南京理工大学紫金学院

计算机 学院

《机器学习课程设计》实验指导

内部教材仅供紫金学院使用

主编：谢玲

南京理工大学紫金学院教务处制

前　言

《机器学习课程设计》是在完成《机器学习》课程的学习之后，为系统掌握相关的知识而进行的实践环节。目的是让学生更加深入了解机器学习的概念、构成及主流的关键技术及使用。培养学生利用机器学习技能结合实际环境进行实践及其独立解决问题的能力。通过任务驱动，学生综合运用以前学习的机器学习算法和 Python 编程方法相关知识，使得所学知识得到进一步巩固和深化，锻炼学生独立查阅资料获取信息、分析问题、解决问题的能力。

南京理工大学紫金学院

计算机学院实验指导书

课程编号：**0332401**

课程名称：机器学习课程设计

课程学分：**1**

实验学时：**40**

适用专业：计算机科学与技术、软件工程

学生实验守则

第一条　学生进入实验室从事实验、科研活动，必须严格遵守实验室的各项规章制度。

第二条　实验前必须做好预习，明确实验的目的、内容、方法和步骤，未经预习或无故迟到 15 分钟以上者，指导人员有权取消其实验资格。

第三条　保持实验室的严肃、安静，不得在实验室内大声喧哗、嬉闹，不准在实验室内进食、吸烟和乱吐乱丢杂物。

第四条　学生必须在实验指导老师规定实验开始时，方可动手实验，严禁未经许可就开始实验。学生必须以实事求是的科学态度进行实验，严格遵守操作规程，服从实验教师或实验技术人员的指导(对有特殊要求的实验，必须按要求穿戴安全防护用具后方可进行实验),如违反操作规程或不听从指导而造成仪器设备损坏等事故者，按学校有关规定进行处理。

第五条　学生应备有专用实验记录本，实验记录是原始性记录，是撰写实验报告的主要依据，内容要求真实、客观地反映实际情况，实验结果须经实验教师或实验技术人员认可。

第六条　严防事故，确保实验室的安全。发现异常情况，及时报告实验教师或实验技术人员，并采取相应的措施，减少事故造成的损失。

第七条　实验完成后，应将仪器、工具及实验场地等进行清理、归还，经实验教师或实验技术人员同意后，方可离开实验室。

第八条　独立完成实验报告，按时交给实验教师，不得抄袭或臆造。

实验报告是实验完成后的全面总结，它主要包括：实验名称、实验目的、实验原理、实验仪器设备、实验条件、实验数据、结果分析和问题讨论。实验报告一律用钢笔或圆珠笔书写，统一采用国家标准所规定的单位与符号，要求文字书写工整，不得潦草；作图规范，不得随手勾画。

目　录

实验一 运营商流失客户分析

一、实验目的

1. 了解数据的预处理。

2. 掌握机器学习算法在该案例中的应用（重点）。

3. 了解最优模型的选择（难点）。

二、实验原理及说明

随着业务的快速发展、移动业务市场的竞争愈演愈烈。如何最大程度地挽留在网用户、吸取新客户，是电信企业最关注的问题之一。竞争对手的促销、公司资费软着陆措施的出台和政策法规的不断变化，影响了客户消费心理和消费行为，导致客户的流失特征不断变化。对于电信运营商而言，流失会给电信企业带来市场占有率下降、营销成本增加、利润下降等一系列问题。在发展用户每月增加的同时，如何挽留和争取更多的用户，是一项非常重要的工作。

随着大数据挖掘技术的不断发展和应用，移动运营商希望能借助数据挖掘技术识别哪些用户可能流失，什么时候会发生流失。而通过建立流失预测模型，分析用户的历史数据和当前数据，提取辅助决策的关键性数据，并从中发现隐藏关系和模式，进而预测未来可能发生的行为，就可以帮助移动运营商实现这些要求。

利用用户的短信、流量、通话、消费的使用情况以及客户基本信息的数据，采用数据挖掘技术对用户进行分群。

分析不同群体用户的使用规律，识别各群体客户流失的重要特征。

建立不同群体用户流失模型，结合结果为运营商提供差异化的意见和建议。

三、实验内容

1.数据预处理

（1）重复数据处理

由于使用重复数据会对下面的分析和建模产生不利影响，因此，将所有重复数据视

为异常值剔除。

（2）降维

降维处理原始数据的变量。

（3）数据的提取和整合

□ 在原始的建模数据中，每个用户有三行的数据，对应着同一个 ID 三个月的信息。这种数据格式不便于进行建模和分析。因此，本文对数据进行提取和整合，将每个用户信息处理为一行数据。具体的处理方式如下:

a 移动客户基本信息：由于同一个用户的基本信息在三个月内是相同的。因此，用户 ID、性别、星座、年龄以及终端硬件类型只取 1 月的数据。

b 在网时长：由于每个月的在网时长等于上个月的在网时长加一（为零的除外），故第三个月的在网时长的数据就包含了这三个月在网时长的全部信息。因此，每个 ID 只提取第三个月的数据。

c 是否合约有效：将这三个月的合约有效的情况规整为一个数据，处理如下：当三个月不全为 1 时，用第三个月的值减去前两个的均值；当三个月的值都是为 1 时，取值为 1.5。所有取值情况为-1、-0.5、0、0.5、1、1.5。

d 合约计划到期时间:取第三个月的数据作为合约计划到期时长,将空值赋值为-1，不是空值的到期时间以 201603 为基准 0,每增加一个月,数据为 1,例如日期为 201607，那么该月份的到期时间为 4，以此类推。

e 信用等级：取 3 个月的平均值。

f VIP 等级：将空值赋值为 0 ，同一个 ID 值，三个月的数值相等，则取第三个月的数值；三个月的数值都不相等，则第三个月的数据减去前两个月的均值。这样的处理可以突出用户在第三个月的变化情况。

g 本月费用：取三个月的平均值。

h 平均每次通话时长：将各种通话时长除以通话次数，得到各类的平均通话时长。

i 其余的变量处理方式相同：同一个 ID 下，取三个月的平均值。

经上述规则提取和整合后，得到的新数据集。

（4）缺失值的处理

用 python 对新数据集进行缺失值查找，缺失值集中在用户的基本信息——星座、年龄和性别上。对缺失的数据进行赋值，性别缺失的赋值为 3，年龄缺失的赋值为 0，

星座缺失的赋值为 0。

（5）异常值的处理

对新数据集进行异常值的查找，查找出在网时长小于 0 的数据，费用大于 4 万元的数据。这些异常数据有很大的可能是因为人为录入时出现失误，直接利用异常数据建模会导致模型结果与实际情况出现严重的偏差。

2.流失用户的特征分析

对性别、年龄、星座、在网时长、合约是否有效、合约到期时间、信用等级、VIP 等级等进行分析吗，进一步对数据进行降维。接着，使用 K-Means 进行聚类分析，对人群进行分类。

3.特征值的提取

特征抽取是数据挖掘任务最为重要的一个环节，一般而言，它对最终结果的影响要高过数据挖掘算法本身。只有先把现实用特征表示出来，才能借助数据挖掘的力量找到问题的答案。特征选择的另一个优点在于：降低真实世界的复杂度，模型比现实更容易操纵。

特征选择的原因有：(1)降低复杂度；(2)降低噪音；(3)增加模型可读性。

单个特征和某一类别之间相关性的计算方法有很多，比较有效的有卡方检验（chi2）以及互信息和信息熵，本文我们选择基于信息熵的方法来选取特征变量。

信息熵是在决策树中广泛使用的一个变量，用以获取最优划分的节点。基于树的预测模型能够用来计算特征的重要程度，因此能用来去除不相关的特征。

因此选择基于树的特征选择(Tree-based feature selection)来获取特征变量。

4. 模型的建立与求解

分别对分类后的每个数据集使用神经网络模型、决策树模型、朴素贝叶斯模型和支持向量机模型进行建模，使用测试数据集计算出模型的错误率，选择错误率最低的模型作为最优模型，并使用最优的模型对该类别的预测数据集进行预测。

分类后的五个数据集作为建模数据集，将各个建模数据集按照 8:2 的比例随机抽取 80%的数据作为待处理训练数据集，20%的建模数据作为测试数据集，待处理训练测试集和测试数据集中均包含流失和非流失用户的信息。

5. 最优模型的选择和预测

对整合数据的本月消费进行 K-Means 聚类，将数据分为 5 类，分别为高费用、中

高费用、一般费用、中低费用、低费用，结合选出的关于每类的最优模型，对预测数据的每一类进行流失预测。

四、实验安全事项

实验过程中注意用电安全。

五、实验提交方式

√ 实验报告　　□ 现场打分　　□ 线上平台提交