第 5 章 鸢尾花分类项目作业要求

构建一个鸢尾花分类模型，可以从这些已知特征和类别的鸢尾花测量数据中进行分类学习，从而能够根据一朵鸢尾花的四个特征预测这朵花所属的类别。

参考算法—K 最邻近分类算法（也可以使用其他分类算法）

K 最邻近（KNN，K-NearestNeighbor）分类算法思路在分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别，算法流程如下：

（1） 数据预处理，例如数据集打乱、种类维度的数据变换。

（2） 将数据集分训练集和测试集。

（3） 计算测试样本点即待分类点到训练接中每个样本点的距离，使用前 4 个维度计算。采用欧式距离计算两个样本点的距离，距离公式为：

dis=$\sqrt{\sum_{i=0}^{DIMENSION-1}(x.data[i]-y.data[i])*(x.data[i]-y.data[i])}$，其中 DIMENSION=4，x 为测试集中的某个样本，y 为训练集中的某个样本，data 数组中存放了样本的四个属性特征值。

（4） 对每个距离进行排序，选择出与测试样本点距离最小的 K 个点。

（5） 排序后对距离最小的 K 个点所属的类别进行统计，根据少数服从多数的原则，将测试样本点归入在 K 个点中类比占比最高的那一类。

（6） 最后计算测试样本被分类的准确度。使用以下公式计算分类的准确度：

准确率=（预测准确的个数/总测试样本数）。

数据集

列数	属性名	解释
1	sepalLength	花萼长度，单位是 cm;
2	sepalWidth	花萼宽度，单位是 cm;
3	petalLength	花瓣长度，单位是 cm;
4	petalWidth	花瓣宽度，单位是 cm;
5	species	种类，1 代表 Iris Setosa（山鸢尾），2 代表 Iris Versicolor（变色鸢尾），3 代表 Iris Virginica（维尔吉尼亚鸢尾）

评分要求：

1、完成数据读入（20 分）；

2、完成数据集预处理（20 分）；

3、完成数据集切分，训练集和测试集（10 分）；

4、完成测试集分类计算（40 分）

5、完成预测准确度评估（10 分）