**实验KNN**

姓名: 李昀哲 学号： 20123101

**1 题目和数据：**

题目：KNN

基于kNN算法改进约会网站的配对效果

基于kNN算法实现手写体数字识别

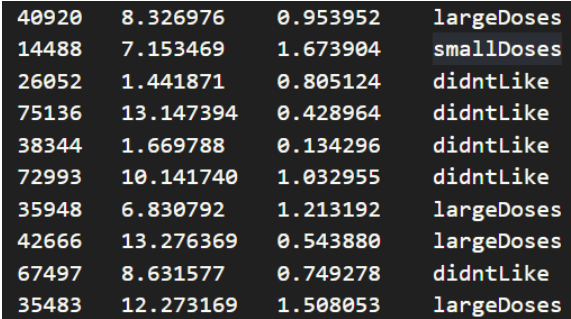
数据描述:

* **Helen女士约会网站数据**

海伦收集约会数据已经有了一段时间，她把这些数据存放在文本文件datingTestSet.txt中，每个样本数据占据一行，总共有1000行。

海伦收集的样本数据主要包含以下3种特征：

* 每年获得的飞行常客里程数
* 玩视频游戏所消耗时间百分比
* 每周消费的冰淇淋公升数



* **手写数字识别数据**

数据集目录下有两个子目录：trainingDigits和testDigits

* trainingDigits为训练数据集的文件，文件数量为1935；
* testDigits为测试数据集的文件，文件数量为947。

每个文件的命名格式都为“真实数字\_编号.txt”。为了简便处理，实验中，用txt文本文件表示图片。原图片中像素值为黑色（0,0,0）的像素点在txt中对应的用0表示，像素值为白色（255,255,255）的像素点用1表示。所以，只需要处理这些文本文件即可，不用再去解析图片格式。每个txt文件中，数据共有32行和32列，这是由于原图片的大小为32X32。

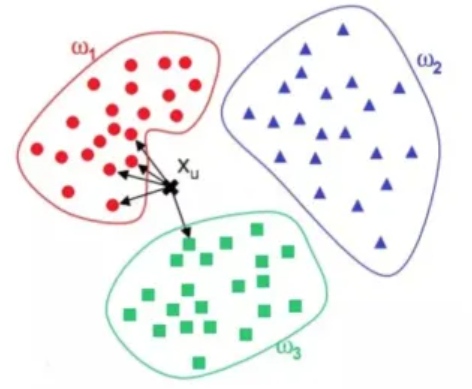
**2 算法：**

KNN（K-Nearest Neighbor）算法是机器学习算法中最基础、最简单的算法之一。它既能用于分类，也能用于回归。KNN通过测量不同特征值之间的距离来进行分类。

KNN算法的思想非常简单：对于任意n维输入向量，分别对应于特征空间中的一个点，输出为该特征向量所对应的类别标签或预测值。

KNN算法是一种非常特别的机器学习算法，因为它没有一般意义上的学习过程。它的工作原理是利用训练数据对特征向量空间进行划分，并将划分结果作为最终算法模型。存在一个样本数据集合，也称作训练样本集，并且样本集中的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。

输入没有标签的数据后，将这个没有标签的数据的每个特征与样本集中的数据对应的特征进行比较，然后提取样本中特征最相近的数据（最近邻）的分类标签。



由于KNN是基于距离度量的算法，因此距离计算公式常用欧几里得距离、曼哈顿距离和闵可夫斯基距离，三种距离公式如下所示

Euclidean Distance = 

Manhattan Distance = 

Minkowski Distance = 

KNN的优势在于：

* 易于实现：鉴于算法的简单性和准确性，它是新数据科学家将学习的首批分类器之一。
* 轻松适应：随着新训练样本的增加，算法会根据任何新数据进行调整，因为所有训练数据都存储在内存中。
* 很少的超参数：KNN 只需要 a k 值和距离度量，与其他机器学习算法相比，所需的超参数很少；

KNN的劣势在于：

* 不能很好地扩展：由于 KNN 是一种惰性算法，因此与其他分类器相比，它占用了更多的内存和数据存储。 从时间和金钱的角度来看，这可能是昂贵的。 更多的内存和存储将增加业务开支，而更多的数据可能需要更长的时间来计算。 虽然已经创建了不同的数据结构（例如 Ball-Tree）来解决计算效率低下的问题，但分类器是否理想可能取决于业务问题；
* 容易过拟合：由于"维度的诅咒"，KNN 也更容易过拟合。 虽然利用特征选择和降维技术来防止这种情况发生，但 k 的值也会影响模型的行为。 较小的 k 值可能会过度拟合数据，而较大的 k 值往往会"平滑"预测值，因为它是对更大区域或邻域的值进行平均。 但是，如果 k 的值太高，那么可能会欠拟合数据。

**3 代码及结果**

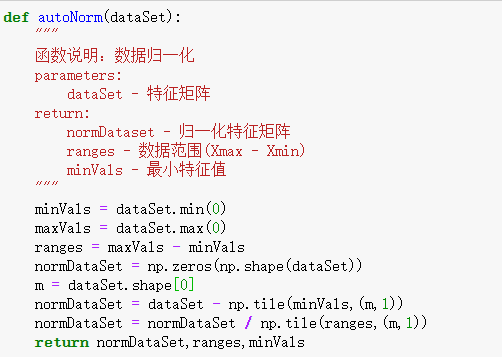
**数据集读入**

通过特征中的类别，分别用1，2，3表示；



**数据归一化**

在处理不同取值范围的特征值时，我们通常采用的方法是将数值归一化，如将取值范围处理为０到１或者-１到１之间。本实验使用如下的公式可以将任意取值范围的特征值转化为０到１区间内的值；

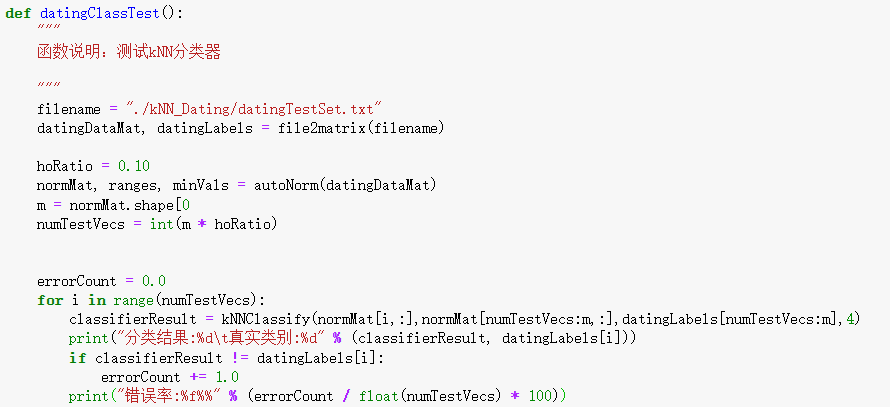


**KNN划分算法**

算法的思想是简单的，但其中涉及较多python字典的处理，较为复杂；

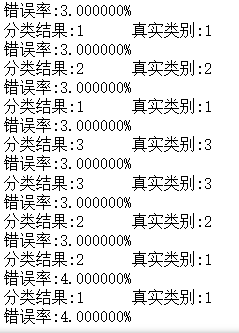


**分类测试**

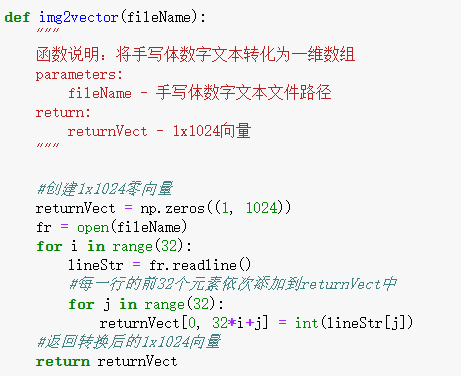


**分类结果**

测试集数据量为100，分类准确率为96%



**手写数字数据加载**





**分类测试**



**测试结果**

共测试946组数据，准确率为99.987%

