Topic 2 分类与回归

Part 1：分类模型

魏准 （eleweiz@zju.edu.cn）
浙江大学，信息与电子工程学院
2022

总体内容

k-近邻算法（k-Nearest Neighbor，KNN）进行分类

主要知识点

- k-近邻算法概述；
- k-近邻算法一般步骤；
- k-近邻算法代码实现；
  - AB分类；
  - 约会网站的配对效果分类；
  - 手写数字识别；

1.1：k-近邻算法概述

k-近邻算法：采用测量不同特征值之间的**距离**方法进行分类。

工作原理：

- 存在一个**样本数据集合**，也称作训练样本集，并且样本集中每个数据都存在**标签**，即我们知道样本集中每一数据与所属分类的对应关系。
- 输入**没有标签的新数据**后，将新数据的每个**特征**与样本集中数据对应的**特征**进行比较，然后算法提取**样本集**中特征**最相似数据（最近邻）**的分类标签。
- 一般来说，我们只选择样本数据集中**前k个最相似**的数据，这就是k-近邻算法中k的出处，通常k是不大于20的整数。最后，选择k个最相似数据中**出现次数最多**的分类，作为新数据的分类。

1.2：一般步骤

- 定义样本数据集合函数（训练样本集），输出样本数据集特征及其标签。
- 定义k-近邻算法函数，输入为需要测试的数据特征，样本数据集特征及其标签、k值。通过计算比较测试的数据与样本数据特征的距离，找到k个最相邻的数据及其标签，k个数据中标签次数最多的标签作为输出测试数据的分类结果。
- 测试上述函数。

1.3：kNN代码实现-AB分类

Step1 样本数据产生函数：创建名为kNN.py的Python模块

```
from numpy import *
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
```

科学计算包NumPy

运算符模块：k-近邻算法执行排序操作时将需要这个模块提供的函数。

创建名为tests.py的Python模块：

```
import kNN

group,labels = kNN.createDataSet()
```

实验1-1：编写程序，利用classify0函数（groups, labels, k=3）；测试[0,0]、[0.8,0.7]等点的类别

附加

1.4：示例：使用 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。

步骤

- 创建函数，以此来实现输入为文件名字符串，输出为训练样本矩阵和类标签向量。
- 分析可视化数据（选做）
- 处理数据（归一化）
- 测试算法

Step 1数据处理：数据存放在文本文件datingTestSet2.txt中，每个样本数据占据一行，总共有1000行。主要包含以下3种特征 （3:极具魅力，2：魅力一般的人，1：不喜欢的人）：

- 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

40920	8.326976	0.953952	3
14488	7.153469	1.673904	2
26052	1.441871	0.805124	1
75136	13.147394	0.428964	1
38344	1.669788	0.134296	1
72993	10.141740	1.032955	1
35948	6.830792	1.213192	3
42666	13.276369	0.543880	3
67497	8.631577	0.749278	1
35483	12.273169	1.508053	3
50242	3.723498	0.831917	1
63275	8.385879	1.669485	1
5569	4.875435	0.728658	2
51052	4.680098	0.625224	1
77372	15.299570	0.331351	1
43673	1.889461	0.191283	1

实验1-2：编写程序，利用classify0函数（k=3）；系统性的实现datingTestSet2.txt中10% 数据的测试，并打印出结果；

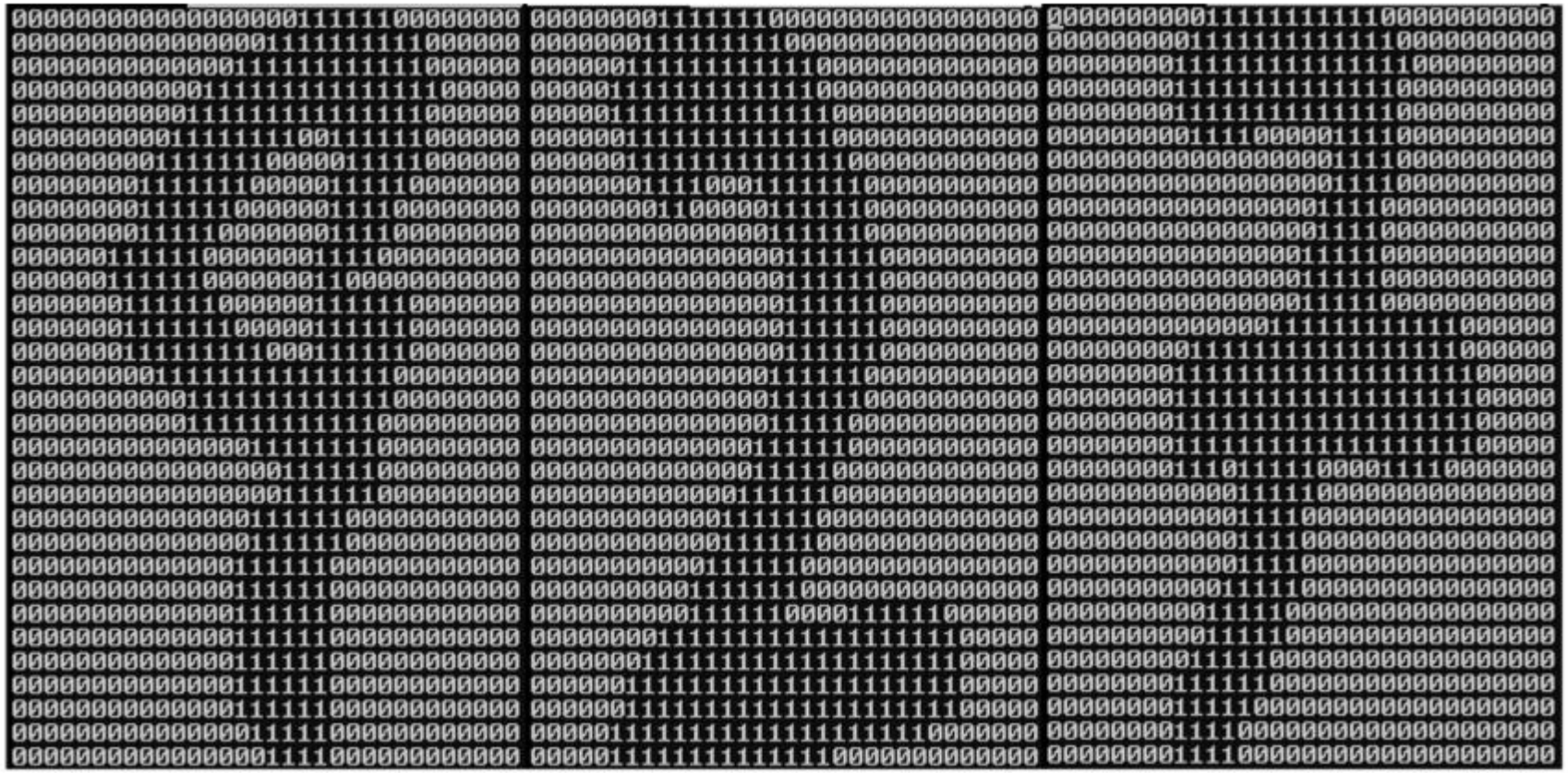
1.5：示例：使用 k-近邻算法进行数字识别

需要识别的数字已经使用图形处理软件，处理成具有相同的色彩和大小：宽高是32像素× 32像素的黑白图像。

步骤

- 创建函数，将一个32×32的二进制图像矩阵转换为1× 1024的向量，以使用前面的分类器。
- 将trainingDigits目录中的文件内容存储在列表中，并解析出数据及标签。
- 将testDigits目录中的文件内容存储在列表中，并解析出数据及标签。
- 利用前述k-近邻算法对testDigits中的数据测试并打印结果。

手写数字数据集的例子：目录trainingDigits中包含了大约2000个例子，每个数字大约有200个样本；目录testDigits中包含了大约900个测试数据。



实验1-3：编写程序，利用classify0函数（k=3）、trainingDigits作为参考数据；系统性的实现testDigits中数据的测试，并打印出结果；