**南京信息工程大学 人工智能**导论 **实验(实习)报告**

实验(实习)名称 人工智能导论 实验(实习)日期 2024.11.07 得分 指导教师 应龙

学院 计算机与网络空间安全学院专业 计算机科学与技术 年级 2022 班次 4 姓名张瑞晨 学号 202283290159

1. 实验目的

1. 掌握 kmeans聚类算法的主要思想和优化目标函数。

2. 掌握 kmeans聚类的算法步骤。

3.了解聚类的评价指标：基于有标签测试数据的评价指标和无标签测试数据的评价指标。

二、实验内容与步骤

1.K-means聚类

K-means聚类是一种非常流行且广泛使用的无监督学习算法，主要用于将数据集中的样本划分成K个聚类，其中每个样本都属于最近的均值（即聚类中心）所代表的聚类。K-means算法的目标是最小化每个点到其聚类中心的距离的总和，这个距离通常是欧几里得距离。

K-means算法的基本步骤：

初始化：

随机选择K个数据点作为初始的聚类中心（质心）。

分配：

将每个数据点分配给最近的聚类中心，形成K个聚类。

更新：

重新计算每个聚类的中心，即每个聚类中所有点的均值。

迭代：

重复步骤2和3，直到满足某个终止条件，如聚类中心的变化小于一个预设的阈值，或者达到预设的迭代次数。

终止：

当满足终止条件时，算法结束，每个数据点都有了所属的聚类。

K-means算法的关键点：

K的选择：K的值需要预先指定，且对结果有很大影响。太小可能导致多个数据点被分到同一个聚类中，太大则可能导致每个数据点都单独成为一个聚类。常用的方法包括肘部法则（Elbow Method）、轮廓分析（Silhouette Analysis）等来估计最佳的K值。

聚类中心的初始化：聚类中心的初始选择对最终结果有影响，因为K-means算法可能会收敛到局部最优解。为了减少这种影响，可以通过多次运行算法并选择最佳结果来提高稳定性。

距离度量：K-means通常使用欧几里得距离，但也可以使用其他距离度量，如曼哈顿距离等。

收敛性：K-means算法是迭代的，并且通常会收敛，但不保证找到全局最优解。

对噪声和异常值的敏感性：K-means对噪声和异常值比较敏感，因为它们可能会对聚类中心产生较大影响。

缩放敏感性：由于K-means依赖于距离计算，因此对数据的缩放敏感。在应用K-means之前，通常需要对数据进行归一化处理。

K-means的应用场景：

K-means聚类因其简单和高效而被广泛应用于市场细分、社交网络分析、图像分割、文档分类等领域。尽管它有一些局限性，但由于其计算效率和实现简单，K-means仍然是许多领域中首选的聚类方法之一。

**代码步骤：**

1. 导入必要的库：matplotlib.cm、matplotlib.pyplot、numpy、sklearn.cluster中的KMeans、sklearn.datasets中的make\_blobs、sklearn.metrics中的silhouette\_samples和silhouette\_score。

2.使用make\_blobs函数生成样本数据，这里生成了500个样本，2个特征，4个中心点，并且设置了一个随机状态以保证结果的可重复性。

3.定义了一个聚类数的范围range\_n\_clusters，包含了从2到6的聚类数。

4.对于每个聚类数，创建一个子图，其中包含两个轴：一个用于绘制轮廓图，另一个用于绘制实际的聚类结果。

5.对于每个聚类数，初始化一个KMeans聚类器，并使用fit\_predict方法对数据进行聚类。

6.计算并打印所有样本的平均轮廓系数。

7.计算每个样本的轮廓分数，并根据聚类标签对它们进行排序。

8.在轮廓图中，为每个聚类绘制填充区域，并在中间位置标记聚类编号。

9.在轮廓图中添加平均轮廓系数的垂直线。

10.在第二个子图中，使用散点图显示聚类结果，并在聚类中心绘制白色的圆圈。

11.设置子图的标题和轴标签。

12.使用plt.show()显示所有的子图。

**用了几个关键的函数和类，每个都有其特定的功能和输入输出。以下是它们的总结：**

make\_blobs：

功能：生成指定数量的样本和特征的随机数据集，这些数据集围绕指定数量的中心点（聚类中心）分布。

输入：样本数量（n\_samples）、特征数量（n\_features）、中心点数量（centers）、中心点的方差（cluster\_std）、中心点的边界（center\_box）、是否打乱数据（shuffle）和随机状态（random\_state）。

输出：生成的数据集（X）和对应的标签（y）。

KMeans：

功能：KMeans聚类算法的实现，用于将数据集分成指定数量的聚类。

输入：聚类数（n\_clusters）、初始化次数（n\_init）和随机状态（random\_state）。

输出：聚类中心（cluster\_centers\_）和每个样本的聚类标签（cluster\_labels）。

silhouette\_score：

功能：计算所有样本的平均轮廓系数，用于评估聚类效果。

输入：数据集（X）和样本的聚类标签（cluster\_labels）。

输出：平均轮廓系数。

silhouette\_samples：

功能：计算每个样本的轮廓系数。

输入：数据集（X）和样本的聚类标签（cluster\_labels）。

输出：每个样本的轮廓系数数组。

plt.subplots：

功能：创建一个图形窗口，并在其中创建子图。

输入：子图的行数和列数。

输出：图形对象和子图的轴对象。

算法功能模块设计的重要思想：

数据可视化：通过轮廓图和散点图直观地展示聚类效果，使得非专业人士也能评估聚类算法的性能。

参数优化：使用轮廓分析来选择最佳的聚类数（n\_clusters），这是一种无监督学习方法中常见的参数优化手段。

模块化设计：代码将数据生成、聚类、评估和可视化分开处理，使得每个部分都可以独立使用和修改，提高了代码的可维护性和可扩展性。

重复性和可复现性：通过设置随机状态（random\_state），确保了结果的可重复性，这对于科学研究和实验是非常重要的。

性能评估：轮廓系数提供了一个量化的指标来评估聚类的质量，这比仅仅依赖于视觉评估要准确得多。

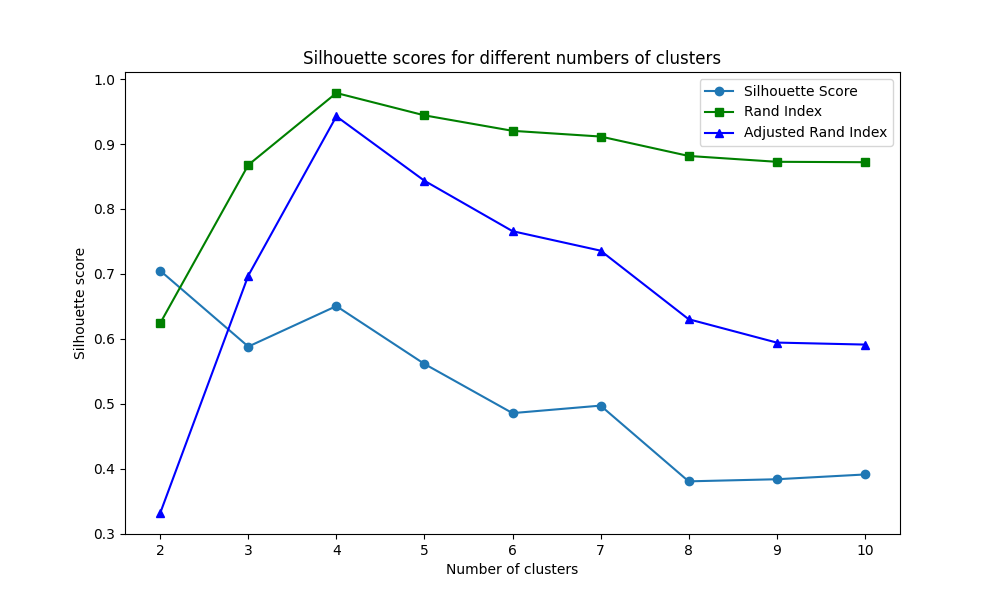
用户交互：虽然这个脚本是自动化运行的，但是它的设计允许用户通过修改输入参数来探索不同的聚类效果，增加了用户与程序的交互性。

**实验结果：**

**1.聚类簇数2-10的结果图像化**

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. **2到10不同 k 值选取的 silhouette\_avg 曲线**



1. **对应的 Rand index 和 adjusted Rand index 的值。**

For n\_clusters = 2, silhouette\_avg: 0.705, Rand index: 0.624, Adjusted Rand index: 0.332

For n\_clusters = 3, silhouette\_avg: 0.588, Rand index: 0.868, Adjusted Rand index: 0.698

For n\_clusters = 4, silhouette\_avg: 0.651, Rand index: 0.979, Adjusted Rand index: 0.943

For n\_clusters = 5, silhouette\_avg: 0.561, Rand index: 0.944, Adjusted Rand index: 0.843

For n\_clusters = 6, silhouette\_avg: 0.486, Rand index: 0.920, Adjusted Rand index: 0.766

For n\_clusters = 7, silhouette\_avg: 0.497, Rand index: 0.912, Adjusted Rand index: 0.736

For n\_clusters = 8, silhouette\_avg: 0.381, Rand index: 0.882, Adjusted Rand index: 0.630

For n\_clusters = 9, silhouette\_avg: 0.384, Rand index: 0.873, Adjusted Rand index: 0.594

For n\_clusters = 10, silhouette\_avg: 0.391, Rand index: 0.872, Adjusted Rand index: 0.591

三、实验心得

在进行了KMeans聚类和轮廓分析的实验后，我有以下几点心得体会：

理论与实践的结合：通过这次实验，我更加深刻地理解了KMeans聚类算法的理论基础和实际操作过程。实验不仅仅是对理论的验证，更是对理论知识的深化和拓展。

参数选择的重要性：在KMeans聚类中，选择合适的聚类数（K值）对于结果至关重要。通过轮廓分析，我们可以更直观地评估不同K值对聚类效果的影响，这在实际应用中非常有用。

数据可视化的力量：轮廓图和散点图的绘制让我直观地看到了数据的分布和聚类结果，这种视觉化的方法对于理解数据和解释聚类效果非常有帮助。

算法评估的多维度：除了轮廓系数，Rand index和adjusted Rand index也是评估聚类效果的重要指标。它们从不同角度衡量聚类的质量，让我对聚类算法的评估有了更全面的认识。

实验操作的挑战：在实验过程中，我遇到了一些挑战，比如参数调整和结果解释。这些挑战让我意识到实际操作中需要灵活运用理论知识，并且要有耐心和细心去调试和优化。

创新思维的培养：在面对实验中的难题时，我学会了如何思考问题的本质，并尝试创新的方法来解决问题，这对于我的科研能力和创新思维的培养非常有益。