数据挖掘
Data Mining

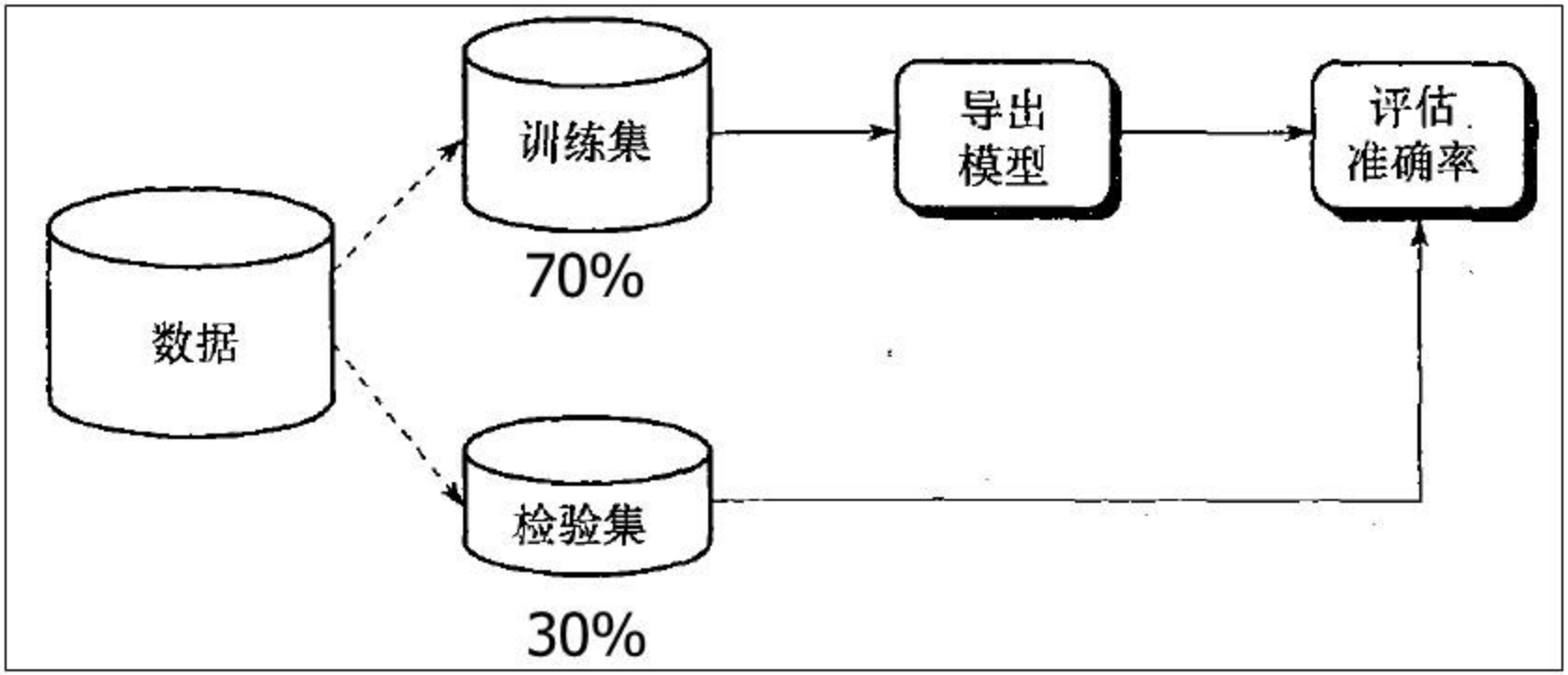
模型的评价

数据挖掘
Data Mining

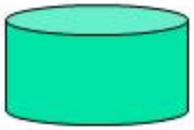
模型的评价

分类问题 Recap

- 数据预处理→模型训练→模型调整→对新数据分类→模型评价



新数据预测



内容提纲

1准确率的局限
2不平衡分类
3过拟合和欠拟合

1准确率的局限

1.1准确率评价

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

$$\text{准确率（Accuracy）} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

填空题 1分

设置

- 考虑一个二分类问题
  - 0类的实例数 = 9990
  - 1类的实例数 = 10

- 如果模型预测每个实例为0类, 则**准确率为** [填空1]

  - 准确率是误导
  - 模型不能正确预测任何1类实例
  - 而在疾病检测中，1类更需要被关心

正常使用填空题需3.0以上版本雨课堂

作答

1.2其它度量

- 混淆矩阵

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

- 真阳历TP，真阳性（True positive rate, *TPR*）或灵敏度（sensitivity）、查全率（recall）

$$TPR = TP/(TP + FN)$$

- 真阴历TN，真阴性（True negative rate, *TNR*）或特指度（specificity）

$$TNR = TN/(TN + FP)$$

- 假阳历FP，假阳性（False positive rate, *FPR*）或 误报率

$$FPR = FP/(TN + FP)$$

- 假阴历FN，假阴性（False negative rate, *FNR*）漏报率（与查全率此消彼长）

$$FNR = FN/(TP + FN)$$

TPR是指

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

A 真阳性

B 灵敏度

C 漏报率

D 查全率（recall）

提交

TNR是指

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

A 真阴性

B 灵敏度

C 特指度

D 查全率（recall）

提交

FPR是指

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

A 假阴性

B 假阳性

C 漏报率

D 误报率

提交

FNR是指

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

A 假阴性

B 假阳性

C 漏报率

D 误报率

提交

1.2其它度量(续)

- 两个广泛使用的度量
  - 召回率（查全率，**recall**）和精确率（查准率，**precision**）

$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

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

填空题 7分

设置

- 假设我们手上有60个正样本，40个负样本，我们要找出所有的正样本，系统查找出50个，其中只有40个是真正的正样本，计算上述各指标。

  - TP: 将正类预测为正类数:[填空1]
  - FN: 将正类预测为负类数:[填空2]
  - FP: 将负类预测为正类数:[填空3]
  - TN: 将负类预测为负类数:[填空4]

  - 准确率(accuracy)= 预测对的/所有 = (TP+TN)/(TP+FN+FP+TN) =[填空5
  - 精确率(precision)= TP/(TP+FP) =[填空6]
  - 召回率(recall)= TP/(TP+FN) = [填空7]

正常使用填空题需3.0以上版本雨课堂

作答

1.2其它度量(续)

- 假设我们手上有60个正样本，40个负样本，我们要找出所有的正样本，系统查找出50个，其中只有40个是真正的正样本，计算上述各指标。

  - TP: 将正类预测为正类数 40
  - FN: 将正类预测为负类数 20 （60-40，剩余没正确分类的正样本）
  - FP: 将负类预测为正类数 10
  - TN: 将负类预测为负类数 30

  - 准确率(accuracy)= 预测对的/所有 = (TP+TN)/(TP+FN+FP+TN) = 70%
  - 精确率(precision) = TP/(TP+FP) = 80%
  - 召回率(recall) = TP/(TP+FN) = 2/3

设置

1.3查全率vs. 查准率

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

- 下面是两个场景：
  - 1. **地震的预测**，对于地震的预测，我们希望的是recall非常高，也就是说每次地震我们都希望预测出来。这个时候我们可以牺牲precision。情愿发出1000次警报，把10次地震都预测正确了

    TP= [填空1] ,FN= [填空2] ,FP= [填空3]

    也不要预测100次，对了8次，漏了2次。

    TP= [填空4] ,FN= [填空5] ,FP= [填空6]
  - 2. **嫌疑人定罪**，基于不错怪一个好人的原则（无罪推定原则，presumption of innocence），对于嫌疑人的定罪我们希望是非常准确的（precision高），及时有时候放过了一些罪犯（recall低），但也是值得的。

$$F_1 = \frac{2rp}{r+p} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

正常使用填空题需3.0以上版本雨课堂

作答

1.4 ROC曲线

- 前面分类器性能评价的局限性：分类器预测结果为离散的1或者0
- 朴素贝叶斯输出？ $p(x|y)$=?
- 其他分类器输出？

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P’	N’	All

1.4 ROC曲线

- 前面分类器性能评价的局限性：分类器预测结果为离散的1或者0
- 朴素贝叶斯输出？ $p(x|y)=?$
- 其他分类器输出？

输出是一个连续的概率值，且同我们仅仅关系“1”类别的概率

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P’	N’	All

1.4 ROC曲线

- 前面分类器性能评价的局限性：分类器预测结果为离散的1或者0
- 朴素贝叶斯输出？ $p(x|y)=?$
- 其他分类器输出？

Instance	P(+|A)
1	0.95
2	0.93
3	0.87
4	0.85
5	0.85
6	0.85
7	0.76
8	0.53
9	0.43
10	0.25

输出是一个连续的概率值，且同我们仅仅关系“1”类别的概率

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P’	N’	All

1.4 ROC曲线

- 前面分类器性能评价的局限性：分类器预测结果为离散的1或者0
- 朴素贝叶斯输出？ $p(x|y)=?$
- 其他分类器输出
- 解决方法：连续的值离散化
- 导致的问题：离散阈值难以确定

输出是一个连续的概率值，且同我们仅仅关系“1”类别的概率

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P’	N’	All

Instance	P(+|A)
1	0.95
2	0.93
3	0.87
4	0.85
5	0.85
6	0.85
7	0.76
8	0.53
9	0.43
10	0.25

1.4 ROC曲线

- 接收者操作特征曲线（Receiver Operating Characteristic Curve，或者叫ROC曲线）是一种坐标图式的分析工具，用于
  - (1) 选择最佳的分类模型、舍弃次佳的模型。
  - (2) 在同一模型中设定最佳阈值。

- 给定一个二元分类模型和它的阈值，就能从所有样本的(阳性 / 阴性)真实值和预测值计算出一个 **(X=FPR, Y=TPR)** 坐标点。

A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P'	N'	All

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(TN + FP)$$

$$TNR = TN/(TN + FP)$$

$$FNR = FN/(TP + FN)$$

单选题 1分

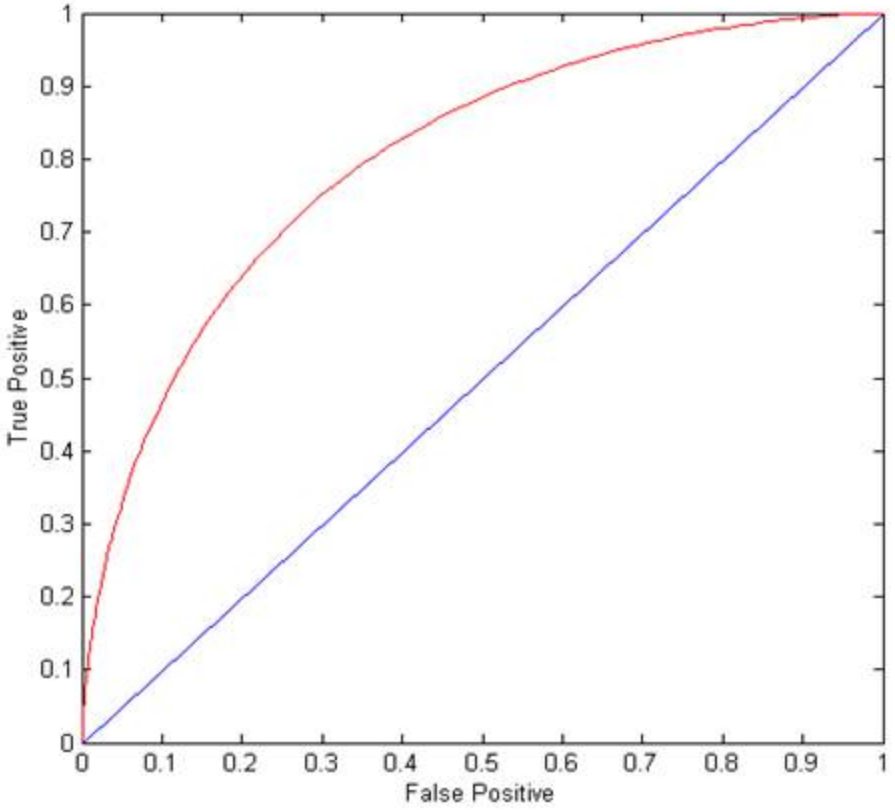
设置

(FPR, TPR):

- (0,0): 任何分类都是阴性
- (1,1): 任何分类都是【选择题】
- (0,1): 理想分类

$$TPR = TP/(TP + FN)$$

- 对角线: $$FPR = FP/(TN + FP)$$
  - 随机猜测结果
  - 对角线以下：
    - 预测结果与真实结果相反



A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P'	N'	All

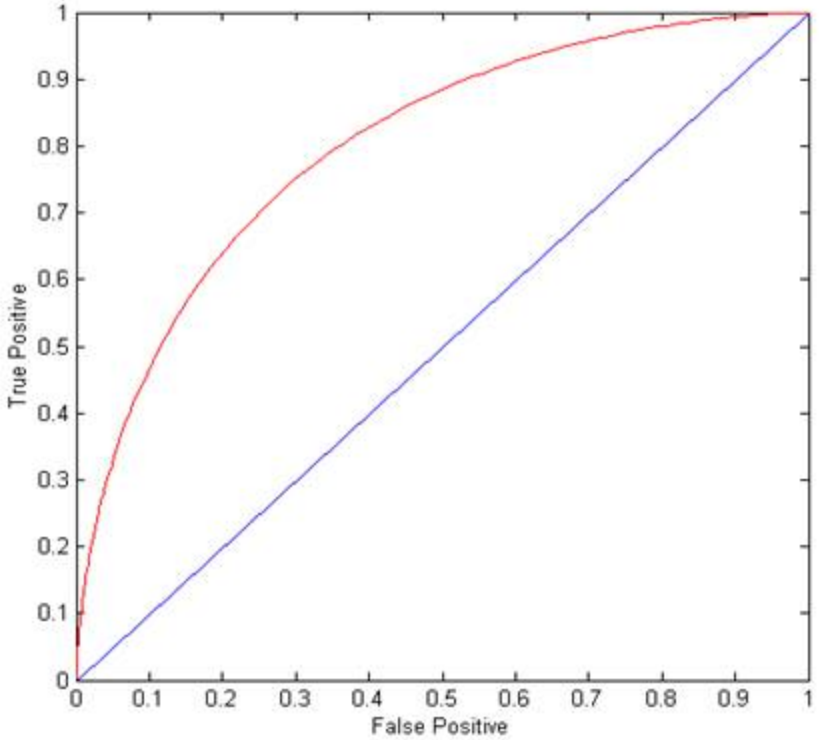
A 阴性

B 阳性

提交

设置

- ROC曲线下方的区域称为AUC，Area Under the ROC curve
  - Ideal:
    - Area = [填空1]
  - Random guess:
    - Area = [填空2]



A\P	C	¬C	
C	TP	FN	P
¬C	FP	TN	N
	P'	N'	All

$$TPR = TP/(TP + FN)$$

$$FPR = FP/(TN + FP)$$

正常使用填空题需3.0以上版本雨课堂

作答

1.4如何构建ROC曲线

- 首先利用分类器计算每个数据记录的后验概率P(+|A)
- 将这些数据记录对应的P(+|A)从高到低排列（如右表）：
  - 由低到高, 对于每个P(+|A)值（threshold，阈值），把对应的记录以及那些值**高于或等于阈值**指派为阳性类positive, 把那些值**低于阈值**指派为阴性类negative
  - 统计 TP, FP, TN, FN
  - 计算TPR = TP/(TP+FN)和

    FPR = FP/(FP + TN)
- 绘出诸点(FPR, TPR)并连接它们

Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	–
4	0.85	–
5	0.85	–
6	0.85	+
7	0.76	–
8	0.53	+
9	0.43	–
10	0.25	+

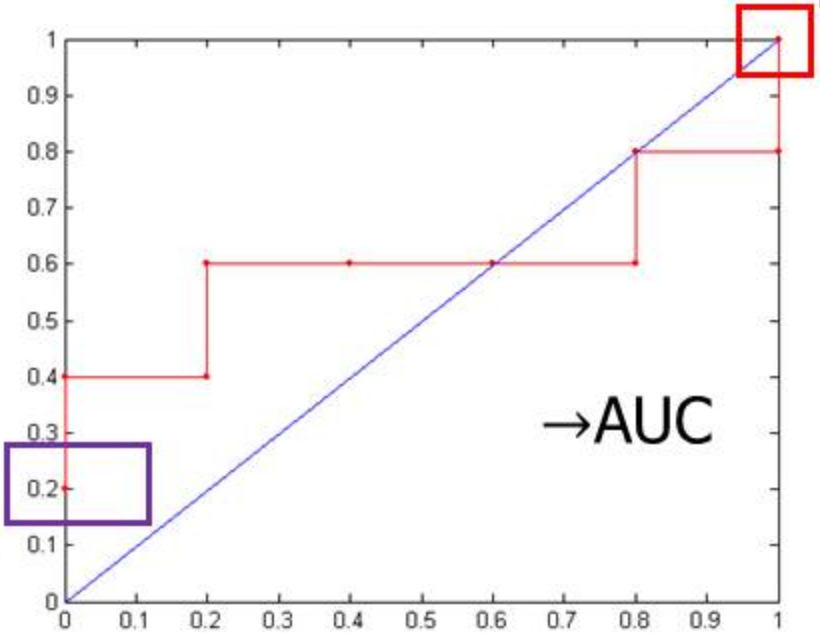
填空题 2分

设置

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	A	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	B	0

A= [填空1]

B= [填空2]



Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	–
4	0.85	–
5	0.85	–
6	0.85	+
7	0.76	–
8	0.53	+
9	0.43	–
10	0.25	+

作答

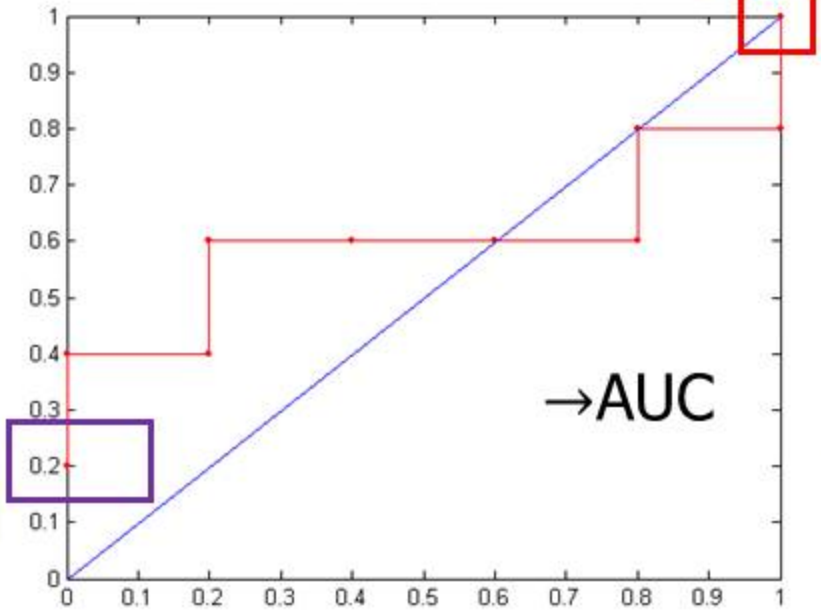
填空题 2分

设置

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	A	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	B	0	0

A= [填空1]

B= [填空2]



Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	–
4	0.85	–
5	0.85	–
6	0.85	+
7	0.76	–
8	0.53	+
9	0.43	–
10	0.25	+

作答

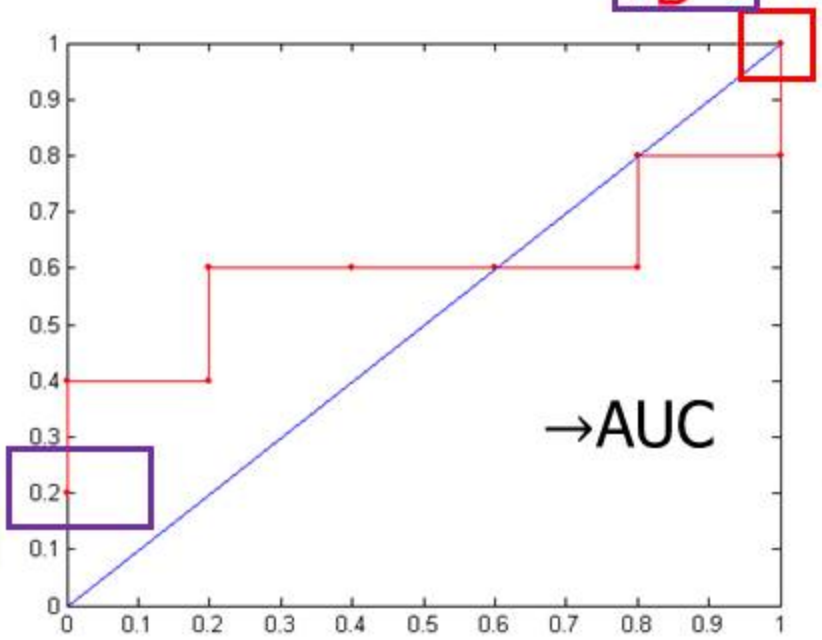
填空题 2分

设置

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	A	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	B	0	0	0

A= [填空1]

B= [填空2]

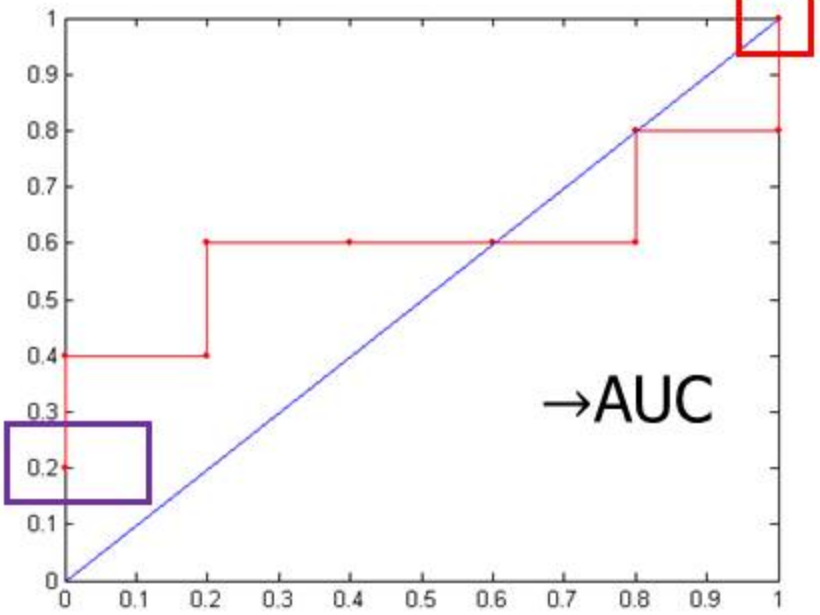


Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	–
4	0.85	–
5	0.85	–
6	0.85	+
7	0.76	–
8	0.53	+
9	0.43	–
10	0.25	+

作答

1.4如何构建ROC曲线

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

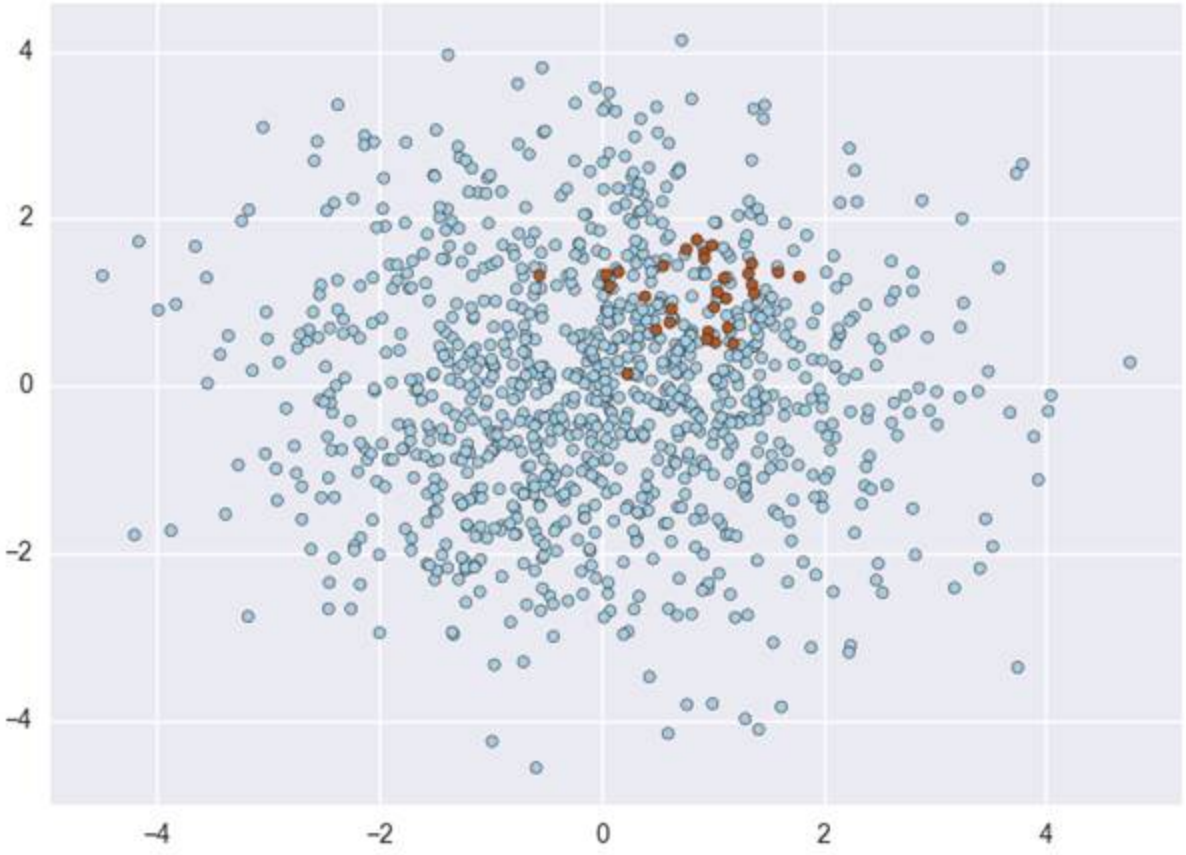


Instance	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	–
4	0.85	–
5	0.85	–
6	0.85	+
7	0.76	–
8	0.53	+
9	0.43	–
10	0.25	+

2不平滑分类

2 Imbalanced Data Mining

- 数据不平衡问题

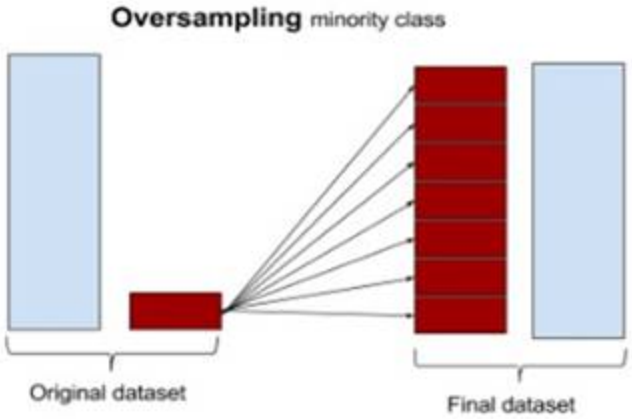


2.1基于抽样的方法

- 基于抽样的方法
  - 考虑一个包含100个正样本和1000个负样本的数据集
  - **Oversampling过采样**
    - 复制正样本, 直到训练集中正样本和负样本一样多
    - 可能导致模型过分拟合，因为一些噪声样本也可能被复制多次
  - **Undersampling欠采样**
    - 随机抽取100个负样本, 与所有的正样本一起形成训练集
    - 问题: 一些有用的负样本可能没有选出来用于训练, 因此导致一个不太优的模型
    - 解决问题的方法：多次执行不充分抽样, 并归纳类似于组合学习方法的多分类器
  - **Oversampling + Undersampling**

2.1基于抽样的方法

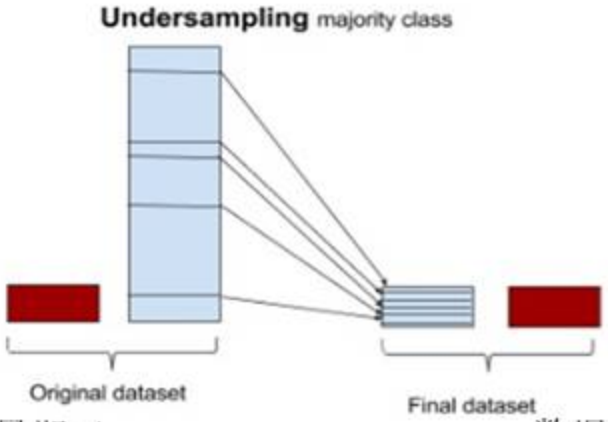
- 基于抽样的方法
  - 考虑一个包含100个正样本和1000个负样本的数据集
  - **Oversampling 过采样**
    - 复制正样本, 直到训练集中正样本和负样本一样多
    - 可能导致模型过分拟合，因为一些噪声样本也可能被复制多次



噪声样本也可能
被复制多次

2.1基于抽样的方法

- 基于抽样的方法
  - 考虑一个包含100个正样本和1000个负样本的数据集
  - **Oversampling过采样**
  - **Undersampling欠采样**
    - 随机抽取100个负样本,与所有的正样本一起形成训练集
    - 问题: 一些有用的负样本可能没有选出来用于训练, 因此导致一个不太优的模型
    - 解决问题的方法：多次执行不充分抽样,并归纳类似于组合学习方法的多分类器



有用的负样本可能没有选出来用于训练

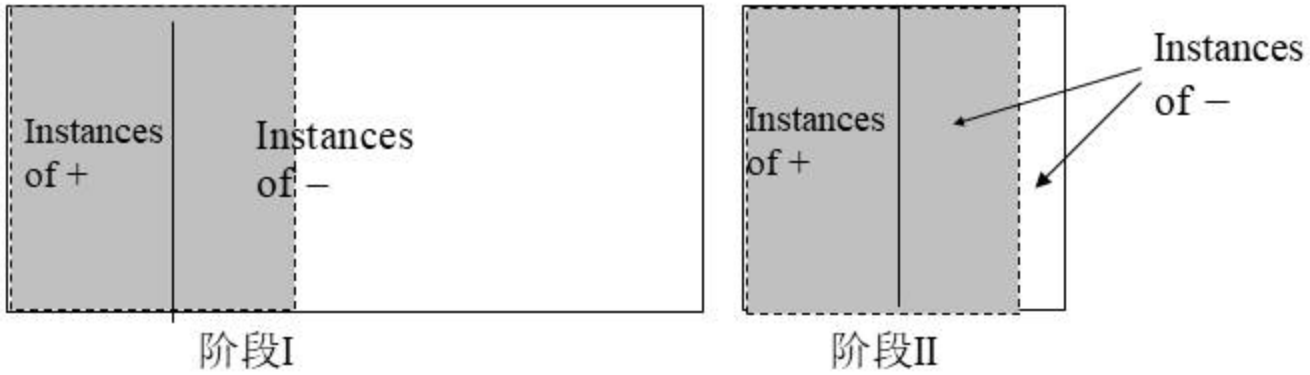
2.1基于抽样的方法

- 基于抽样的方法
  - 考虑一个包含100个正样本和1000个负样本的数据集
  - **Oversampling过采样**
    - 复制正样本, 直到训练集中正样本和负样本一样多
    - 可能导致模型过分拟合，因为一些噪声样本也可能被复制多次
  - **Undersampling欠采样**
    - 随机抽取100个负样本, 与所有的正样本一起形成训练集
    - 问题: 一些有用的负样本可能没有选出来用于训练, 因此导致一个不太优的模型
    - 解决问题的方法：多次执行不充分抽样, 并归纳类似于组合学习方法的多分类器
  - **Oversampling + Undersampling**

2.2两阶段学习

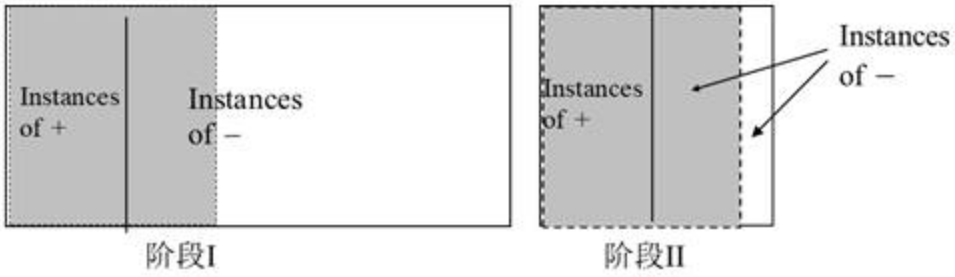
- 两阶段学习：PN-Rules
  - 是基于规则的分类
  - 学习分两个阶段，每个阶段学习一组规则

- 训练
  - 阶段I：学习一组规则，尽可能覆盖正类（少的那一类）
  - 阶段II：使用阶段I覆盖的正类和负类样本+部分其它负类样本，学习一组规则



2.2两阶段学习(续)

- 分类
  - 用第一组规则对x分类，如果分到负类，则x属于负类
  - 否则，用第二组规则确定x所属的类
- **R. Agarwal**, and M. V. Joshi. PNrule: A New Framework for Learning Classifier Models in Data Mining (A Case-Study in Network Intrusion Detection). In Proc. of the First SIAM Conference on Data Mining. Chicago, USA, April 2001



3过拟合和欠拟合

3.1模型过分拟合和拟合不足

- 分类模型的误差大致分为两种：
  - 训练误差：是在训练记录上误分类样本比例
  - 泛化误差：是模型在未知记录上的期望误差

- 一个好的分类模型不仅要能够很好的拟合训练数据，而且对未知样本也要能准确分类。

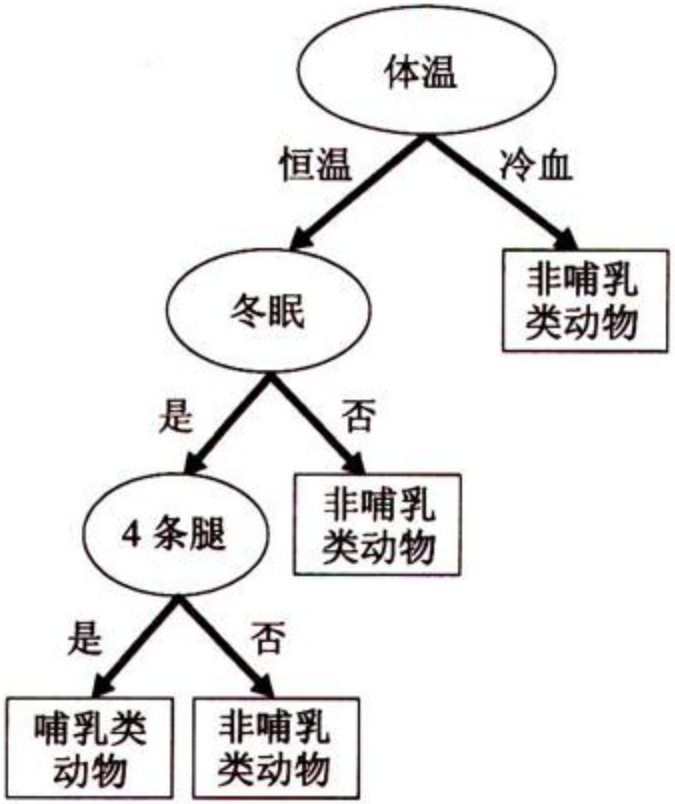
- 换句话说，一个好的分类模型必须具有低训练误差和低泛化误差。

- 当训练数据拟合太好的模型（**较低训练误差**），其**泛化误差**可能比**具有较高训练误差**的模型高，这种情况成为模型**过分拟合**。

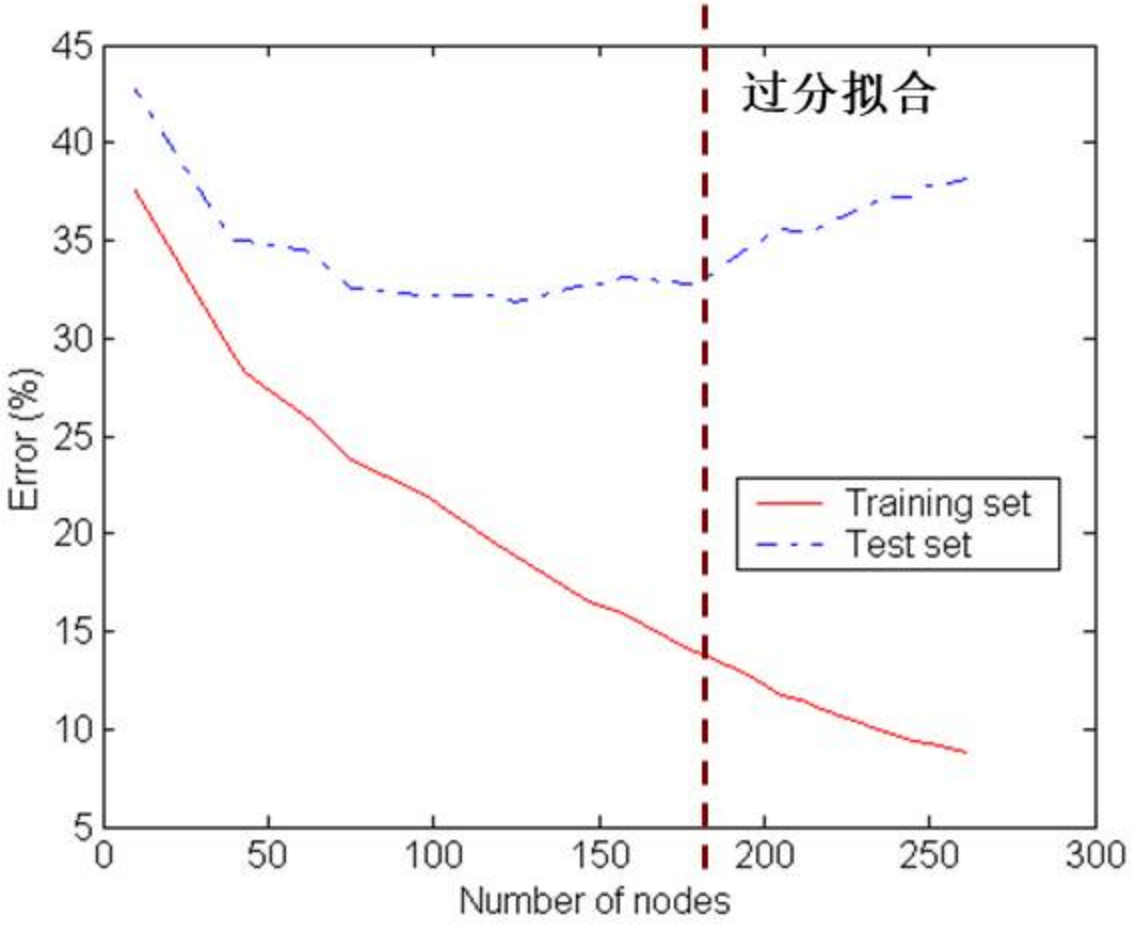
- 数据预处理→模型训练→模型调整→对新数据分类→模型评价

3.1模型过分拟合和拟合不足

- 以决策树算法为例
  - 当决策树很小时，训练和检验误差都很大，这种情况称为模型拟合不足。出现拟合不足的原因是模型尚未学习到数据的真实结构。
  - 随着决策树中结点数的增加，模型的训练误差和泛化误差都会随之下降。
  - 当树的规模变得太大时，即使训练误差还在继续降低，但是泛化误差开始增大，导致模型过分拟合。



3.1模型过分拟合和拟合不足



填空题 2分

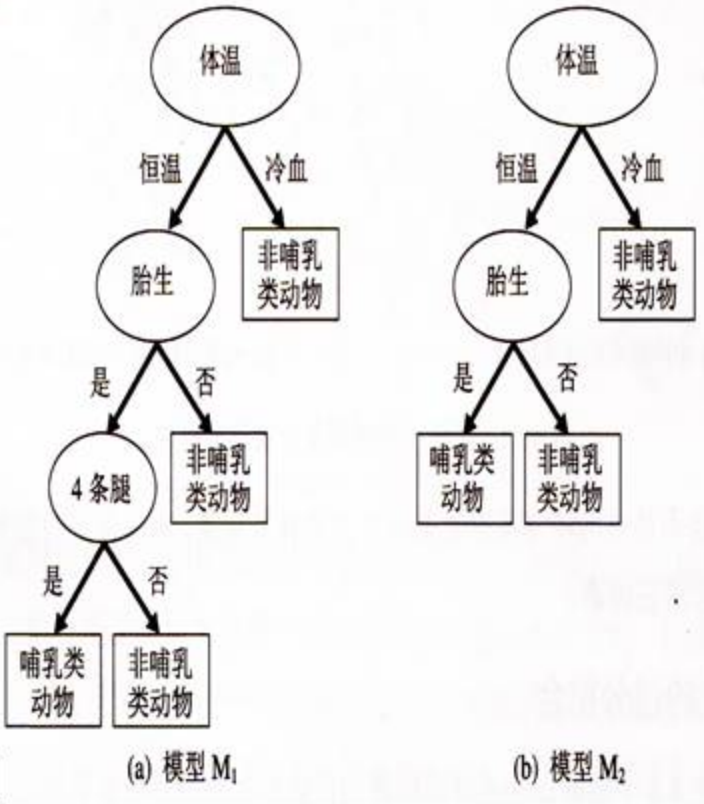
设置

表 4-3 哺乳类动物分类的训练数据集样本。打星号的类标号代表错误标记的记录

名称	体温	胎生	4 条腿	冬眠	类标号
豪猪	恒温	是	是	是	是
猫	恒温	是	是	否	是
蝙蝠	恒温	是	否	是	否*
鲸	恒温	是	否	否	否*
蝾螈	冷血	否	是	是	否
科莫多巨蜥	冷血	否	是	否	否
蟒蛇	冷血	否	否	是	否
鲑鱼	冷血	否	否	否	否
鹰	恒温	否	否	否	否
虹鳉	冷血	是	否	否	否

表 4-4 哺乳类动物分类的检验数据集样本

名称	体温	胎生	4 条腿	冬眠	类标号
人	恒温	是	否	否	是
鸽子	恒温	否	否	否	否
象	恒温	是	是	否	是
豹纹鲨	冷血	是	否	否	否
海龟	冷血	否	是	否	否
企鹅	冷血	否	否	否	否
鳗	冷血	否	否	否	否
海豚	恒温	是	否	否	是
针鼹	恒温	否	是	是	是
希拉毒蜥	冷血	否	是	是	否



决策树M1的训练误差为[填空1]，但它在检验数据上的误差达[填空2]

作答

填空题 2分

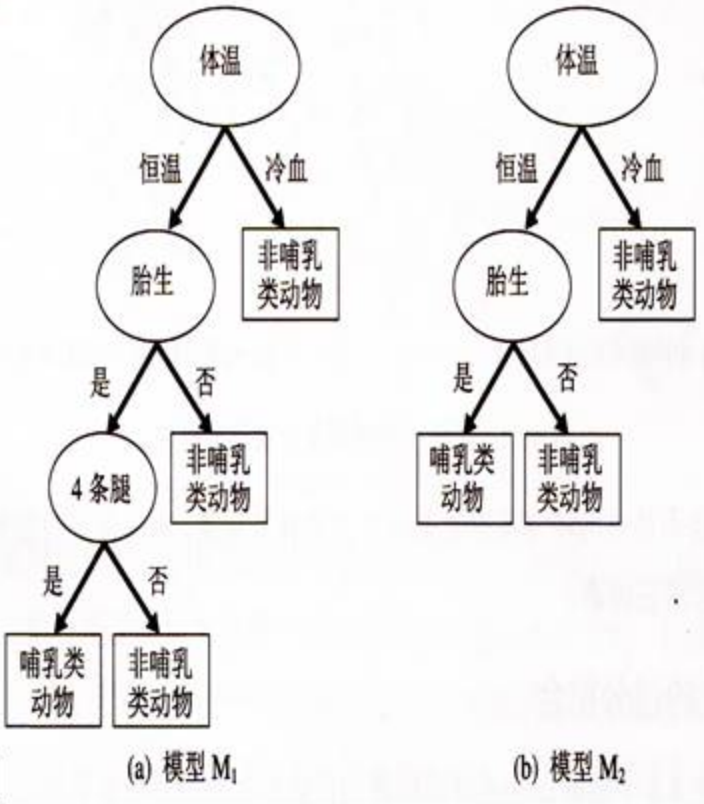
设置

表 4-3 哺乳类动物分类的训练数据集样本。打星号的类标号代表错误标记的记录

名称	体温	胎生	4 条腿	冬眠	类标号
豪猪	恒温	是	是	是	是
猫	恒温	是	是	否	是
蝙蝠	恒温	是	否	是	否*
鲸	恒温	是	否	否	否*
蝾螈	冷血	否	是	是	否
科莫多巨蜥	冷血	否	是	否	否
蟒蛇	冷血	否	否	是	否
鲑鱼	冷血	否	否	否	否
鹰	恒温	否	否	否	否
虹鳉	冷血	是	否	否	否

表 4-4 哺乳类动物分类的检验数据集样本

名称	体温	胎生	4 条腿	冬眠	类标号
人	恒温	是	否	否	是
鸽子	恒温	否	否	否	否
象	恒温	是	是	否	是
豹纹鲨	冷血	是	否	否	否
海龟	冷血	否	是	否	否
企鹅	冷血	否	否	否	否
鳗	冷血	否	否	否	否
海豚	恒温	是	否	否	是
针鼹	恒温	否	是	是	是
希拉毒蜥	冷血	否	是	是	否



决策树M2的训练误差为[填空1]，但它在检验数据上的误差达[填空2]

作答

导致过拟合的原因

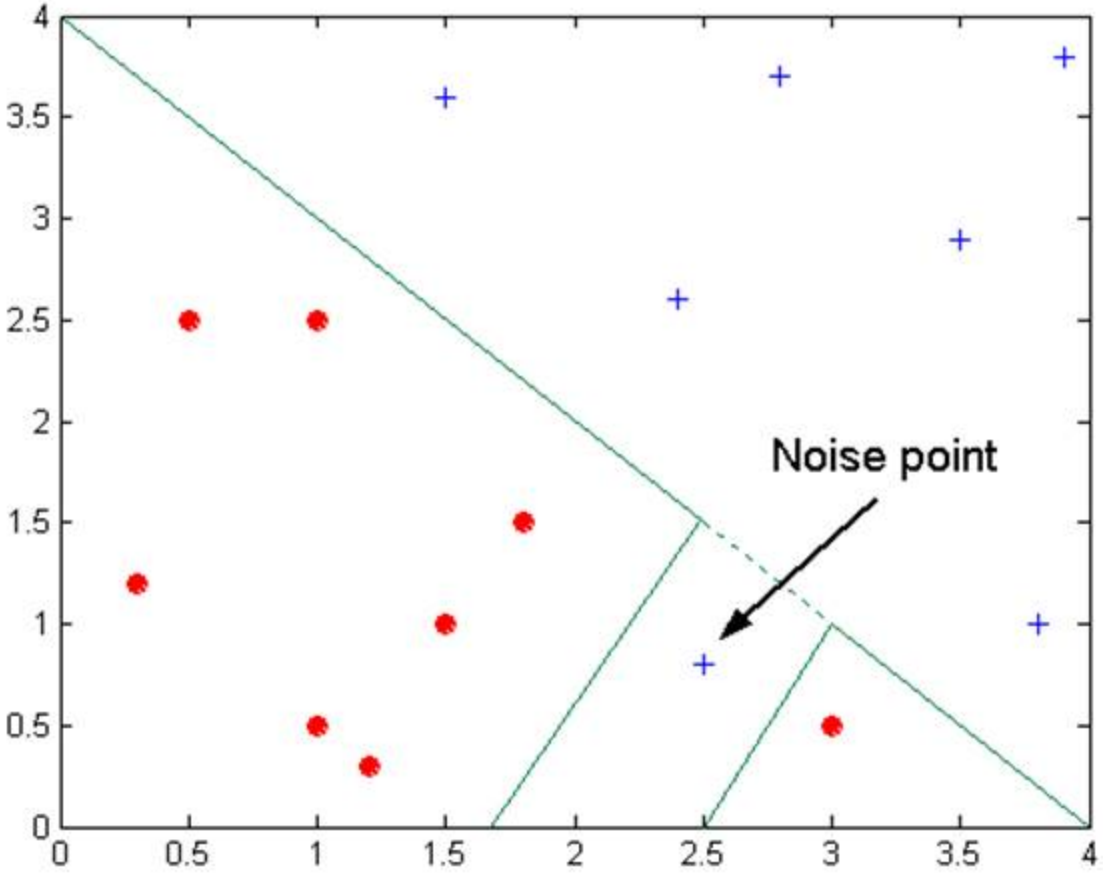
A 训练集规模太大

B 训练集中存在大量噪音数据

C 训练集规模太小，训练模型过于复杂

提交

3.2噪声导致的过分拟合



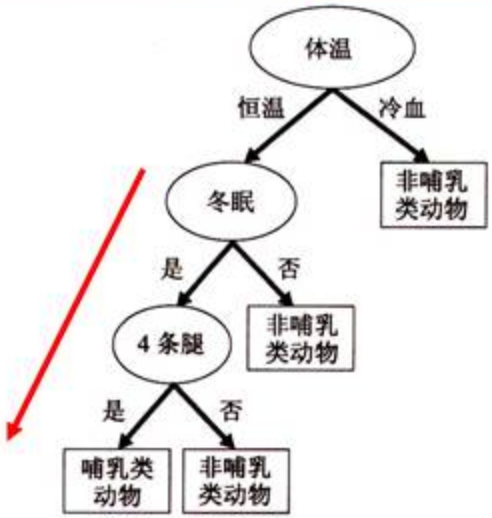
噪声导致决策边界的改变

3.3缺乏代表性样本导致的过分拟合

- 根据少量训练记录做出分类决策的模型也容易受过分拟合的影响。
- 由于训练数据缺乏具有代表性的样本，在没有多少训练记录的情况下，学习算法仍然细化模型就会产生过分拟合。

表 4-5　哺乳动物分类的训练集样本

名称	体温	胎生	4 条腿	冬眠	类标号
蝾螈	冷血	否	是	是	否
虹鳉	冷血	是	否	否	否
鹰	恒温	否	否	否	否
弱夜鹰	恒温	否	否	是	否
鸭嘴兽	恒温	否	是	是	是



训练集太少，模型太复杂

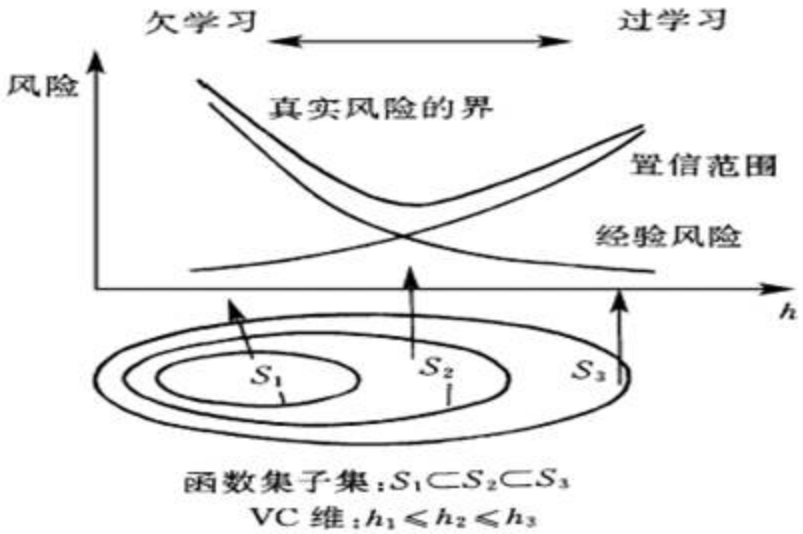
图 4-26　根据表 4-5 中的数据集建立的决策树

3.4减少泛化误差

- 过分拟合的主要原因一直是个争辩的话题，但数据挖掘研究界普遍认为模型的复杂度对模型的过分拟合有影响。

- 如何确定正确的模型复杂度？理想的复杂度是能产生最低泛化误差的模型的复杂度。

- 奥卡姆剃刀定律

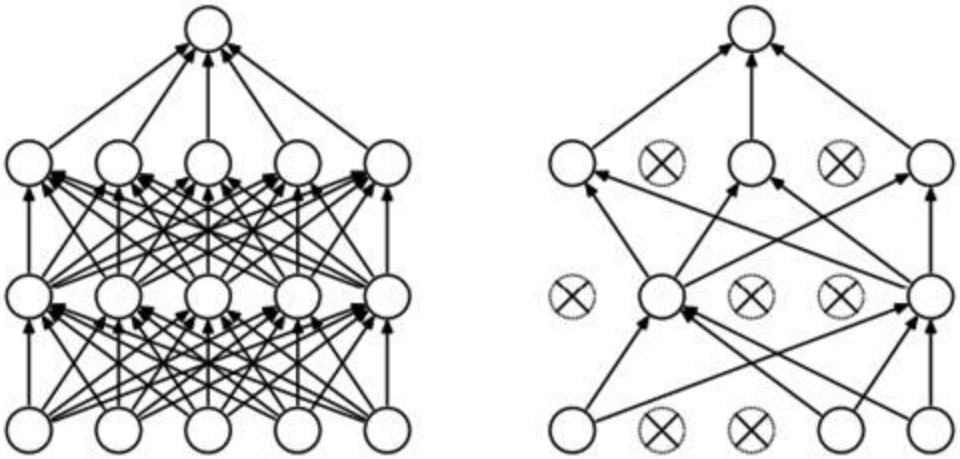


3.4奥卡姆剃刀(Occam's Razor)

- 奥卡姆剃刀（Occam‘s Razor），拉丁文为*lex parsimoniae*，意思是简约之法则。
- 是由14世纪逻辑学家、圣方济各会修士威廉奥卡姆William of Occam（约1287年至1347年）提出的一个解决问题的法则。
- 他在《箴言书注》第2卷15章说“**切勿浪费较多东西，去做：用较少的东西，同样可以做好的事情**”。
- 奥卡姆剃刀定律被广泛运用在多个学科的逻辑定律，它的简单表述：
  - **如无必要，勿增实体**
  - Entities should not be multiplied unnecessarily

3.4.1减少泛化误差

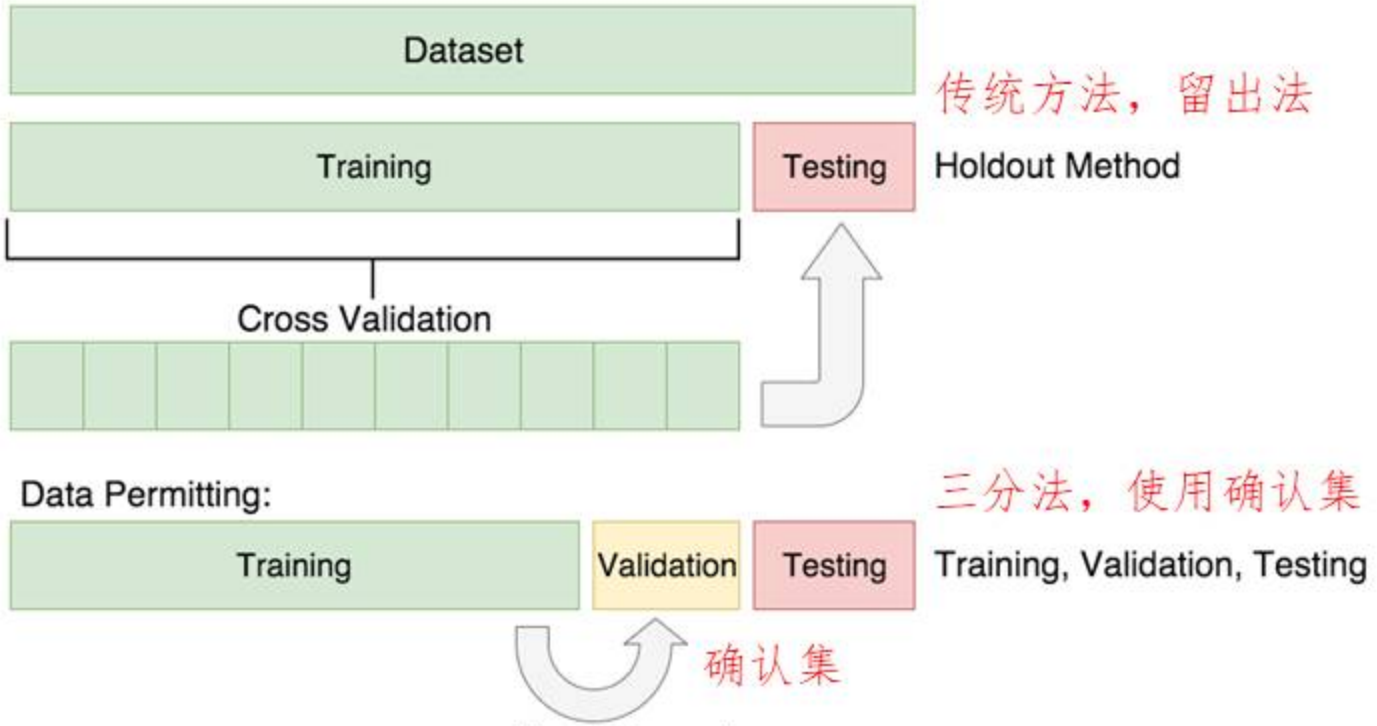
- 根据奥卡姆剃刀原则
  - 引入惩罚项，使较简单的模型比复杂的模型更可取
    - 引入正则项
    - 神经网络中，引入dropout机制



(a) Standard Neural Net　　(b) After applying dropout.

3.4.2减少泛化误差

- 使用确认集
  - 该方法中，不是用训练集估计泛化误差，而是把原始的训练数据集分为两个较小的子集，一个子集用于训练，而另一个称为确认集，用于估计泛化误差。
  - 该方法为评估模型在未知样本上的性能提供了较好办法。



多选题 1分

设置

下列说法正确的是

A 过拟合是由于训练集多，模型过于简单

B 过拟合是由于训练集少，模型过于复杂

C 欠拟合是由于训练集多，模型过于简单

D 欠拟合是由于训练集少，模型过于简单

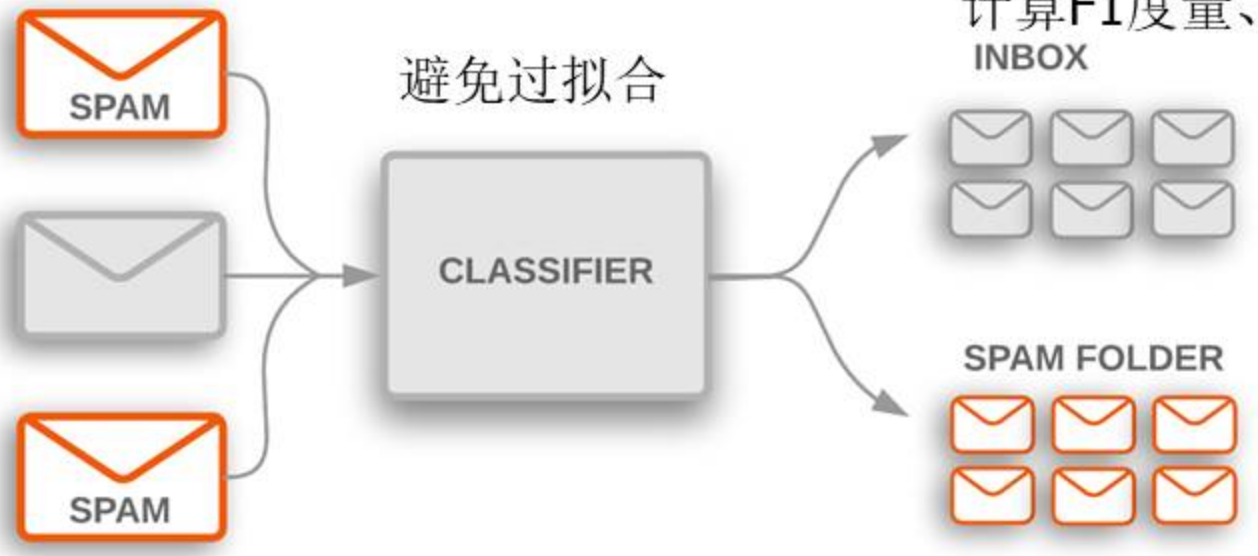
提交

总结

使用确认集，基于抽样的方法

统计查全率vs. 查准率
计算F1度量、ROC、AUC

避免过拟合



Any Questions？

谢谢！

数据挖掘竞赛案例2
<重复购买预测>

01 赛题介绍

02 数据描述

03 数据处理

04 特征提取

05 模型训练

06 模型结果

01 赛题介绍

商家有时会在特定日期（例如“Boxing-day”，“黑色星期五”或“双11”）进行大促销（例如折扣或现金券），以吸引大量新买家。许多吸引的买家都是一次性交易猎人，这些促销可能对销售产生很小的长期影响。为了缓解这个问题，商家必须确定谁可以转换为重复买家。通过瞄准这些潜力忠诚的客户，商家可以大大降低促销成本，提高投资回报率（ROI）。

题目提供了一套商家及其在“双11”日促销期间获得的相应新买家。任务是预测对于指定商家的新买家将来是否会成为忠实客户。即预测这些新买家在6个月内再次从同一商家购买商品的概率。一个包含大约20万用户的数据集用于训练，还有一个类似大小的数据集用于测试。

02 数据描述

数据格式

官方给了数据： data_format1

data_format1: user_log_format1, user_info_format1, test_format1, train_format1

用户行为日志：包含用户ID、商品ID、商品类别、商户ID、商品品牌、时间和用户行为类别7个特征。

用户信息：包含用户ID、用户年龄段和用户性别信息。

训练集和测试集：分别包含用户ID、商户ID和是否为重复买家标签，其中训练集标签为0-1，测试集标签为空，需要预测。

数据量

Name ▲	Type	Size	Value
data1	DataFrame	(54925330, 7)	Column names: user_id, item_id, cat_id, seller_id, brand_id, time_stam ...

03 数据处理

步骤2：数据清洗

进行brand_id缺失值(91015)填充，并使用pickle模块进行序列化，加快速度读写

步骤1：数据压缩

压缩csv中的数据，通过改变扫描每列的dtype，转换成适合的大小。

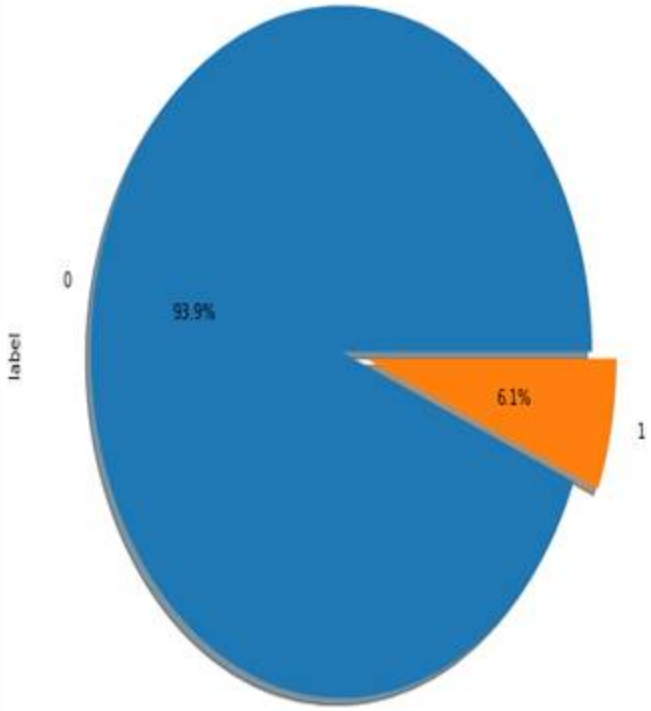
步骤3：数据可视化

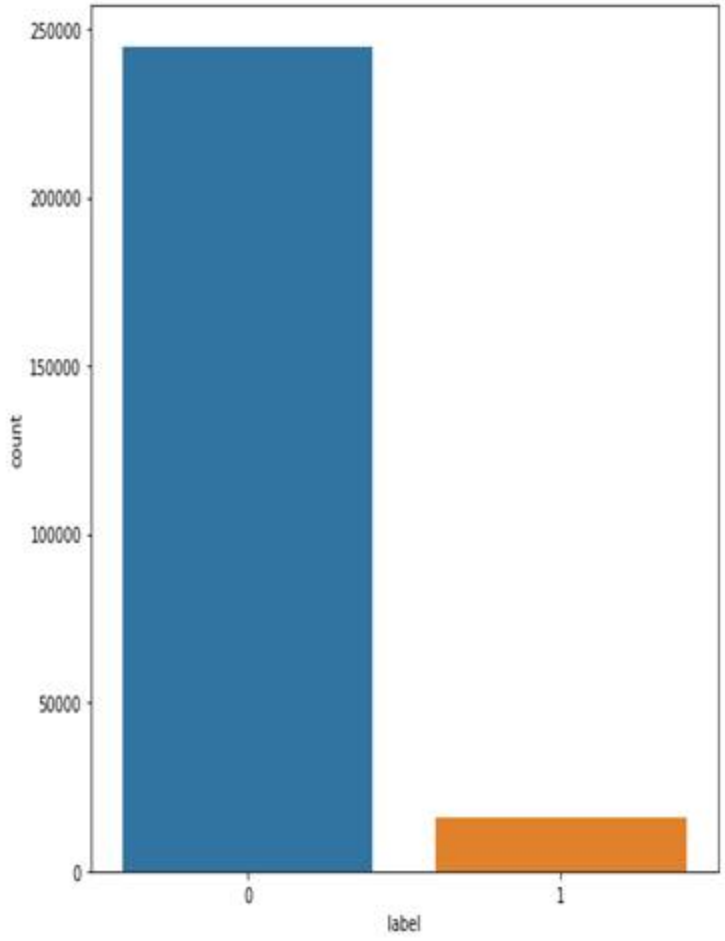
读取训练集，对正负样本、正负样本与性别的比例、正负样本与年龄段的比例进行可视化。

数据可视化

训练集正负样本可视化：

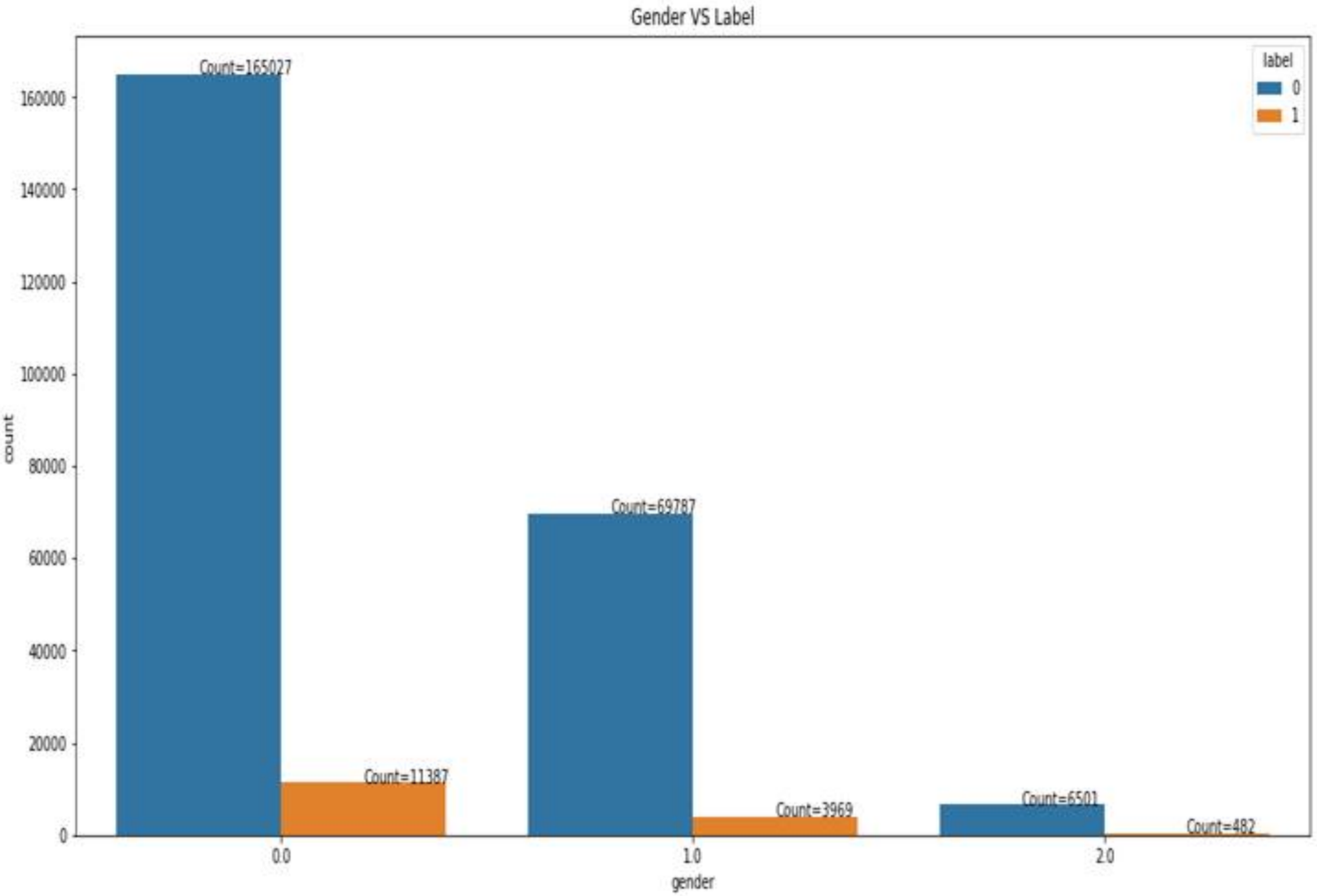
训练集中label取值范围 {0, 1}，1表示重复购买，0 表示非重复购买。





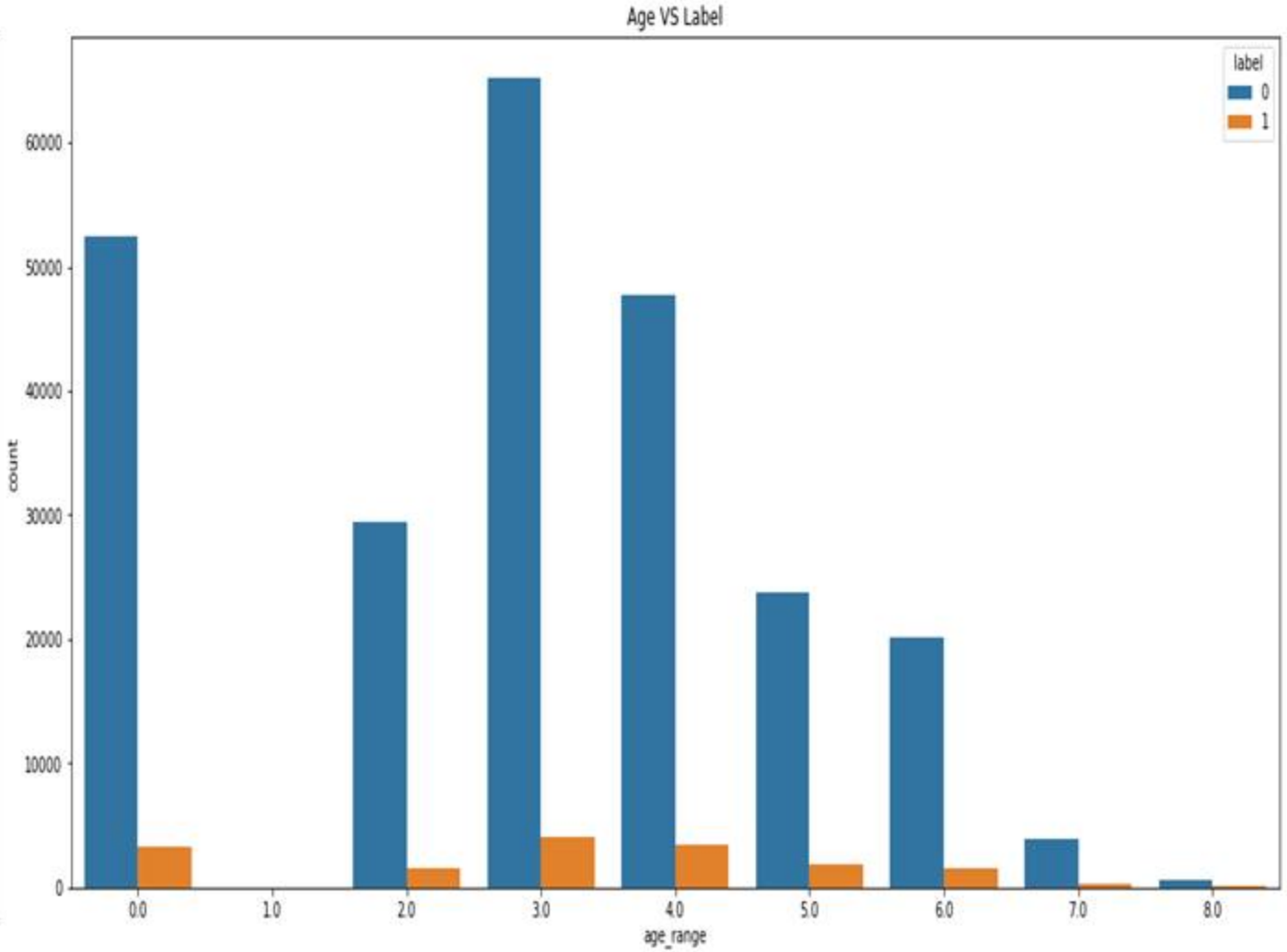
数据可视化

读取用户信息数据，并与训练集数据进行合并；
展示正负样本与用户性别的比例；
顾客性别：0 表示女性，1 表示男性，2 and NULL 表示未知.



数据可视化

展示正负样本与用户年龄段的比例；
顾客年龄范围：1 表示<18；2 表示[18,24]；3 表示[25,29]；4 表示[30,34]；5 表示[35,39]；6 表示[40,49]；7 and 8 表示 >= 50；0 and NULL 表示未知



04 特征提取

1. 基于用户的特征
   - 分析用户自身的偏好及购物习惯
2. 基于商户的特征
   - 分析商户在当前市场的受欢迎程度及商户自身对忠实用户的吸引力
3. 基于用户-商户的特征
   - 分析给定用户及商户之间所形成的关系，以及得到更好的预测效果

```
0   age_0.0 q
1   age_1.0 q
2   age_2.0 q
3   age_3.0 q
4   age_4.0 q
5   age_5.0 q
6   age_6.0 q
7   age_7.0 q
8   age_8.0 q
9   female  q
10  male    q
11  unknown q
12  userTotalAction_0   q
```

```
111 save_days   q
112 item_click_count    q
113 item_add_count  q
114 item_buy_count  q
115 item_save_count q
116 cat_click_count q
117 cat_add_count   q
118 cat_buy_count   q
119 cat_save_count  q
120 brand_click_count   q
121 brand_add_count q
122 brand_buy_count q
123 brand_save_count    q
```

基模型：

LGBM 、 XGBoost 、 MLP 、 GBDT 、 RandomForest

集成学习：

GBM

06 模型结果

	train	test	final
AUC	0.7112	0.6731	0.6775

06 模型结果

43	_sssssyy	浙江大学	0.681079	2018-01-10
44	大西瓜瓜	盒子科技	0.681011	2018-10-07
45	大厉	浙江大学	0.680769	2017-12-21
46	控几我寄几	University of Aberdeen	0.679506	2018-05-31
47	DeepDarkFantasy_j...	其它-上海科技大学	0.679450	2018-06-17
48	小七要读博	天津理工	0.678982	2018-05-31
49	凉口三三	重庆邮电大学	0.678950	2018-05-31
50	lccc0312	某厂	0.678352	2017-04-21
51	美帝掌握核心科技	电子科技大学	0.678300	2018-05-31
52	zweiHasen_rcababitt	其它-上海科技大学	0.678231	2018-06-17
53	downle	Downle	0.678102	2017-03-14
54	zweiHasen_meeto	其它-上海科技大学	0.677829	2018-06-17
55	Texas_2019	University of Toronto	0.677663	2018-06-08
56	丁兆云dm杨凯晶	国防科大	0.677507	2018-11-21