



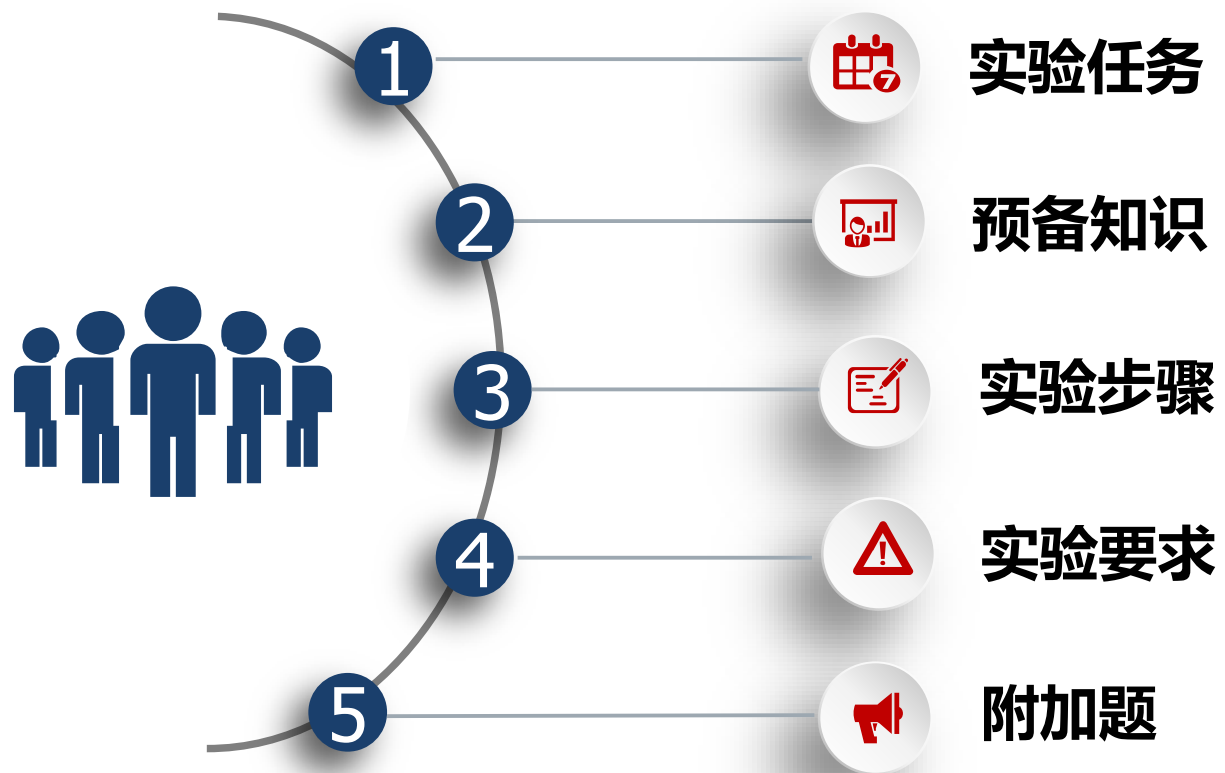
哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

实验三：构建决策树模型实现 银行借贷预测

《统计机器学习》

2024年春

目录



本学期实验总体安排



本学期实验课程共 10 个学时， 5 个实验项目， 总成绩为 30 分。

实验项目	一	二	三	四	五
学时	2	2	2	2	2
实验内容	Python基础实践	感知机模型	决策树模型	K近邻模型	支持向量机模型
分数	4	6	7	6	7
上课时间	第11周	第13周	第14周	第15周	第16周
检查方式	提交实验截图文档		提交实验报告、工程文件		

一、实验任务

- ◆ 银行借贷是基于分析历史按时还款、逾期或不还的用户群体的各自特征建立模型，未来借款用户只要符合符合借款要求，就给予借贷，如果不符合，则拒绝。
- ◆ **任务一**：请使用Python自编程，创建一个决策树模型，进行银行贷款预测。
- ◆ **任务二**：调用Sklearn库分类器，创建一个决策树模型，进行银行贷款预测。

◆ 数据集

 银行借贷数据集test
 银行借贷数据集train

- ① name_id: 姓名id
- ② profession: 职业，1-企业工作者，2-个体经营户，3-自由工作者，4-事业单位，5-体力劳动者
- ③ education: 教育程度，1-博士及以上，2-硕士，3-本科，4-专科，5-高中及以下
- ④ house_loan: 是否有房贷，1-有，0-没有
- ⑤ car_loan: 是否有车贷，1-有，0-没有
- ⑥ married: 是否结婚，1-是，0-否
- ⑦ child: 是否有小孩，1-有，0-没有
- ⑧ revenue: 月收入
- ⑨ **approve**: 是否予以贷款，1-贷款，0-不贷款

nameid	profession	education	house_loan	car_loan	married	child	revenue	approve
1	5	1	0	0	1	1	8204	1
2	3	1	1	1	0	0	5674	0
3	2	3	1	0	1	0	10634	1
4	2	2	0	0	0	0	43551	1
5	4	2	0	1	0	1	14065	0

二、预备知识1



数据集划分



银行借贷数据集test
银行借贷数据集train

- ◆ 训练集 (Training Dataset) 是用来训练模型使用的。
- ◆ 验证集 (Validation Dataset) 来看看模型在新数据 (验证集和测试集是不同的数据) 上的表现如何。
- ◆ 测试集 (Test Dataset) 来做最终的评估。

说明:

- 1、验证集不像训练集和测试集，它是**非必需的**。如果不需要调整**超参数**，就可以不使用验证集，直接用测试集来评估效果。
- 2、验证集评估出来的效果并非模型的最终效果，主要是用来调整超参数的，模型最终效果以测试集的评估结果为准。



二、预备知识1

数据集划分

◆ **基本准则：**保持训练集、验证集、测试集之间的**互斥性**。

◆ **参考原则：**

- 1、对于小规模样本集（几万量级），常用的分配比例是 **60%** 训练集、**20%** 验证集、**20%** 测试集。
- 2、对于大规模样本集（百万级以上），只要验证集和测试集的数量足够即可，例如有 100w 条数据，那么**留 1w 验证集，1w 测试集**即可。
1000w 的数据，同样留 1w 验证集和 1w 测试集。
- 3、超参数越少，或者超参数很容易调整，那么可以**减少验证集的比例**，更多的分配给训练集。

 银行借贷数据集test
 银行借贷数据集train



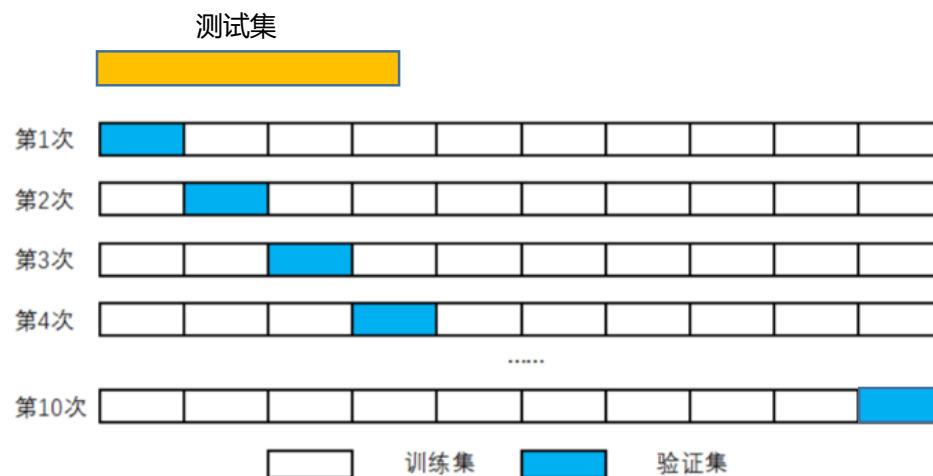
二、预备知识1

数据划分方法

◆ **划分方法：K折交叉验证**（一种动态验证的方式，这种方式可以降低数据划分带来的影响）

以10折交叉验证为例，具体步骤如下：

- 1、将数据集分为训练集和测试集，**将测试集放在一边**；
- 2、将训练集平均分成不相交的10个子集；
- 3、每一次挑选其中的1份作为验证集，其余的9份作为训练集进行模型训练，得到模型以及评价指标；
- 4、重复第3步10次，通过 10 次训练后，得到了 10个不同的模型；
- 5、将10个模型的评价指标取平均值，作为交叉验证的评估指标；
- 6、使用不同的超参数，重复以上2-5步，根据最好的交叉验证评估指标，挑选出最优的超参数；
- 7、使用最优的超参数，将数据全部作为训练集重新训练模型；
- 8、最后使用测试集测试**评估模型**，计算**评价指标**。



二、预备知识1

数据集划分方法

◆ **实现方法：** 比如 sklearn中的model_selection.KFold函数，格式如下：

```
sklearn.model_selection.KFold(n_splits=3, shuffle=False, random_state=None)
```

参数说明	含义
n_splits	分为几折交叉验证
shuffle	在每次划分时，是否进行洗牌。 若为Falses时，其效果等同于random_state等于整数，每次划分的结果相同； 若为True时，每次划分的结果都不一样，表示经过洗牌，随机取样的
random_state	随机种子数（设置了这个参数之后，每次生成的结果是一样的，而且设置了random_state之后就没必要设置shuffle)

代码示例：

```
# 导入包
from sklearn.model_selection import KFold
import numpy as np
# 构建数据集
X = np.arange(24).reshape(12,2)
Y = np.arange(12).reshape(12,1)

#调用k折交叉验证方法
kf = KFold(n_splits=3,shuffle=False)
for train_index,valid_index in kf.split(X):
    print("TRAIN:", train_index, "VALID:", valid_index)
    X_train, X_valid = X[train_index], X[valid_index]
    Y_train, Y_valid = Y[train_index], Y[valid_index]
```

运行结果：

```
TRAIN: [ 4  5  6  7  8  9 10 11] VALID: [0 1 2 3]
TRAIN: [ 0  1  2  3  8  9 10 11] VALID: [4 5 6 7]
TRAIN: [0 1 2 3 4 5 6 7] VALID: [ 8  9 10 11]
```



二、预备知识2



评价模型

在二分类任务中，各指标的计算基础都来自于对正负样本的分类结果，用混淆矩阵表示为：

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

◆ **精确率**：分类正确的正样本个数占分类器判定为正样本的样本个数的比例。

分类正确的正样本个数：即真正例(TP)。

分类器判定为正样本的个数：包括真正例(TP)和假正例(FP)

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

◆ **召回率**：分类正确的正样本个数占真正的正样本个数的比例。

分类正确的正样本个数：即真正例(TP)。

真正的正样本个数：包括真正例(TP)和假负例(FN)

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

◆ **F1-score**：精确率和召回率的调和均值。

$$F1 = \frac{2TP}{2TP + FP + FN}$$

二、预备知识3

欠拟合、适度拟合、过拟合

◆ 欠拟合：

定义：训练集和测试集上的准确率都不高，且相差不大，这种情况称为欠拟合（Under-Fitting）。如：一个为80%，另一个为82%。

解决办法：添加其他特征项，模型出现欠拟合的时候是因为特征项不够导致的，可以添加其他特征项来很好地解决。

◆ 适度拟合：

训练集和测试集的准确率都很高，且相差不大，这种情况称为适度拟合，**这是我们想要的结果**。如：一个为99%，另一个为98%。

◆ 过拟合：

定义：训练集准确率 远远大于 测试集准确率，这种情况称为过拟合（Over-Fitting）。如：一个为99%，另一个为88%。

解决办法：正则化、随机失活、逐层归一化、提前终止、Bagging

更多学习 https://blog.csdn.net/weixin_39852647/article/details/111095814

二、预备知识5



ID3 算法

输入：训练数据集 D ，特征集 A ，阈值 ϵ

输出：决策树 T

Step1: 若 D 中所有实例属于同一类 C_K ，则 T 为单结点树，并将类 C_K 作为该节点的类标记，返回 T ；

Step2: 若 $A=\emptyset$ ，则 T 为单结点树，并将 D 中实例数最大的类 C_K 作为该节点的类标记，返回 T ；


Step3: 否则，计算 A 中每个特征对 D 的 **信息增益**，选择 **信息增益** 最大的特征；

Step4: 如果 A_g 的 **信息增益** 小于阈值 ϵ ，则 T 为单节点树，并将 D 中实例数最大的类 C_K 作为该节点的类标记，返回 T

Step5: 否则，对 A_g 的每一种可能值 a_i ，依 $A_g=a_i$ 将 D 分割为若干非空子集 D_i ，将 D_i 中实例数最大的类作为标记，构建子结点，由结点及其子树构成树 T ，返回 T ；

Step6: 对第 i 个子节点，以 D_i 为训练集，以 $A - \{A_g\}$ 为特征集合，递归调用**Step1~step5**，得到子树 T_i ，返回 T_i 。

以“课本样例数据”为例，老师已给出示例代码，同学们请自行学习

 lab3-示例代码.ipynb

三、实验步骤

◆ 实验步骤

1、准备数据

- ✓读取数据，提取特征；
- ✓将数据分割为训练集和验证集

2、配置模型

3、训练模型


4、预测模型


5、评估模型

- ✓计算模型在测试集上的精确率、召回率和F1值

6、绘出决策树（只对调用Sklearn库有要求）

四、实验要求及注意事项

 银行借贷数据集test

 银行借贷数据集train

- ◆ **任务一：**使用**Python自编程**，请基于老师给的示例代码，编写**C4.5或CART算法**，来构造决策树模型。（或不参考示例代码，独立完成也行）

实验要求：自编程定义核心算法（重要代码处要有注释），以及自定义精确率P、召回率R和F1值指标来评价模型。

- ◆ **任务二：**使用**Sklearn库**完成决策树模型预测银行借贷与否。

实验要求：要用交叉验证思想，调参过程，评价指标，绘制出决策树。

四、实验要求及注意事项

- 1、数据集中nameid列要去除，如；

```
df= df.drop(['nameid'], axis=1)
```

- 2、数据集中revenue列数据要进行离散化，如；

```
re = [0,10000,20000,30000,40000,50000]  
df['revenue']=pd.cut(df['revenue'],re,labels=False)
```

- 3、计算信息增益比时，注意分母不能为0

- 4、绘制决策树时，如果遇到：

InvocationException: Program terminated with status:

1. stderr follows: Format: "png" not recognized. Use one of:

解决：可用管理员身份运行cmd，执行 **dot -c**

```
Microsoft Windows [版本 10.0.18363.418]  
(c) 2019 Microsoft Corporation。保留所有权利。  
  
C:\Users\lenovo>cd C:\Program Files\Graphviz 2.44.1\bin  
C:\Program Files\Graphviz 2.44.1\bin>dot -c  
C:\Program Files\Graphviz 2.44.1\bin>
```

五、提交方式

实验报告提交至平台 <http://labgrader.hitsz.edu.cn:8000/#/courses>

注意：

- 1、用户名、密码默认均为学号（若之前有修改过密码的，请用新密码登陆）；
- 2、请提交到相应的条目「2024春-统计机器学习-数学1&2」课程 - 实验三；
- 3、提交截止时间：下周二 6月11号 晚24点；
- 4、文件夹&压缩包命名要求：学号_姓名_统计机器学习实验三
- 5、提交内容：实验报告(.pdf文件)+代码(.py/ipynb文件)，一起打包为zip格式压缩包。

其他：

- 1) 数学1&2班 作业提交至课程「2024春-统计机器学习-数学1&2」
- 2) 数学3&4班 作业提交至课程「2024春-统计机器学习-数学3&4」
- 3) 计算机/通信/机械/自动化/光电/电气等专业 作业提交至课程「2024春-统计机器学习-综合班」
- 4) 每位同学都只会显示一个统计机器学习课程的，对上实验几提交即可。

统计机器学习实验

同学们，请开始实验吧！