­­­­

数据挖掘实验报告

实验名称： 数据预处理

姓 名：

学 号：

实验日期：

# 一、实验目的及要求

1. 掌握数据预处理概念

2. 掌握基本的数据预处理方法

3. 将源代码及实验结果的截图，粘贴到实验步骤的每一小问中

## 二、实验所需Python库

|  |  |
| --- | --- |
| 名称 | 版本 |
| SKlearn | 0.22.1 |
| Numpy | 1.19.2 |
| Pandas | 0.25.0 |

## 数据集介绍

## Price-label数据集

此数据集由20条购物车记录组成，有2个属性，分为别“price”商品价格；“label”购物车标签，其中为“1”代表在购物车中，为“0”代表不在购物车中。

## Newsgroups20

此数据集是一个新闻组文档的集合。20个新闻组的已成为机器学习技术文本应用（如文本分类和文本聚类）实验的热门数据集。

**注**：本实验仅用到**二十**类中的**两**类，即：'alt.atheism','comp.graphics'。

## 实验内容

1. 读取price-label数据集，完成任务1，基于信息熵做数据离散化处理，并打印出处理结果。
2. 读取 Newsgroups20 文件中的数据，完成任务2中的习题。

## 实验步骤

**任务1：**

基于信息熵的数据离散化算法是由监督学习算法，在使用该方法对数据进行离散化时，需要数据有对应的标签。

本次实验使用的是一份用户最近点击的20个商品的价格与是否加入购物车对应关系。

数据文件为压缩包中“price-label.csv”文件。

**信息熵**

信息熵的定义:信息量度量的是一个具体事件发生了所带来的信息，而熵则是在结果出来之前对可能产生的信息量的期望——考虑该随机变量的所有可能取值，即所有可能发生事件所带来的信息量的期望。

信息熵就是信息量的数学期望。

公式：



P(xi)表示随机事件X为 xi 的概率。

**步骤：**

1. 准备数据

初始化相关函数，并加载数据。对应如下基于信息熵的数据离散化--设置参数并加载数据。

（2）计算数据的信息熵

该步骤是计算数据的信息熵，是为下一步分割数据集做准备。

（3）分割数据集

寻找一组数据最佳分割点的方法是：遍历所有属性值，数据按照该属性分割，使得平均熵最小。

（4）数据离散化

按照上面基于信息熵分组的内容，对数据做离散化处理并打印出处理结果。

**任务2：**

对Newsgroups20文本信息预处理。利用词袋模型，将文本数据向量化。

1. 对文本进行分词，得到词表，即一个包含不同单词，且无重复单词的集合，这里需要考虑：去除标点符号、单词均转化为小写。

例如：data=['I am a student.','I am a teacher.']，分词后得到集合('i', 'am', 'a', 'student', 'teacher')

2. 统计每篇文章出现的词的次数，得到二维数组

例如：上面的例子中得到二维数组[[1, 1, 1, 1, 0],[1, 1, 1, 0, 1]]

3. 计算`每列`非零元素的比例

4. 对数据进行规范化，使得数据每一列均值为0，方差为1

5. 对每行数据进行范数规范化，

其中，为第行的向量，表示数组中位置为的元素。

## 六、个人感想