《大数据分析方法》课程实验报告

学号：

姓名：

专业：

班级：

队名：

**实验一：关联规则算法实验**

**实验目的：**在掌握关联规则算法的原理的基础上，能够应用关联规则算法解决实际问题。

**实验内容：**根据实验数据，采用Apriori等关联规则发现算法，给出相关关联规则。

**实验要求：**给出数据预处理过程、关联规则发现算法及发现关联规则，并对关联规则结果进行分析说明。

**实验题目：**蔬菜价格相关性分析

蔬菜的价格会受季节、天气等多方面因素的影响，但许多会出现同涨或者同跌等现象，请根据给出的蔬菜价格数据，采用关联规则发现算法，发现哪些蔬菜之间具有同涨、同跌或者涨跌不同步的现象。

数据格式如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 日期 | 蔬菜名 | 价格 | 肉食禽蛋 | 批发价格 |
| 2009/6/11 | 大白菜 | 0.7 | 精猪肚 | 28 |
| 2009/6/11 | 青萝卜 | 1.2 | 猪口条 | 28 |
| 2009/6/11 | 胡萝卜 | 2 | 猪心 | 24 |
| 2009/6/11 | 水萝卜 |  | 猪肝 | 7 |
| 2009/6/11 | 葱 | 2.5 | 猪大肠 | 22 |
| 2009/6/11 | 芽葱 |  | 猪蹄 | 15 |
| 2009/6/11 | 元葱 | 0.8 | 猪大排 | 20 |
| 2009/6/11 | 生姜 | 3 | 猪肋排 | 25 |
| 2009/6/11 | 大蒜 | 3 | 精猪肉 | 17 |
| 2009/6/11 | 土豆 | 1.3 | 精牛肉 | 30 |
| 2009/6/11 | 菠菜 | 1.4 | 白条肉 | 16 |
| 2009/6/11 | 韭菜 | 1 | 羊肉片 | 32 |
| 2009/6/11 | 芹菜 | 1.5 | 西装鸡 | 12 |
| 2009/6/11 | 油菜 | 1 | 鸡大腿 | 12 |
| 2009/6/11 | 苔菜 |  | 鸡翅 | 18 |
| 2009/6/11 | 香菜 | 2 | 鸡爪 | 16 |
| 2009/6/11 | 大头菜 | 0.6 | 鸡胸肉 | 14 |
| 2009/6/11 | 芸豆 | 1.2 | 翅根 | 14 |
| 2009/6/11 | 豆角 | 2 | 翅中 | 26.2 |
| 2009/6/11 | 茄子 | 1.2 |  |  |
| 2009/6/11 | 青椒 | 1.7 | 鸡心 | 11 |
| 2009/6/11 | 西红柿 | 1.4 | 松花蛋 |  |
| 2009/6/11 | 黄瓜 | 1.1 | 咸鸭蛋 |  |
| 2009/6/11 | 茭瓜 | 0.7 | 鸡蛋 | 6.6 |
| 2009/6/11 | 冬瓜 | 1.6 |  |  |

**实验过程：**

**实验二：分类回归算法的基础实验**

**实验目的：**在学习线性回归、逻辑回归、KNN、决策树、贝叶斯分类、神经网络等算法的原理的基础上，掌握已有的数据分类算法软件包中数据分类回归算法的使用。

**实验内容：**学习scikit-learn包中的[Classification](http://scikit-learn.org/stable/supervised_learning.html" \l "supervised-learning)，网络地址：http://scikit-learn.org/

**实验要求：**至少实验2种算法的使用

**实验题目：**根据scikit-learn中的算法要求，自行选择实验。

以下是可选择的范围：建议着重练习1.1,1.6,1.9,1.10,1.17：

1.1. Generalized Linear Models

1.2. Linear and Quadratic Discriminant Analysis

1.3. Kernel ridge regression

1.4. Support Vector Machines

1.5. Stochastic Gradient Descent

1.6. Nearest Neighbors

1.7. Gaussian Processes

1.8. Cross decomposition

1.9. Naive Bayes

1.10. Decision Trees

1.11. Ensemble methods

1.12. Multiclass and multilabel algorithms

1.13. Feature selection

1.14. Semi-Supervised

1.14.1. Label Propagation

1.15. Isotonic regression

1.16. Probability calibration

1.17. Neural network models (supervised)

注意：1.17中的神经网络部分，建议学习Keras软件包的深度学习网络构建方法。

**实验三：分类算法应用实验**

**实验目的：**在掌握分类算法基础实验应用的基础上，能够应用分类算法解决实际问题。

**实验内容：**根据实验数据，采用分类回归算法，解决实际应用问题。

**实验要求：**给出实际问题的解决方案，数据预处理过程、分类模型的建立过程、参数调优、结果分析。

**实验题目：**猫狗大战

问题源于Kaggle(www.kaggle.com)猫狗图片识别比赛，数据集中共包含猫、狗的图片各12500张，要求设计一个分类算法，实现对图片为“猫”或者是“狗”的图片的分类。

数据集下载地址：https://www.kaggle.com/c/dogs-vs-cats/data

**实验过程：**

**实验四：聚类算法实验**

**实验目的：**在掌握聚类算法基础原理基础上，掌握应用聚类算法解决实际问题。

**实验内容：**根据给定的实验数据，采用层次聚类、基于划分的聚类、密度聚类等算法对数据聚类。

**实验要求：**给出实际问题的解决方案，数据预处理过程、聚类算法的建立过程、结果分析。

**实验题目：《**红楼梦**》作者分析**

《红楼梦》是我国著名的四大名著之一，一般的认为《红楼梦》的前八十回为曹雪芹撰写，后四十回为高鹗续写，但也有学者对此并不认可。能否利用聚类分析的方法来对《红楼梦》的作者进行分析。

提示思路：一般来说，不同的作者往往会具有不同的写作风格，这些风格可以通过在文中的虚词的频率进行衡量，因此，可以考虑统计各章中虚词出现频率，并以此作为基础数据来聚类分析，对《红楼梦》章节进行划分，从而分析章节与作者之间的关系。

数据：《红楼梦》完整版.txt

参考文献：李贤平.《红楼梦》成书新说.复旦学报(社会科学版).1987年第5期.

**实验过程：**