机器学习聚类实验报告

1 课题综述

1.1 课题说明

成员名单及分工（按学号排序）：

学号：2151531　姓名：栾佳浩　任务：手写实现 K-Means、实现均值漂移

学号：2152486　姓名：刘翌帆　任务：手写实现层次聚类、实现高斯混合模型

学号：2152496　姓名：郭桉齐　任务：数据集选择、数据分析与预处理

学号：2154312　姓名：郑博远　任务：手写实现 DBSCAN、实现亲和传播

1.2 课题目标

本实验旨在运用传统机器学习方法，对 Country Data 数据集进行聚类任务。选取了六种传统聚类算法，包括 K-Means、DBSCAN、层次聚类（Agglomerative Clustering）、均值漂移（Mean Shift）、亲和传播（Affinity Propagation）以及高斯混合模型（Gaussian Mixture），采用了 Min Max 归一化和 PCA 主成分分析进行数据降维。

本次实验的主要目标是对上述聚类算法在"Country Data"数据集上的表现进行分析与评估，并比较它们在无监督学习任务中的优劣。在本次聚类实验中，我们还将关注算法在数据集中的聚类效果、对数据集结构的适应能力，以及对异常值和噪声的鲁棒性。我们希望通过对比实验结果，深入分析各聚类算法在"Country Data"上的性能差异，为无监督学习任务的模型选择和解释提供实际可行的建议，从而为未来在类似任务上的研究和应用提供有益的经验和指导。

1.3 课题数据集

Country Data 是一个用于聚类任务的公共数据集（https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data）。该数据集有 10 列数据，包括国家、儿童死亡率、出口、卫生支出、进口、人均收入、通货膨胀率、新生儿预期寿命、总生育率和人均 GDP。出口、进口、通货膨胀率和人均 GDP 提供了国家经济信息的估计，儿童死亡率、卫生支出、新生儿预期寿命和总生育率与国家社会信息相关，收入是一个同时包含社会和经济信息的参数。我们的任务是基于这些特征找出国家的社会经济状况。我们希望能够找到一种模式，从而较好地将相似的国家进行聚类，并将需要 HELP 国际援助的国家进行分类。

2 实验报告设计

2.1 数据准备

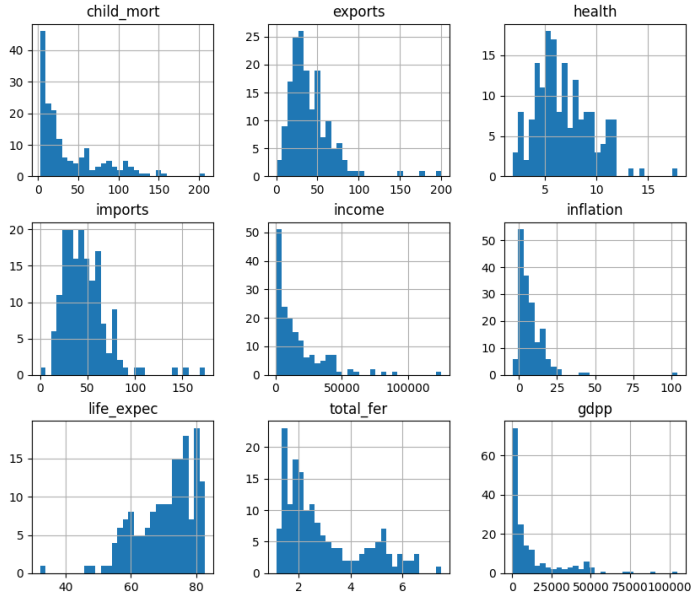
首先访问 KAGGLE（https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-data），进行数据集获取下载。数据集共 10 列，167 行，通过 info()的方法输出数据特征，发现没有空缺项，且除 country 特征外，其余特征均是数值特征。

2.2 数据预处理

首先通过输出数据集各特征的基本信息来检查缺失项等，然后通过分析我们 drop 了 country（国家名字）这一特征，将剩余数值特征做 MinMaxScaler 的特征缩放，最后通过 PCA 主成分分析，将数据集特征降维为五维。

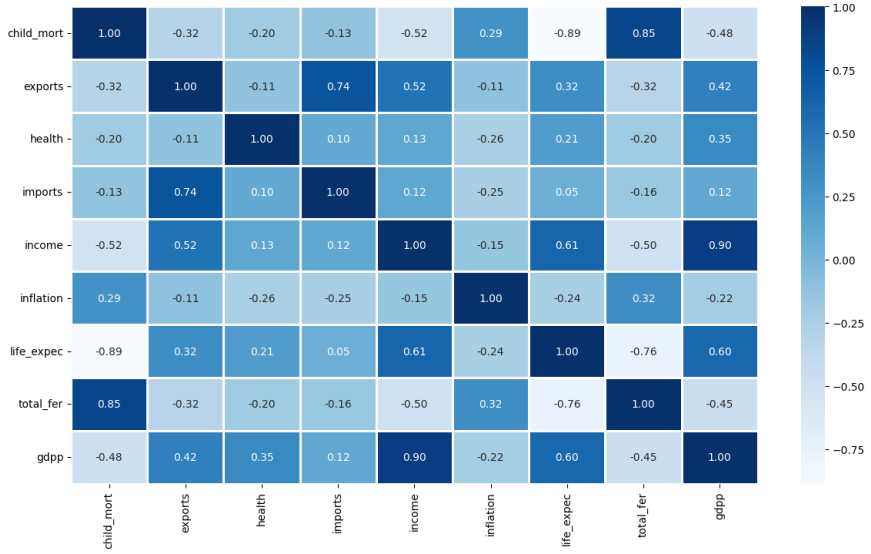
2.2.1 数据分析

首先通过分析，只有 country 特征是非数值特征；且根据特征含义分析，这一特征对聚类任务的帮助不大，因此选择 drop 此特征的列。然后我们对其余九个数值特征做分析，输出如下图所示的条形图，发现大部分数据呈现偏斜分布。唯一一个呈左偏的特征变量是寿命预期，这表明大多数国家已经解决了这个问题，除此之外，大多数特征变量的分布右偏，表示其中大部分存在右侧的异常值。因此可以考虑对其做标准化处理。



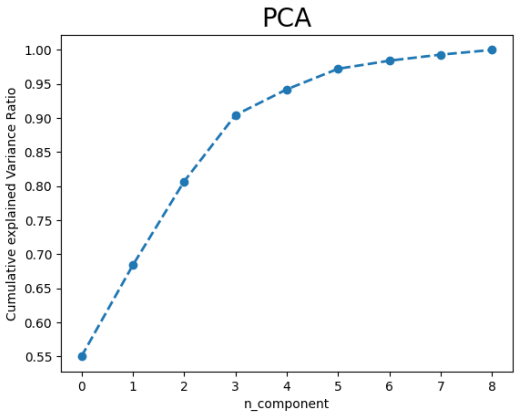
接下来输出九个数值特征的热力图，结果显示，变量之间存在着一系列相关性。儿童死亡率与总生育率（预期）、出口与进口、人均 GDP 与收入（预期）之间呈现出明显的正相关关系。而儿童死亡率与预期寿命（预期）、总生育率与预期寿命（预期）之间表现出较强的负相关性。此外，其他一些正相关关系包括出口与收入、出口与人均 GDP、卫生支出与人均 GDP（富裕国家在卫生支出上更为充裕）、收入和预期寿命（收入较高时，儿童的预期寿命较长），表明了这些因

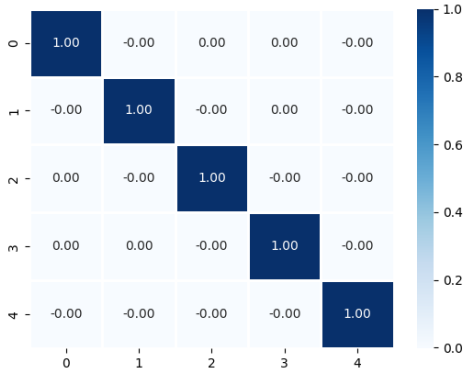
素之间存在着一定的关联。相反，儿童死亡率与收入（贫困国家儿童死亡率较高）、儿童死亡率与人均 GDP、出口与儿童死亡率、通货膨胀率与卫生支出（通货膨胀较高的国家，即经济受到影响，卫生支出较少）等方面呈现负相关性。针对以上分析的相关性，我们采用 PCA 的方法降维。



2.2.2 数据标准化

使用 sklearn.preprocessing 中的 MinMaxScaler 库。这是一种常用的特征缩放方法，用于将数据缩放到[0,1]的范围内。这样的缩放操作有助于确保不同特征的数值范围相对一致，避免某些特征对模型产生不必要的影响。特别是在使用基于距离的算法时，特征的数值范围对模型的性能影响较大，因此缩放是一个常见的预处理步骤。在这里，我们希望对数据分布范围有明确的要求，并尽可能保留原始数据的分布形状，因此选择 MinMaxScaler 而不是 StandardScaler 的方法。





2.2.3 PCA 主成分分析

我们采用 PCA 主成分分析对特征进行降维，旨在保留数据中的主要变化，并减少数据的维度。首先我们做出累计解释方差比关于主成分数量变化的曲线图（如上图左所示），发现当主成分数量是 5 时，已可以保留 97%的特征，如果采用更多的主成分数量对于精度的提高不大，但会

增加数据维度，消耗更多计算资源，综合考虑选择主成分数量为 5。采用 IPCA 的方式对新数据进行在线更新，画出新数据集特征的热力图（如上图右所示），很好地解决了数据相关性的问题。

2.3 模型搭建

我们使用了 6 种传统聚类机器学习模型，包括 K-Means、DBSCAN、层次聚类（Agglomerative Clustering）、均值漂移（MeanShift）、亲和传播（Affinity Propagation）以及高斯混合模型（Gaussian Mixture），下面具体进行介绍：

2.3.1 K-Means

我们实现了手写的的 K-Means 聚类算法。算法的核心步骤具体如下：首先，从数据点中随机选择初始簇中心。主循环迭代执行以下步骤：先将样本分配给簇，计算每个数据点与每个簇中心之间的距离；然后计算每个簇的新中心，即分配给该簇的所有样本的平均值。如果簇中心不再变化则提前结束迭代，K-Means 聚类结束。

2.3.2 DBSCAN

我们手写实现了 DBSCAN 算法。DBSCAN 的算法主体是用于聚类的 fit 函数，其负责执行聚类的核心逻辑，大致如下：遍历数据集中的每个数据点，并对于每个未被访问过的点将其标记为已访问，然后找到其邻域内的数据点。邻域内点的数量时如果小于 min_samples，将当前点标记为噪声（-1）；如果邻域内的点数量足够多，则调用_expand_cluster 方法进行聚类扩展。

在_expand_cluster 方法中，将当前点标记为当前聚类标签，并使用循环来遍历邻域内的点。对于每个未被访问的点，我们将其标记为已访问，并寻找其新的邻域内的点。如果新邻域内的点数量满足条件，则将该点加入核心点下标，并将新邻域内的点合并到当前邻域内。为了寻找邻域内的点，我们还实现了_find_neighbors 方法，从而实现计算给定数据点到其他数据点的距离，并找到距离小于等于 eps 的邻域内的点。

2.3.3 层次聚类

我们手写实现了 Agglomerative Clustering 算法。在初始化方法中，我们允许用户指定簇的数量，如果不提供，默认为 2。同时，我们初始化了存储每个样本所属簇的标签（labels_），样本间距离的矩阵（distances），以及最终形成的簇的标识（clusters）。

聚类方法（fit_predict）的实现过程为：接受样本矩阵作为输入，初始时将每个样本视为一个独立的簇，然后通过迭代合并最近的簇，不断更新簇的标签，直到达到指定的簇数量。在每次迭代中，找到距离最近的两个簇，将其中一个合并到另一个，然后更新距离矩阵。最终，每个样本被分配到一个最终形成的簇中，返回样本所属的最终簇的标识。

在算法中还涉及到更新距离矩阵的方法（update_distances），算法在每次合并簇后，调用该方法更新距离矩阵。流程为遍历距离矩阵的行，将合并后的簇的距离更新为与其他簇的最小距离，然后将合并前的簇在距离矩阵中的行和列的距离设为无穷大，表示它们之间的距离不再考虑。

2.3.4 均值漂移

均值漂移是一种非参数化的聚类算法，用于发现数据中的潜在簇结构。与 K-Means 不同，均值漂移不需要预先指定聚类的数量，而是根据数据分布自动确定簇的数量。均值漂移的关键思想是通过在数据分布中寻找局部密度最大的区域来找到簇。它的优势在于对于不规则形状和大小的簇能够更好地适应，并且不需要事先知道簇的数量。算法的输出结果是每个数据点所属的簇标签。

2.3.5 亲和传播

亲和传播聚类是一种无需预先设定聚类数量的算法，它通过数据点之间的消息传递来自适应地发现聚类结构。亲和传播聚类具有独特的自适应性，即它不需要预设聚类数量，而是通过数据点之间的动态消息传递发现数据集的内在结构。

在迭代过程中，亲和传播的每个数据点根据其他点的可用性和自身与其他点的相似度更新责任信息。同时，利用其他点对其的责任信息和自身的可用性信息更新可用性矩阵。这种消息传递过程不断迭代，直到达到停止准则，例如最大迭代次数或聚类结果收敛。最后，根据更新后的责任和可用性矩阵，确定每个数据点的聚类结果。被选为聚类中心的点形成聚类，而聚类中心由具有较大正责任值的数据点确定。

2.3.6 高斯混合模型

高斯混合模型（Gaussian Mixture Model）通常简称 GMM，使用高斯分布作为参数模型，并使用了期望最大（Expectation Maximization，简称 EM）算法进行训练。当样本数据 X 是多维数据（Multivariate）时，高斯分布遵从下方概率密度函数：

$$P(x|\theta) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum|^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu)^T\sum^{-1}(x-\mu)}{2}\right)$$

其中，μ为数据均值（期望），$\sum$为协方差（Covariance），D 为数据维度。

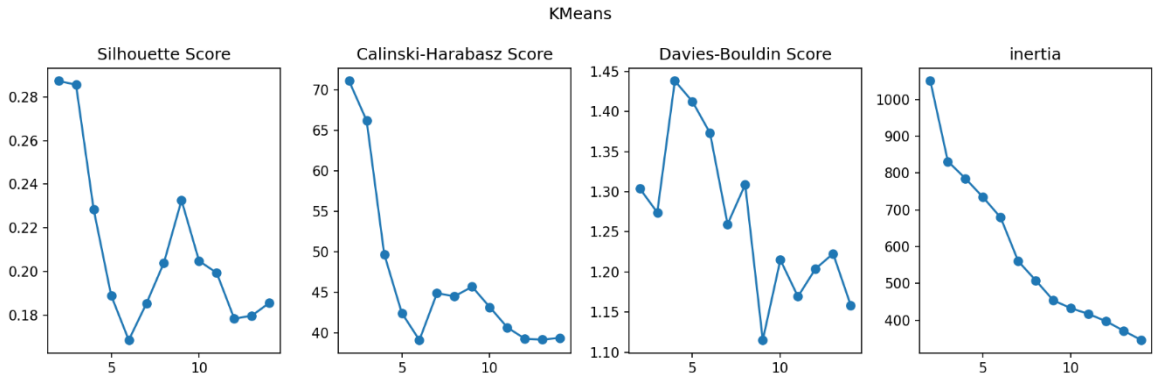
混合模型是一个可以用来表示在总体分布（distribution）中含有 K 个子分布的概率模型，其不要求观测数据提供关于子分布的信息，来计算观测数据在总体分布中的概率。高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型，这 K 个子模型是混合模型的隐变量（Hidden variable），其 Log-Likelihood 函数为：

$$\log L(\theta) = \sum_{j=1}^{N} \log \mathrm{P}\left(x_j|\theta\right) = \sum_{j=1}^{N} \log\left(\sum_{k=1}^{K} \alpha_k \phi(x|\theta_k)\right)$$

2.4 模型训练测试

2.4.1 K-Means

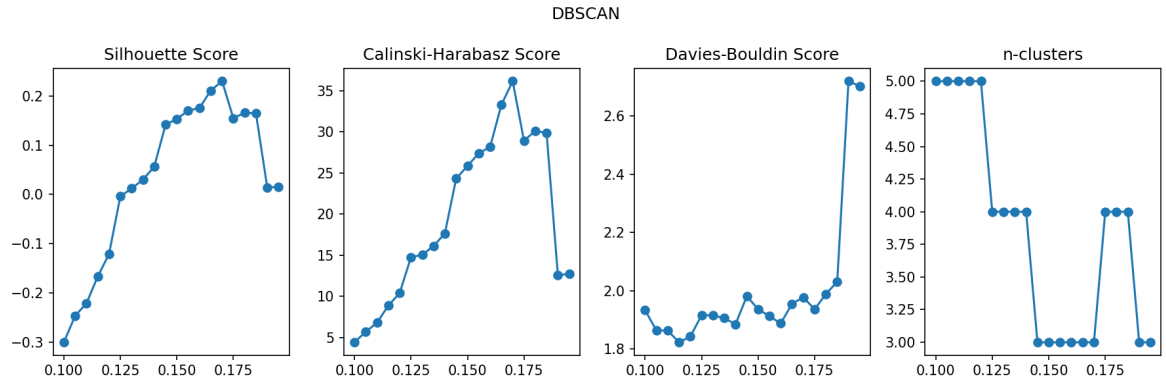
K-Means 比较重要的超参为聚类数量，根据题意，设计了从 2-15 数量的类别，并分别计算了 Silhouette Score,Calinski-Harabasz Score, Davies-Bouldin Score 以及 inertia，结果如下:



从图像中可以看出，k=3 时 Davies-Bouldin Score 表现最佳，同时 Silhouette Score、Calinski-Harabasz Score 均没有明显衰减，interia 与 k=2 时有明显提升，因此选择 k=3 作为聚类数量。

2.4.2 DBSCAN

DBSCAN 中有两个重要的参数，分别是邻域半径 eps 和最小样本数 min_samples。首先，我们选择不同的 min_samples 数，在此基础上绘制各个指标随领域半径 eps 的变化曲线。由于篇幅限制，此处仅展示 min_samples=4 时的变化曲线：



根据上述图像曲线可得，当 eps 小于 0.18 时，Davies-Bouldin Score 都维持在较低值。而当 eps 为 0.17 时，Silhouette Score、Calinski-Harabasz Score 都达到最高值，此时聚类数目为 3。

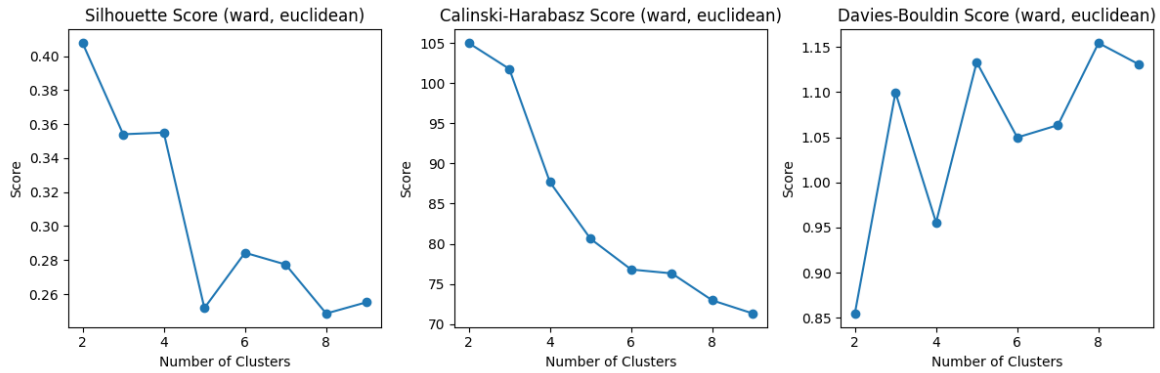
2.4.3 层次聚类

Agglomerative Clustering 算法中有两个参数可以用于调整，即链接方法（linkage methods）和相似性度量方法（affinity methods）。

linkage_methods 定义了三种链接方法，分别是 'ward'、'complete' 和 'average'。在层次聚类中，链接方法用于确定簇与簇之间的距离如何计算，以便在每一步中合并最近的两个簇。'ward' 使用 ward 方差最小化方法，'complete' 使用最远点法，'average' 使用平均链接法。

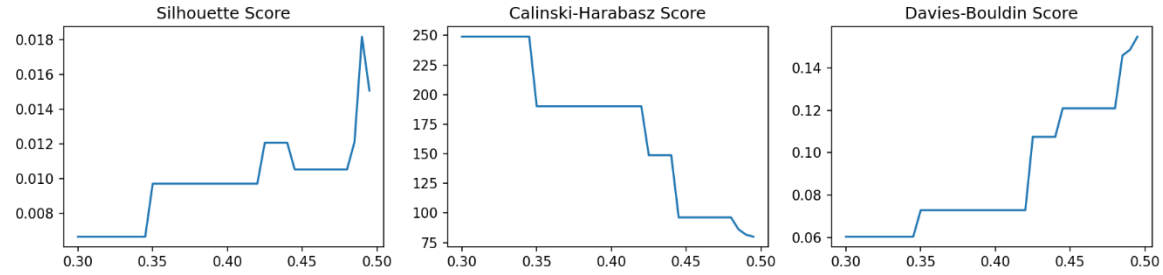
affinity_methods 定义了三种相似性度量方法，分别是 'euclidean'、'manhattan' 和 'cosine'。相似性度量方法用于计算两个样本之间的相似性。'euclidean' 使用欧几里德距离，'manhattan' 使用曼哈顿距离，'cosine' 使用余弦相似度。

根据每种参数的相互匹配，综合在不同聚类数目下三个指标的数据，我们选择聚类数目为 3，使用'ward'作为链接方法，'euclidean'作为相似性度量方法。



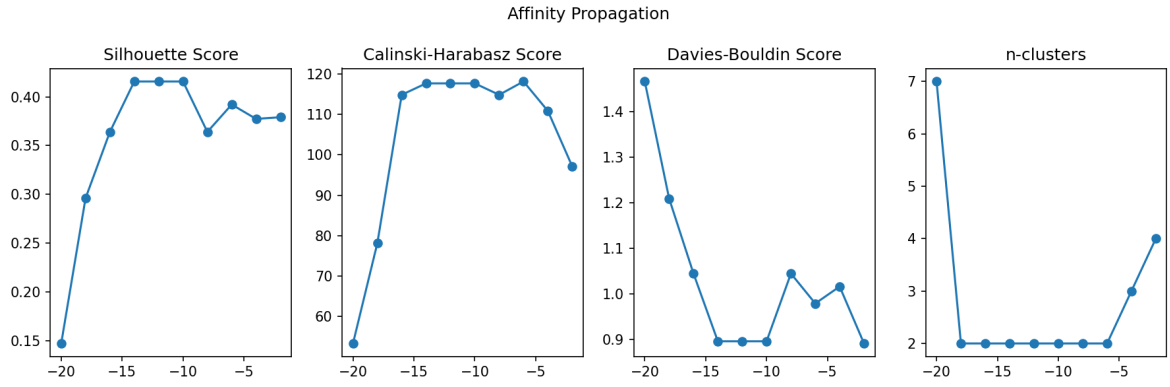
2.4.4 均值漂移

带宽是均值漂移算法的一个关键参数，决定了用于计算均值的邻域的大小。它实际上定义了数据点在寻找均值时考虑的范围。大带宽可能导致生成的簇过于平滑，而小带宽可能导致生成的簇过于细致。选择合适的带宽通常需要通过交叉验证或其他优化方法进行调优。实验中将带宽范围选择 0.3-0.5，步长为 0.01 进行测试，得出下图结果：



当带宽选择 0.375 时，综合三个参数都能获得相对不错的效果，所以选择 0.375 作为带宽。

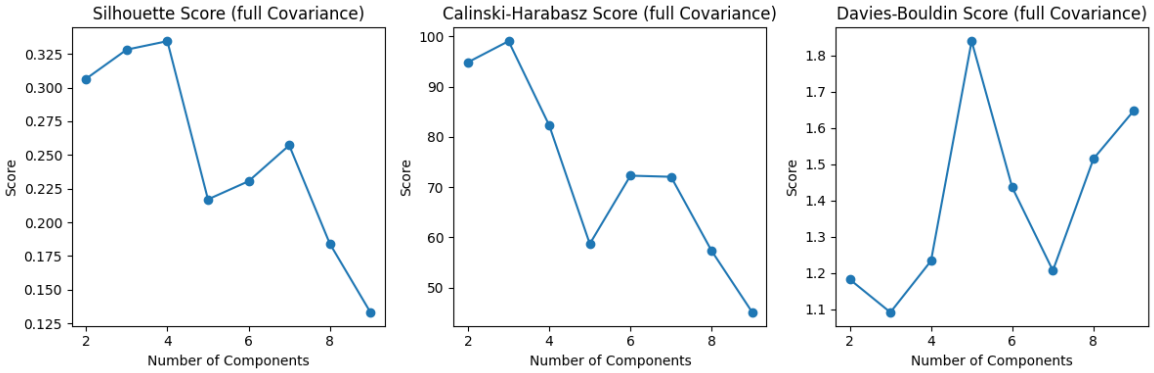
2.4.5 亲和传播



亲和传播中，主要有两个参数对聚类效果影响较大，分别是初始参考度（preference）和阻尼系数（damping factor）。阻尼系数主要用于避免在算法更新过程中出现振荡现象，因此不纳入训练测试的调参范围内，选用模型默认值 0.5。可以观察到，当 preference 取-10 时，Silhouette Score 与 Calinski-Harabasz Score 都维持在较高值，且此时 Davies-Bouldin Score 都维持在较低值。故选取 preference 为 10。

2.4.6 高斯混合模型

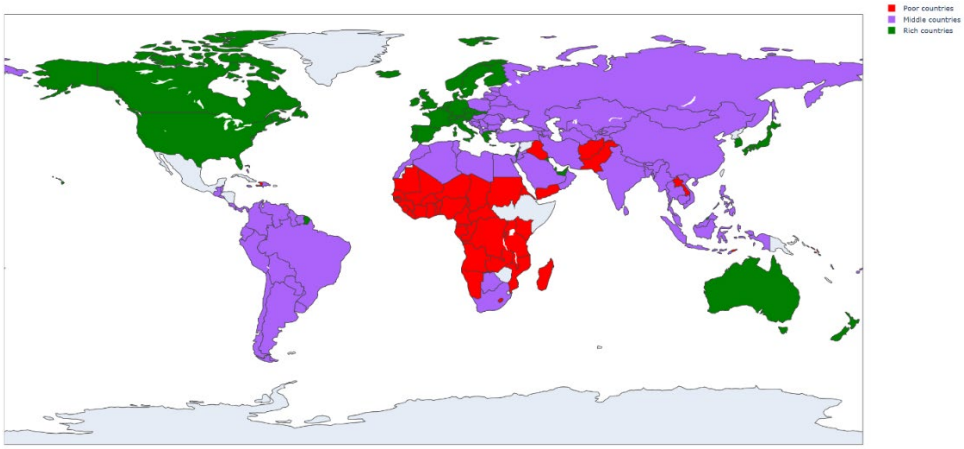
在高斯混合模型（Gaussian Mixture）中，我们主要针对不同聚类数量（n_components）和协方差类型（covariance_type）进行了评估，其中协方差类型包括'full'、'tied'、'diag'和'spherical'。通过保存每个参数组合下的评估指标，可以比较在不同配置下的效果，从而选择最优的模型配置。



根据每种参数的相互匹配，综合在不同聚类数目下三个指标的数据，我们选择聚类数目为 3，使用'full'作为协方差类型。

2.5 结果可视化

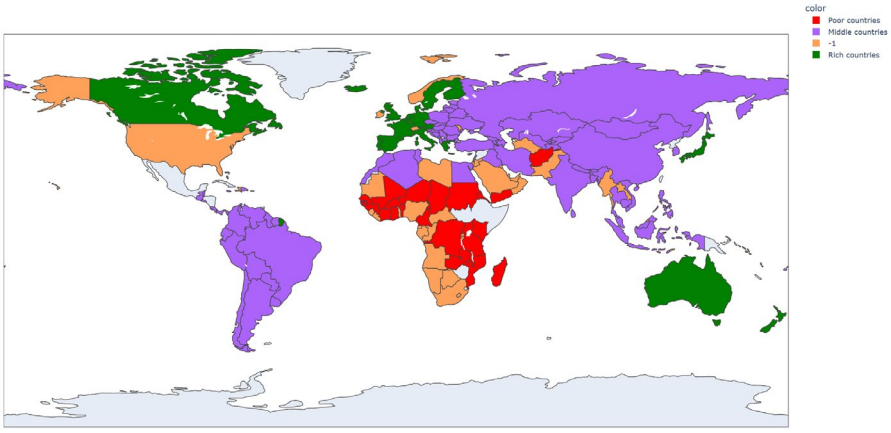
2.5.1 K-Means



由于聚类只会对数据进行分类，无法根据标签直观地体现效果，我们根据分出的几个类别，基于事实人为地将其打上发达国家、中等国家、贫穷国家等标签，并在世界地图上进行可视化。可以看出当 k=3 时 K-Means 的聚类效果较好，基本符合常识上的认知。

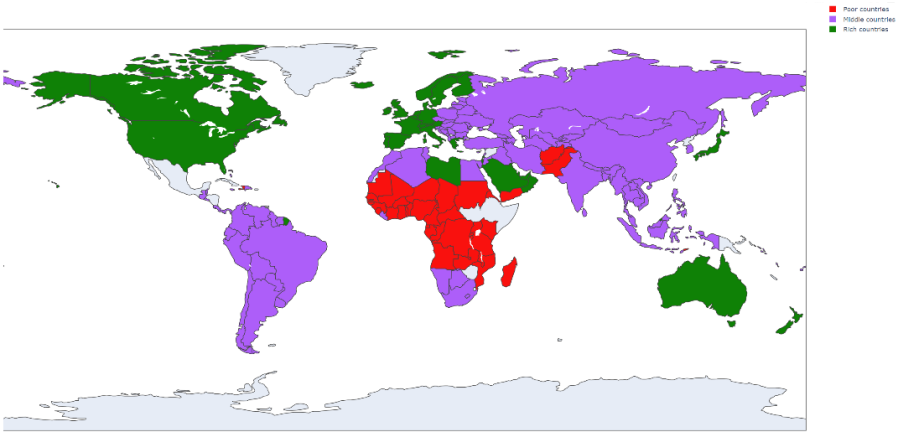
2.5.2 DBSCAN

通过下图，可以看到 eps=7，min_samples=4 时 DBSCAN 的聚类效果。其中，橙色部分（对应标签为-1）表示未被聚类的噪声点。可以观察到，对于国家的聚类效果整体较好，较为贴合如今各个国家的发展与经济状况。美中不足的是，对于美国以及非洲的部分国家，都直接纳入噪声点的范畴，没有成功地对这些国家进行聚类。



2.5.3 层次聚类

根据上述分析，选择聚类数目为 3，使用'ward'作为链接方法，'euclidean'作为相似性度量方法，可以得到如下图所示的各个国家发展情况。

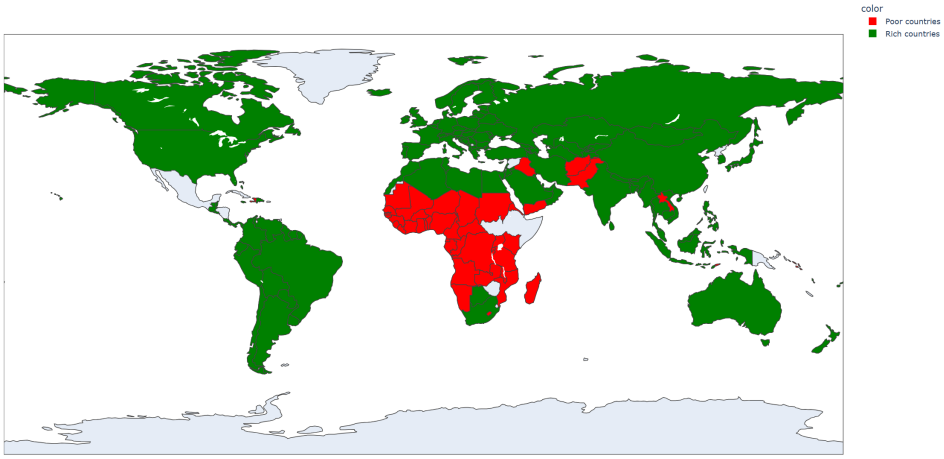


2.5.4 均值漂移

带宽选择 0.375 可以发现，基于密度的 Meanshift 算法虽然杂类较多、偶有错误，但也能较为准确地分出正确大类别。此处由于报告篇幅所限，不展示可视化图片。

2.5.5 亲和传播

根据上述的分析，选取 preference 参数为-10，此时聚类数目为 2。可以观察到，聚类的划分较为自然，比较符合各个国家的真实发展情况。



2.5.6 高斯混合模型

根据上述分析，选择聚类数目为 3，使用'full'作为协方差类型，可以得到各个国家发展情况，且较为符合现实情况。此处由于报告篇幅所限，不展示可视化图片。

2.6 分析和优化

通过上述实验，我们对多种聚类算法进行了深入分析与优化。我们综合考虑了如 Silhouette Score、Calinski-Harabasz Score 和 Davies-Bouldin Score 在内的各项指标，确定了每种算法的最优参数配置。例如，在 K-Means 中选择聚类数量为 3，DBSCAN 中确定 eps 为 0.17 和 min_samples 为 4，以及在高斯混合模型中选择聚类数目为 3 且协方差类型为'full'。上述优化结果能够很好地提升模型的准确性和效率。此外，我们也注意到如 DBSCAN 对噪声点处理欠佳、Meanshift 的聚类杂项较多等问题，可以通过进一步的数据处理与调参优化取得更好的结果。

3 结语

本实验的主要目标是应用传统机器学习方法进行聚类任务。在实验的数据准备阶段，我们使用了包含了国家社会与经济信息的"Country Data"数据集。在数据预处理中，我们分析并删除了非数值特征"Country"，随后进行了 MinMax 归一化和 PCA 主成分分析以降维数。

在模型搭建阶段，我们选择了六种传统聚类算法，包括手写实现的 K-Means、手写实现的 DBSCAN、手写实现的层次聚类、均值漂移、亲和传播和高斯混合模型。在模型训练和测试阶段中，我们对每个算法进行了详尽的训练和测试，并进行了参数优化。实验结果通过可视化展现，以直观呈现各算法在数据集上的聚类效果。通过分析可视化结果，我们对不同聚类算法的性能有了更为深入的理解，并为无监督学习任务的模型选择进行了优化与分析。

总体而言，我们通过本次实验，不仅对六种聚类算法在"Country Data"数据集上的性能进行了全面评估，还通过可视化结果深入分析了它们在不同情境下的表现。我们小组的同学通过应用上述传统聚类算法，很好地将机器学习课堂中学习到的理论知识融入到了实践之中。