实验五：层次聚类

Hierarchical Clustering

基本概念

- 定义：聚类就是对大量未知标注的数据集，按数据的内在相似性将数据集划分为多个类别，使**类别内的数据相似度较大**而**类别间的数据相似度较小**。

- 对象：观测数据或样本集合

- 核心概念：相似度 (Similarity) /距离 (Distance)

- 场景：图片检索：图片内容相似度
  图片分割：图片像素（颜色）相似度
  网页聚类：文本内容相似度
  社交网络聚类：（被）关注人群，喜好，喜好内容
  电商用户聚类：点击 / 加购 / 购买商品

相似性计算方法总结

- Pearson 相关系数: $$\rho_{ij} = \frac{\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{m}(x_{ki} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{m}(x_{kj} - \bar{x}_j)^2}}$$

- 余弦相似度 (Cosine similarity): $$s_{ij} = \frac{\sum_{k=1}^{m} x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{m} x_{ki}^2}\sqrt{\sum_{k=1}^{m} x_{kj}^2}}$$

越接近于 1，表示样本越相似；越接近于 0，表示样本越不相似。

用相似性度量时，相关系数越大样本越相似。

距离计算方法总结

样本集合为 $X = [x_{ij}]_{m\times n}$，$x_i = (x_{1i}, x_{2i}, \ldots, x_{mi})^T$，$x_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$

- 闵可夫斯基距离 (Minkowski)：$d_{ij} = \left(\sum_{k=1}^{m} |x_{ki} - x_{kj}|^p \right)^{\frac{1}{p}}$

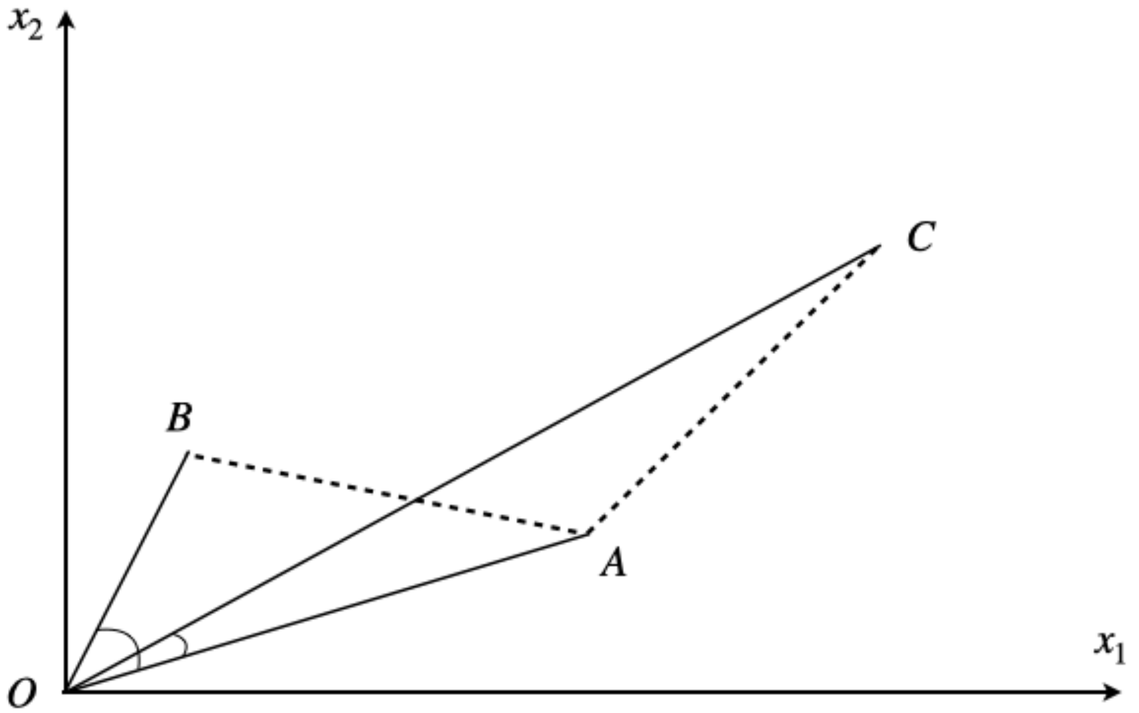
  p=2，欧式距离；p=1，曼哈顿距离；p=∞，切比雪夫距离

- 马哈拉诺比斯距离 (Mahalanobis)：$d_{ij} = \left[(x_i - x_j)^T S^{-1} (x_i - x_j) \right]^{\frac{1}{2}}$

  X 的协方差矩阵记作 S。S 为单位矩阵时，即样本数据的各个分量互相独立且各个分量的方差为 1 时，马氏距离就是欧式距离。

用距离度量时，距离越大相似度越小，距离越小相似度越大。

距离与相关系数



如果从距离的角度看，A 和 B 比 A 和 C 更相似；但是从相关系数的角度看，A 和 C 比 A 和 B 更相似。所以进行聚类时，选择合适的距离或相似度非常重要!

层次聚类

层次聚类假设类别之间存在层次结构，将样本聚到层次化的类中。层次聚类属于硬聚类，因为每个样本只属于一个类别。

- 聚合聚类 (bottom-up)

采用自底向上的策略，开始将每个样本各自分到一个类；之后将相距最近的两类合并，建立一个新的类，重复操作直到满足停止条件；得到层次化的类别

- 分裂聚类 (top-down):

采用自顶向下的策略，开始将所有样本分到一个类；之后将已有类中相距最远的样本分到两个新的类，重复操作直到满足停止条件；得到层次化的类别

聚合聚类

输入：n 个样本组成的样本集合及样本之间的距离；

输出：对样本集合的一个层次化聚类。

① 计算 n 个样本两两之间的欧式距离 $\{d_{ij}\}$，记作矩阵 $D = [d_{ij}]_{n\times n}$

② 构造 n 个类，每个类只包含一个样本

③ 合并类间距离最小的两个类，其中最短距离为类间距离，构建一个新类

④ 计算新类与当前各类的距离。若类的个数为 1，终止计算，否则回到步骤3

时间复杂度为 $O(n^3m)$，m 是样本的维数，n 是样本个数

聚合聚类

例：给定 5 个样本集合，样本之间的欧式距离由 $D=[d_{ij}]_{n\times n}$ 表示：

$$D=\begin{bmatrix}0&7&2&9&3\\7&0&5&4&6\\2&5&0&8&1\\9&4&8&0&5\\3&6&1&5&0\end{bmatrix}$$

① 用 5 个样本构建 5 个类，$G_i=\{x_i\}, i=1,2,\ldots,5$

② $D_{35}=D_{53}=1$ 为最小，把 G_3 和 G_5 合并为一个新类 $G_6=\{x_3,x_5\}$

③ 计算 G_6 与另 3 类之间的最短距离：$D_{61}=2, D_{62}=5, D_{64}=5$

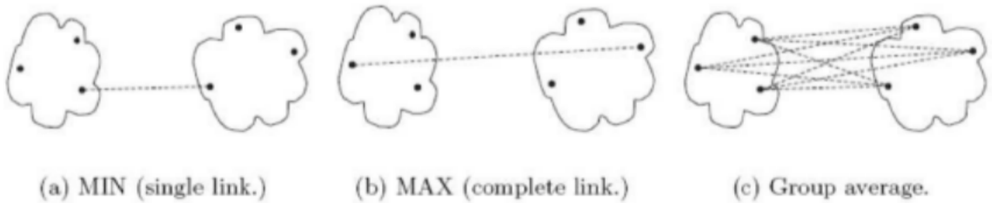
其余两类之间距离是：$D_{12}=7, D_{14}=9, D_{24}=4$

由于 $D_{61}=2$ 最小，所以将 G_1 和 G_6 合并成一个新类，记作 $G_7=\{x_1,x_3,x_5\}$

④ 计算 G_7 与另 2 类之间的最短距离：$D_{72}=5, D_{74}=5$，又有 $D_{24}=4$，将 G_2 和 G_4 合并成一个新类，记作 $G_8=\{x_2,x_4\}$

⑤ 将 G_7 和 G_8 合并成一个新类，记作 $G_9=\{x_1,x_2,x_3,x_4,x_5\}$，即将全部样本聚成一类，聚类终止。

类与类之间的距离



类 G_p 与类 G_q 之间的距离 $D(p,q)$，也称为连接，设类 G_p 包含 n_p 个样本，类 G_q 包含 n_q 个样本，连接有以下定义：

- 最短距离 / 单连接 (single linkage)：类 G_p 的样本与 G_q 的样本之间的最短距离

$$D_{pq} = min\{d_{ij} \mid x_i \in G_p, x_j \in G_q\}$$

- 最长距离 / 全连接 (complete linkage): 类 G_p 的样本与 G_q 样本之间的最长距离

$$D_{pq} = max\{d_{ij} \mid x_i \in G_p, x_j \in G_q\}$$

- 平均距离 (average linkage): 类 G_p 与 G_q 任意两个样本之间距离的平均值

$$D_{pq} = \frac{1}{n_p n_q} \sum_{x_i \in G_p} \sum_{x_j \in G_q} d_{ij}$$

实验要求

以“学号+姓名(5)”的命名格式打包代码.py文件和实验报告.pdf文件发送至邮箱：jipeng.huang@mail.nankai.edu.cn

np.random.multivariate_normal(mean, cov, temp_num)

- 数据集：生成 2000 个样例，每个样例的前 3 列表示特征，第 4 列表示标签
- 基本要求(4’)：绘制聚类前后样本分布情况

  (1) 实现 single-linkage 层次聚类算法；

  (2) 实现 complete-linkage 层次聚类算法。
- 中级要求(1’)：实现 average-linkage 层次聚类算法，绘制样本分布图。
- 提高要求(1’)：对比上述三种算法，给出结论。
- 拓展要求：通过变换聚类簇的个数，测试上述三种算法的性能，并给出分析。