《数据挖掘》PBL作业

【项目名称】 分类算法的应用

【实验目的】

1.掌握数据的预处理技术

2.了解分类算法理论基础

3. 利用分类算法实现数据类别划分和预测

【实验原理】

有监督学习

【实验步骤】

复习分类算法的简单实现：

通过一个小案例回顾如何构建一个最简单的分类器：

（1）导入数据

```
from sklearn import datasets
iris = datasets.load_iris()
```

（2）创建分类器（以朴素贝叶斯为例）

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
```

（3）训练

```
y_pred = gnb.fit(iris.data, iris.target).predict(iris.data)
```

（4）分类器的效果

```
print("Number of mislabeled points out of a total %d points : %d" %
(iris.data.shape[0],(iris.target != y_pred).sum()))
```

题目：人群的收入预测

Adult数据集（即“人口普查收入”数据集），由美国人口普查数据集库抽取而来，其中共包含48842条记录，年收入大于50k美元的占比23.93%，年收入小于50k美元的占比76.07%，并且已经划分为训练数据32561条和测试数据16281条。 该数据集类变量为年收入是否超过50k美元，属性变量包括年龄、工种、学历、职业等 14类重

要信息，其中有8类属于类别离散型变量，另外6类属于数值连续型变量。该数据集是一个分类数据集，用来预测年收入是否超过50k美元。

样本属性及含义，具体请参考adult.name。

属性名	类型	含义
age	continuous	年龄
workclass	discrete	工作类别
fnlwgt	continuous	人口普查员序号
education	discrete	受教育程度
education-num	continuous	受教育时间
marital-status	discrete	婚姻状况
occupation	discrete	职业
relationship	discrete	社会角色
race	discrete	种族
sex	discrete	性别
capital-gain	continuous	资本收益
capital-loss	continuous	资本支出
hours-per-week	continuous	每周工作时间
native-country	discrete	国籍

STEP1:原始数据属性有离散和连续，属性值类型包括字符，数值，请根据所选择分类算法，进行数据转换和编码。

STEP2：选择合适的分类器（多个），将STEP1处理好的数据进行训练，并使用交叉验证评测各分类器的效果。

STEP3：对测试集采用同样的编码处理，预测其收入类别。

说明：

网络资料丰富，可参考，但请注意梳理，力求整体逻辑清晰，编程风格统一，切勿东拼西凑。

期待大家的作品！