黑色的仙人掌

中度可信度描述已自动生成

**计算机与信息学院**

**《人工智能》课程作业**

**2024年秋季学期**

|  |  |
| --- | --- |
| 课程类型： | 专业核心课 |
| 学 号： | 202210120518 |
| 姓 名： | 胡国昌 |
| 专 业： | 计算机科学与技术 |
| 授课教师： | 臧兆祥 |

完成日期：2024年 10月 26日

**K-means聚类**

# 案例内容与要求

## 实验内容与任务

现在有一批鸢尾花的数据，共包含150个样本，每个样本有四个属性，Sepal Length（花萼长度），Sepal Width（花萼宽度），Petal Length（花瓣长度），Petal Width（花瓣宽度）。同时，每个样本所属类别也已经标出，一共有3个类别：Iris Setosa（山鸢尾）、Iris Versicolour（杂色鸢尾），以及Iris Virginica（维吉尼亚鸢尾）。要求学生根据样本的属性数据将鸢尾花用K-Means算法进行聚类，获得3个类别，并将每个样本分到一个类别中。然后将聚类所得的样本类别分布情况与原始数据中的样本类别分布情况进行对比，分析K-Means算法的性能。

## 实验过程及要求

1. 实验环境要求：Windows/Linux操作系统，Python编译环境，numpy，matplotlib等程序库。
2. 加载鸢尾花数据集，观察数据集特征。
3. 实现K-Means算法,运行并观察聚类结果。
4. 研究初始聚类中心的设置对K-Means算法收敛性的影响。
5. 研究参数K对聚类结果的影响。

# 原理论述及解决方法

## 原理概述

聚类是人类挖掘知识的重要手段，例如对自然界的生物进行类别和群体的划分。在商业活动中，对客户群体进行划分，能对客户特点进行分析并对不同群体进行针对性营销。在机器学习中，聚类属于无监督学习，直接在数据中挖掘类别关系。聚类跟有监督学习中的分类的区别是缺乏有标记的训练数据。聚类有两个任务，首先要确定将数据划分多少类，其次要将每个样本分到一个类别中。例如图1中的数据点，我们仅根据数据点的分布情况，可以考虑将其划分4个类，坐标比较相似的点处于同一个类中。完成聚类要基于两个原则：

1. 不同类别的样本之间相似性很小。
2. 同一类别的样本之间相似性很大。

图表, 散点图

描述已自动生成