



本课件仅用于教学使用。未经许可，任何单位、组织和个人不得将课件用于该课程教学之外的用途(包括但不限于盈利等)，也不得上传至可公开访问的网络环境

1

# 新媒体大数据分析

## New Media Big Data Analysis

### 第三章 数据建模

黄振亚，朱孟潇，张凯

课程主页：

<http://staff.ustc.edu.cn/~huangzhy/Course/NM2025.html>

助教：齐畅，朱家骏

[bigdata\\_2025@163.com](mailto:bigdata_2025@163.com)

11/26/2025



# 分类与预测

91

- 有监督学习：分类与预测
- 常用方法
  - 规则方法
  - 决策树
  - 最近邻方法
  - 感知机，支持向量机 (SVM)
  - 集成方法
- 分类的评价指标
- 类不平衡问题



# 分类：集成学习

92

## □ 分类——集成学习

- 思想：集成多个模型的能力，得到比单一模型更好的效果
- 为什么能够提升效果？

- 增强模型的表达能力

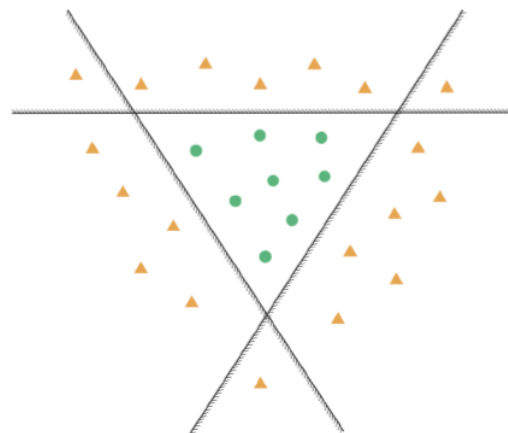
- 单个感知机无法正确分类数据能用三个感知机完成

- 降低误差

- 假设单个分类器误差 $p$ ，有 $T$ 个独立的分类器采用投票进行预测，得到集成模型  $H$
  - 集成分类器误差为

$$Error_H = \sum_{k \leq \frac{T}{2}} C_T^k \cdot p^{T-k} \cdot (1-p)^k$$

- $T = 5, p = 0.1$ 时,  $Error_H < 0.01$





# 分类：集成学习

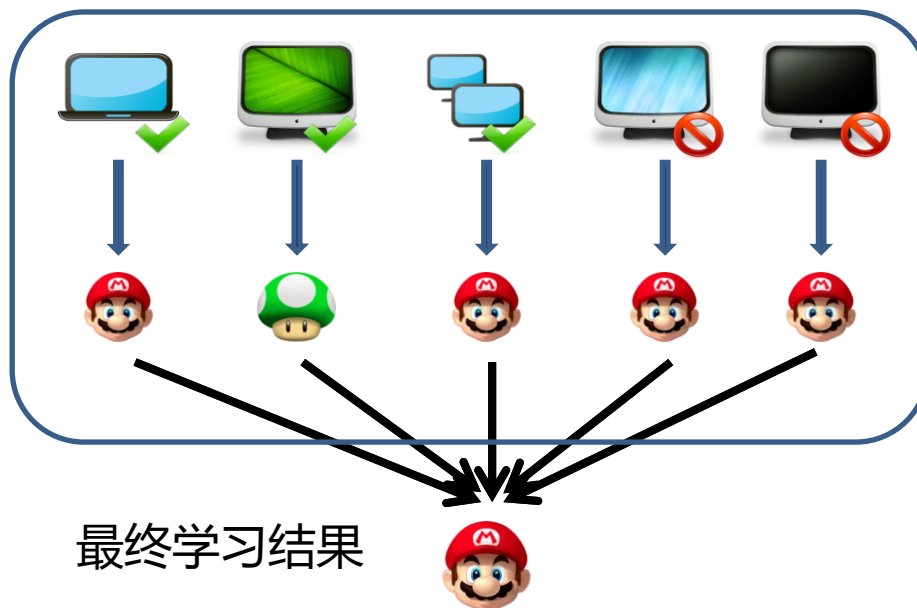
93

## □ 分类——集成学习

### □ 集成过程

学习器

各自学习  
结果



投票

最终学习结果

### □ 此类方法经常用在数据挖掘竞赛中(如KDD CUP, CCF-BDCI)

- KDD Cup2021 MAG240M-LSC赛道第一名集成了30个模型
- KDD Cup2021 WikiKG90M-LSC赛道第三名集成了15个模型



# 分类：集成学习

94

## ▣ 常见集成学习方法

### ▣ Bagging (Bootstrap Aggregating)

- 对样本或特征随机取样，学习产生多个独立的模型，然后平均所有模型的预测值
- 主要减小方差
- 典型代表：随机森林

### ▣ Boosting

- 串行训练多个模型，后面的模型是基于前面模型的训练结果（误差）
- 主要减小偏差
- 典型代表：AdaBoost



# 随机森林

95

## 集成学习算法：随机森林

- 最典型的Bagging算法：算法思想

- 随机：每棵树，保证各棵树之间的独立性，采用两到三层的随机性

  - 随机有放回的抽取样本作为训练集

样本和特征尽可能不同

  - 随机选取m个特征作为树节点的划分特征

  - 随机选择特征取值进行分割 (不遍历特征所有取值)

- 森林：多颗决策树集成

  - 假设使用三棵决策树组合成随机森林，每各不相同且预测结果相互独立，每棵树的预测错误率为 40%。那么两棵树以及上预测错误的概率下降为：三三棵全部错误+两棵树错误一个正确 =  $0.4^3 + 3 * 0.4^2 * (1 - 0.4) = 0.352$

`sklearn.ensemble.RandomForestClassifier`



# AdaBoost

96

## 集成学习算法：AdaBoost

- 最有代表性的Boosting算法

- 算法思想**：利用同一训练样本的不同加权版本，训练一组弱分类器，然后把这些弱分类器以加权的形式集成起来，形成一个最终的强分类器：

- 在每一步迭代过程中，会给训练集中的样本赋予一个权重  $w_1, w_2, \dots, w_n$
- 样本的初始权重都一样，设置为  $\frac{1}{n}$ ；
- 在每一步迭代过程中，
  - 被当前弱分类器**分错的样本**的权重会相应得到**提高**
  - 被当前弱分类器**分对的样本**的权重则会相应**降低**；
- 弱分类器的权重则根据当前分类器的加权错误率来确定。

```
from sklearn.ensemble import AdaBoostClassifier
```



# 分类与预测

97

- 有监督学习：分类与预测
- 常用方法
  - 规则方法
  - 决策树
  - 最近邻方法
  - 支持向量机 (SVM)
  - 集成方法
- 分类的评价指标





# 分类模型的评价

98

- 如何评价分类模型的效果？——以二分类为例
  - 基本概念
    - T/F: True or False, 表示二分类结果的正确与否
    - P/N: Positive or Negative, 表示算法对样本的判断
  - 四种简写的含义:
    - 真正(True Positive, TP): 样本为正例, 预测为正, (正确)
    - 假负(False Negative, FN): 样本为正例, 预测为负, (错误)
    - 假正(False Positive, FP): 样本为负例, 预测为正, (错误)
    - 真负(True Negative, TN): 样本为负例, 预测为负, (正确)



# 分类模型的评价

99

- 如何评价分类模型的效果？——指标Accuracy
  - 通常用混淆矩阵表示：TP、FN、FP、 TP

ACTUAL (真实) CLASS	PREDICTED (预测) CLASS	
	Class=Yes	Class=No
	Class=Yes a (TP) b (FN)	Class=No c (FP) d (TN)

- 评价指标1:  $Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$



# 分类模型的评价

100

- 如何评价分类模型的效果？——指标Accuracy
  - 举例：假设一共有200个测试数据样本，以下是使用分类模型得到的分类结果，请问Accuracy是多少？

ACTUAL (真实) CLASS	PREDICTED (预测) CLASS		
		Class=Yes	Class=No
		60 (TP)	40 (FN)
	Class=No	20 (FP)	80 (TN)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{60 + 80}{60 + 80 + 20 + 40} = 0.7$$



# 分类模型的评价

101

## □ 如何评价分类模型的效果？ — Accuracy的局限性

□ 样本不均衡时，评估结果可能不合理

□ 例子：考虑2分类问题

■ 假设10000个数据样本中，类别0的样本数为 9990，类别1的样本数为 10

■ 假设模型将所有样本均预测为0

■ 计算  $\text{Accuracy} = 9990/10000 = 99.9\%$

Accuracy很高，  
表明模型很好？

结论：模型不好

原因：10个正例均预测错误

分析：这个例子中，显然我们关注类别为1的样本，但Accuracy被类别为0的样本影响了

Count	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	a	b
ACTUAL CLASS	Class=No	c	d



# 分类模型的评价

102

## 如何评价分类模型的效果？——分类问题的常用指标

□ 准确率(查准率)  $\text{Precision}(p) = \frac{a}{a+c} = \frac{TP}{TP+FP}$ ：预测为Yes中正确的比例

■ 正确预测的个体总数 / 预测出的个体总数

□ 召回率(查全率)  $\text{Recall}(r) = \frac{a}{a+b} = \frac{TP}{TP+FN}$ ：真实为Yes中被预测正确的比例

■ 正确预测的个体总数 / 测试集中Yes的个体总数

Count	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
ACTUAL CLASS	Class=No	c (FP)	d (TN)



# 分类模型的评价

103

## □ 分类模型的其他指标——分类问题的常用指标

□ F值 =  $\frac{2PR}{P+R} = \frac{2a}{2a+b+c}$  : 正确率和召回率的调和平均值

□ 意义: 不同应用场景中, 对于准确率和召回率有着不同的侧重

- 邮件分类: 宁愿放过一些垃圾邮件, 也不能错杀正常邮件
  - 牺牲召回率, 保证较高准确率
- 智慧医疗: 宁愿多判断一些疑似患者, 不能漏掉一个病人
  - 牺牲准确率, 保证较高召回率

Count	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	a	b
ACTUAL CLASS	Class=No	c	d



# 分类模型的评价

104

## □ 分类问题的常用指标——课堂练习

- 在一次垃圾邮件检测中，使用某分类模型认为有100篇邮件是垃圾邮件，后经过专家判定，其中真是垃圾邮件的为60篇，其余的40篇为误分类，那么请问本次分类的准确率Precision就等于\_\_\_\_\_。
- 假如专家发现邮件样本集里还有90篇垃圾邮件，由于各种原因而未被检出（漏检），那么按照上述公式，本次分类的查全率Recall就等于\_\_\_\_\_, F1值等于\_\_\_\_\_。

$$\text{Precision}(P) = \frac{a}{a+c}$$

$$\text{Recall}(R) = \frac{a}{a+b}$$

$$F\text{值} = \frac{2PR}{P+R} = \frac{2a}{2a+b+c}$$

	PREDICTED CLASS	
	Class=Yes	Class=No
	Class=Yes	Class=No
ACTUAL CLASS	a (TP)	b (FN)
	c (FP)	d (TN)



# 分类模型的评价

105

## □ 分类模型的其他指标—ROC与AUC

### □ ROC(Receiver Operating Characteristic)与AUC(Area Under the ROC curve)

- 背景：发展于20世纪50年代的信号检测理论，用于分析噪声信号
- 两者的关系：ROC曲线的面积就是AUC

### □ 基本概念

- **真正例率TPR**(True Positive Rate) =  $TP/(TP+FN)$ 
  - 预测为正且实际为正的样本占有所有正样本的比例
- **假正例率FPR**(False Positive Rate) =  $FP/(TN+FP)$ 
  - 预测为正但实际为负的样本占有所有负样本的比例

	PREDICTED CLASS		
		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
ACTUAL CLASS	Class=No	c (FP)	d (TN)





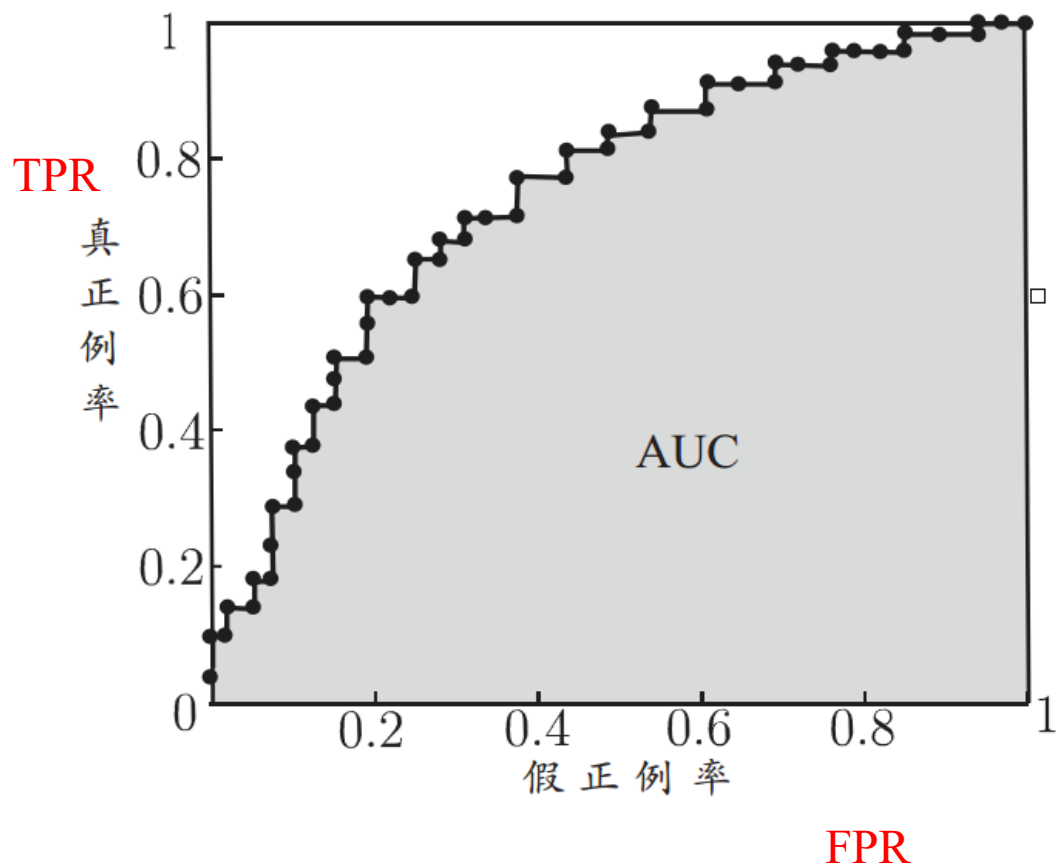
# 分类模型的评价

106

## □ ROC曲线

特点:

- 对角线表示区分能力为0，即随机猜测
- 在对角线上端越远，效果越好
- 低于对角线的结果无意义（无区分度）



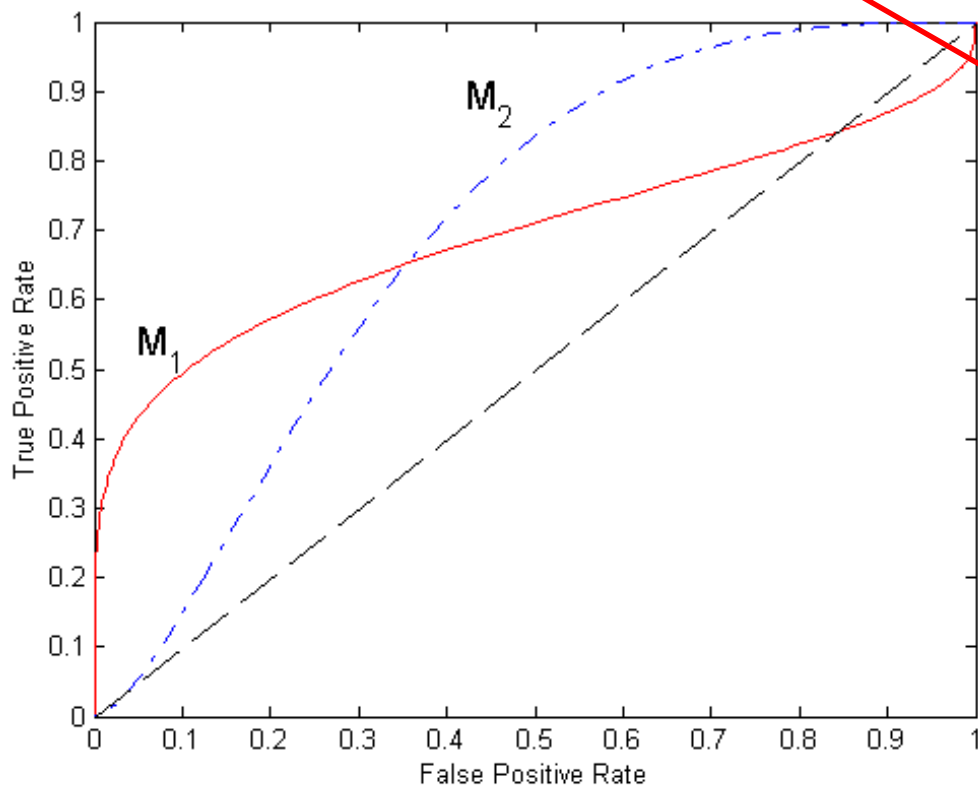


# 分类模型的评价

107

## 如何基于ROC曲线比较模型好坏？

- 一般而言，模型A的ROC曲线将模型B的完全包住，则模型A更好
- 但往往并不会完全包住



对比ROC曲线发现，两个模型都不是一直表现得好

- $M_1$ 在FPR较小时表现好
- $M_2$ 在FPR较大时表现好



# 分类模型的评价

108

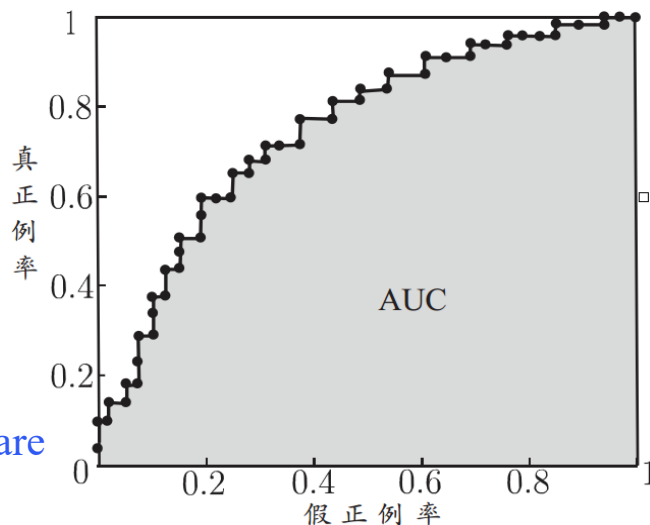
## □ ROC不能量化——AUC量化指标

- AUC定义为ROC曲线的面积，可以直接计算如下
- 假设ROC曲线由 $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m), x_1 = 0, x_m = 1\}$ 的点按序连接而形成，则AUC为：

$$AUC = \frac{1}{2} \sum_{i=1}^{m-1} (x_{i+1} - x_i) \cdot (y_{i+1} + y_i)$$

**AUC越大，结果越好**

**AUC衡量了样本预测的排序质量**





# 分类模型的评价

109

- 如何评价分类模型的效果？——分类模型的比较方法
  - 关于性能比较：
    - 测试性能并不等于泛化性能
    - 测试性能随着测试集的变化而变化
    - 很多机器学习算法本身有一定的随机性
      - 例如，对两个模型：
        - M1: accuracy = 85%, 在30个样本上测试
        - M2: accuracy = 75%, 在5000个样本上测试
  - 常常进行**假设检验**，判断不同模型的性能差别是否具有统计意义
    - 假设检验为学习器性能比较提供了重要依据，基于其结果我们可以推断出：若在测试集上观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握有多大。

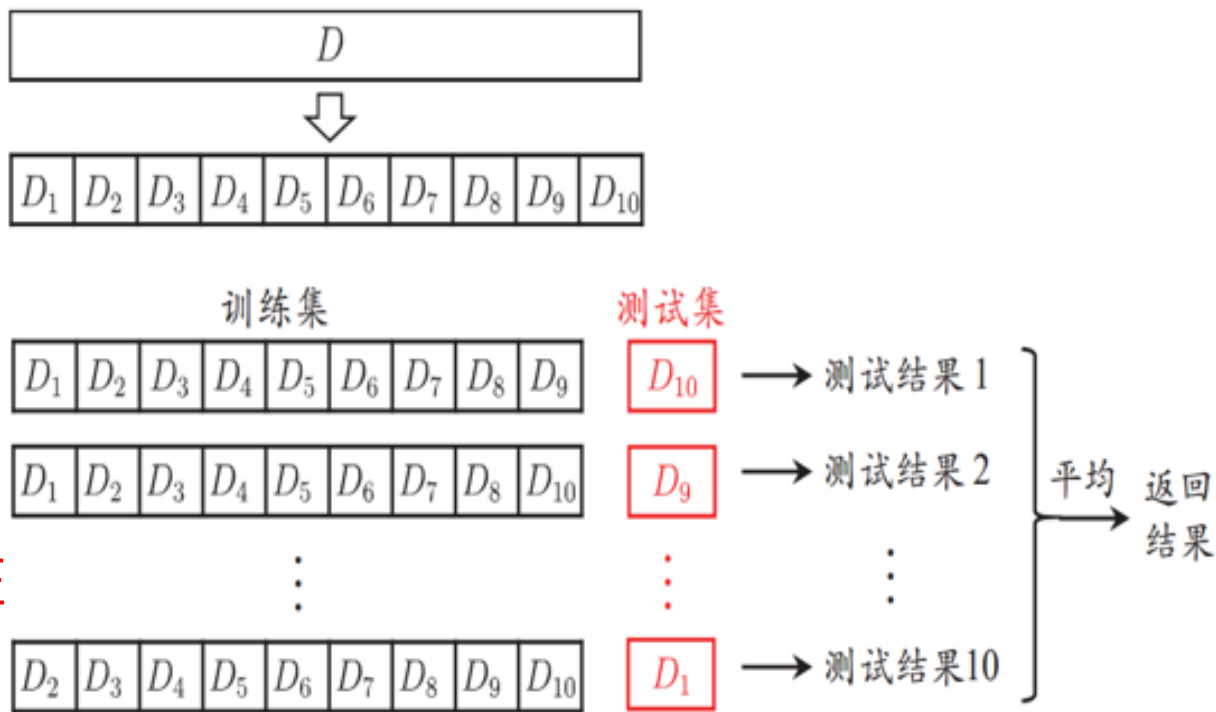


# 分类模型的验证

110

## 分类模型的验证方法 —— 增加结果的可靠性

- 交叉验证法 (Cross Validation) : 将数据集分层采样划分为k个大小相似的互斥子集, 每次用k-1个子集的并集作为训练集, 余下的子集作为测试集, 最终返回k个测试结果的均值, k最常用的取值是10.



10折交叉验证



# 分类与预测

111

- 有监督学习：分类与预测
- 常用方法
  - 规则方法
  - 决策树
  - 最近邻方法
  - 支持向量机 (SVM)
  - 集成方法
- 分类的评价指标