实验八 ：KNN算法

**【实验目的】**

K最近邻 (k-Nearest Neighbors，KNN) 算法是一种分类算法，1968年由 Cover 和 Hart 提出，可以应用于字符识别、文本分类、图像识别等领域。该算法的思想是：一个样本与数据集中的k个样本最相似，如果这k个样本中的大多数属于某一个类别，则该样本也属于这个类别。是最简单易懂的机器学习算法。本实验通过解决iris数据集分类，帮助学生更好的熟悉和掌握KNN算法。

【实验内容】

使用iris数据集（可网上自行下载）进行KNN实验。

iris数据集的中文名是安德森鸢尾花卉数据集，英文全称是Anderson’s Iris data set。iris包含150个样本，对应数据集的每行数据。每行数据包含每个样本的四个特征和样本的类别信息，所以iris数据集是一个150行5列的二维表。

通俗地说，iris数据集是用来给花做分类的数据集，每个样本包含了花萼长度、花萼宽度、花瓣长度、花瓣宽度四个特征（前4列），我们需要建立一个分类器，分类器可以通过样本的四个特征来判断样本属于山鸢尾、变色鸢尾还是维吉尼亚鸢尾（这三个名词都是花的品种）。

iris的每个样本都包含了品种信息，即目标属性（第5列，也叫target或label）。

【实验原理】

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。

一般而言，我们只选择样本数据集中前k个最相似的数据，这就是KNN算法中K的由来，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类别，作为新数据的分类。步骤如下：

1、计算测试数据与各个训练数据之间的距离

2、按照升序（从小到大）对距离（欧氏距离）进行排序

3、选取距离最小的前k个点

4、确定前k个点所在类别出现的频率

5、返回前k个点中出现频率最高的类别作为测试数据的分类

**关于k值的选取**

临近数，即在预测目标点时取几个临近的点来预测，K值得选取非常重要。

1、当K的取值过小时，一旦有噪声得成分存在们将会对预测产生比较大影响，例如取K值为1时，一旦最近的一个点是噪声，那么就会出现偏差，K值的减小就意味着整体模型变得复杂，容易发生过拟合。

2、如果K的值取的过大时，就相当于用较大邻域中的训练实例进行预测，学习的近似误差会增大。这时与输入目标点较远实例也会对预测起作用，使预测发生错误。K值的增大就意味着整体的模型变得简单。

3、K的取值尽量要取奇数，以保证在计算结果最后会产生一个较多的类别，如果取偶数可能会产生相等的情况，不利于预测。

4、常用的方法是从k=1开始，估计分类器的误差率。重复该过程，每次K增值1，允许增加一个近邻，直到产生最小误差率的K。一般k的取值不超过20，上限是n的开方，随着数据集的增大，K的值也 要增大。

【**实验要求**】

1.搜索资料掌握KNN算法的基本原理。

2.使用KNN算法编程解决iris数据集分类问题。

3.撰写实验报告。