

目录

精通数据和管

ONE 非均衡数据集

从编馆回的初深覆管

精通数据科学 从给收益回的歌作春度管型

准确度悖论

精通数据测污彩港

糖通数据科学

TWO 解决办法 调整类别权重

超通数源和学

THREE 代码实现。这种"

指标定义

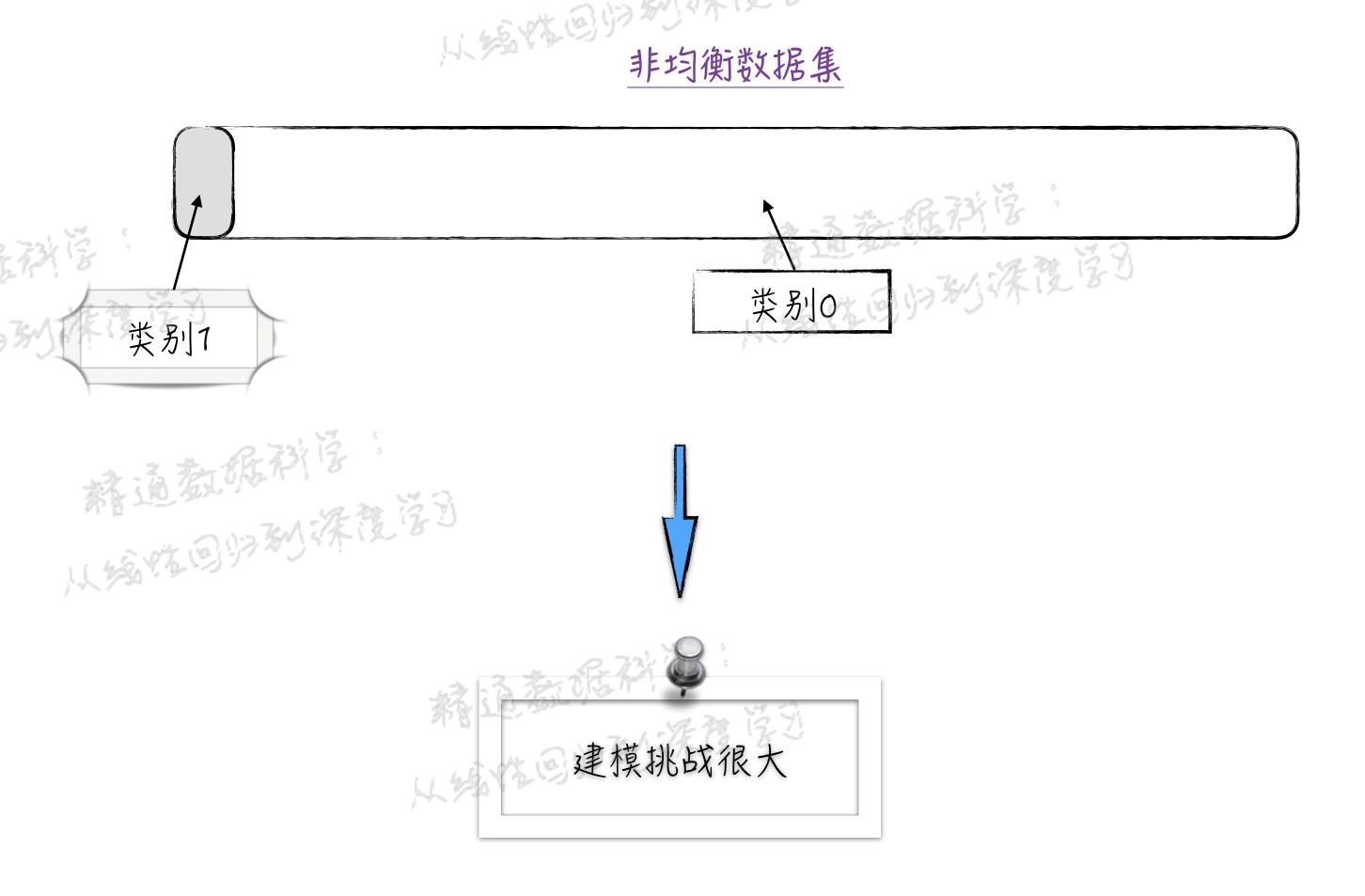
精通数据科学

各类别占比差别很大的数据集就是非均衡数据集

- · 在实际的应用中, 会经常遇到非均衡数据集, 比如信贷、反欺诈、广告预测等
- ·对于多元分类问题,如果使用OvR策略,也容易引发"潜在的"非均衡数据集

非均衡数据集会给模型搭建带来困难





准确度定义

精通数据科学

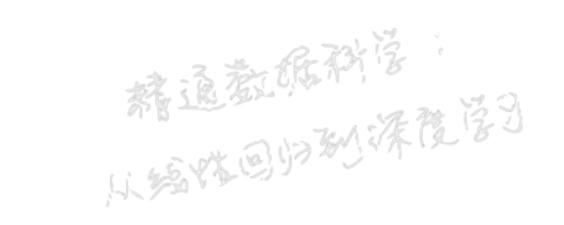
准确度 = 预测准确的样本数 / 全体样本数

这个看似很直观的评估指标在面对非均衡 分类问题时会严重失真 糖通数概不管

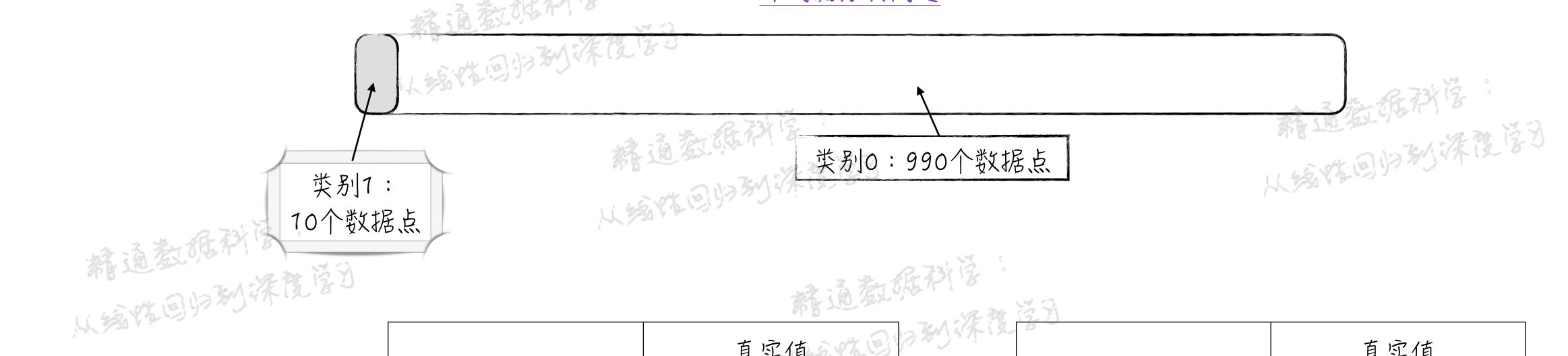
		真实值		
8 2 m 7 7		1	20世间的形体度量的	
7 . VIII / +		真阳性 (true positive) TP	伪阳性 (false positive) FP	
预测值	MORI	伪阴性 (false negative) FN	真阴性(true negative) TN	

$$ACC = \frac{TP + TN}{TP + FP + FN + TN}$$

准确度悖论



非均衡分类问题



准确度悖论: 模型A比模型B更好?

	模型A		真实值	
悖论:			1	0
型B更好?	33 加 (古	1	0	0
1. 25 PM	预测值		10	990
320 1.10				

$$ACC(A) = \frac{990+0}{990+10+0+0} = 99\%$$

大芒 -	型B	真实值		
√关 <u>·</u>	至D	1	0	
预测值	的粉練質	Z 9	90	
	0	1	900	

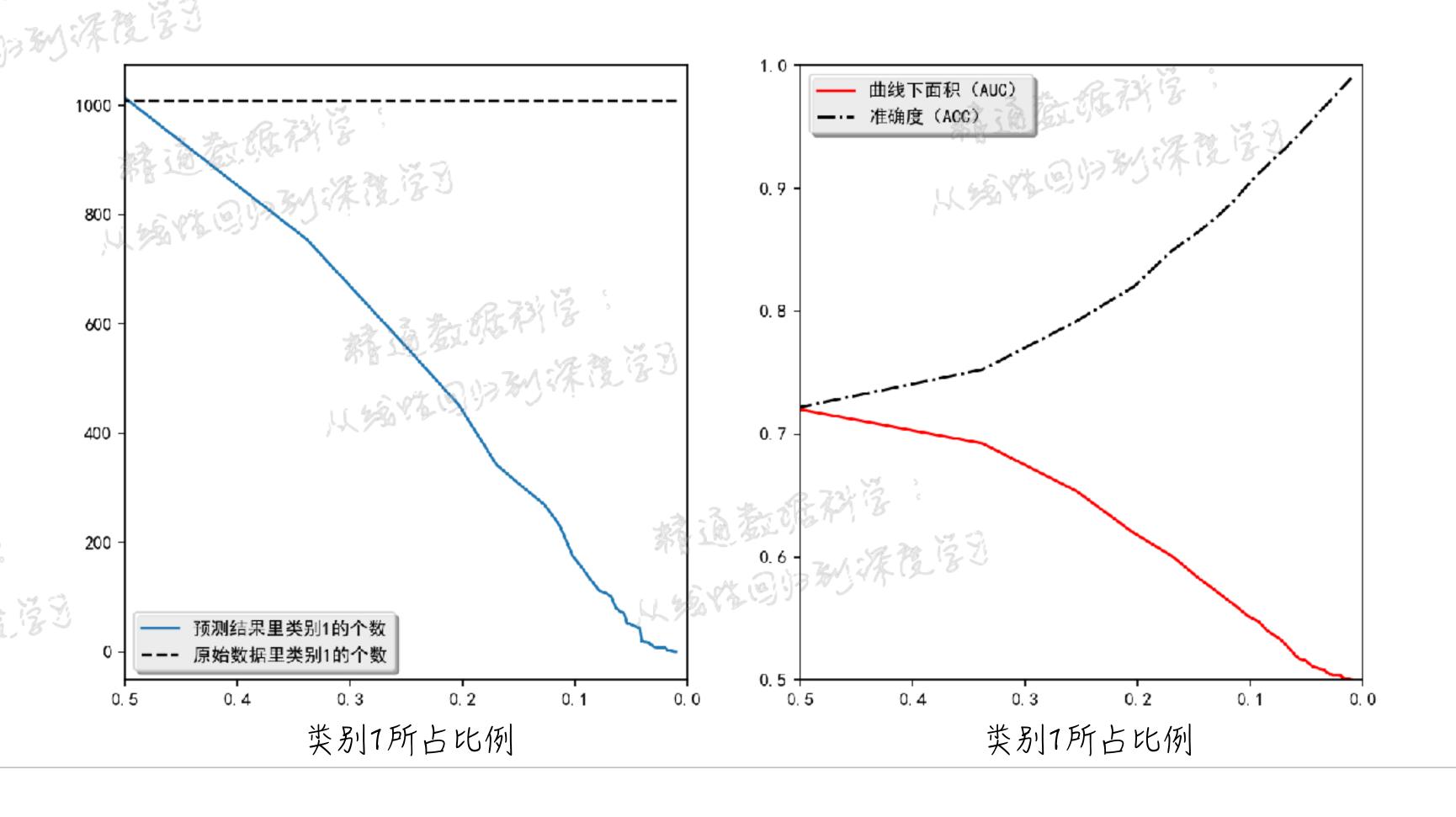
$$ACC(B) = \frac{900+9}{900+90+9+1} = 90.9\%$$

对建模的影响

 $y = \begin{cases} 1, x_1 - x_2 + \varepsilon > 0 \\ 0, else \end{cases}$ $\longrightarrow \mathcal{E}$ 是随机扰动项, $\longrightarrow \mathcal{E}$ 辑回归模型的假设

使用逻辑回归模型对数据进行建模

- · 虽然y, x1, x2之间完美服从逻辑回归模型的假设,但数据越不均衡,模型效果越差
- · 当面对非均衡数据集时,准确度这个 评估指标会严重失真



目录

精通数据科学 所述 E 非均衡数据集

准确感染色态。

糖通数概念清楚

精通查证程度等

糖通数据科学

TWO 解决办法 调整类别权重

楼通数流彩道:

THREE 代码实现。阿科·森德等

解决办法

数学原因

4(Xi)靠近7, 预测结果为类别7; 反之预测结果为0

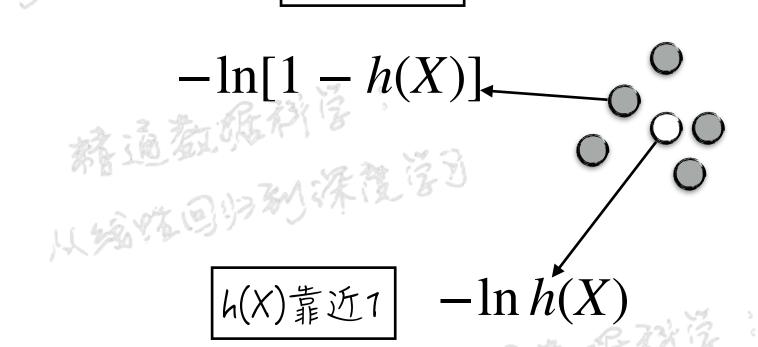
逻辑回归的参数估计公式:

$$h(X_i) = \frac{1}{(1 + e^{-X_i \beta})}$$

$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{i} -y_i \ln h(X_i) - (1 - y_i) \ln[1 - h(X_i)]$$

类别1少,类别0多

数据分布:



h(X)靠近0

每个点的权重都等于

〇 类别0

〇 类别1

 $X_1 \approx X_2 \approx \ldots \approx X_6 \approx X$

模型训练结果:

 $h(X) \approx 0$

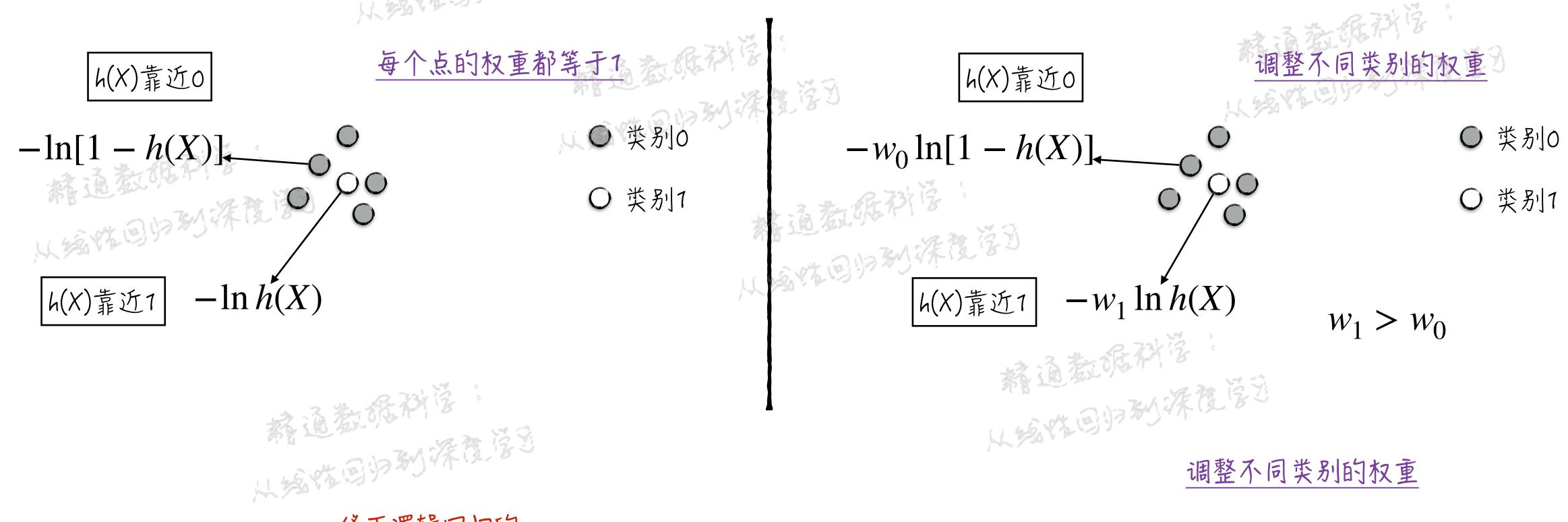
从给你回归和深度管理 "牺牲"类别7,"迁就"类别0; 模型预测结果几乎都为类别O

解决办法

调整类别权重

每个点的权重都等于1

$$\widehat{\beta} = \operatorname{argmin}_{\beta} \sum_{i} -y_{i} \ln h(X_{i}) - (1 - y_{i}) \ln[1 - h(X_{i})]$$



调整不同类别的权重

修正逻辑回归的 损失函数

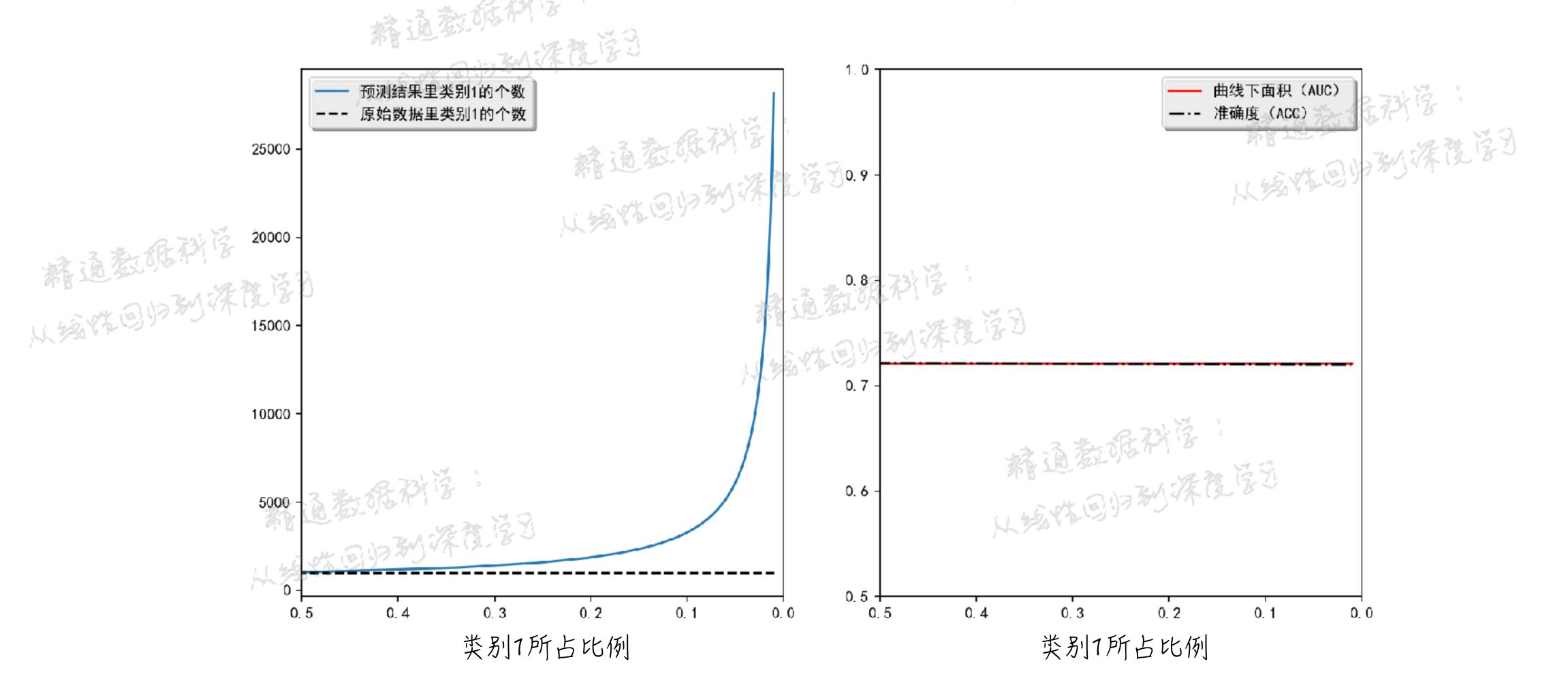
$$\hat{\beta} = \operatorname{argmin}_{\beta} \sum_{i} -w_{1}y_{i} \ln h(X_{i}) - w_{0}(1 - y_{i}) \ln[1 - h(X_{i})]$$

解决办法

调整类别权重

糖通数原料管

平衡类别权重后的模型效果



目录

精通数据和资

准确感染态态。 从编馆回的到海德湾到 精通查验系科管,

精通数概称管

从给收益回的秘证不管管的

糖通数据科学

TVV 解决数据

楼通数流彩道:

THREE 代码实现。为为"体质"。

精通数据科学: 从始级回的秘证

THANK等

从给性回归那样度管的

精通数据科语: 从给您回归到深度管理

超通数混彩道: 从给你国的秘证

糖通数概称管: 从给你回的秘证不随管的

> 精通数据科学 从绝对回归对流汗度管的

精通数据科学: 从给你回的那样随管型