



# 统计机器学习实验

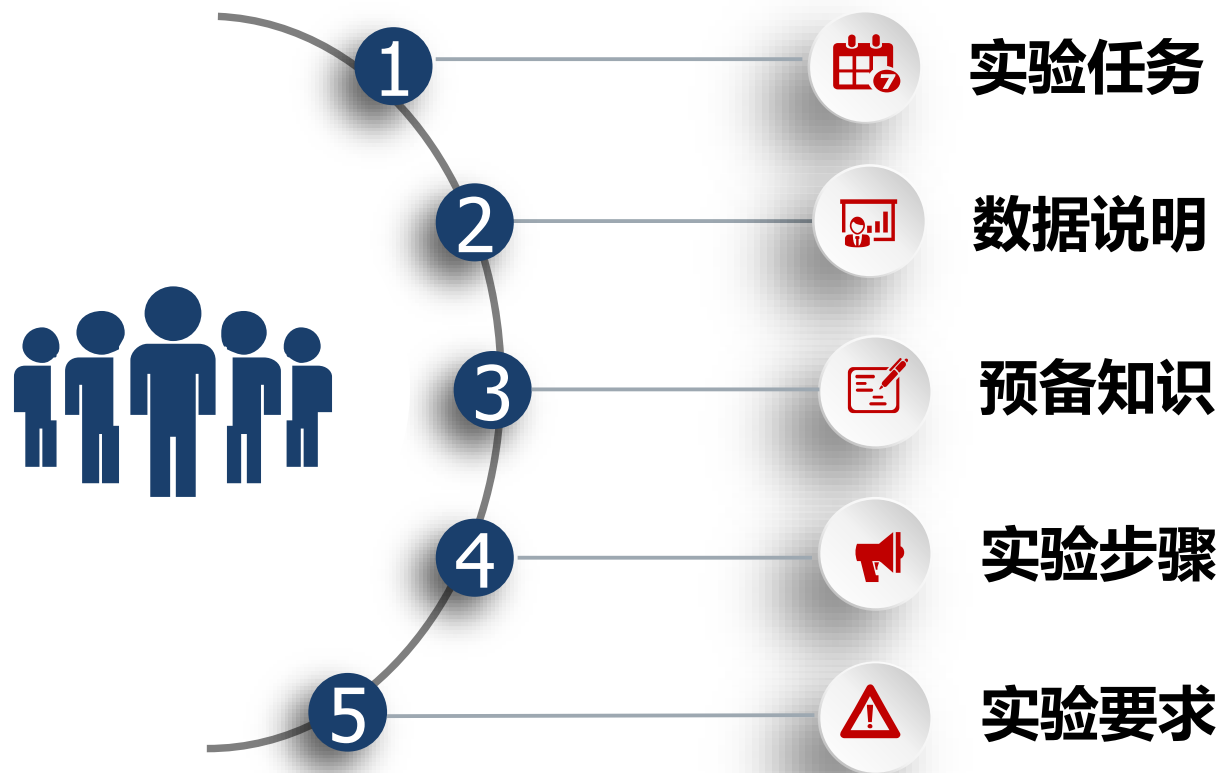
---

## 实验二：构建决策树模型实现 银行借贷预测

主讲教师：严资林

实验教师：匡慈维

# 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 20 分。

实验项目	一	二	三	四	五
学时	2	2	2	2	2
实验内容	感知机模型	决策树模型	K近邻模型	支持向量机模型	聚类模型
分数	3	4	4	4	5
上课时间 (地点)	第11周 周四 (T2102)	第12周 周六 (T2102)	第14周 周四 (T2102)	第16周 周二 (T2102)	第17周 周四 (T2102)
检查方式	提交实验截图文档	提交实验报告、工程文件			

5-6节 3&4班； 7-8节 1&2班

线上腾讯会议：848-8762-6539

# 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。
- ◆ 本地实验将根据自建的一份包含借款人信息及银行是否借贷的数据集，**创建一个决策树模型，并进行预测。**
- ◆ **任务：**使用Python自编程实现银行贷款预测；
- ◆ **附加题：**调用Sklearn库实现银行借贷预测（**选做**）。

注：附加内容有20%的加分，但总分不超过该次实验满分。

# 数据说明

## ◆ 数据集

 银行信贷数据集

- 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, 2-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

# 预备知识

## 数据集划分



- ◆ 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。
- 3、**这次实验不需要验证集。**

# 预备知识

## 评分模型

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。

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

# 预备知识

---

## ❖ 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](https://blog.csdn.net/weixin_39852647/article/details/111095814)



# 预备知识

## ❖ 决策树构建过程



- ◆ 特征选择表示从众多的特征中选择一个特征作为当前节点分裂的标准，如何选择特征有不同的量化评估方法，从而衍生出不同的决策树，如ID3、C4.5、CART。
- ◆ 根据选择的特征评估标准，从上至下递归地生成子节点，直到数据集不可分则停止决策树停止生长。
- ◆ 决策树容易过拟合，一般需要剪枝来缩小树结构规模、缓解过拟合。

# 预备知识

## ID3 / C4.5算法

输入：训练数据集 $D$ ，特征集 $A$ ，阈值 $\epsilon$

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 / 信息增益比**，选择 **信息增益 / 信息增益比** 最大的特征 ；

**Step4:** 如果 $A_g$ 的 **信息增益 / 信息增益比** 小于阈值 $\epsilon$ ，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 $T_i$ 。

# 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测试集

### 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）

# 实验要求

---

- ◆ **任务：**使用Python自编程，实验要求选用一种合适的算法，来构造决策树模型，结合精确率P、召回率R以及F1值来评价模型；
- ◆ **附加题：**使用Sklearn库编程完成决策树模型预测银行借贷与否。实验要求要有调参过程，要评价模型，绘制出决策树。

# 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>
```

# 提交方式

---

实验报告提交至平台 <http://grader.tery.top: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「2022春统计机器学习」课程 - 实验二；
- 3、提交截止时间：下周六 晚24点前；
- 4、文件夹&压缩包命名要求：学号\_姓名\_统计机器学习实验二
- 5、提交内容：实验报告(.pdf文件)+代码(.py文件)，一起打包为zip格式压缩包。

# 统计机器学习实验

---

同学们，请开始实验吧！