云南大学数学与统计学院

《数据挖掘与决策支持实验》上机实践报告

|  |  |  |
| --- | --- | --- |
| **课程名称**：运筹学数据挖掘与决策支持实验 | **年级**：2015级 | **上机实践成绩**： |
| **指导教师**：彭程 | **姓名**：刘鹏 | **专业：**信息与计算科学 |
| **上机实践名称**：对iris数据进行贝叶斯分类 | **学号**：20151910042 | **上机实践日期**：2018-07-04 |
| **上机实践编号**：05 | **组号**： |  |

# 一、实验目的

学习使用R语言进行变量选择。

# 二、实验内容

如下表：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 产品编号 | 加工时间 | 机台类型 | 加工时间 | 机台类型 | 良率 |
| 1 | 28 | A01 | 48 | B03 | 0.53 |
| 2 | 27 | A01 | 42 | B03 | 0.62 |
| 3 | 31 | A03 | 43 | B21 | 0.84 |
| 4 | 42 | A02 | 33 | B02 | 0.91 |
| 5 | 46 | A02 | 28 | B03 | 0.85 |
| 6 | 50 | A01 | 27 | B03 | 0.68 |
| 7 | 35 | A02 | 24 | B01 | 0.83 |
| 8 | 24 | A03 | 36 | B02 | 0.69 |
| 9 | 28 | A02 | 25 | B01 | 0.88 |
| 10 | 44 | A03 | 37 | B03 | 0.92 |

将给定数据进行变量选择，从而实现维归约。

# 三、实验平台

Windows 10 Pro 1803；

Microsoft© Visual Studio 2017 Enterprise。

Version 1.1.442 – © 2009-2018 RStudio, Inc.

# 四、算法设计

背景介绍：在进行数据挖掘时，我们并不需要将所有的自变量用来建模，而是从所有的变量中选择最重要的变量，这称为变量选择（feature selection）。一种算法是向后选择，即现将所有的变量都包括在模型中，再次计算效能，反复迭代，找出合适的自变量数目。

通常，如果一个变量描述了不同种类的实体，则可以检查不同种类的样本。用变量的方差对变量的均值进行标准化，然后比较不同种类的标准化值。如果标准化均值相差很大，说明这个特征就很重要，反之说明两者的互信息比较大，两者所含信息量重叠很多，可以去掉其中之一。这种思想代表的是一种试探性的、非优化的特征选择方法。不过这种方法符合很多将数据挖掘技术应用于特征分类的实际经验。下面是检验公式：

从样本数据中可以看到，良率全都在百分之五十之上，不妨取0.80作为阈值进行检测。

# 五、程序代码

5.1 程序描述

5.2 程序代码

程序代码 1

# 六、运行结果

运行结果 1 （经过了反相处理）

代码分析

# 七、实验体会

# 八、参考文献