 引入二阶特征的决策边界

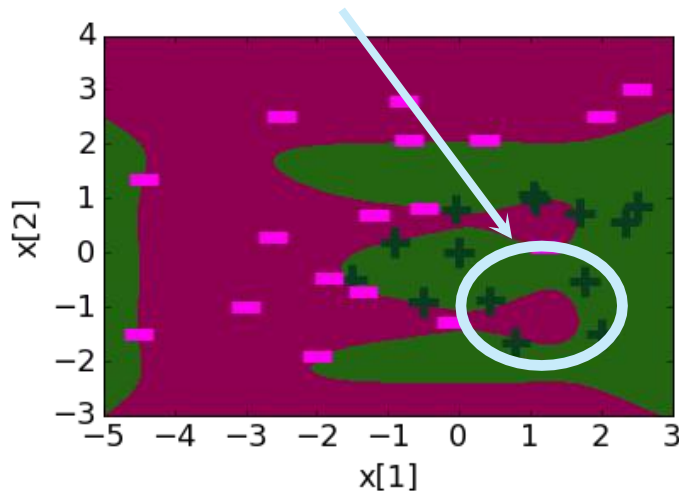
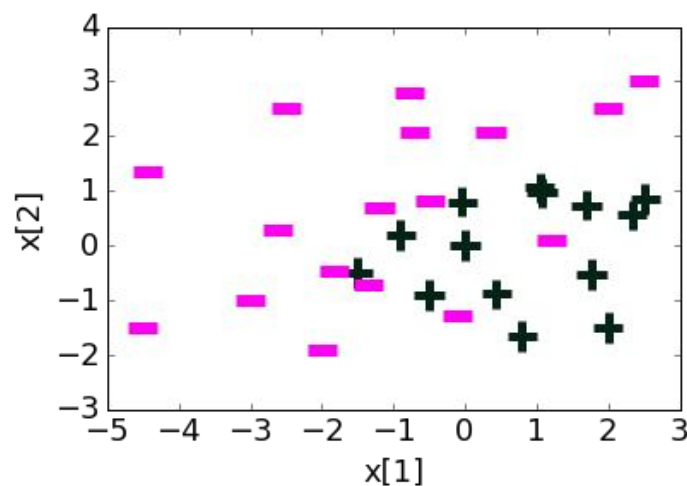
- 2阶特征（二维）

| 特征          | 值          | 学到的系数 |
|-------------|------------|-------|
| $\phi_0(x)$ | 1          | 1.68  |
| $\phi_1(x)$ | $x[1]$     | 1.39  |
| $\phi_2(x)$ | $x[2]$     | -0.59 |
| $\phi_3(x)$ | $(x[1])^2$ | -0.17 |
| $\phi_4(x)$ | $(x[2])^2$ | -0.96 |



# 引入多阶特征的决策边界

- 6阶特征（二维）



系数值越来越大

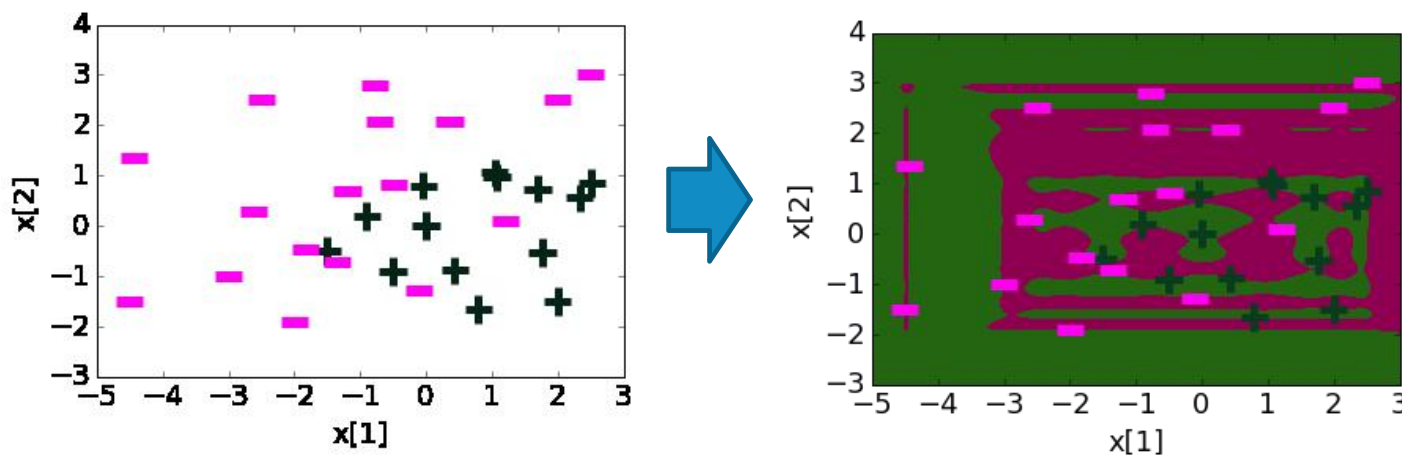
$Score(x) < 0$

| 特征             | 值          | 学到的系数 |
|----------------|------------|-------|
| $\phi_0(x)$    | 1          | 21.6  |
| $\phi_1(x)$    | $x[1]$     | 5.3   |
| $\phi_2(x)$    | $x[2]$     | -42.7 |
| $\phi_3(x)$    | $(x[1])^2$ | -15.9 |
| $\phi_4(x)$    | $(x[2])^2$ | -48.6 |
| $\phi_5(x)$    | $(x[1])^3$ | -11.0 |
| $\phi_6(x)$    | $(x[2])^3$ | 67.0  |
| $\phi_7(x)$    | $(x[1])^4$ | 1.5   |
| $\phi_8(x)$    | $(x[2])^4$ | 48.0  |
| $\phi_9(x)$    | $(x[1])^5$ | 4.4   |
| $\phi_{10}(x)$ | $(x[2])^5$ | -14.2 |
| $\phi_{11}(x)$ | $(x[1])^6$ | 0.8   |
| $\phi_{12}(x)$ | $(x[2])^6$ | -8.6  |

# 引入更高阶特征的决策边界

- 20阶特征（二维）

过度拟合时，常常出现非常大的估计系数 $w$

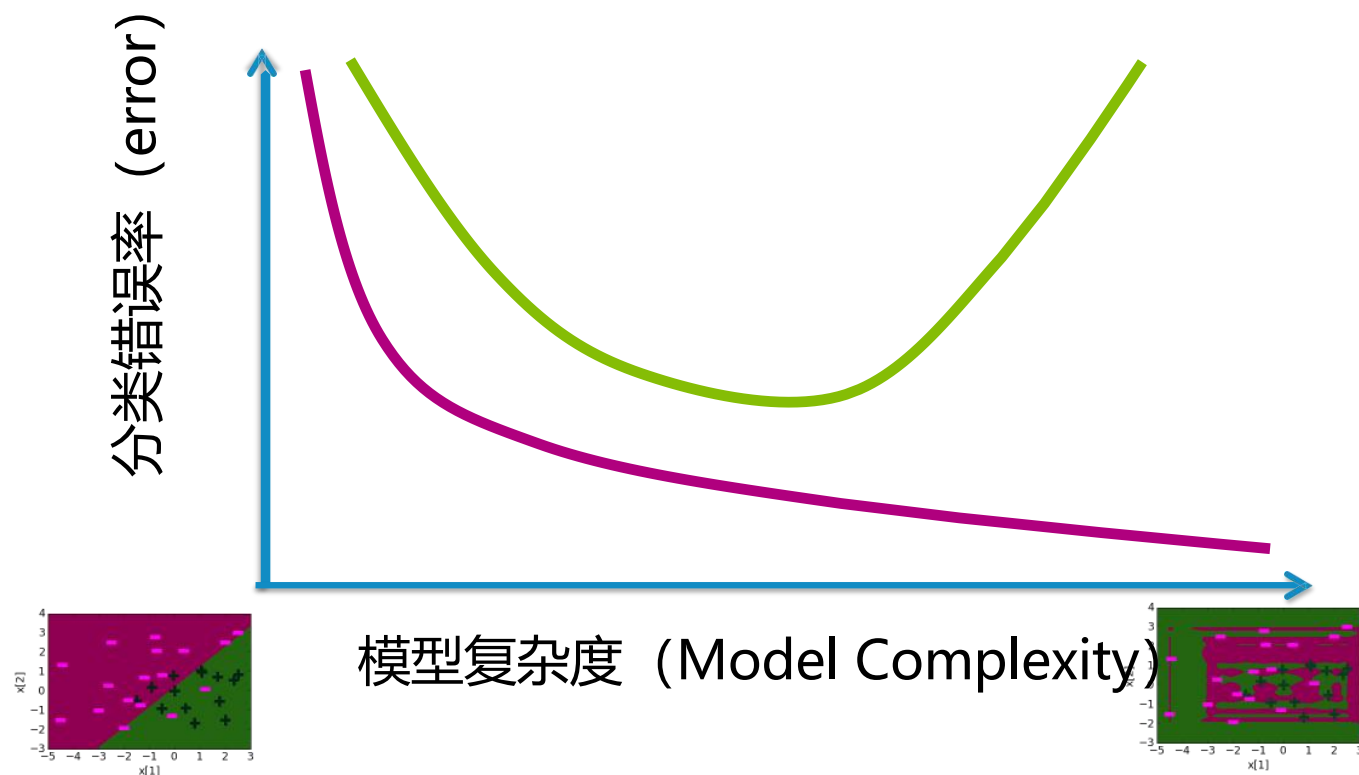


| 特征             | 值             | 学到的系数          |
|----------------|---------------|----------------|
| $\phi_0(x)$    | 1             | 8.7            |
| $\phi_1(x)$    | $x[1]$        | 5.1            |
| $\phi_2(x)$    | $x[2]$        | 78.7           |
| ...            | ...           | ...            |
| $\phi_{11}(x)$ | $(x[1])^6$    | -7.5           |
| $\phi_{12}(x)$ | $(x[2])^6$    | 3803           |
| $\phi_{13}(x)$ | $(x[1])^7$    | 21.1           |
| $\phi_{14}(x)$ | $(x[2])^7$    | -2406          |
| ...            | ...           | ...            |
| $\phi_{37}(x)$ | $(x[1])^{19}$ | $-2 * 10^{-6}$ |
| $\phi_{38}(x)$ | $(x[2])^{19}$ | -0.15          |
| $\phi_{39}(x)$ | $(x[1])^{20}$ | $-2 * 10^{-8}$ |
| $\phi_{12}(x)$ | $(x[2])^{20}$ | 0.03           |

# 分类中的过拟合

- 如果存在  $w^*$  满足：

•

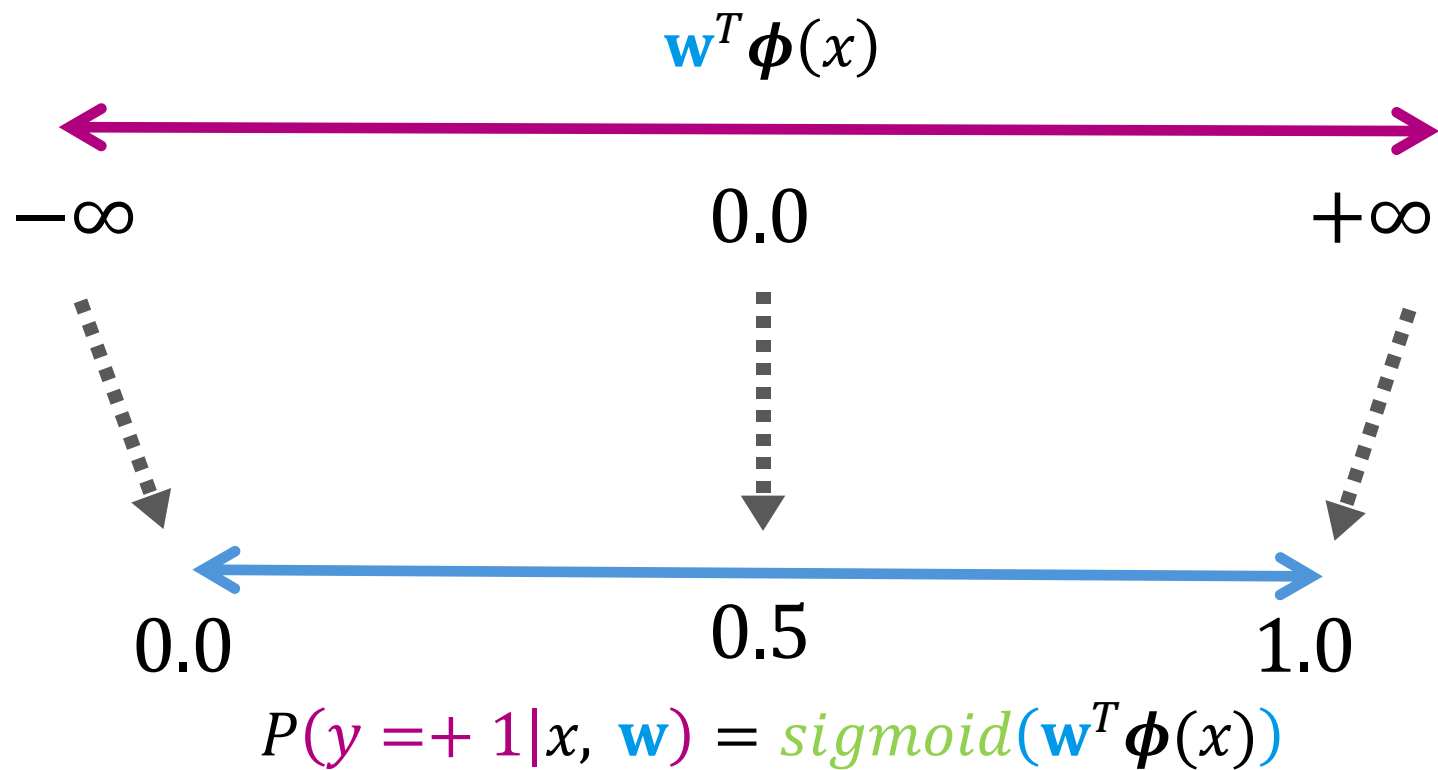


# 提纲

---

- Logistic Regression 框架
- **过拟合问题**
  - 过度自信预测
  - 大值系数偏好
- 正则化

# Recap: 逻辑回归

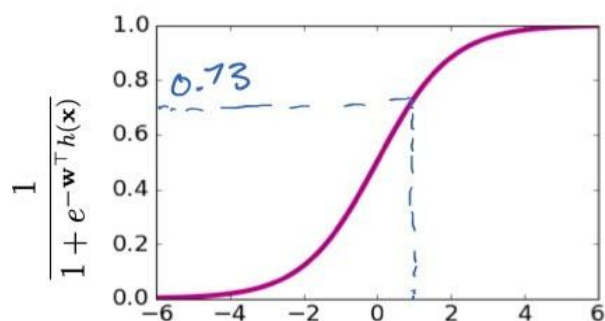


# 过拟合 = 过度自信

- $\text{Score}(\mathbf{x}) = w_0 + w_1 * \text{贷款金额} + w_2 * \text{年收入} = \mathbf{w}^T \phi(\mathbf{x})$

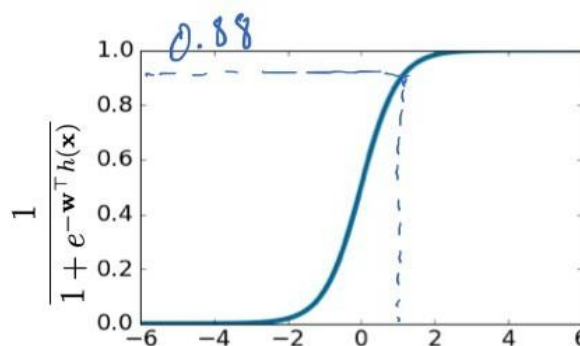
输入: #贷款金额=2, #年收入=1

|       |    |
|-------|----|
| $w_0$ | 0  |
| $w_1$ | +1 |
| $w_2$ | -1 |



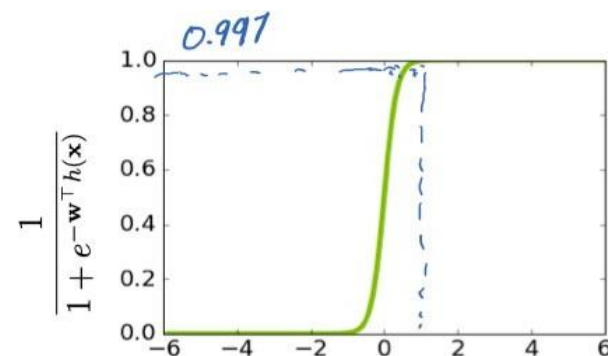
#贷款金额 - #年收入

|       |    |
|-------|----|
| $w_0$ | 0  |
| $w_1$ | +2 |
| $w_2$ | -2 |



#贷款金额 - #年收入

|       |    |
|-------|----|
| $w_0$ | 0  |
| $w_1$ | +6 |
| $w_2$ | -6 |



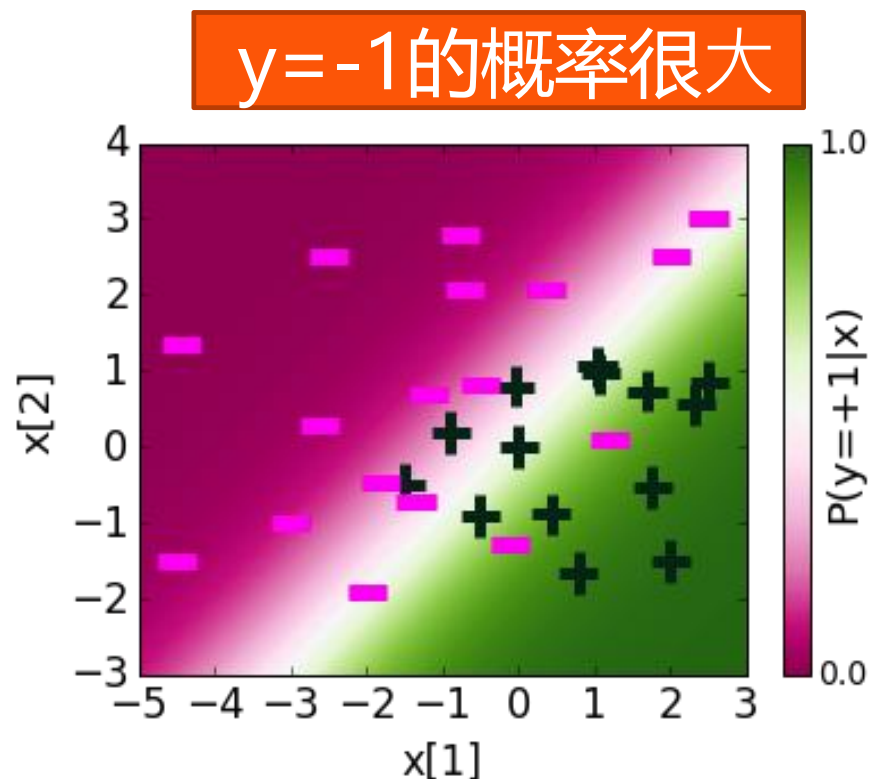
#贷款金额 - #年收入

权重越大，曲线越陡峭，模型输出的概率越极端地逼近0或1。

# 基于一阶特征的概率分布

$$\bullet P(y = +1 | x, \mathbf{w}) = \frac{1}{1 + e^{-\mathbf{w}^T \phi(x)}}$$

| 特征          | 值      | 学到的系数 |
|-------------|--------|-------|
| $\phi_0(x)$ | 1      | 0.23  |
| $\phi_1(x)$ | $x[1]$ | 1.12  |
| $\phi_2(x)$ | $x[2]$ | -1.07 |

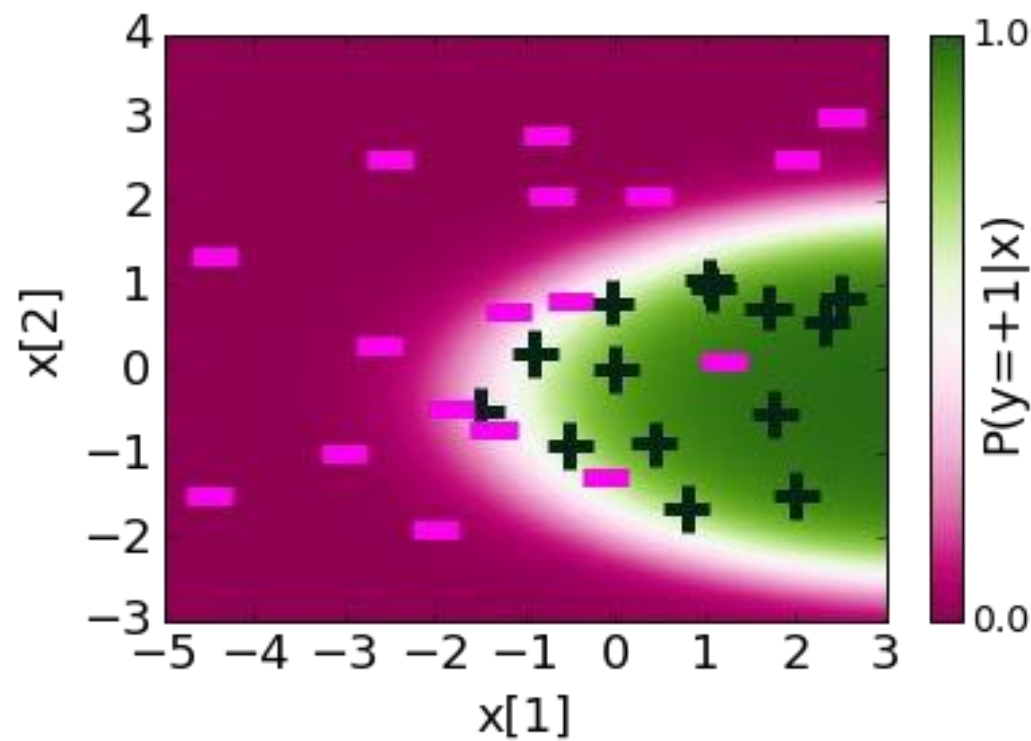


不确定性区域较宽

# 引入二阶特征的概率分布

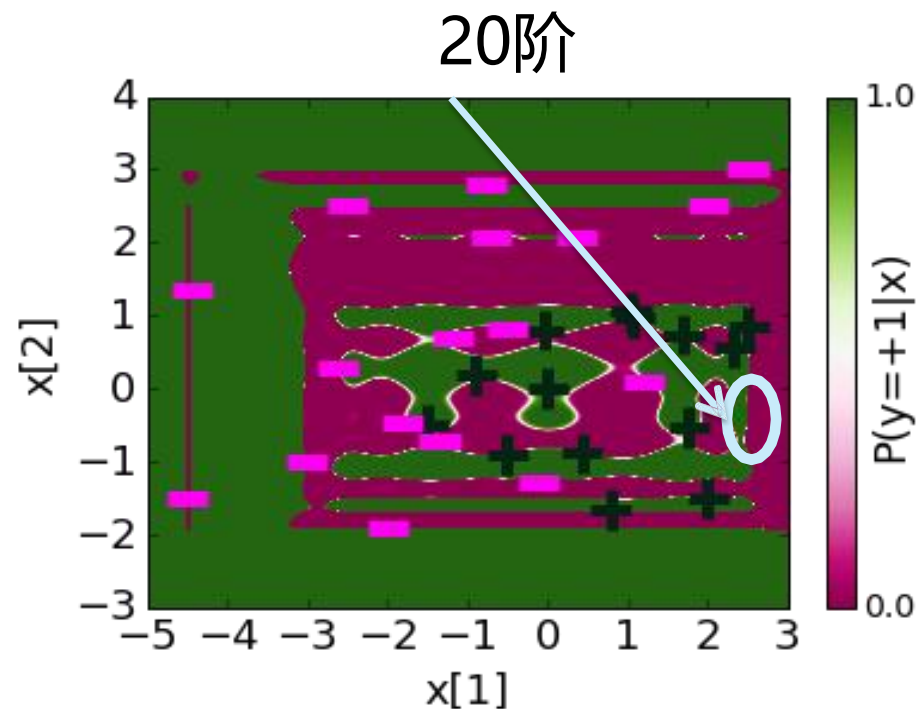
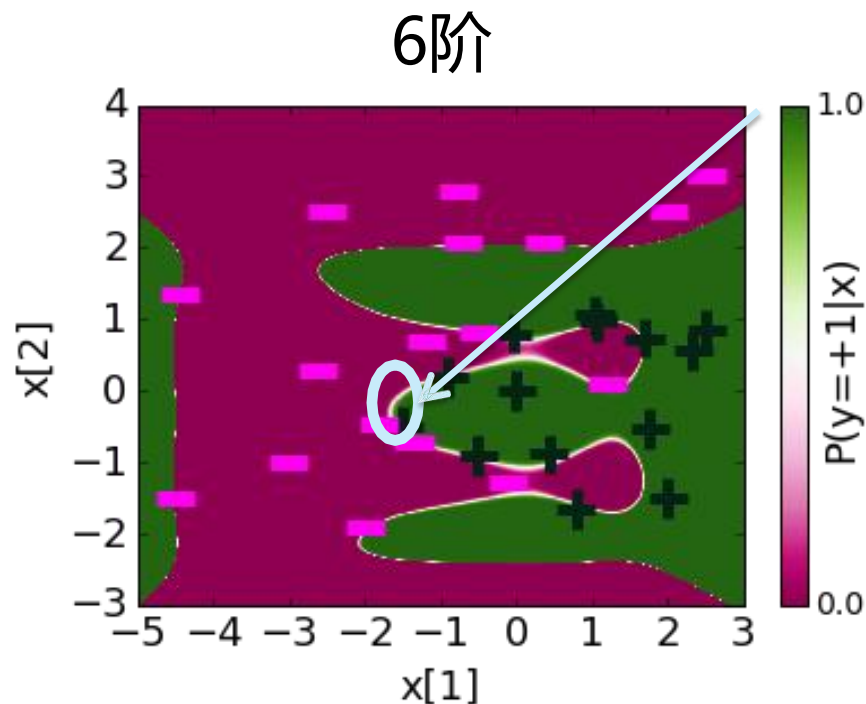
| 特征          | 值          | 学到的系数 |
|-------------|------------|-------|
| $\phi_0(x)$ | 1          | 1.68  |
| $\phi_1(x)$ | $x[1]$     | 1.39  |
| $\phi_2(x)$ | $x[2]$     | -0.59 |
| $\phi_3(x)$ | $(x[1])^2$ | -0.17 |
| $\phi_4(x)$ | $(x[2])^2$ | -0.96 |

较好地拟合数据



不确定性区域较窄

# 引入更高阶特征的概率分布



- 过拟合情况下，决策边界周围的不确定区域变得极窄；
- 模型几乎对所有样本都给出接近0或1的概率；
- 结果： **错误预测也会带着极高置信度输出 -> “错得非常自信”**

# 提纲

---

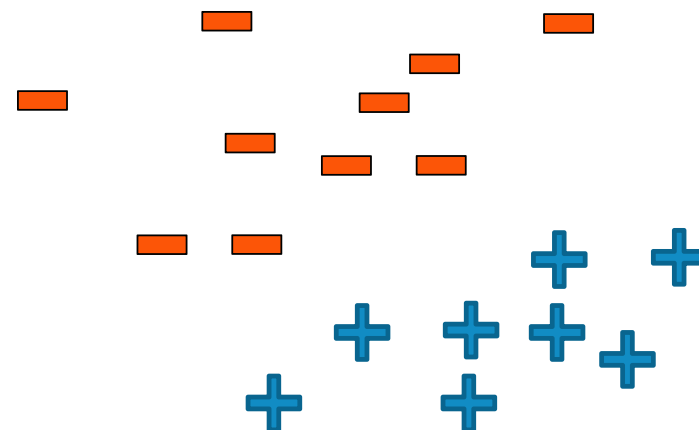
- Logistic Regression 框架
- **过拟合问题**
  - 过度自信预测
  - 大值系数偏好
- 正则化

# 线性可分数据

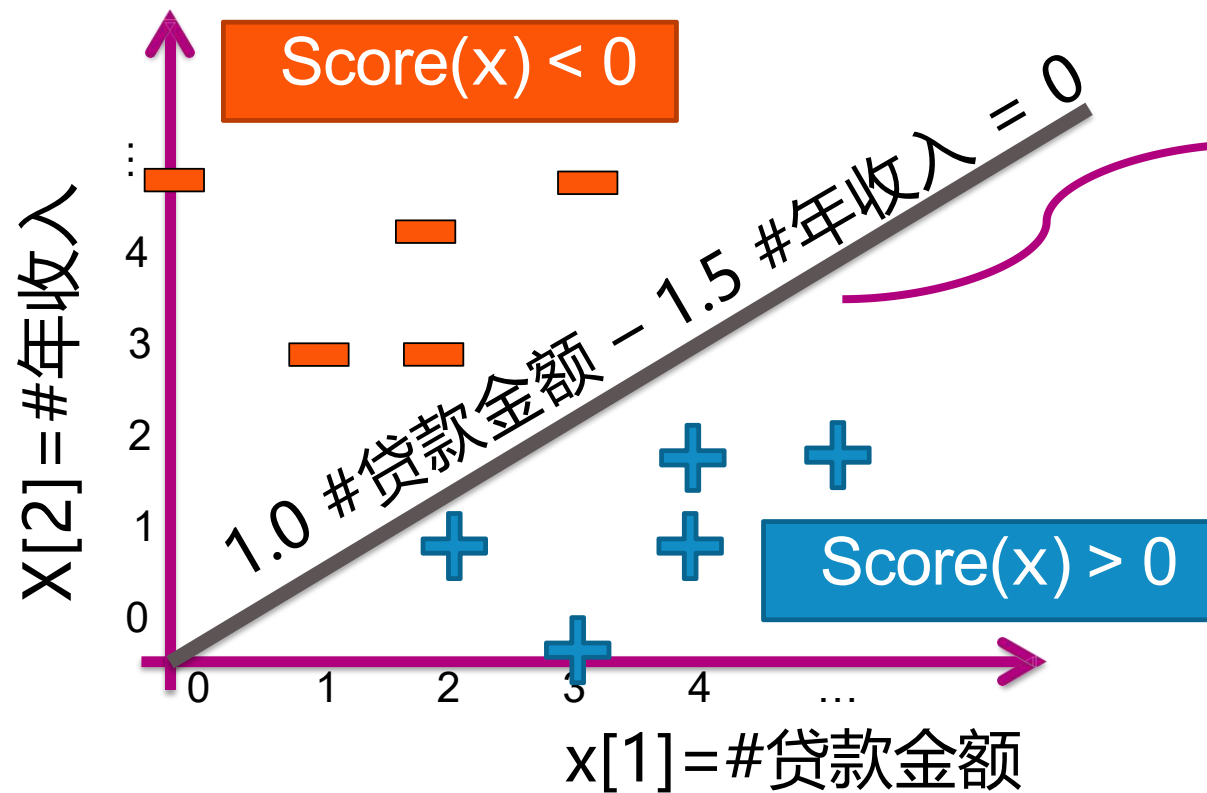
- 如果满足以下条件，则数据是线性可分的
  - 存在系数 $\hat{\mathbf{w}}$ 使得:
    - 对于训练数据中所有正样本:  $\text{Score}(\mathbf{x}) = \hat{\mathbf{w}}^T \boldsymbol{\phi}(\mathbf{x}) > 0$
    - 对于训练数据中所有负样本:  $\text{Score}(\mathbf{x}) = \hat{\mathbf{w}}^T \boldsymbol{\phi}(\mathbf{x}) < 0$

$$\text{training\_error}(\hat{\mathbf{w}}) = 0$$

- 注1: 如果使用D个特征，线性可分性发生在D维空间中。
- 注2: 如果有足够多的特征，数据（几乎）总是线性可分的。



# 同一决策边界，可能有不同系数



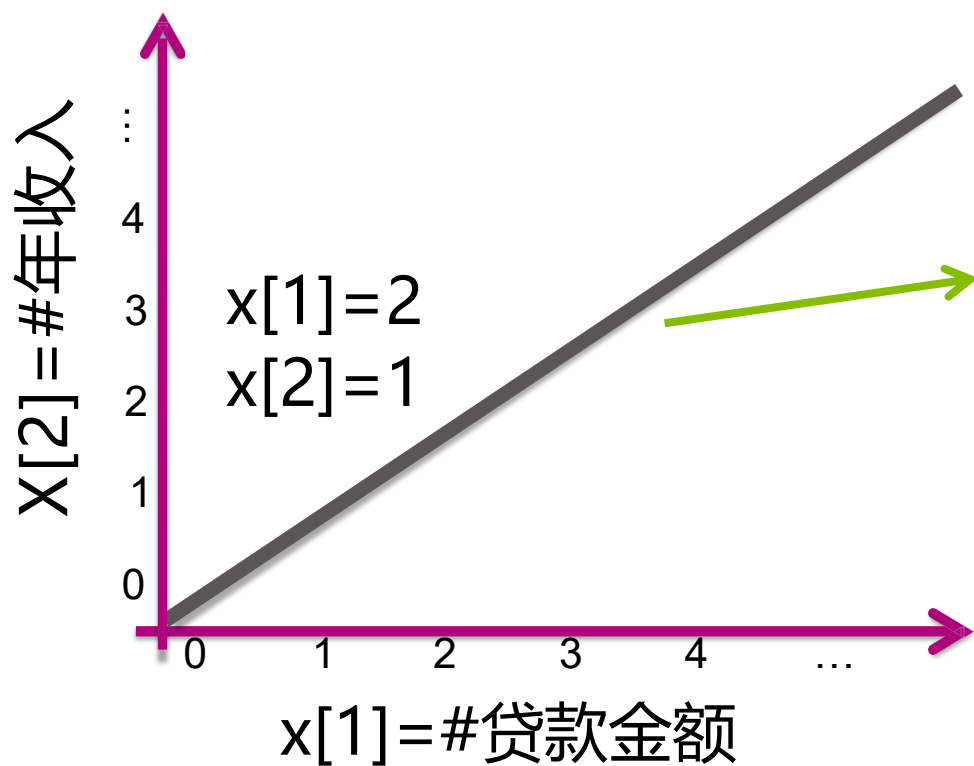
数据对  $\hat{w}_1 = 1.0$ ,  
 $\hat{w}_2 = -1.5$  线性可分

数据对  $\hat{w}_1 = 10$ ,  
 $\hat{w}_2 = -15$  同样线性可分

数据对  $\hat{w}_1 = 10^9$ ,  
 $\hat{w}_2 = -1.5 \times 10^9$  同样线性可分

# 最大似然估计倾向于选择较大的系数

- 最大似然估计偏好特定模型
- 对于线性可分离数据，系数趋于无穷大！



| $P(y = +1   \mathbf{x}, \hat{\mathbf{w}})$       |
|--------------------------------------------------|
| $\frac{1}{1 + e^{-(2-1.5)}} = 0.62$              |
| $\frac{1}{1 + e^{-(15-10)}} = 0.99$              |
| $\frac{1}{1 + e^{-(0.5 \times 10^9)}} \approx 1$ |

数据对  $\hat{w}_1 = 1.0$ ,  
 $\hat{w}_2 = -1.5$  线性可分

数据对  $\hat{w}_1 = 10$ ,  
 $\hat{w}_2 = -15$  同样线性可分

数据对  $\hat{w}_1 = 10^9$ ,  
 $\hat{w}_2 = -1.5 \times 10^9$  同样线性可分

# 两种视角的合并

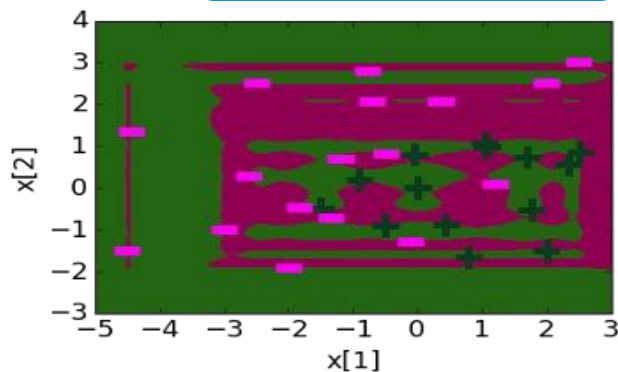
模型试图找到区分  
数据的决策边界



如果数据是  
线性可分的

过于复杂的  
边界

系数趋于  
无穷大



$$\hat{w}_1 = 10^9$$
$$\hat{w}_2 = -1.5 \times 10^9$$

# 提纲

---

- Logistic Regression 框架
- 过拟合问题
- **正则化**

# 理想的总成本形式

---

- 目标函数中的权衡项目：
  - 函数对数据的拟合程度（衡量模型能多好地解释训练数据）；
  - 回归系数的大小（避免系数过大导致过拟合）。
- Total quality = **measure of fit** - **measure of magnitude of coefficients**
  - **measure of fit**: 数据似然 (data likelihood)
    - 大数值 → 表示模型对训练数据拟合得好；
    - 小数值 → 表示模型无法很好拟合训练数据。
  - **measure of magnitude of coefficients**
    - 数值过大 → 过拟合

# 提纲

---

- Logistic Regression 框架
- 过拟合问题
- **正则化**
  - L2 正则化
  - L1 正则化

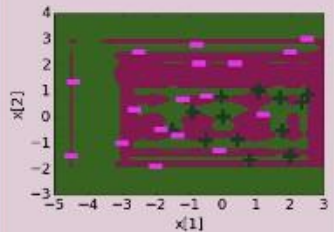
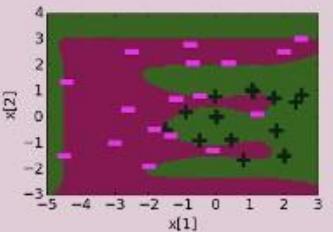
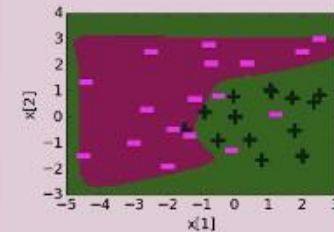
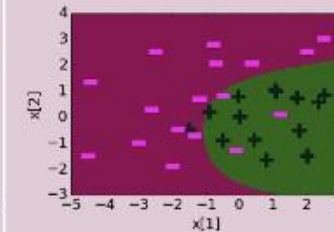
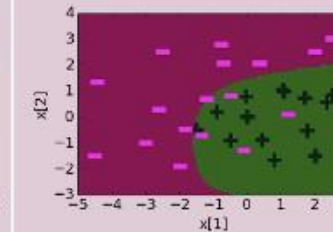
# L2正则化

---

- L2正则的逻辑回归
- 选择  $\hat{\mathbf{w}}$  以最大化  $\ell(\mathbf{w}) - \lambda \|\mathbf{w}\|_2^2$ 
  - 调整参数  $\lambda$  在拟合与参数规模之间平衡
- $\lambda$  选择方式
  - 验证集（适用于大型数据集）
  - 交叉验证（适用于较小的数据集）

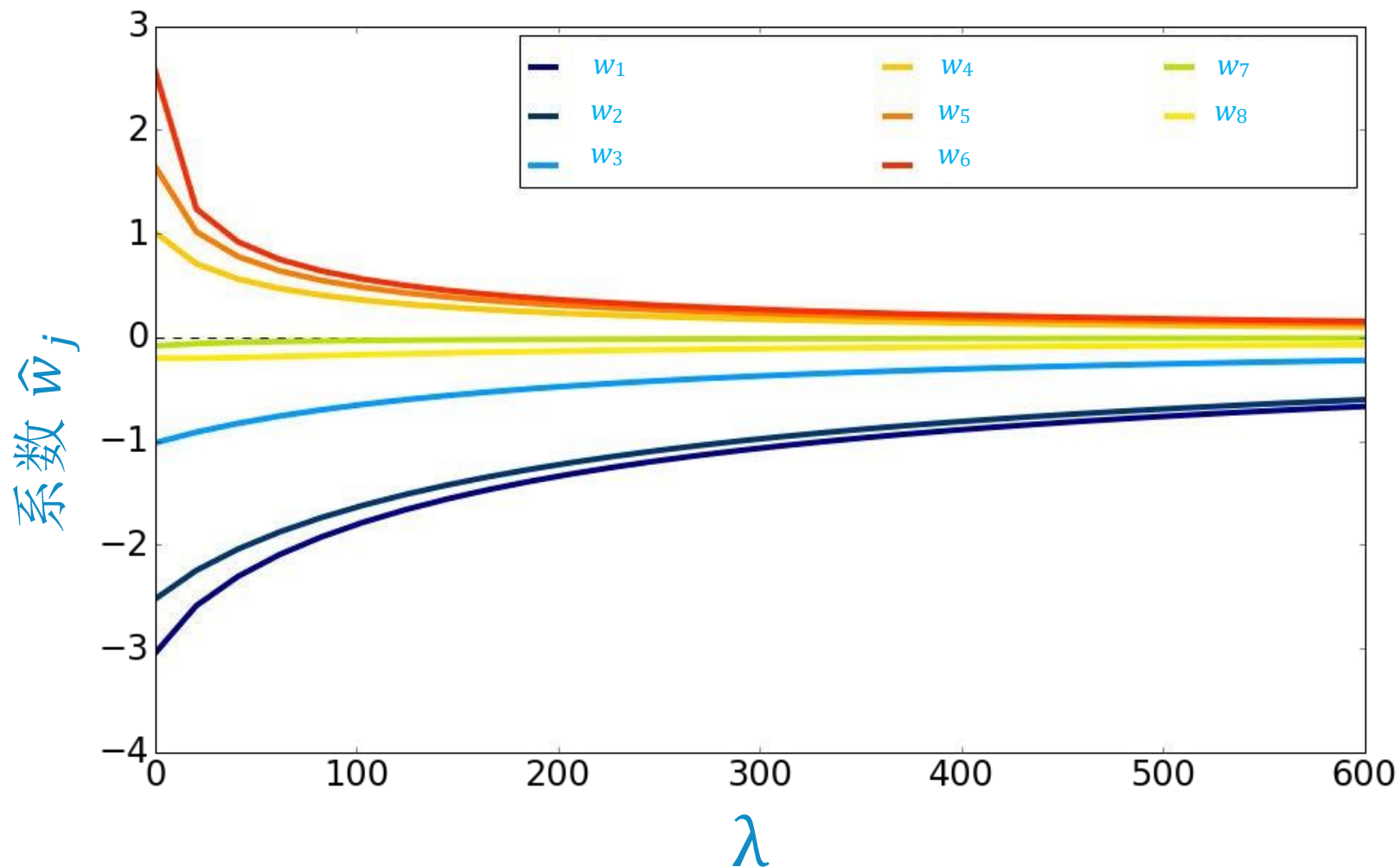
# 惩罚系数 $\lambda$ 对决策边界的影响

- 20阶特征下的决策边界

| 正则化  | $\lambda = 0$                                                                     | $\lambda = 0.00001$                                                                | $\lambda = 0.001$                                                                   | $\lambda = 1$                                                                       | $\lambda = 10$                                                                      |
|------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|
| 系数范围 | -3170 ~ 3803                                                                      | -8.04 ~ 12.14                                                                      | -0.70 ~ 1.25                                                                        | -0.13 ~ 0.57                                                                        | -0.05 ~ 0.22                                                                        |
| 决策边界 |  |  |  |  |  |

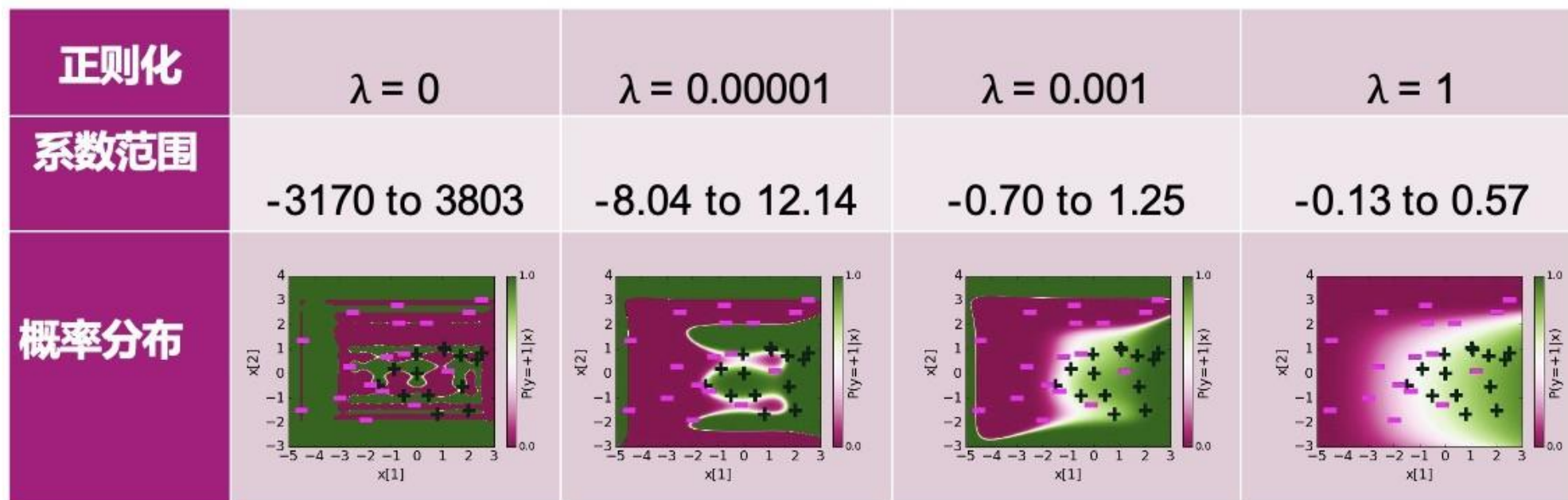
# 系数轨迹

- L2正则下的系数轨迹



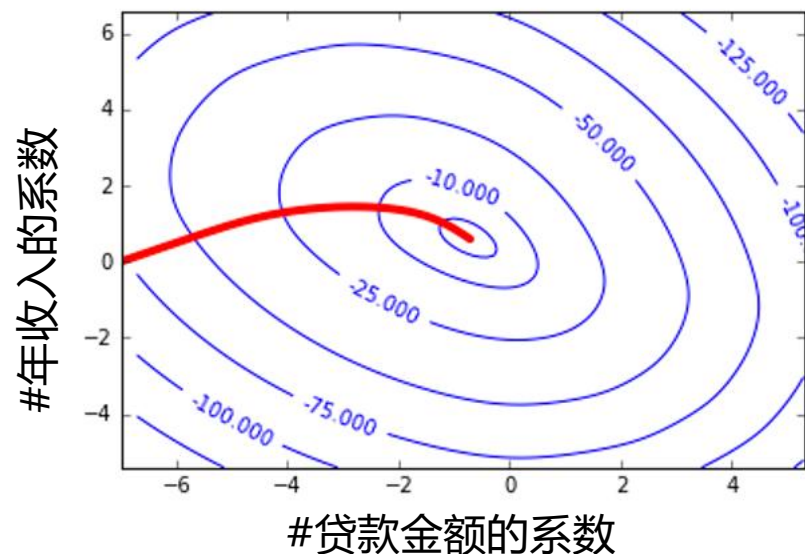
# 正则化可以降低模型的“过度自信”

- 20阶特征下的概率分布



# Recap: 岭回归/L2正则化回归

- L2正则逻辑回归的梯度上升



```
init  $w^{(1)} = 0, t = 1$   
while  $\|\nabla \ell(w^{(t)})\| > \varepsilon$   
  for  $j = 0, \dots, D$   
    partial[j] =  $\sum_{i=1}^N \phi_j(x_i) (\mathbf{1}[y_i = +1] -$   
       $P(y_i = +1 | x_i, w^{(t)}))$   
     $w_j^{(t+1)} \leftarrow w_j^{(t)} + \eta (partial[j] - 2\lambda w_j^{(t)})$   
   $t \leftarrow t + 1$ 
```

# 提纲

---

- Logistic Regression 框架
- 过拟合问题
- **正则化**
  - L2 正则化
  - L1 正则化

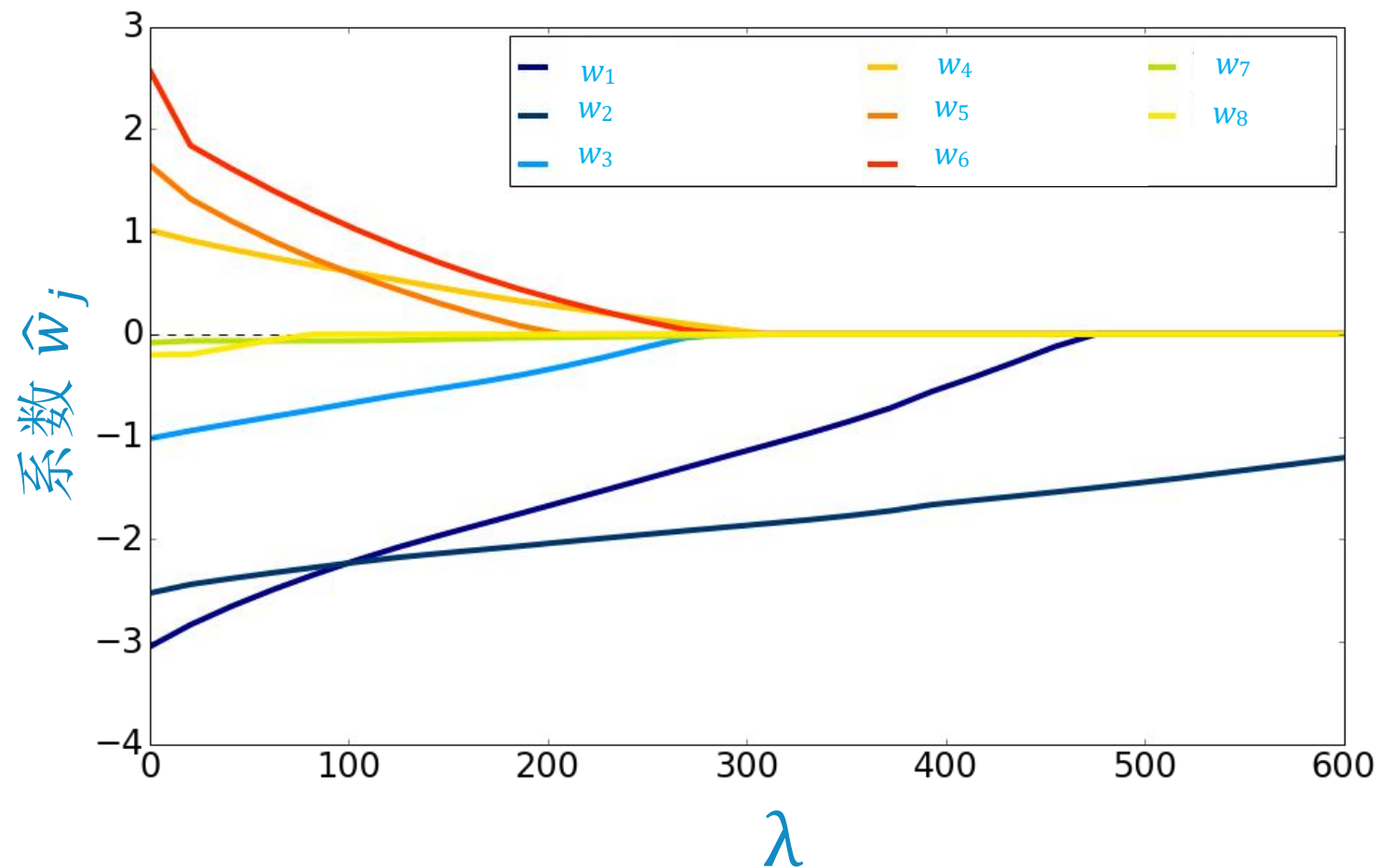
# L1正则化

---

- L1正则惩罚进行稀疏逻辑回归
- 选择  $\hat{\mathbf{w}}$  以最大化  $\ell(\mathbf{w}) - \lambda \|\mathbf{w}\|_1$ 
  - 调整参数  $\lambda$  在拟合与参数规模之间平衡
- $\lambda$  选择方式
  - 验证集（适用于大型数据集）
  - 交叉验证（适用于较小的数据集）
    - 岭回归（Ridge Regression）
    - 套索回归（Lasso Regression）

# 系数轨迹

- L1正则下的系数轨迹



# Questions

---

