**《数据挖掘》**

**实验报告2**

**班 级 数212**

**学 号 21011171**

**姓 名 范鸿玥**

**批阅教师 钱夕元**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**评分标准：**

1. 成绩分为A/B/C/D四个等级（A为最高）
2. A：实验结果正确；过程进行了清晰的描述；作图清晰美观；对结果进行了详细的分析、讨论
3. B：实验结果基本正确；过程进行了部分描述；作图基本正确；对结果进行了简单的分析、讨论
4. C：实验结果部分正确；过程描述很少；作图不够正确；对结果的分析不够正确
5. D：实验结果出现明显的错误；过程混乱；没有结果呈现；对结果没有分析或分析完全错误
6. 没有按时提交作业、补交作业，成绩降一档
7. 抄袭作业成绩为零分。

**教师评语：**

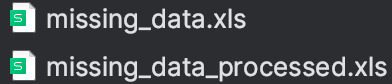
|  |
| --- |
|  |

**实验成绩：**

|  |  |
| --- | --- |
| 实验名称：数据挖掘—分类 | 实验地点：八教611 |
| 所使用的工具软件及环境：Anaconda | |
| 1. 实验目的：    1. 掌握拉格朗日插值法进行缺失值处理    2. 掌握CART决策树构建分类模型    3. 掌握Logistic回归模型用于分类    4. 了解LM神经网络分类 | |
| 1. 实验内容    1. 用户的用电数据存在缺失值，数据见“/data/missing\_data.xls”，利用拉格朗日插值算法补全数据。    2. 对所有窃漏电用户及正常用户的电电量、告警及线损数据和该用户在当天是否窃漏电的标识，按窃漏电评价指标进行处理并选取其中291个样本数据，得到专家样本，数据见“/data/model.xls”，分别使用CART决策树、Logistic回归和LM神经网络实现分类预测模型，利用混淆矩阵和ROC曲线对模型进行评价 | |
| 1. 操作步骤   1)把“test/data/missing\_data.xls”数据放入当前工作目录。  2)使用Pandas把数据读入当前工作目录。  3)针对读入的数据每一列，进行编程。编程主要参考第4章的拉格朗日插值算法，具体主要步骤如下：  a)针对每列数据的每一个缺失值，逐个进行补数（这样可以在连续两个缺失值的情况下，使用前面一个已经补数的值来再次补数后面的一个值）。  b)针对一个缺失值，构造参考组。选取前面5个作为前参考组，后面5个为后参考组。如果前参考组或后参考组不足5个则按实际个数构造参考组。  c)确认缺失值在参考组中的相对位置，然后使用拉格朗日插值进行缺失值插值。  d)根据插值后的值更新原始数据中相应位置的值。  4)把经过预处理的专家样本数据“test/data/model.xls”数据放入当前工作目录，并使用Pandas读入当前工作空间。  5)把工作空间的建模数据随机分为两部分，一部分用于训练，一部分用于测试。  6)使用Scikit-Lrean库的sklearn.tree的DecisionTreeClassifier函数以及训练数据构建CART决策树模型、Logistic回归模型（参见授课ppt），使用predict函数和构建的CART决策树模型、Logistic回归模型分别对训练和测试数据进行分类，并与真实值进行对比，得到模型正确率，同时使用sklearn.metrics的confusion\_matrix 和roc\_curve函数画混淆矩阵和ROC曲线图  7）尝试使用LM神经网络对上述训练数据及测试数据进行分类操作，评价其分类效果.  **注：内容从下页开始。** | |

* **用电数据的分类预测模型**
  + **利用拉格朗日插值算法补全用电数据**

1)把“missing\_data.xls”数据放入当前工作目录。2)使用Pandas把数据读入当前工作目录。3)针对读入的数据每一列，进行拉格朗日插值算法的编程。具体主要步骤如下：a)针对每列数据的每一个缺失值，逐个进行补数（这样可以在连续两个缺失值的情况下，使用前面一个已经补数的值来再次补数后面的一个值）。b)针对一个缺失值，构造参考组。选取前面5个作为前参考组，后面5个为后参考组。如果前参考组或后参考组不足5个则按实际个数构造参考组。c)确认缺失值在参考组中的相对位置，然后使用拉格朗日插值进行缺失值插值。d)根据插值后的值更新原始数据中相应位置的值。



* + **得到专家样本**

对所有窃漏电用户及正常用户的电电量、告警及线损数据和该用户在当天是否窃漏电的标识，按窃漏电评价指标进行处理并选取其中291个样本数据，得到专家样本“model.xls”。



* + **构建并测试CART决策树模型**

4)把经过预处理的专家样本数据“model.xls”数据放入当前工作目录，并使用Pandas读入当前工作空间。5)把工作空间的建模数据随机分为两部分，分别用于训练和测试。6)使用Scikit-Lrean库的sklearn.tree的DecisionTreeClassifier函数以及训练数据构建CART决策树模型、Logistic回归模型，使用predict函数和构建的CART决策树模型、Logistic回归模型分别对训练和测试数据进行分类。与真实值进行对比，得到模型正确率，同时画出混淆矩阵和ROC曲线图。



****

如图所示，此分类模型具有非常好的分类和预测性能。

* + **构建并测试LM神经网络模型**

7）尝试使用LM神经网络对上述训练数据及测试数据进行分类操作，评价其分类效果。如图所示，LM神经网络最终的分类效果与CART决策树模型相差不大。但是从速度上相比，LM神经网络的速度会差于CART决策树模型。



