实验四：构建分类模型实现银行客户流失预测

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	5	4
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

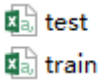
- 客户对于银行而言是重要的资产，对银行的收益以及市场占有率起着决定性作用。但是银行每年都要面对严重的客户流失问题，相较留住一个客户，获取一个新客户所需的成本往往是其数倍。因此分析出一个客户是否可能是潜在的易流失客户对于银行而言具有极大价值。
- 大多数银行对于客户流失问题关注度很高，但研究相对较少，目前只有少部分银行开始对真实案例进行建模分析。通过研究客户的历史行为来捕捉流失客户的特点，分析客户流失原因，从而可以在客户真正流失之前做出相应的营销干预，对客户进行挽留。

- **任务一**：对给出的数据集，**利用Sklearn库建立SVM模型**，对银行客户流失做预测。
- **任务二**：对给出的数据集，构建**其他分类模型（任选两种）**，对银行客户流失做预测。

数据说明

◆ 数据集

- 包含**训练集train**（共9000条数据），**测试集test**（1000条数据）



- 每一条数据由12个特征值，1个目标值（Exited）组成。
- 每个特征值的含义见文件 数据说明

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
2	1	15634602	Hargrave	619	France	Female	42	2	1	1	1	101348.88	1
3	2	15647311	Hill	608	Spain	Female	41	1	1	0	1	112542.58	0
4	3	15619304	Onio	502	France	Female	42	8	3	1	0	113931.57	1
5	4	15701354	Boni	699	France	Female	39	1	2	0	0	93826.63	0
6	5	15737888	Mitchell	850	Spain	Female	43	2	1	1	1	79084.1	0
7	6	15574012	Chu	645	Spain	Male	44	8	2	1	0	149756.71	1
8	7	15592531	Bartlett	822	France	Male	50	7	2	1	1	10062.8	0
9	8	15656148	Obinna	376	Germany	Female	29	4	4	1	0	119346.88	1
10	9	15792365	He	501	France	Male	44	4	2	0	1	74940.5	0
11	10	15592389	H?	684	France	Male	27	2	1	1	1	71725.73	0
12	11	15767821	Bearce	528	France	Male	31	6	2	0	0	80181.12	0
13	12	15737173	Andrews	497	Spain	Male	24	3	2	1	0	76390.01	0
14	13	15632264	Kay	476	France	Female	34	10	2	1	0	26260.98	0
15	14	15691483	Chin	549	France	Female	25	5	2	0	0	190857.79	0
16	15	15600882	Scott	635	Spain	Female	35	7	2	1	1	65951.65	0

预备知识

1、数据预处理

- 通常获得的数据集会存在冗余属性、噪音或非数值类属性等，无法直接使用，因此需要预先对于数据进行处理加工，得到较高质量的数据集后再将对其训练。
- 常见的数据预处理的方法有**数据清理、数据集成、数据变换以及数据规约**等等，如右图所示，详细内容请自行找资料学习。
- 对本次实验任务已给出数据预处理的代码，同学们能读懂加注释即可。

预备知识

2、Sklearn库内SVM分类器的参数详解

```
from sklearn.svm import SVC

svm_classifier = SVC(C=1.0, kernel=‘rbf’,\
decision_function_shape='ovo', gamma=0.01)
svm_classifier.fit(X_train, Y_train)
print(“准确率:", svm_classifier.score(X_test, Y_test))
```

误差项惩罚系数

C为误差项的惩罚系数

（1）C越大即对分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低；

（2）float参数，默认为1。

kernel

表示采用的核函数类型，可选的参数有：

‘linear’：线性核函数

‘rbf’：径向基核函数/高斯核函数

‘sigmoid’：sigmoid核函数等

‘poly’：多项式核函数

决策函数

decision_function_shape

表示决策函数，可选值：

ovo：用于二分类

ovr：用于多分类

gamma

（1）float参数，默认为'auto'；

（2）'rbf'，'poly'和'sigmoid'的核系数。当前默认值为'auto'，它使用1 / n_features

更多参数详解见：https://www.cnblogs.com/solong1989/p/9620170.html

预备知识

3、网格搜索

◆ **目的：**是为了让模型准确性更高。

◆ **基本思想：**通常情况下，有很多参数是需要手动指定的（如K近邻中的k值， SVM算法中的C以及gamma值等），这种叫超参数。但是手动过程繁杂，所以需要对模型预设几种超参数组合。每组超参数都采用交叉验证来进行评估，最后选出最优参数组合建立模型。

预备知识

4、Sklearn中网格搜索和交叉验证集成API

sklearn.model_selection.GridSearchCV(estimator, param_grid=None,cv=None)

其中参数含义为：

estimator：选择使用的分类器

param_grid：需要最优化的参数的取值，值为字典或者列表

cv：整数类型，指定K折交叉验证。

还包含常用的2个Methods和4个Attributes：

GridSearchCV的相关信息		
（1）Methods (方法-函数)		
1	**fit**	输入训练数据
2	**score**	准确率
（2）Attributes (属性-变量)		
1	**best_score_**	交叉验证中测试的最好的结果
2	**best_estimator_**	交叉验证中测试的最好的参数模型
3	**best_params_**	交叉验证中测试的最好的参数
4	**cv_results_**	每次交叉验证的结果

其他参数说明见：https://blog.csdn.net/weixin_41988628/article/details/83098130

预备知识

示例如下：

```
from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC

svc=SVC(decision_function_shape='ovo')
param_grid={'kernel':['linear','sigmoid],
            'C':[0.01,0.1],
            'gamma':[0.01,0.1]
            }
algo=GridSearchCV(estimator=svc,param_grid=param_grid,cv=10)
algo.fit(X_train,Y_train)
print("训练集:", algo.score(X_train, Y_train))

# 查看最好的参数模型
print(“最好的参数模型：\n”，algo. best_params_)
```

实验步骤

◆ 实验步骤

1、准备数据

✓读取数据、数据预处理、提取合适特征

2、配置模型

✓利用交叉验证和网格搜索法，导入GridSearchCV库，配置分类模型

3、训练模型

4、评估模型

✓使用测试集评估模型

5、比较多个分类模型的结果

实验要求

- 1、对数据预处理部分，重要代码读懂并添加注释；
- 2、要求用到**交叉验证和网格搜索方法**；
- 3、不同模型参数，调参过程和结果要记录；
- 4、使用加权平均 精确率、召回率、F值来评价模型，要求加权F值指标>0.85 ；
- 5、分析对比不同分类算法的预测结果。

附加题

➢ 优化数据样本不均衡问题

本次实验训练数据集中 0类样本 7142个，1类样本1858个，数据类别**特别不均衡**，可考虑对数据进行重采样处理，改变非平衡数据的正负类分布。重采样方法分为过采样和欠采样，请查阅资料学习，任选一种方法并实现。

注：附加内容有10%的加分， 但总分不超过该次实验满分。

提交方式

实验作业提交至平台 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 - 实验四；
- 3、提交截止时间：6月7号（下周二） 晚24点前；
- 4、文件夹&压缩包命名要求： 学号_姓名_统计机器学习实验四
- 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。

统计机器学习实验

同学们，请开始实验吧！