实验五：构建支持向量机模型

实现餐饮客户流失预测

《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

- 餐饮行业作为我国第三产业中的一个传统服务性行业，始终保持着旺盛的增长势头，取得了突飞猛进的发展，展现出繁荣兴旺的新局面，某餐饮企业正面临着房租价格高、人工费用高、服务工作效率低等问题。如何在保证产品质量的同时提高企业利润，成为某餐饮企业急需解决的问题。
- 某餐饮企业的系统数据库中积累了大量的与客户用餐相关的数据，通过对某餐饮企业的数据进行分析，最终为餐饮企业提出改善的建议。

- **任务一**：构建**支持向量机模型**，构建客户流失预测模型，并对模型进行评价。
- **任务二**：构建**已学习过的分类模型（任选两种）**，构建客户流失预测模型，并对模型进行评价。

二、数据说明

◆ 数据集

- 包含**训练集train**（共4605条数据），**测试集test**（1400条数据）
- 每一条数据由5个特征值，1个目标值（type）组成。
- 每个特征值的含义如下：

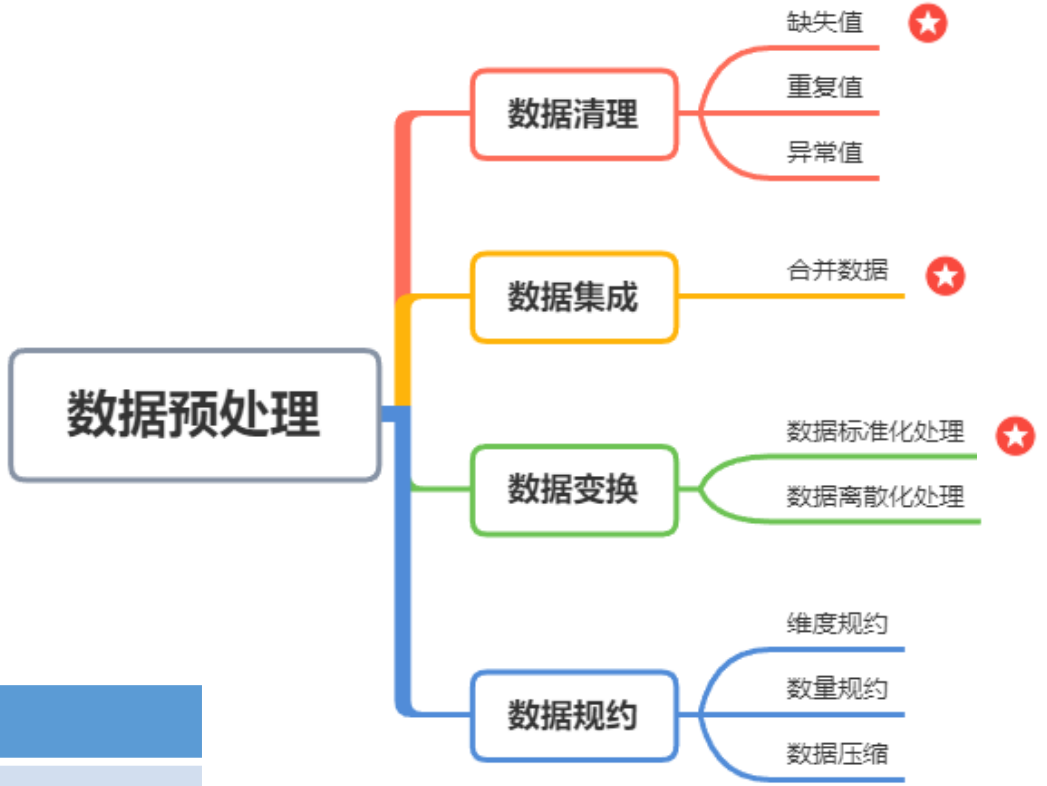
test
train

客户ID	客户账号	客户最近一次用餐的时间	是否流失	消费人数	消费金额
USER_ID	ACCOUNT	LAST_VISITS	type	number_consume	expenditure
983	邓彬彬	2016/6/20 13:15	准流失	4	753
983	邓彬彬	2016/6/20 13:15	准流失	10	1215
986	莫子建	2016/7/30 13:46	非流失	3	356
986	莫子建	2016/7/30 13:46	非流失	7	1146
986	莫子建	2016/7/30 13:46	非流失	3	605
988	郭仁泽	2016/3/15 17:34	非流失	10	1169
988	郭仁泽	2016/3/15 17:34	非流失	2	436
988	郭仁泽	2016/3/15 17:34	非流失	8	1406
988	郭仁泽	2016/3/15 17:34	非流失	7	1377
989	唐莉	2016/7/21 12:57	准流失	10	1269
989	唐莉	2016/7/21 12:57	准流失	4	516
991	麦凯泽	2016/6/11 11:25	非流失	10	1852
991	麦凯泽	2016/6/11 11:25	非流失	5	725
991	麦凯泽	2016/6/11 11:25	非流失	10	1677
994	刘乐瑶	2016/6/15 12:42	准流失	3	452
994	刘乐瑶	2016/6/15 12:42	准流失	7	1364

三、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。
- 对右图做标记的三种数据处理场景做代码示例：



场景	示例代码
删除有缺失值的行记录	info_user_new = info_user_new.dropna(axis=0)
删除numbers为0的客户	info_user_new = info_user_new[info_user_new['numbers'] != 0]
合并两张表	info_user_new = pd.merge(info_user1, info_user2, left_on='USER_ID', right_on='USER_ID', how='left')
数据标准化（Z-Score）	from sklearn.preprocessing import StandardScaler X1 = StandardScaler().fit_transform(X)

三、预备知识

2、支持向量机中常用的核函数

输入	含义	适用场合	核函数表达式	gamma	degreee	coef0
linear	线性核	线性	$K(x,y)=x^Ty=x\cdot y$	无	无	无
poly	多项式核	偏线性	$K(x,y)=(\gamma(x\cdot y)+r)^{\mathrm{d}}$	有	有	有
rbf	高斯径向核	偏非线性	$K(x,y)=e^{-\gamma\|x-y\|^2},\gamma>0$	有	无	无
sigmoid	双曲正切核	非线性	$K(x,y)=\tanh(\gamma(x\cdot y)+r)$	有	无	有

gamma，核函数系数
degree，多项式核函数的阶数n
coef0，核函数中的独立项

三、预备知识

3、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

‘poly’ ：多项式核函数

‘rbf’ ： 径向基核函数/高斯核函数

‘sigmoid’ ：双曲正切核函数

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

三、预备知识

4、网格搜索

实验四内的要求：

使用交叉验证思想，划分训练集和验证集；

除了实验三介绍的sklearn中的model_selection.Kfold方法外，还需要调研其他方法来实现交叉验证思想，总结分析对比。

◆ **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。

三、预备知识

5、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

三、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_)
```

四、实验数据处理

◆数据处理

在本案例中，客户流失的特征主要体现在以下4个方面。

- 用餐次数越来越少。
- 很长时间没有来店里消费。
- 平均消费水平越来越低。
- 总消费金额越来越少。

客户ID	客户账号	客户最近一次用餐的时间	是否流失	消费人数	消费金额
USER_ID	ACCOUNT	LAST_VISITS	type	number_consume	expenditure
983	邓彬彬	2016/6/20 13:15	准流失	4	753
983	邓彬彬	2016/6/20 13:15	准流失	10	1215
986	莫子建	2016/7/30 13:46	非流失	3	356
986	莫子建	2016/7/30 13:46	非流失	7	1146
986	莫子建	2016/7/30 13:46	非流失	3	605
988	郭仁泽	2016/3/15 17:34	非流失	10	1169
988	郭仁泽	2016/3/15 17:34	非流失	2	436
988	郭仁泽	2016/3/15 17:34	非流失	8	1406
988	郭仁泽	2016/3/15 17:34	非流失	7	1377
989	唐莉	2016/7/21 12:57	准流失	10	1269
989	唐莉	2016/7/21 12:57	准流失	4	516
991	麦凯泽	2016/6/11 11:25	非流失	10	1852
991	麦凯泽	2016/6/11 11:25	非流失	5	725
991	麦凯泽	2016/6/11 11:25	非流失	10	1677
994	刘乐瑶	2016/6/15 12:42	准流失	3	452
994	刘乐瑶	2016/6/15 12:42	准流失	7	1364

四、实验数据处理

◆数据处理

基于这4个方面，本案例需要构造4个相关客户流失特征。

- **总用餐次数**（frequence）。即观测时间内每个客户的总用餐次数。
- 客户最近一次用餐的时间距离观测窗口结束（2016-7-31 0点）的**天数**（recently）。
- 客户在观测时间内用餐人**均销售额**（average）。即客户在观察时间内的总消费金额除以用餐总人数。
- 客户在观测时间内的**总消费金额**（amount）。

USER_ID	ACCOUNT	LAST_VISITS	type	number_consume	expenditure
983	邓彬彬	2016/6/20 13:15	准流失	4	753
983	邓彬彬	2016/6/20 13:15	准流失	10	1215
986	莫子建	2016/7/30 13:46	非流失	3	356
986	莫子建	2016/7/30 13:46	非流失	7	1146
986	莫子建	2016/7/30 13:46	非流失	3	605
988	郭仁泽	2016/3/15 17:34	非流失	10	1169
988	郭仁泽	2016/3/15 17:34	非流失	2	436
988	郭仁泽	2016/3/15 17:34	非流失	8	1406
988	郭仁泽	2016/3/15 17:34	非流失	7	1377
989	唐莉	2016/7/21 12:57	准流失	10	1269
989	唐莉	2016/7/21 12:57	准流失	4	516
991	麦凯泽	2016/6/11 11:25	非流失	10	1852
991	麦凯泽	2016/6/11 11:25	非流失	5	725
991	麦凯泽	2016/6/11 11:25	非流失	10	1677
994	刘乐瑶	2016/6/15 12:42	准流失	3	452
994	刘乐瑶	2016/6/15 12:42	准流失	7	1364

原数据

USER_ID	ACCOUNT	frequence	amount	average	recently	type
983	邓彬彬	2	1968	140.57	40	准流失
986	莫子建	3	2107	162.08	0	非流失
988	郭仁泽	4	4388	162.52	137	非流失
989	唐莉	2	1785	127.5	9	准流失
991	麦凯泽	3	4254	170.16	49	非流失
994	刘乐瑶	2	1816	181.6	45	准流失
1000	邱泊君	2	1985	152.69	139	准流失
1002	李孟夏	1	775	129.17	85	准流失
1004	陈明杰	1	362	90.5	85	准流失
1009	袁田田	2	1346	103.54	15	准流失
1010	袁家蕊	3	3514	146.42	54	非流失
1011	柴德馨	3	4112	178.78	41	非流失
1012	柴鸿飞	1	873	174.6	35	准流失
1017	卢俊恒	3	2950	155.26	64	非流失
1018	卢子翰	3	3377	168.85	97	非流失
1023	魏巍	2	2392	132.89	48	非流失

处理后的数据样例

五、实验要求

- 1、按照上一页PPT数据样例**对原数据做处理**；
- 2、要求用到**交叉验证和网格搜索方法**；
- 3、使用**准确率**指标来评价模型，要求最好参数的评价指标>0.9；

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- 4、对任务一，还需探究支持向量机参数对预测结果的影响，并分析做出相应结论：

(1) 比较4种核函数的分类准确率；
(2) 数据标准化对支持向量机分类结果的影响；
(3) 高斯核函数、多项式核函数的参数调整；
(4) 松弛系数惩罚项C的调整。

- 5、对任务二，还需要用其他两种分类模型做预测，做好调参记录，分析做出相应结论。

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 - 实验五；
- 3、提交截止时间：下周二 6月25号 晚24点；
- 4、文件夹&压缩包命名要求： 学号_姓名_统计机器学习实验五
- 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！