中国科学技术大学 University of Science and Technology of China

《人工智能数学原理与算法》

第2章：机器学习基础

2.6 逻辑回归（二）

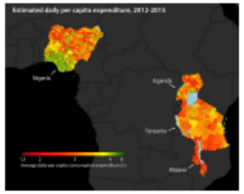
王翔

中国科学技术大学
数据科学实验室LDS

回顾：机器学习中的回归问题（Regression）

$$x \longrightarrow \boxed{f} \longrightarrow y \in \mathbb{R}$$

响应（response）



贫困地图：　卫星图像 → 资产财富指数

房价估计：　房屋信息（位置，面积） → 房价

到达时间：　目的地，天气，时间 → 到达时间

回顾：机器学习中的分类问题（Classification）

分类器（classifier）

$$x \longrightarrow \boxed{f} \longrightarrow y \in \{+1, -1\}$$

标签（label）

欺诈检测: 信用卡交易信息 → 是否欺诈

评论检测: 评论信息 → 是否有害

粒子对撞: 测量到的粒子对撞信息 → 粒子衰变还是背景噪音

问：分类和回归之间的关键区别是什么？

- 分类有**离散的输出**
- 回归有**连续的输出**

扩展：多分类问题 $y \in \{1, \ldots, K\}$

回顾：线性回归（Linear Regression）

模型向量表示：$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x)$ $\mathbf{w} = [w_1, w_2]$ $\phi(x) = [1, x]$

参数向量/模型参数 特征提取器 特征向量

假设类：$\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^2\}$ （预测器 f 的集合）

损失函数：$\text{Loss}(x, y, \mathbf{w}) = (f_{\mathbf{w}}(x) - y)^2$ 平方损失（squared loss）

$$\text{TrainLoss}(\mathbf{w}) = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \text{Loss}(x, y, \mathbf{w})$$

均方误差（MSE, mean squared error）

$$\arg\max_{\mathbf{w}} \ln p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \arg\min_{\mathbf{w}} \sum_{(x,y)\in\mathcal{D}} (y - \phi(x)\mathbf{w})^2$$

高斯假设下的最大似然估计 = 最小化平方误差

回顾：逻辑回归（Logistic Regression）

模型向量表示：$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x)$　　$\mathbf{w} = [w_1, w_2]$　　$\phi(x) = [1, x]$

参数向量/模型参数　特征提取器　特征向量

假设类：$\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathbb{R}^2\}$ （预测器f的集合）

损失函数：$\hat{P}(y = +1|x) = \text{sigmoid}(\text{Score}(\boldsymbol{x})) = \frac{1}{1+e^{-w^{\top}\phi(x)}}$　平方损失（squared loss）

$$\max_{w} l(w) = \prod_{i=1}^{N} P(y_i|\boldsymbol{x}_i, w)$$

最大似然估计（ Maximum Likelihood Estimation）

$$\max_{w} ll(w) = \sum_{i=1}^{N} \log(P(y_i|\boldsymbol{x}_i, w))$$

对数似然（Log-likelihood ）

目录

01 多分类问题

02 过拟合的定义

03 过拟合的两个视角

04 过拟合的解决方案

目录

01 多分类问题

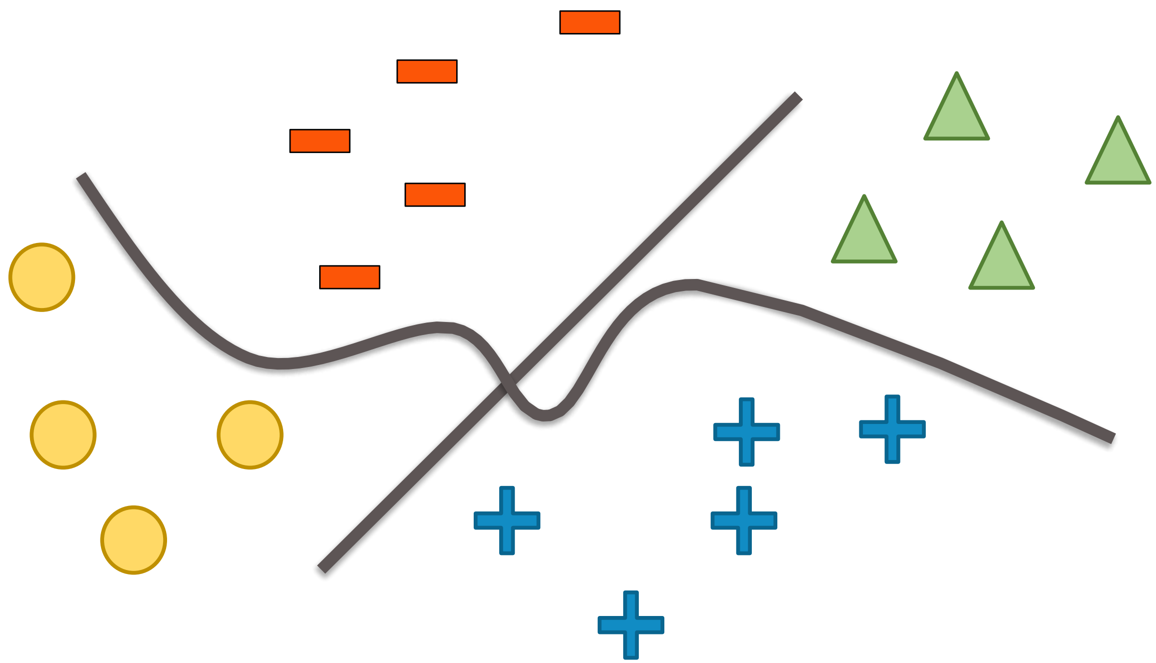
02 过拟合的定义

03 过拟合的两个视角

04 过拟合的解决方案

多分类问题

有时候，样本点的类别不止两个



使用逻辑回归解决多分类问题

解决方案1：One-vs-All

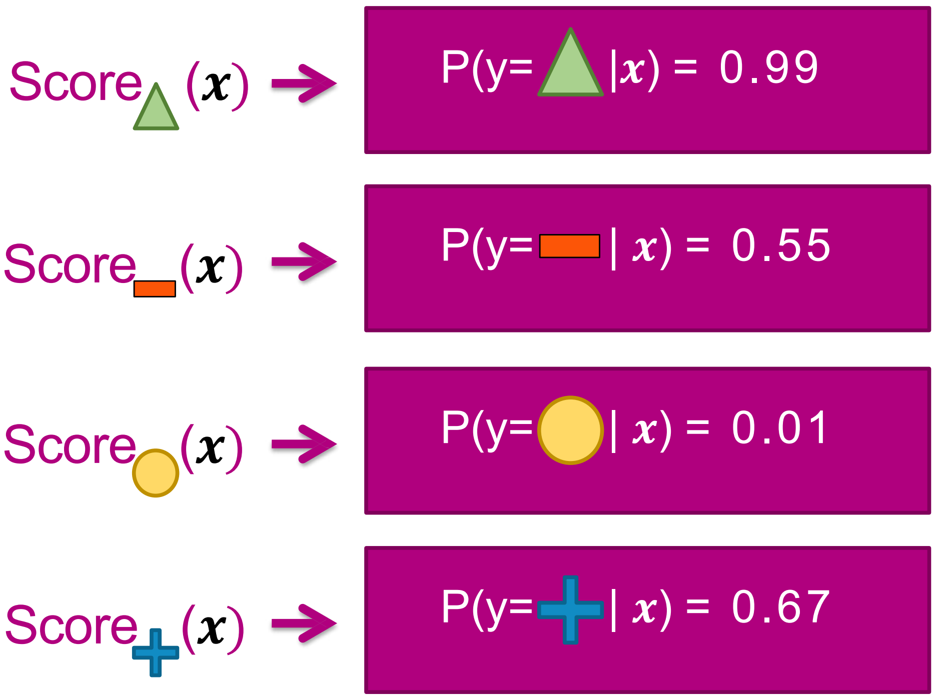
- 核心策略：将K分类任务分解为K个二分类问题
- 每个分类器像一位"专业裁判"，只负责判断是否属于特定类别

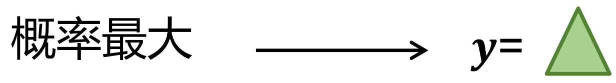
解决方案2：交叉熵损失

- 核心策略：将K分类任务建模为K类别上的概率分布
- 一个分类器，判断一个样本在K个类上的概率分布

方案1：One-vs-All方法

训练K个独立的二分类器，分别判断数据是否属于每个类别





选择其中概率最大的作为最终的多分类预测。

方案2：交叉熵损失

- 线性变换：$\text{Score}(\boldsymbol{x}) = \boldsymbol{w}^{\top}\boldsymbol{\phi}(\boldsymbol{x}) \in \boldsymbol{R}^{K}$, $\boldsymbol{w} \in \boldsymbol{R}^{D\times K}$, $\boldsymbol{\phi}(\boldsymbol{x}) \in \boldsymbol{R}^{D}$
- Softmax函数：

$$P(y|\boldsymbol{x},\boldsymbol{w}) = \text{softmax}(\text{Score}(\boldsymbol{x}))$$

$$P(y=k|\boldsymbol{x},\boldsymbol{w}) = \frac{e^{Score_k(x)}}{\sum_{k'=1}^{K} e^{Score_{k'}(x)}}, \qquad \sum_{k=1}^{K} P(y=k|\boldsymbol{x},\boldsymbol{w}) = 1$$

方案2：交叉熵损失

- 最大化对数似然

$$\max_{w} ll(w) = \sum_{i=1}^{N} \log(P(y_i = k|\boldsymbol{x}_i, w))$$

- 最小化交叉熵损失：

$$\min_{w} \text{cross\_entropy}(w) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \mathbb{1}(y_i = k) \log P(y_i = k|\boldsymbol{x}_i, \boldsymbol{w})$$

两种方案的对比

	One-vs-All	交叉熵损失
概率解释	独立计算	联合归一化
优化方式	分别优化	整体优化
优点	• 实现简单 • 并行训练优势 • 增量友好	• 概率一致（保证 $\sum_{k=1}^{K} P(y=k\|\boldsymbol{x},w)=1$）
缺点	• 概率不一致（可能 $\sum_{k=1}^{K} P(y=k\|\boldsymbol{x},w)\neq 1$） • 决策边界模糊	• 类别不平衡敏感

目录

01 多分类问题

02 过拟合的定义

03 过拟合的两个视角

04 过拟合的解决方案

更复杂的模型往往偏差更少……

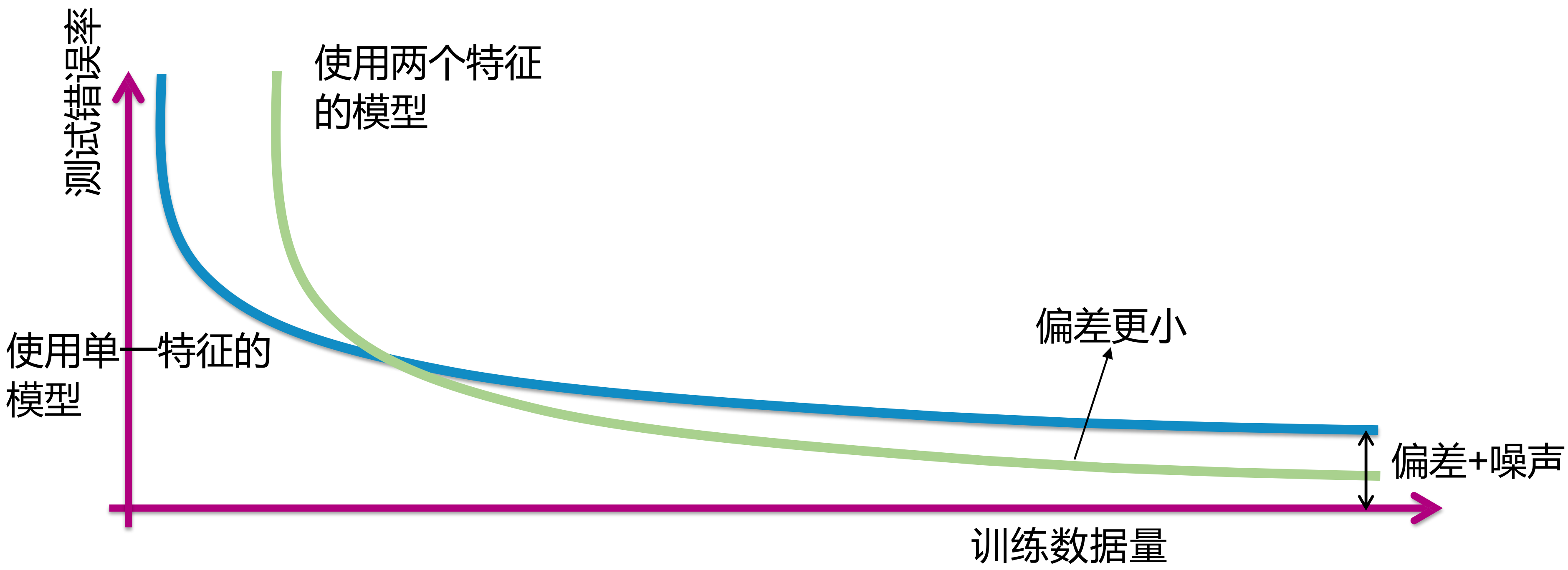
使用一个特征（如年收入）判断是否违约，可能达到预期效果，但…

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？

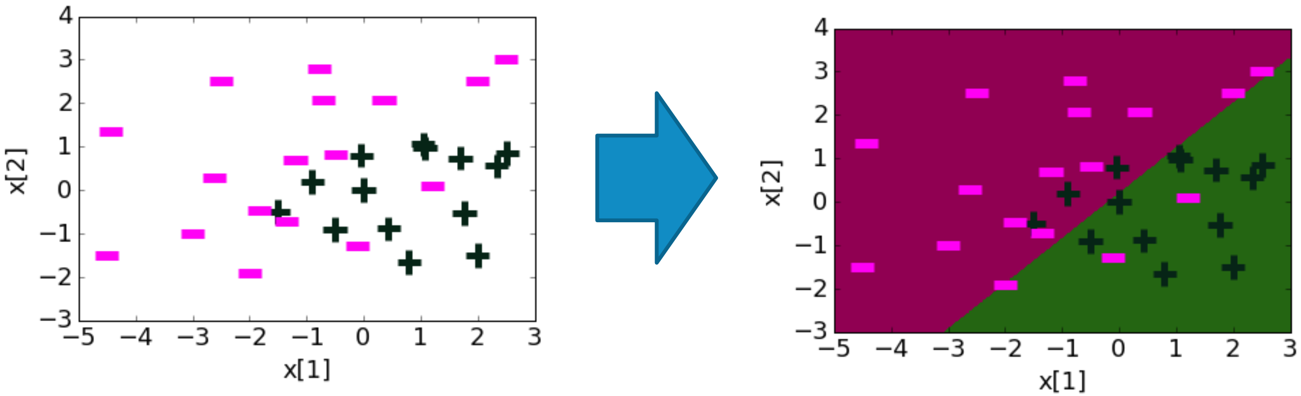
更少的偏差 ➔
可能更准确，需要更多数据来学习

数据更多，效果更好



基于线性特征的决策边界

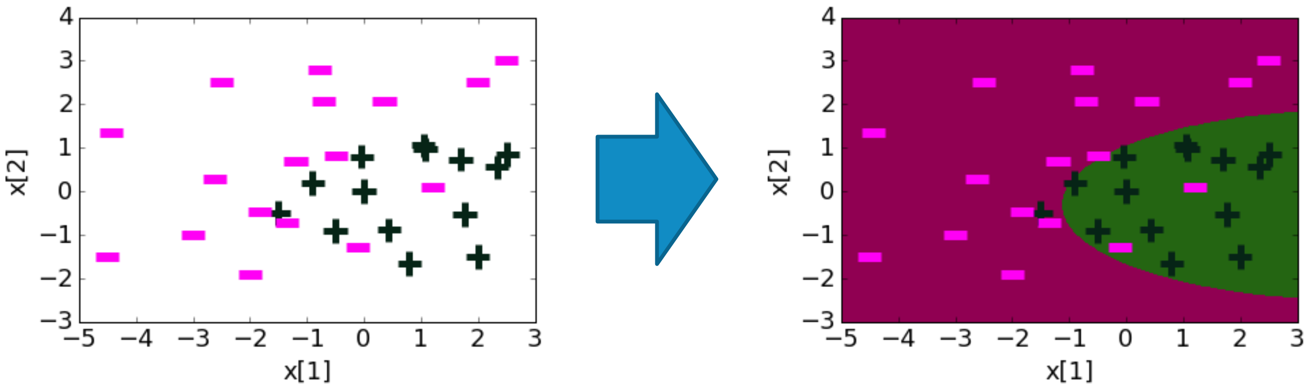
特征	值	学到的系数
$\phi_0(x)$	1	0.23
$\phi_1(x)$	x[1]	1.12
$\phi_2(x)$	x[2]	- 1.07



引入二次特征时的决策边界

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

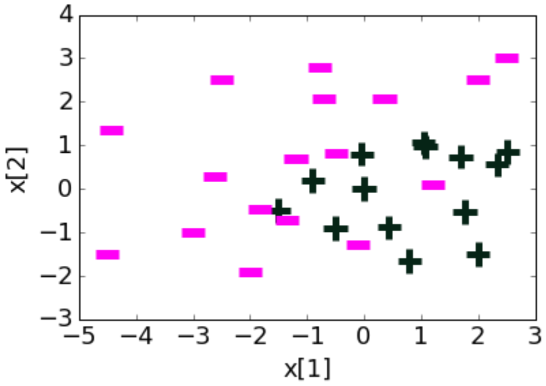
2阶特征（二维）



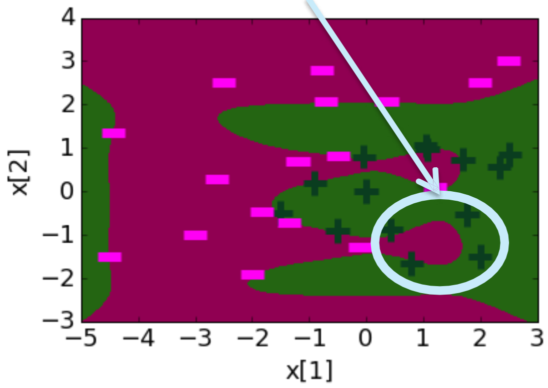
引入多阶特征时的决策边界

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

系数值越来越大



Score (x) <0



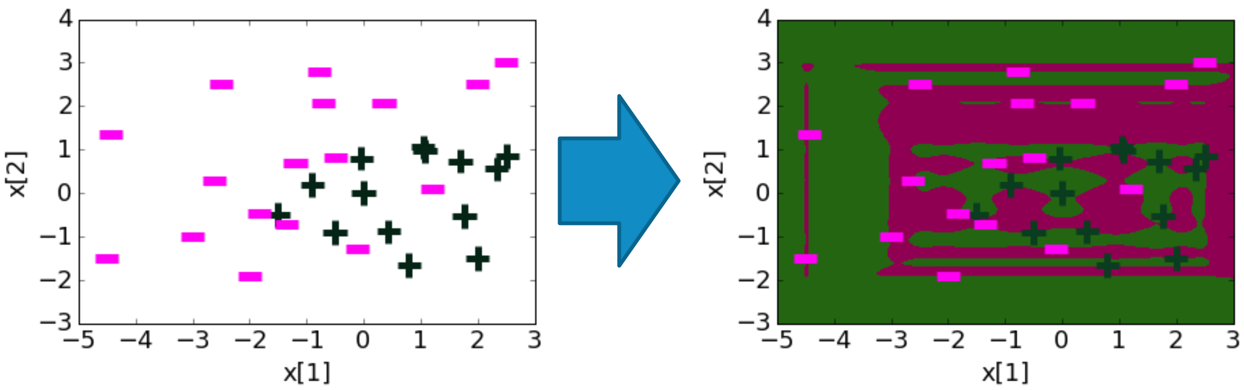
6阶特征（二维）

引入更高阶特征的决策边界

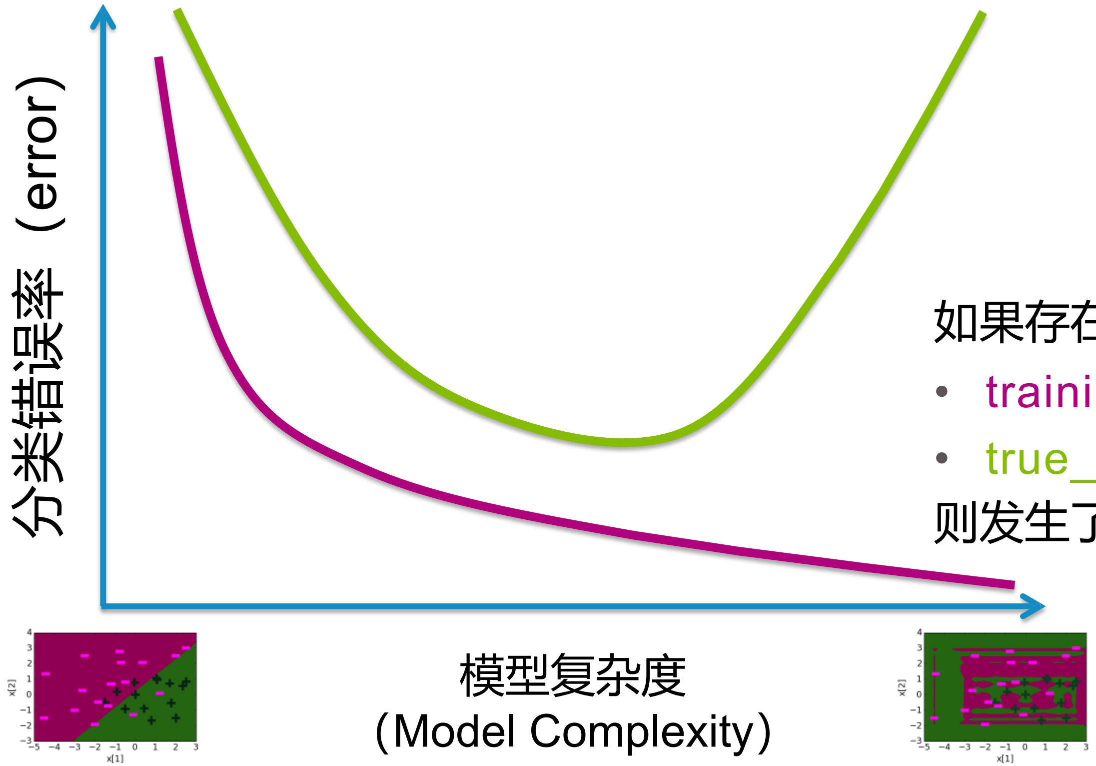
特征	值	学到的系数
$\phi_0(\mathbf{x})$	1	8.7
$\phi_1(\mathbf{x})$	x[1]	5.1
$\phi_2(\mathbf{x})$	x[2]	78.7
…	…	…
$\phi_{11}(\mathbf{x})$	$(x[1])^6$	-7.5
$\phi_{12}(\mathbf{x})$	$(x[2])^6$	3803
$\phi_{13}(\mathbf{x})$	$(x[1])^7$	21.1
$\phi_{14}(\mathbf{x})$	$(x[2])^7$	-2406
…	…	…
$\phi_{37}(\mathbf{x})$	$(x[1])^{19}$	$-2*10^{-6}$
$\phi_{38}(\mathbf{x})$	$(x[2])^{19}$	-0.15
$\phi_{39}(\mathbf{x})$	$(x[1])^{20}$	$-2*10^{-8}$
$\phi_{40}(\mathbf{x})$	$(x[2])^{20}$	0.03

20阶特征（二维）

过度拟合时，常常出现非常大的估计系数w



过拟合（Overfitting）



如果存在w*满足:

- training_error(w*) > training_error(ŵ)
- true_error(w*) < true_error(ŵ)

则发生了过拟合

目录

01 多分类问题

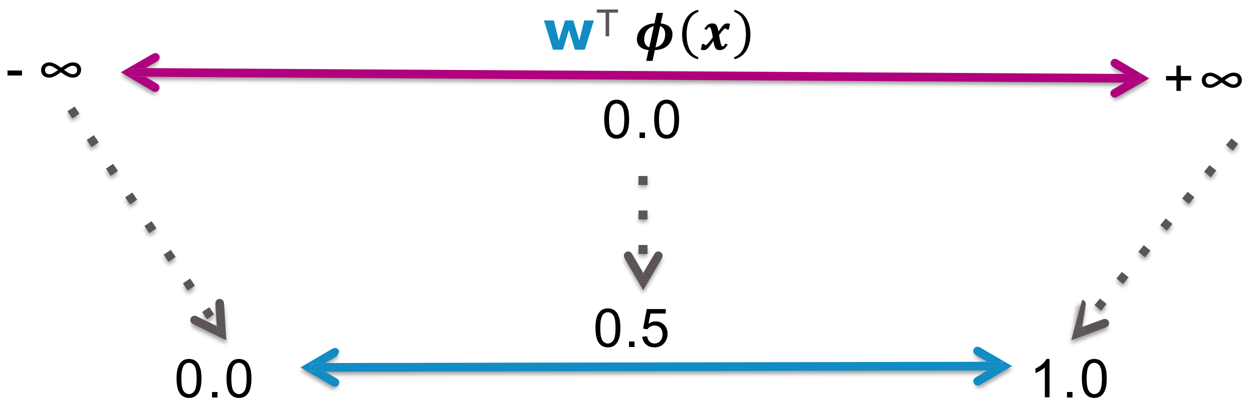
02 过拟合的定义

03 **过拟合的两个视角**

04 过拟合的解决方案

分类器的过拟合：
对预测的过度自信

回顾：逻辑回归



$$P(y=+1 \mid \boldsymbol{x}, \mathbf{w}) = \mathrm{sigmoid}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}))$$

逻辑回归：过拟合=过度自信

过拟合 ➔ 系数过大

$\hat{w}^T \boldsymbol{\phi}(x)$ 远大于零（或远小于零) ➔

sigmoid($\hat{w}^T \boldsymbol{\phi}(x)$) 趋于 1 (或 0)

模型对预测变得过度自信

逻辑回归：模型系数的影响

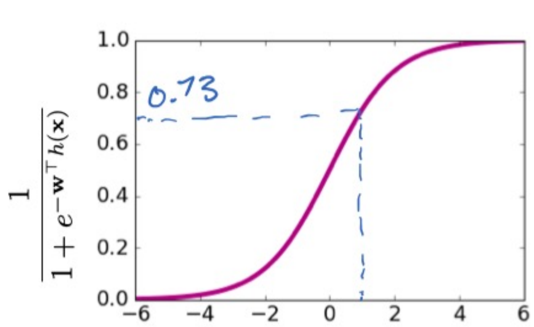
$$\text{Score}(\boldsymbol{x}) = w_0 + w_1 * \text{贷款金额} + w_2 * \text{年收入} = \boldsymbol{w}^\intercal \boldsymbol{\phi}(\boldsymbol{x})$$

输入: #贷款金额=2, #年收入=1

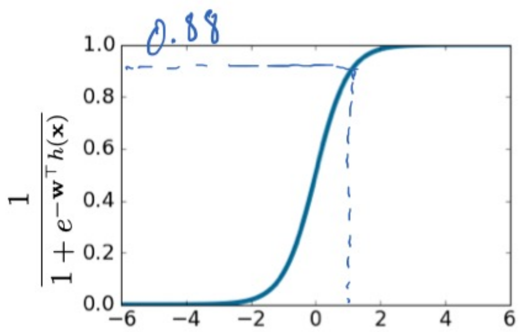
w_0	0
w_1	+1
w_2	−1

w_0	0
w_1	+2
w_2	−2

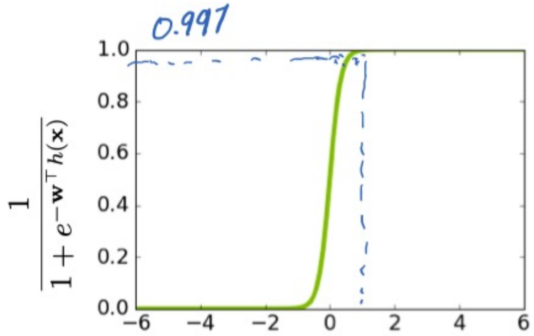
w_0	0
w_1	+6
w_2	−6



#贷款金额 - #年收入



#贷款金额 - #年收入

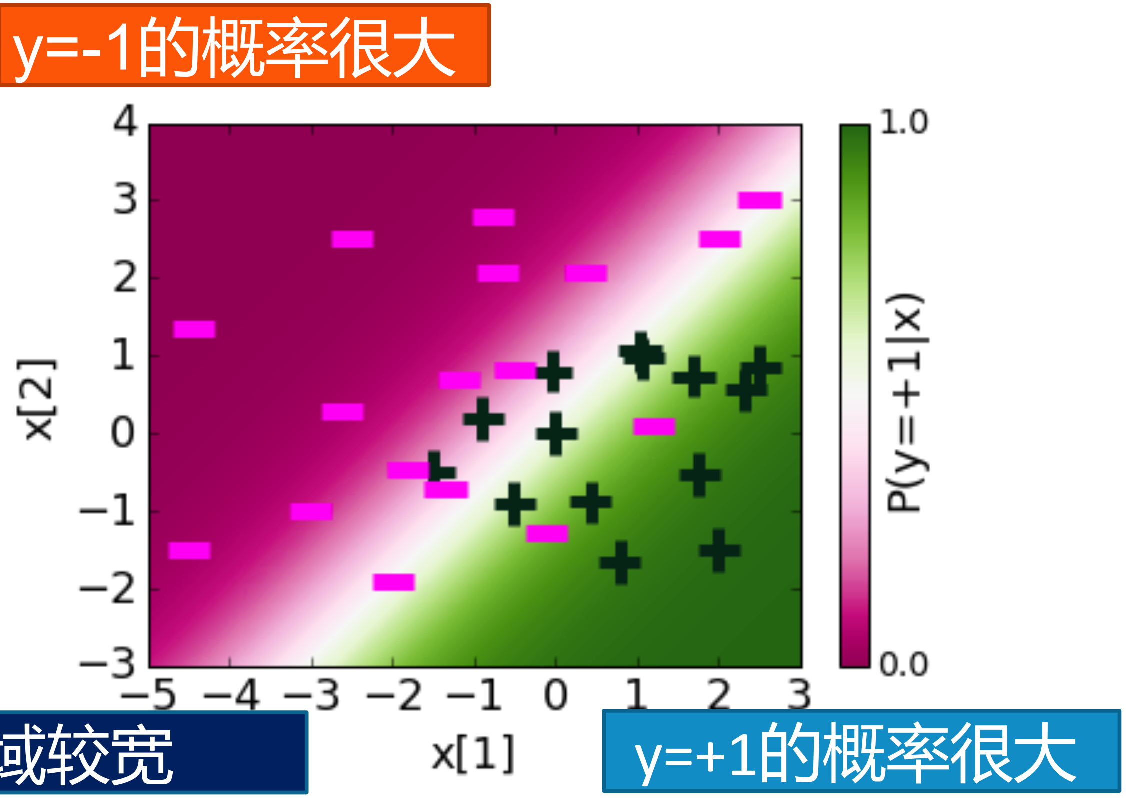


#贷款金额 - #年收入

基于线性特征的概率分布

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

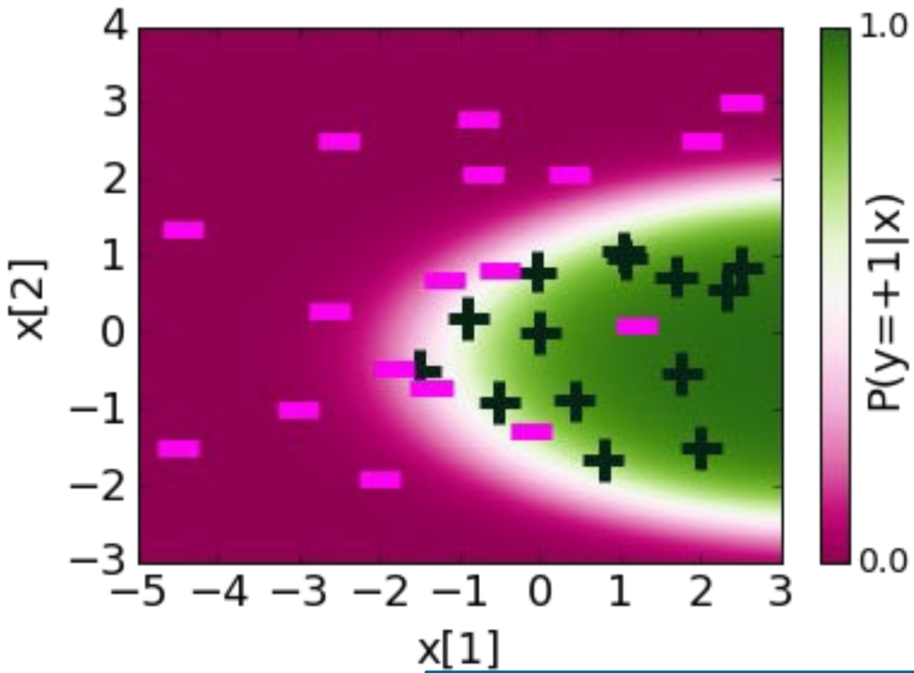
$$P(y=+1 \mid \boldsymbol{x},\boldsymbol{w}) = \frac{1}{1+e^{-\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x})}}$$



引入二次特征时的概率分布

特征	值	学到的系数
$\phi_0(\mathbf{x})$	1	1.68
$\phi_1(\mathbf{x})$	x[1]	1.39
$\phi_2(\mathbf{x})$	x[2]	-0.58
$\phi_3(\mathbf{x})$	$(\mathbf{x}[1])^2$	-0.17
$\phi_4(\mathbf{x})$	$(\mathbf{x}[2])^2$	-0.96

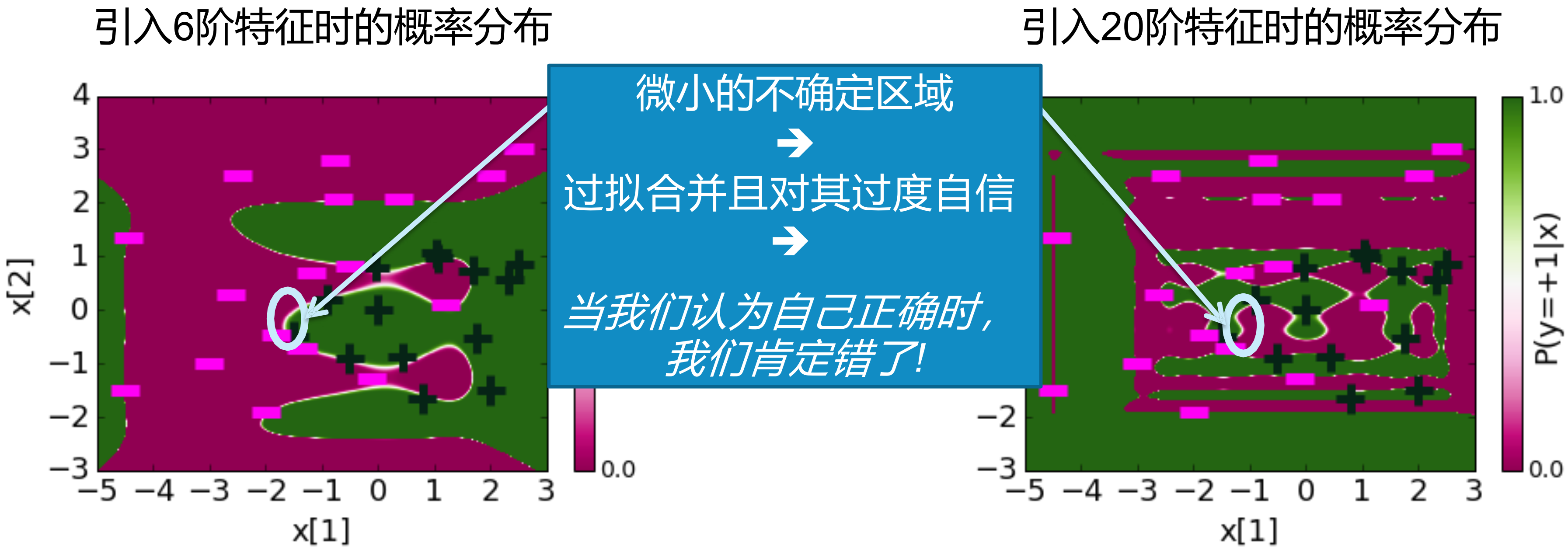
$$P(y = +1 \mid \boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + e^{-\mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})}}$$



较好地拟合数据

不确定性区域较窄

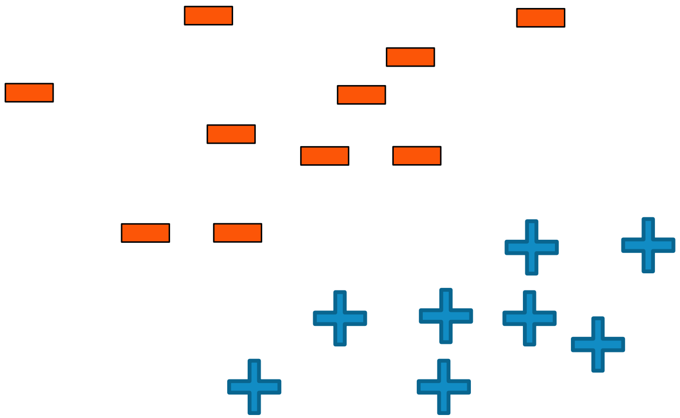
过拟合 ➔ 过分自信的预测



引入6阶特征时的概率分布
x[2]
x[1]
0.0
微小的不确定区域
➔
过拟合并且对其过度自信
➔
当我们认为自己正确时，
我们肯定错了!
引入20阶特征时的概率分布
x[1]
1.0
0.0
P(y=+1|x)

逻辑回归中的过拟合：
另一个视角

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 ŵ 使得:
  - 对于训练数据中所有正样本：

$$Score(x) = \hat{\boldsymbol{w}}^T \boldsymbol{\phi}(\boldsymbol{x}) > 0$$

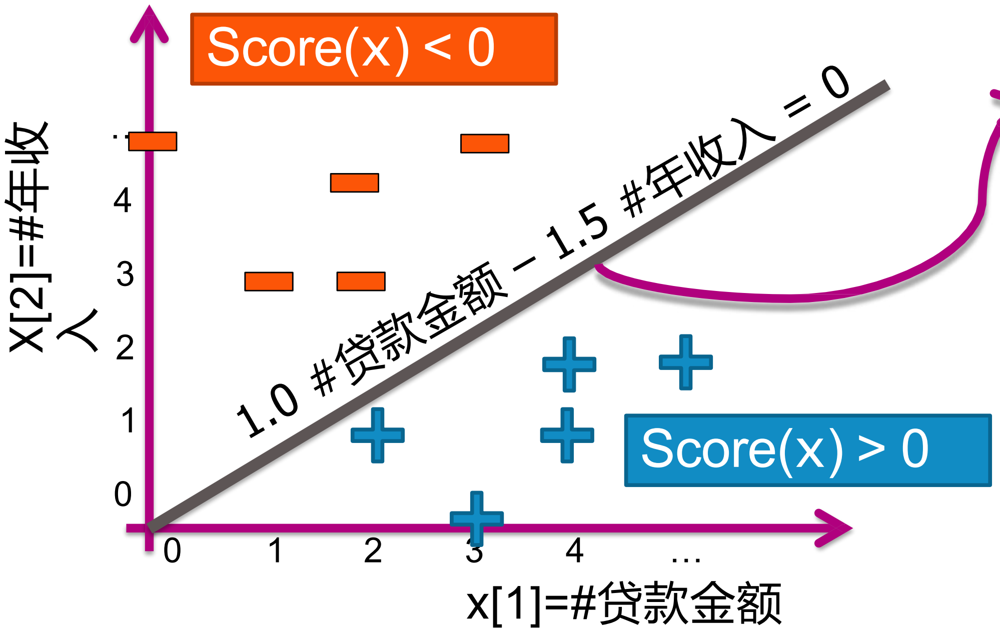
  - 对于训练数据中所有负样本：

$$Score(x) = \hat{\boldsymbol{w}}^T \boldsymbol{\phi}(\boldsymbol{x}) < 0$$

training_error(ŵ) = 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



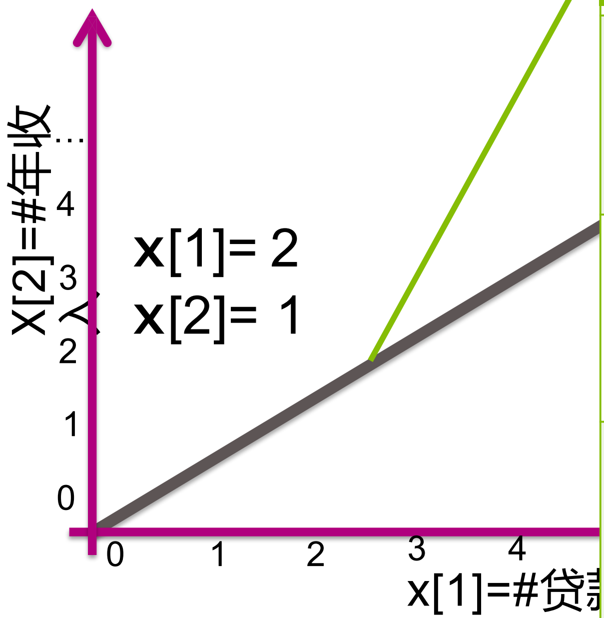
数据对 $\hat{w}_1$=1.0, $\hat{w}_2$=-1.5 线性可分

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

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

最大似然估计倾向于选择较大的系数

最大似然估计（MLE) 偏好特定模型→
对于线性可分离数据，系数趋于无穷大!



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

$$\frac{1}{1+e^{-(2-1.5)}} = 0.62$$

$$\frac{1}{1+e^{-(15-10)}} = 0.99$$

$$\frac{1}{1+e^{-(0.5\times10^{9})}} \approx 1$$

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

↓

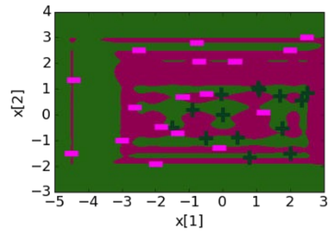
数据对 $\hat{w}_1$=10^9, $\hat{w}_2$=-1.5x10^9 同样线性可分z

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

$\hat{w}_1 = 10^9$

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

目录

01 多分类问题

02 过拟合的定义

03 过拟合的两个视角

04 过拟合的解决方案

加入系数惩罚项以缓解过拟合

目标损失形式

□ 总损失的形式

我们希望平衡：

i. 拟合程度

ii. 系数大小

我们希望平衡

总体质量 =拟合程度的度量（measure of fit） - 系数大小的度量

(数据似然度)
大# ➔ 在训练数据上
拟合较好

大 # ➔ 过拟合

加入系数惩罚项以缓解过拟合

L2正则化的逻辑回归

L2正则化

选择 $\hat{w}$ 以最大化：

$$\ell(w) - \lambda \|w\|_2^2$$

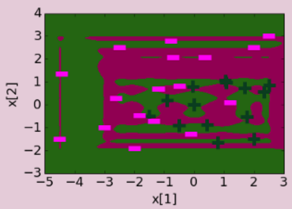
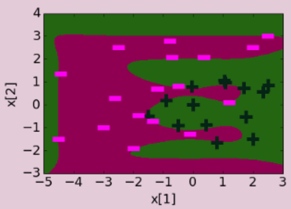
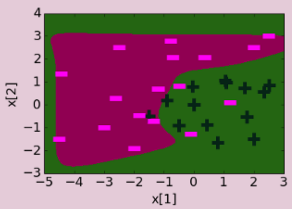
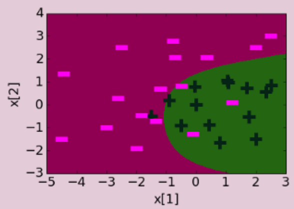
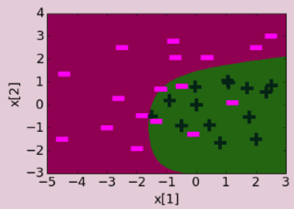
调整参数λ = 在拟合与参数规模之间平衡

L2正则的逻辑回归

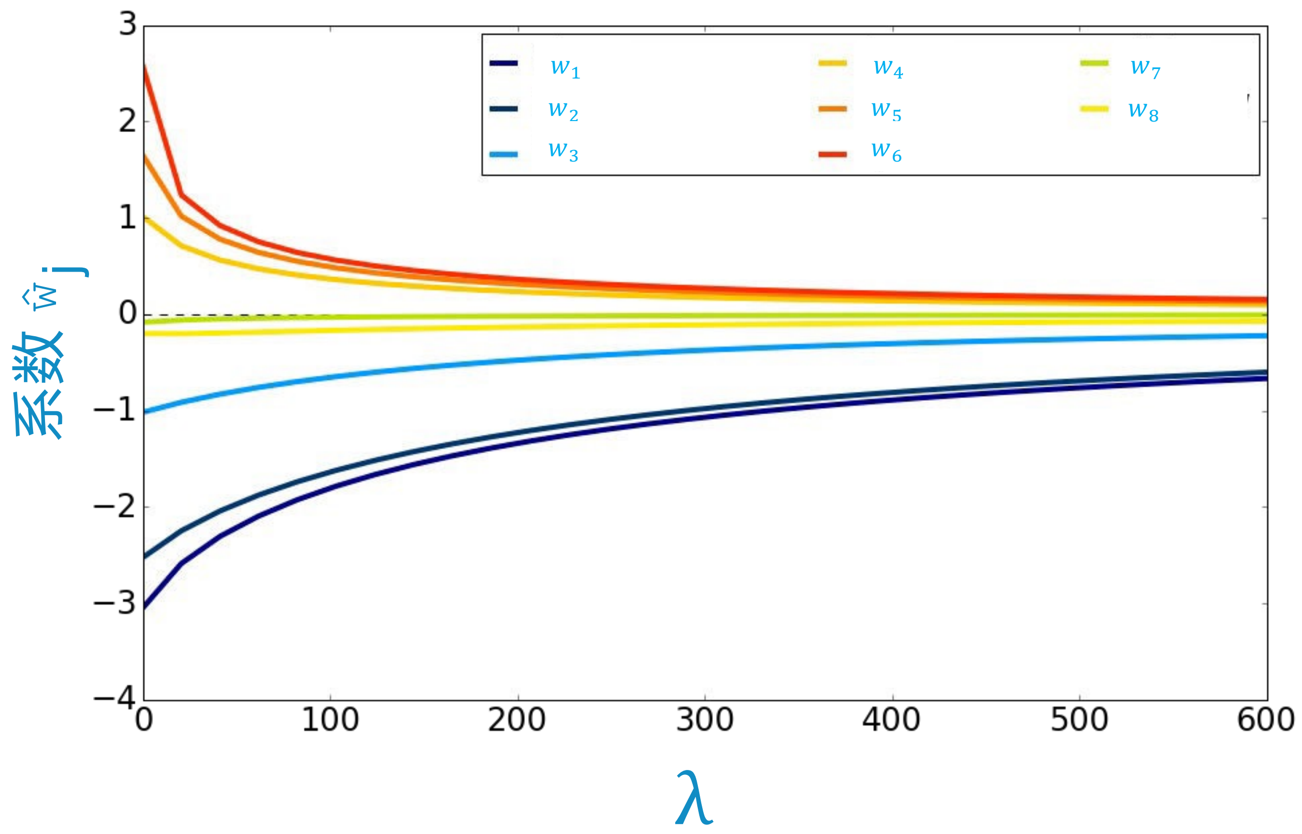
使用以下方式选择 λ：
- 验证集（适用于大型数据集）
- 交叉验证（适用于较小的数据集）

惩罚系数λ对决策边界的影响

☐ 20 阶特征下的决策边界

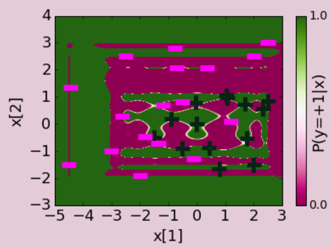
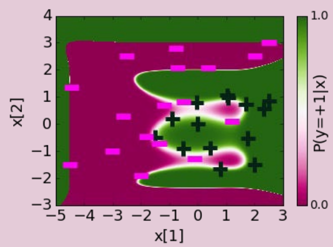
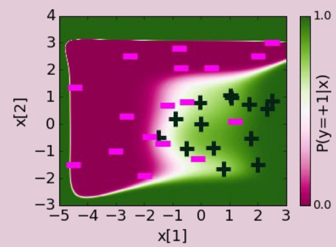
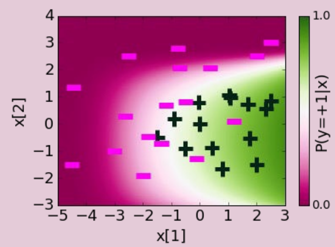
正则化	λ = 0	λ = 0.00001	λ = 0.001	λ = 1	λ = 10
系数范围	-3170 ~ 3803	-8.04 ~ 12.14	-0.70 ~ 1.25	-0.13 ~ 0.57	-0.05 ~ 0.22
决策边界					

系数路径（Coefficient path）



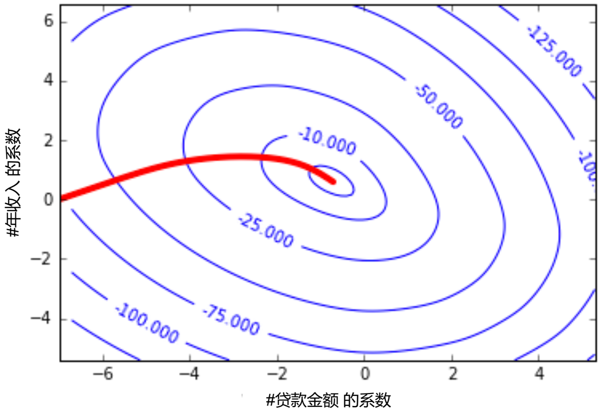
惩罚系数λ对概率分布的影响

□ 20 阶特征下的概率分布

正则化	λ = 0	λ = 0.00001	λ = 0.001	λ = 1
系数范围	-3170 to 3803	-8.04 to 12.14	-0.70 to 1.25	-0.13 to 0.57
概率分布				

L2正则化的逻辑回归的梯度上升

□ L2正则逻辑回归的梯度上升



init $\mathbf{w}^{(1)}=0$, $t=1$

while $\|\nabla\ell(\mathbf{w}^{(t)})\| > \varepsilon$

  for $j=0,\ldots,D$

    $$\text{partial}[j] = \sum_{i=1}^{N} \boldsymbol{\phi}_j(\mathbf{x}_i)\Big(\mathbb{1}[y_i=+1] - P(y=+1 \mid \mathbf{x}_i, \mathbf{w}^{(t)})\Big)$$

    $$\mathbf{w}_j^{(t+1)} \leftarrow \mathbf{w}_j^{(t)} + \eta\,(\text{partial}[j] - 2\lambda\,\mathbf{w}_j^{(t)})$$

  $t \leftarrow t+1$

加入系数惩罚项以缓解过拟合

L1正则化的逻辑回归

L1正则化

□ 使用L1惩罚进行稀疏逻辑回归

选择 $\hat{w}$ 以最大化：

$$\ell(w) - \lambda||w||_1$$

调整参数λ = 在拟合与参数规模之间平衡

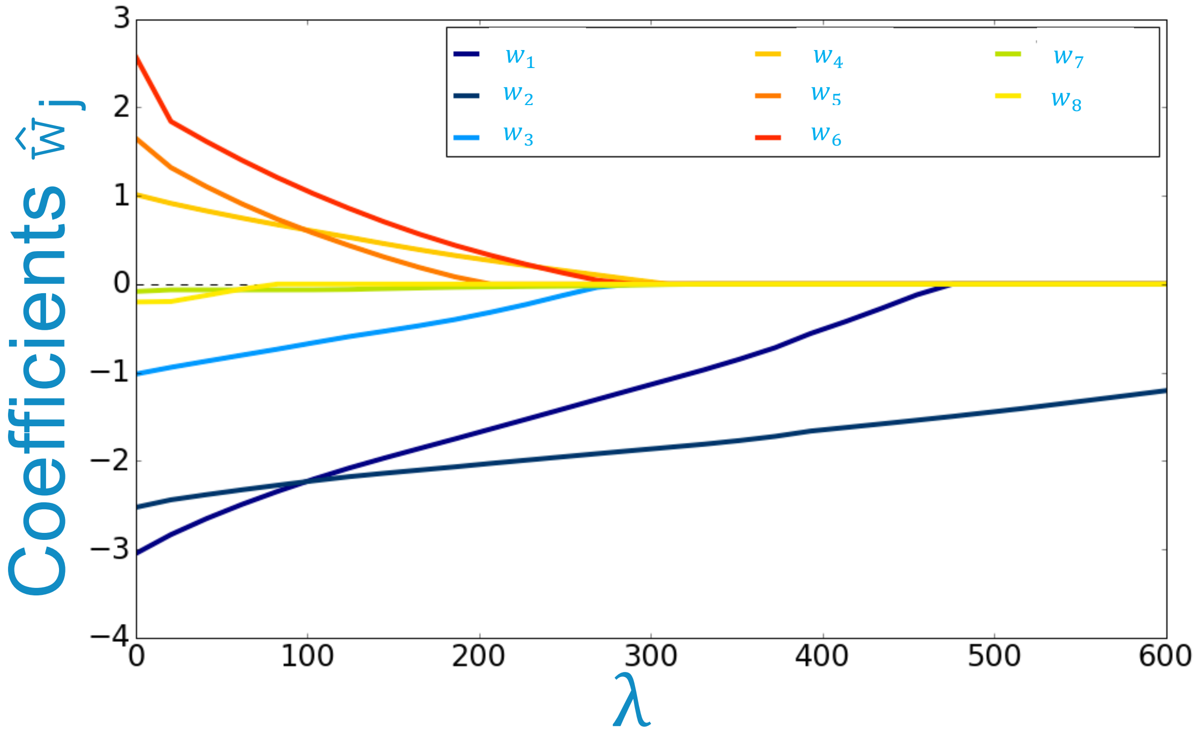
L1正则的逻辑回归

使用以下方式选择 λ：

- 验证集（适用于大型数据集）
- 交叉验证（适用于较小的数据集）（如岭/套索回归）

逻辑回归中的过拟合

☐ L1正则下的系数路径



逻辑回归中的过拟合总结

逻辑回归中的过拟合

- 描述分类任务中过拟合的表现与影响
  - 识别过拟合的发生时机
  - 较大的学习系数与过拟合现象
  - 分析过拟合对线性分类器决策边界及预测概率的影响

- 利用正则化缓解过拟合问题
  - 阐述L2正则化逻辑回归质量指标的构建动机
  - 采用L1正则化获得稀疏逻辑回归解
  - 分析调节参数λ变化对估计系数的影响规律
  - 使用梯度上升法估计L2正则化逻辑回归系数
  - 解读系数路径图

目录

01 多分类问题

02 过拟合的定义

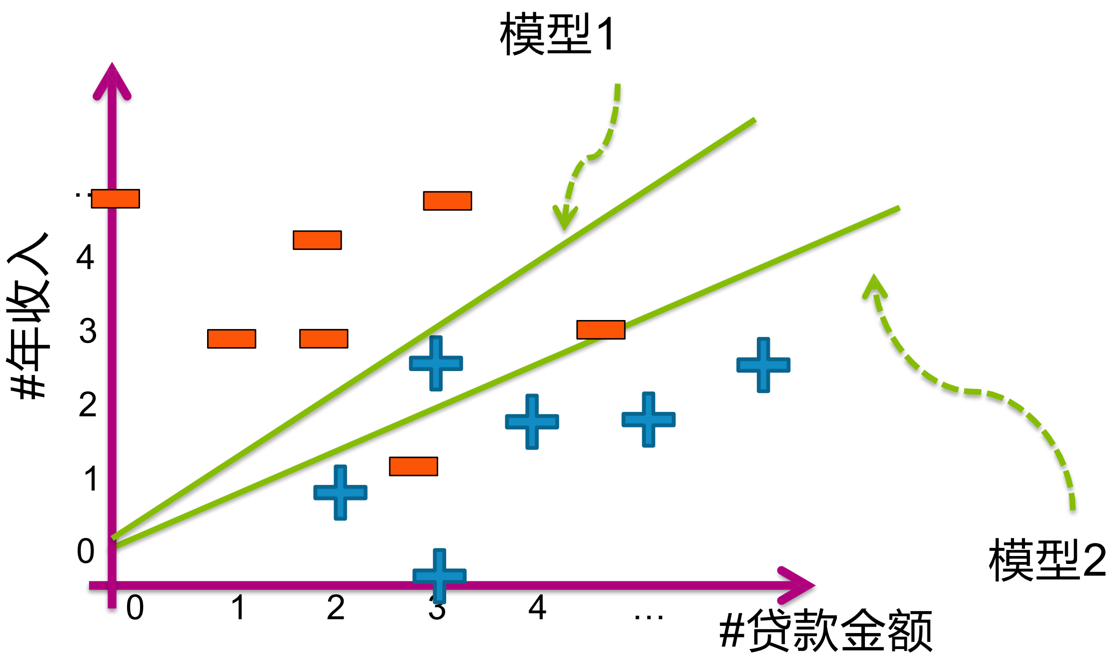
03 过拟合的两个视角

04 过拟合的解决方案

05 评估方法（选）

分类算法评估

- 没有测量就没有改进



两个模型谁更好，谁更差？

分类算法评估

□ 混淆矩阵

混淆矩阵		真实值	
		Positive	Negative
预测值	Positive	TP	FP
	Negative	FN	TN

行代表真实，
列代表预测。

真实值为 Positive，预测值为 Positive，标记为 TP — 真阳性
真实值为 Positive，预测值为 Negative，标记为 FN — 假阴性
真实值为 Negative，预测值为 Positive，标记为 FP — 假阳性
真实值为 Negative，预测值为 Negative，标记为 TN — 真阴性

分类算法评估

混淆矩阵

准确率

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

适合类别均衡场景

精确率

$$\text{Precision} = \frac{TP}{TP + FP}$$

预测为正的样本中有多少真实为正，反欺诈场景关键指标

召回率

$$\text{Reall} = \frac{TP}{TP + FN}$$

真实为正的样本中有多少被找出，医疗检测关键指标

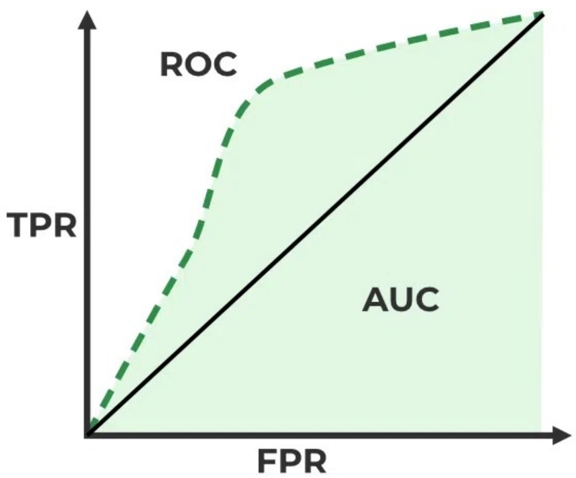
F1分数

$$F_1 = \frac{2TP}{2TP + FP + FN}$$

协调精确率与召回率的调和平均

分类算法评估

ROC曲线下的面积（ROC-AUC）



真阳性率

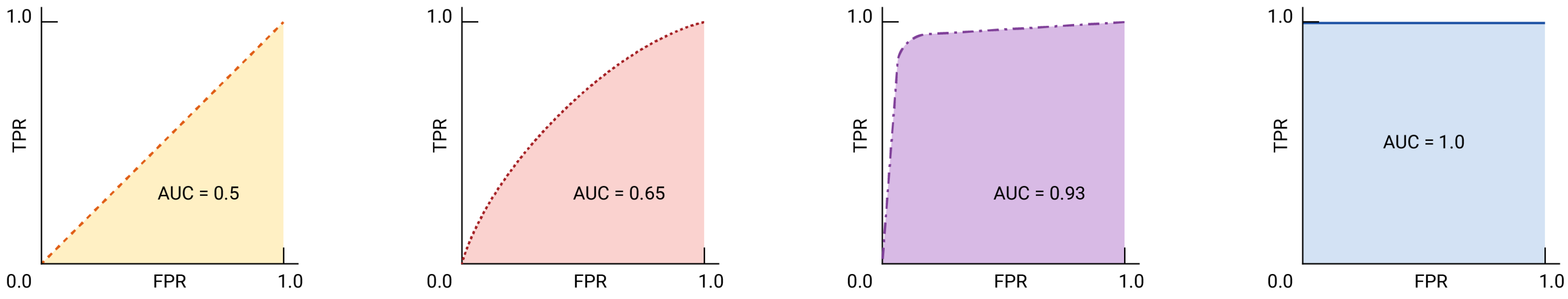
$$\mathrm{TPR} = \frac{TP}{TP + FN} = \mathrm{recall}_{positive}$$

假阳性率

$$\mathrm{FPR} = \frac{FP}{FP + TN} = \frac{FP + TN - TN}{FP + TN} = 1 - \frac{TN}{FP + TN} = 1 - \mathrm{recall}_{negative}$$

分类算法评估

◻ ROC曲线下的面积（ROC-AUC）



随机分类器　　较弱的分类器　　较好的分类器　　完美的分类器