

人工智能导论课程实验报告

|  |  |
| --- | --- |
| 实验名称： | 实验八  KNN算法 |
| 实验日期： | 2024-6-15 |
| 实验地点： | 西部片区4号楼301 |
|  | |
| 学号： | 33920212204567 |
| 姓名： | 任宇 |
| 专业年级： | 软工2021级 |
| 学年学期： | 2023-2024学年第二学期 |

1. **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。

1. **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。

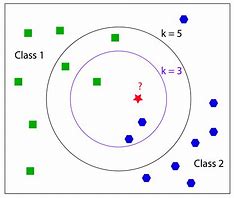
通俗地说，iris数据集是用来给花做分类的数据集，每个样本包含了花萼长度、花萼宽度、花瓣长度、花瓣宽度四个特征（前4列），我们需要建立一个分类器，分类器可以通过样本的四个特征来判断样本属于山鸢尾、变色鸢尾还是维吉尼亚鸢尾（这三个名词都是花的品种）。iris的每个样本都包含了品种信息，即目标属性（第5列，也叫target或label）。

1. **实验过程**
2. 搜索资料掌握KNN算法的基本原理：

k近邻法 (k-nearest neighbor, k-NN）是一种基本分类与回归方法。是数据挖掘技术中原理最简单的算法之一，核心功能是解决有监督的分类问题。KNN能够快速高效地解决建立在特殊数据集上的预测分类问题，但其不产生模型。

在分类任务中，算法的工作原理如下：

* 确定参数K的值：K是指算法中考虑的最近邻居的数量。
* 计算距离：对于每个预测的数据点，计算其与训练集中所有点之间的距离。常用的距离计算方法包括欧氏距离、曼哈顿距离等。
* 找到最近的K个邻居：从训练集中选择与新数据点距离最近的K个点作为最近邻。
* 进行投票或平均：在分类任务中，根据这K个最近邻居的标签进行多数投票，最常见的类别标签被赋予新数据点。



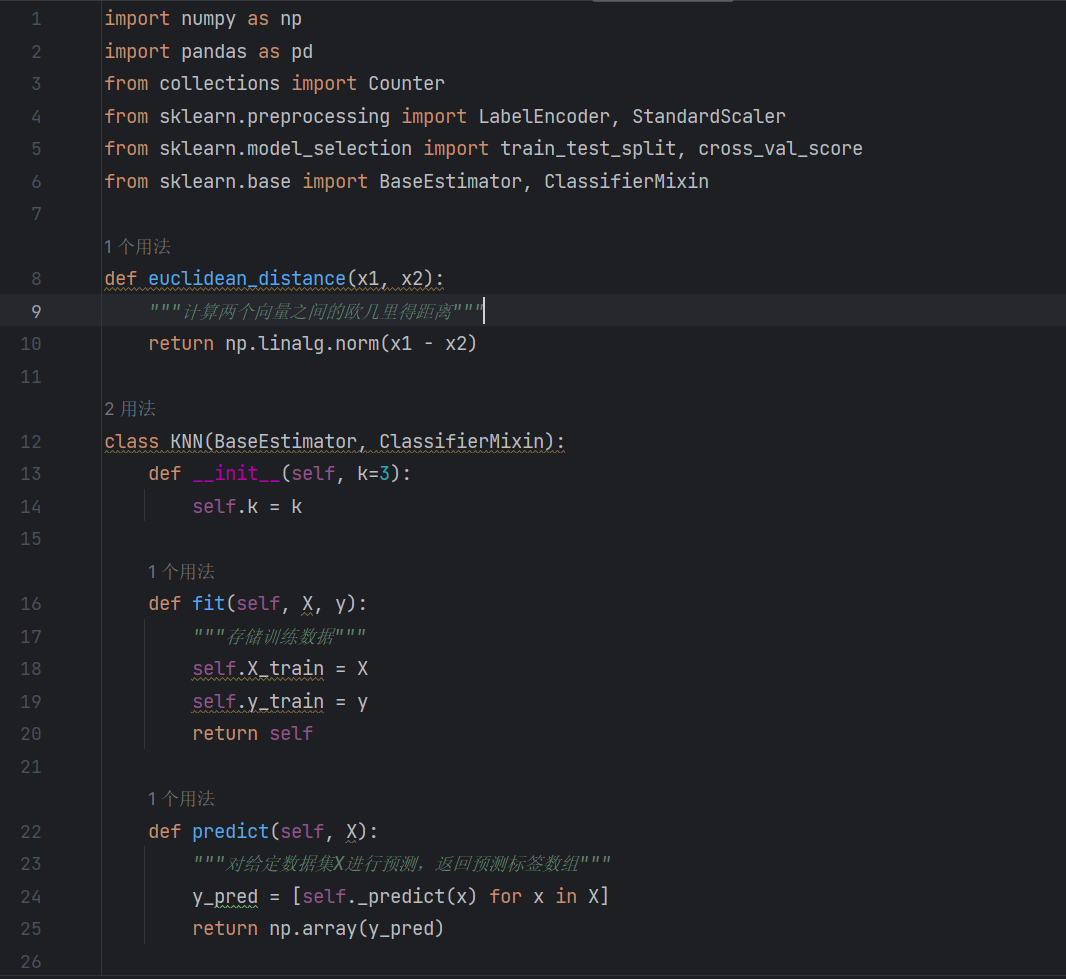
1. 使用KNN算法编程解决Iris数据集分类问题：

Iris数据集使用的是实验六中分割好的测试集和训练集，分别有75个样本。代码首先对数据进行加载和预处理，包括标准化和标签编码。然后通过继承sklearn的类，自定义了一个KNN模型，能够存储训练数据并对新数据点进行预测。为了提高模型的性能，代码使用交叉验证来找到最佳的K值。

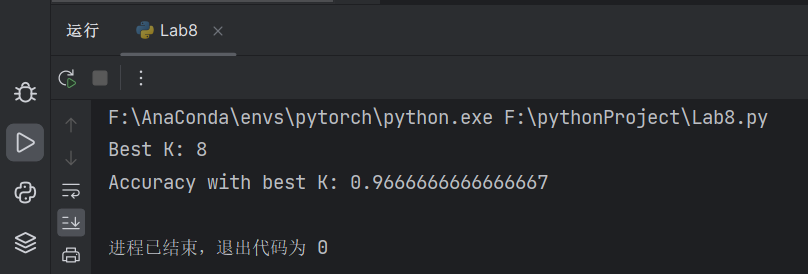
KNN类：

* \_\_init\_\_：初始化方法，设置邻居的数量k。
* fit：存储训练数据，为后续的预测做准备。
* predict：对输入的数据集X进行预测，通过遍历调用\_predict来处理每个单独的数据点。
* \_predict：计算一个数据点与所有训练数据点的距离，找到最近的k个邻居，并通过多数投票原则确定最终的预测标签。

具体代码如下所示：



1. 运行结果：



1. **实验思考及心得**

通过这次实验，我深刻体会到了K-最近邻（KNN）算法的原理及其在实际分类问题中的应用。通过对Iris数据集的操作，我不仅学习了如何实现和调试KNN算法，还对如何处理和预处理数据有了更深的理解。KNN算法的核心在于找出一个新样本的最近邻居，并根据这些邻居的类别信息来预测新样本的类别。这种基于实例的学习方法简单而直观，使我对分类任务有了更加直接的理解。