**Introduction**

本次实验中，我首先使用高斯混合模型（GMM），k-means和层次聚类在经过处理的样本空间中对样本进行聚类拟合，使用网格搜索优化聚类超参数，并使用调整兰德系数（ARI）评估聚类结果。在得到较好的聚类结果后，再对每个簇中的样本分布进行研究。我以此评估聚类簇与学生所在Programme的关联，同时提出相关科学假设。为确保实验的可复现性，相关结果与代码保存在以下仓库中：<https://github.com/MushihimePepsi/XJTLU_ICS_Y2S2_Course-notes_23-24/tree/main/INT104-%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD>

**Task0- 在聚类之前——对聚类指标，可视化和输入特征的申明**

**考虑使用多于标签数的聚类簇数——子群体**

尽管我们往往使用等同于样本类别的数量作为聚类算法的簇数，以期得到可以与样本真实类别一一对应的聚类簇。然而，在实际的样本空间中，同一类别的样本可能存在更细化的子群体，存在不同的特征分布。例如，如果有 4 个标签表示4 种类型的水果，但每种水果中有多个不同的品种，使用多于标签数的聚类簇可能将这些品种分别识别来，并在最后使用多数投票法（Majority Voting）将每个簇分配给频次最高的类别标签，或使用匈牙利算法的分配成本最小化（Cost Minimization）对每个簇的所属类别进行分配。通过这一方式，我们可能获得更准确的聚类结果。但也要注意，这种情况下的聚类效果评估不应使用Accuracy或兰德指数，因为即使随机分配标签，这些评估值也会随着簇数量的增加而增加。作为替代，我们应当使用调整兰德系数（ARI）。

**对聚类评估指标的选择**

**-为什么选择ARI而不是Accuracy**

在先前的监督学习作业中，我们常常使用Accuracy作为评判模型性能的指标；但在属于无监督学习的聚类算法中，由于簇数可以不等于实际类别数，而且算法不能直接确定每个簇所属的类别标签，此时使用混淆矩阵或Accuracy作为评判标准显得不合时宜。此时我们可以选择归一化互信息（NMI）或调整兰德系数（ARI）评判聚类结果与真实标签的匹配程度。然而，随着聚类簇数量的增加，即使对每个簇随机分配标签，Accuracy，NMI等评估值也会随着簇数量的增加而增加。为惩罚随机分类和类别不平衡导致的性能虚高，我在此使用ARI这种考虑了随机分配标签期望值的指数，更可靠地反映了聚类结果的质量。

#RAI公式

越接近1，聚类结果越能反映真实标签。

**-轮廓系数，肘方法的失效**

在大多数情况下，一个簇应当拥有两个良好的性质：内聚和分离。这意味着在通常意义上，同一个簇中的样本都尽可能地靠近，同时尽可能地远离其他簇中的元素，这是因为我们默认相同类别的样本具有相似的特征分布。在这一启发式的影响下，轮廓系数等系数通过评判一个聚类生成的簇的内聚度和分离度，隐式地期望得到更能反映真实标签的聚类。然而在本次任务中，除了Programme=3的学生，其余样本在距离和密度上都是极其接近的。使用这一启发式的系数都无法有效评判本次聚类任务的性能。

#原始类别分布可视化

**-在无监督学习中划分训练集和测试集的必要性**

尽管在某些无监督学习任务中划分训练集和测试集和交叉验证的使用有助于评估模型的泛化能力，但在本次任务中，划分训练集和测试集的意义不大，尤其是在已经使用ARI评估聚类能力的情况下。此外，本实验中总样本数过少且面临类别不平衡的问题，即使使用欠采样或Synthetic Minority Over-sampling Technique (SMOTE)也会导致测试集无法全面反映聚类真实效果。总的来说，无监督学习更注重从数据中发现潜在的结构和模式，利用全部数据进行训练往往能更好地实现这一目标，提供更稳定和全面的评估。

**对聚类可视化和输入特征的选择**

由于本次实验使用到了基于距离的聚类方法，输入特征的规模不同会导致聚类算法对不同特征的重要性产生偏见。因此，在进行聚类之前，输入特征统一进行了Z-Score标准化以消除数据规模产生的影响。在本次实验中，为了方便研究不同特征对聚类结果的影响，输入特征均使用原始特征组，没有采用来自Locally Linear Embedding（LLE）或Principal Component Analysis（PCA）的数据。在可视化的过程中，我使用了PCA降维并简单地使用方差最大的两个方向作为X和Y轴。

**Task1- GMM (Gaussian Mixture Model)**

Gaussian Mixture Model的正确性来自于假设数据是由多个高斯分布的数据组成的，其中每个高斯分布作为一个簇（组件），拥有自己的均值向量和协方差矩阵。我通过调整超参数‘covariance\_type’控制每个簇的协方差矩阵的状态，发现在covariance\_type= {‘full’, ‘tied’, ‘diag’ }中， ‘diag’的聚类结果最好ARI: 0.16760，此时协方差矩阵是对角矩阵，即各特征间相互独立。这与先前在cw1，cw2中得到的结论相同，也就是说，每个特征对聚类的贡献是相对独立的，特征之间只有很小的线性相关性，而不同特征的分布不同。

#折线图

接下来通过网格搜索，我使用最好的超参数得到的聚类可视化与每个簇中的分布

#柱状图+饼图

**Task2- k-means**

K-means是一种基于距离的聚类方法，通过将每个点分配给最近的质心并迭代得到最终的聚类簇。本次实验中我使用欧氏距离作为聚类的距离定义，优化目标是最小化Sum of Squared Errors（SSE）直至收敛。鉴于这一方法对初始质心的选择很敏感，我使用了k-means++的方法初始化聚类中心，使得初始点更加分散。

#柱状图+饼图

K-means最高的的兰德系数在n=3时取到，但这是由于Programme=2的人过少的类别不平衡导致的，聚类放弃了对这一类别的指派。

#n图

**Task3- Hierarchical Clustering**

本次实验中我使用自底向上的凝聚型层次聚类。它也是基于距离的聚类方法，通过逐步合并聚类簇得到聚类树，从而允许生成不同粒度的聚类结果。

#柱状图+饼图

可以调整距离的定义，其中‘ward’表现最好；‘average’在n>8时的score上升是由于数据集的类别不平衡导致的，此聚类放弃了对Programme=2的样本的分类。

#n图

**Task4- Association between clusters and Programme**

**Conclusion & Scientific hypothesis**

经过上述多种聚类方式对数据集的聚类，可以发现，除了Programme=3的样本被较为准确地聚类，其余样本均难以准确地分入独立的簇中。这是由于Programme=3的样本具有Grade=3的特性，而其余样本的标签与数据分布的线性关系较小，这在CW2中已经被详细论述，同时，GMM最优参数的选择也反映出了这一点。从聚类结果上观察，除了Programme=1的样本有轻微内聚的趋势，Programme=2&4的样本并不具有显著的分布差异。

**Limitation**

在本次实验中，输入特征只使用标准化后的原始特征，使用单纯基于距离或密度的聚类算法虽然能提取高维的局部线性特征，但并不能提取出数据集中的非线性特征或非线性组合产生的新特征与标签的相关性。应当考虑通过可复现的非线性流形分析得到的特征进行聚类或使用Self-Organizing Map（SOM）发掘特征非线性组合与样本类别的深层关系。

**致谢**