浙江大学

本科实验报告

k-近邻算法

课程名称：	人工智能实验
姓　　名：	姚桂涛
学　　院：	信息与电子工程学院
专　　业：	信息工程
学　　号：	3190105597
指导老师：	胡浩基、魏准

2022 年 2 月 27 日

一、 实验题目

1. kNN 代码实现-AB 分类

采用测量不同特征值之间的距离方法进行分类，用所给的函数创建具有两个特征与一个标签类型的数据作为训练集，编写 classify0 函数对所给的数据进行 AB 分类。

2. k-近邻算法改进约会网站的配对效果

k-近邻算法改进约会网站的配对效果
通过收集的一些约会网站的数据信息，对匹配对象的归类：不喜欢的人、魅力一般的人、极具魅力的人。
数据中包含了 3 种特征：
每年获得的飞行常客里程数、玩视频游戏所耗时间百分比、每周消费的冰淇淋公升数

二、 实验代码

1. kNN 代码实现-AB 分类

kNN

```
from http.client import ImproperConnectionState
from numpy import *
from collections import Counter
import operator
def createDataSet():
    group = array([[1.0, 1.1], [1.0, 1.0], [0, 0], [0, 0.1]])
    labels = ['A', 'A', 'B', 'B']
    return group, labels

def classify0(inX, group, labels, k = 3):
    res1 = (inX - group)**2
    dist = res1[:,0] + res1[:,1]
    dic = argsort(dist)
    dic = dic[0:k:1]
    newdic = []
    for i in range(k):
        newdic.append(labels[dic[i]])
    c = Counter(newdic).most_common(1)
    return c[0][0]
```

tests

```
import kNN
group, lables = kNN.createDataSet()
print('分类结果')
print('[0, 0] %c' %(kNN.classify0([0, 0], group, lables, k = 3)))
print('[0.8, 0.7] %c' %(kNN.classify0([0.8, 0.7], group, lables, k = 3)))
```

2. k-近邻算法改进约会网站的配对效果

```
import pandas as pd
import kNN
from sklearn.model_selection import train_test_split

df = pd.read_table('datingTestSet2.txt',sep='\s+', names = ['A', 'B', 'C', 'Y'])
# 对特征进行归一化处理
df2 = df.iloc[:, :3]
df2 = (df2-df2.mean())/df2.std()
lable =df.iloc[:,3:4]
df2.loc[:, 'Y'] = lable
# 对数据集进行测试集和训练集划分，90%作为训练集，10%作为测试集
X_train, X_test, Y_train, Y_test = train_test_split(df2.iloc[:, :3], df2.Y, train_size
    =.90)
# 将DataFrame格式转化为numpy格式处理
group = X_train.values
label = Y_train.values
length = len(X_test)
X_test.iloc[0:1,:]
# res以储存测试结果
res = []
# 设置错误正确数count以计算正确率
Tnum = 0
Fnum = 0
for i in range(length):
    inX = X_test.iloc[i:i+1 , :].values
    res.append(kNN.classify0(inX, group, label, k = 3))
    if(kNN.classify0(inX, group, label, k = 3) == Y_test.values[i]):
        Tnum += 1
    else:
        Fnum += 1
res1 = pd.DataFrame(data = res, columns=['TestResult'])
Y_test.reset_index(inplace=True,drop=True)
res1.loc[:, 'OriginTest'] = Y_test

print('前20个数据测试结果和原数据比较')
print('————————————————————————————')
print(res1.head(20))
print('————————————————————————————')
print('正确率%.2f%%' %(100*Tnum/(Tnum+Fnum)))
```

三、 实验结果及分析

1. kNN 代码实现-AB 分类

```
    分类结果
    [0, 0] B
    [0.8, 0.7] A
```

2. k-近邻算法改进约会网站的配对效果

```
前20个数据测试结果和原数据比较
----------------------——---
    TestResult OriginTest
0            2          2
1            3          3
2            1          3
3            2          2
4            2          2
5            3          3
6            3          3
7            2          2
8            1          1
9            1          1
10           1          1
11           3          3
12           2          2
13           2          2
14           1          1
15           2          2
16           1          1
17           2          2
18           1          1
19           3          3
----------------------——---
正确率97.00%
```

从实验结果可以看出，通过 k-近邻算法改进后的约会网站的配对效果比较显著，多次随机划分测试集和训练集后发现正确率基本可以达到 90% 以上。