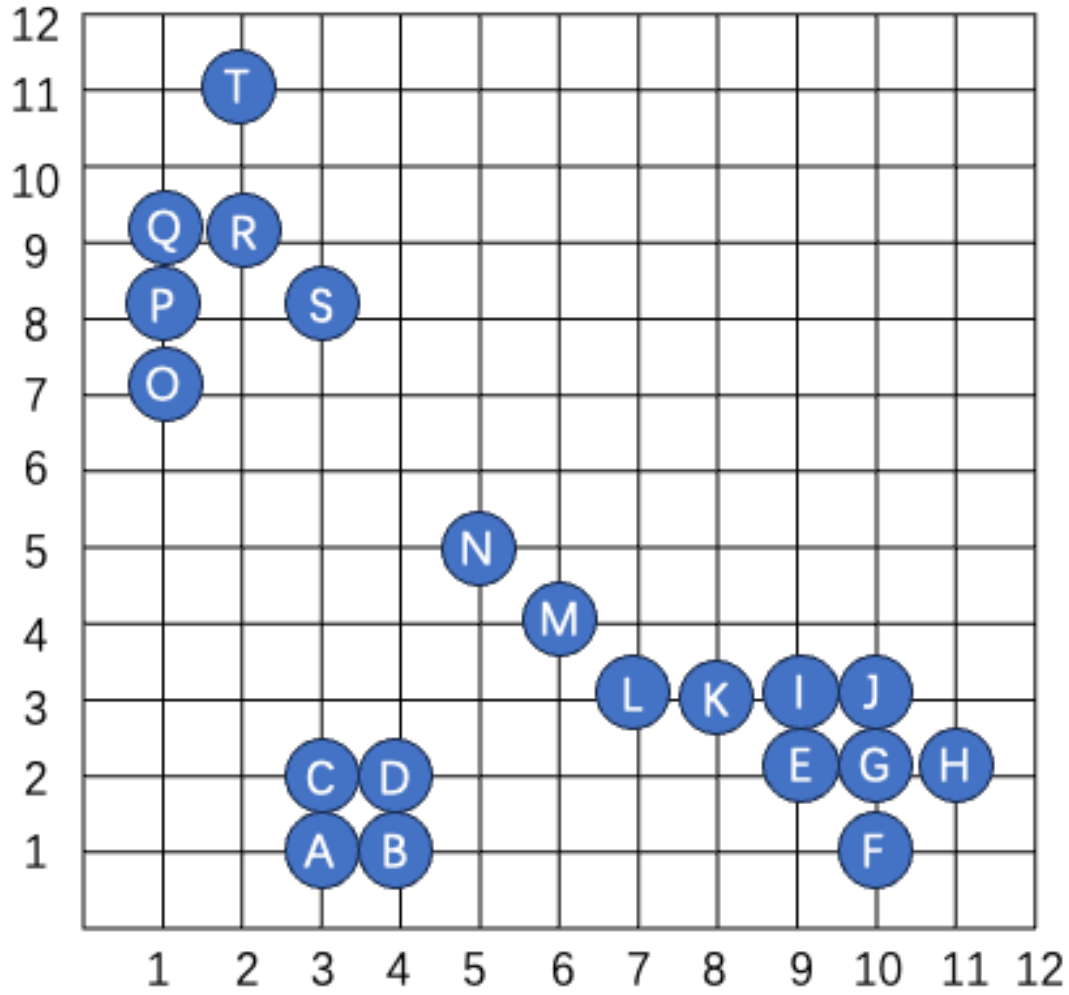
《计算机应用数学第二次作业-2025 秋》

截止时间：2025.12.22

计算题（30 分，每题 15 分）

1、基于如下数据集（二维），分别使用 Manhattan 和 Euclidean 距离，计算 DBSCAN 聚类结果，并标出 core points、border points 和 noise points。使用如下参数$\varepsilon = 1.1$，$minPts = 2$。并讨论两种距离的优劣。



2、在课上讨论的 Gaussian Mixture Model 和 Multinomial Mixture Model 的基础上，写出 Bernoulli Mixture Model 的定义和其对应 EM 算法的详细过程。

编程题 （70 分，20+20+30）

说明：建议使用开源工具包，例如 scikit-learn 中有朴素贝叶斯、高斯混合模型等函数实现， sknetwork 中有 PageRank 函数实现……

1、高斯混合模型与 EM 算法

数据集：Iris 数据集

数据描述：https://www.kaggle.com/datasets/uciml/iris，可通过 sklearn 直接导入数据集

```
from sklearn import datasets
iris = datasets.load_iris()
```

任务描述：使用高斯混合模型与 EM 算法对数据进行分类计算，mixture components 设置为 3。

要求输出：不同高斯分布的 mean 和 variance，每个高斯分布对应的权重，plot 出分布的图。EM 算法可以参考 https://scikit-learn.org/stable/modules/generated/sklearn.mixture.GaussianMixture.html。

Optional：尝试不同的 covariance structures，包括 spherical、diagonal、tied 与 full。

2、Node2vec

数据集：Node2vec_Dataset.csv。

任务描述：利用 Node2vec 计算每个节点的 embedding 值。

要求输出：1）每个节点的 embedding 值列表（csv 文件）；2）随机挑选 10 个 node pair，对比他们在 embedding 上的相似度和在 betweenness centrality 上的相似度（使用 Jaccard similarity）。

3、Clustering

数据集：使用 Make_blobs 生成数据不少于 1000 个 data points，以 3-5 个 cluster 为宜。https://scikit-learn.org/dev/modules/generated/sklearn.datasets.make_blobs.html#sklearn.datasets.make_blobs

任务描述：利用 DBSCAN 算法计算数据的聚类

要求输出：原始数据 plot 的图像和聚类后的结果。尝试不少于三组ε，$minPts$的参数组合。