计算机学院实验报告

|  |  |  |
| --- | --- | --- |
| 实验题目：kNN算法原理及实现 | | 学号：201600181073 |
| 日期： | 班级： 智能16 | 姓名： 唐超 |
| **Email：[1984386166@qq.com](mailto:1984386166@qq.com)** | | |
| 实验目的：  掌握kNN算法的原理，熟悉kNN算法。 | | |
| 实验环境介绍：  Python3.6 & jupyter notebook | | |
| 解决问题的主要思路：      kNN算法的核心思想是如果一个样本在特征空间中的k个最相邻的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，并具有这个类别上样本的特性。kNN方法在类别决策时，只与极少量的相邻样本有关。由于kNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，kNN方法较其他方法更为适合。  kNN算法不仅可以用于分类，还可以用于回归。通过找出一个样本的k个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。更有用的方法是将不同距离的邻居对该样本产生的影响给予不同的权值(weight)，如权值与距离成反比。 | | |
| 实验步骤：  kNN算法流程如下：计算已知类别数据集中的点与当前点之间的距离🡪按照距离递增次序排序🡪选取与当前点距离最小的k个点🡪确定前k个点所在类别的出现频率🡪返回前k个点所出现频率最高的类别作为当前点的预测分类.  接下来用例子验证一下kNN算法，随机挑选的6位高中生，分别让他们做文科综合试卷的分数和理科综合试卷的分数，下表为分数以及分类信息。    基于这些信息，我们利用kNN算法判断成绩为（105,210）所属的分类，运行结果如下显示“文科生”，输出结果比较符合预期。  关键实验代码见实验指导书。 | | |
| 实验结果展示及分析：  kNN算法的分析：  **1. 距离**   * 距离越大，表示两个点越不相似。距离的选择有很多，包括欧氏距离、曼哈顿距离、切比雪夫距离、闵可夫斯基距离、标准化欧氏距离、马氏距离、夹角余弦、汉明距离、杰卡德距离& 杰卡德相似系数、信息熵，通常用比较简单的欧式距离。   **2. 类别的判定**      投票决定：少数服从多数，近邻中哪个类别的点最多就分为该类。  加权投票法：根据距离的远近，对近邻的投票进行加权，距离越近则权重越大（权重为距离平方的倒数）  **3. k值**      k值选择过小，得到的近邻数过少，会降低分类精度，同时也会放大噪声数据的干扰；而如果k值选择过大，并且待分类样本属于训练集中包含数据数较少的类，那么在选择k个近邻的时候，实际上并不相似的数据亦被包含进来，造成噪声增加而导致分类效果的降低。k值通常是采用交叉检验来确定（以k=1为基准）。经验规则：k一般低于训练样本数的平方根。 | | |