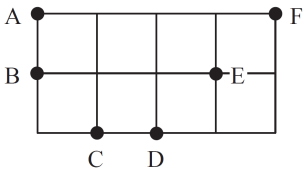
作业 **8** 参考答案　　聚类分析

1. 考虑下图平面方格上的 6 个点 A-F，计算它们之间的 Manhattan 距离矩阵，应用单连结 层次聚集聚类法进行聚类。画出树图。



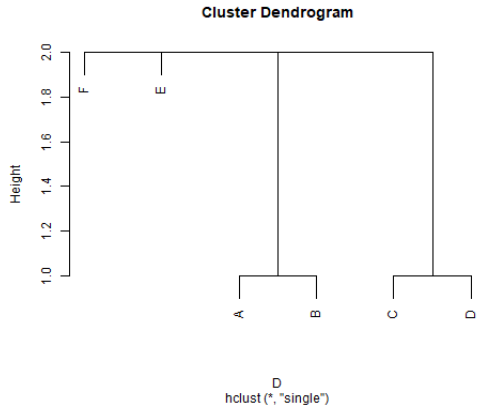
解：假设方格子边长为 1，Manhattan 距离矩阵

$$D = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccccc} A & B & C & D & E & F \end{array} \\ \begin{pmatrix} 0 & 1 & 3 & 4 & 4 & 4 \\ 1 & 0 & 2 & 3 & 3 & 5 \\ 3 & 2 & 0 & 1 & 2 & 5 \\ 4 & 3 & 1 & 0 & 2 & 4 \\ 4 & 3 & 2 & 2 & 0 & 2 \\ 4 & 5 & 5 & 4 & 2 & 0 \end{pmatrix} \end{array}$$

将 6 个点作为 6 个类，它们之间最小距离为 $d_{AB} = d_{CD} = 1$，将它们分别聚集为一类得到 4 个类（数图上标记合并时的距离 1）：$(AB), (CD), E, F$, 它们的单连结距离矩阵为

$$D = \begin{array}{c} \\ AB \\ CD \\ E \\ F \end{array} \begin{array}{c} \begin{array}{cccc} AB & CD & E & F \end{array} \\ \begin{pmatrix} 0 & 2 & 3 & 4 \\ 2 & 0 & 2 & 4 \\ 3 & 2 & 0 & 2 \\ 4 & 4 & 2 & 0 \end{pmatrix} \end{array}$$

四个类的最小距离为 $d_{(AB)(CD)} = d_{(CD)E} = d_{EF} = 2$, 分别合并为 $(ABCD), (CDE), (EF)$ 因为前两个类共有 (CD), 后两个类共有 E, 所以这三个类可以合并为一个类即 $(ABCDEF)$ (注意：数图上需标记 Height, 合并时的距离 2).



2. 假设物件 a-e 的距离矩阵如下：

$$D = \begin{array}{c} \\ a \\ b \\ c \\ d \\ e \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 0 & 1 & 3 & 4 & 3 \\ 1 & 0 & 2 & 3 & 2 \\ 3 & 2 & 0 & 1 & 2 \\ 4 & 3 & 1 & 0 & 1 \\ 3 & 2 & 2 & 1 & 0 \end{pmatrix} \end{array}$$

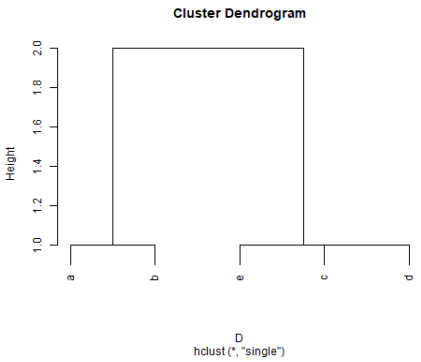
试应用单连结以及完全连结的层次聚集聚类方法进行聚类，画出树图，比较两种方法得到的结果。

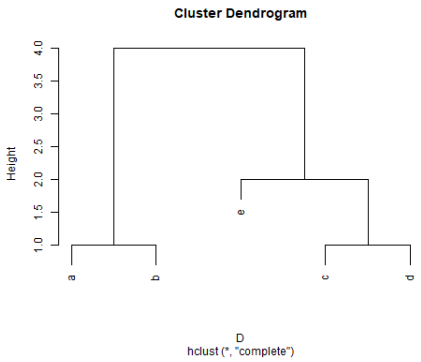
解：(1) 单连结. 将 5 个点作为 5 个类，因为 $d_{ab}=1, d_{cd}=d_{de}=1$ 最小，分别合并 a 和 b, c 和 d, d 和 e，得到三类 $(ab),(cd),(de)$ 【注意 d 同时属于后两类, 单连结时可以直接合并，但其它连结时不可直接合并 】, 它们之间的单连结距离:

$$\begin{aligned} d_{(ab)(cd)} &= \min\{d_{ac}, d_{ad}, d_{bc}, d_{bd}\} = \min\{3,4,2,3\} = 2, \\ d_{(ab)(de)} &= \min\{d_{ad}, d_{ae}, d_{bd}, d_{be}\} = \min\{4,3,3,2\} = 2, \\ d_{(cd)(de)} &= \min\{d_{cd}, d_{ce}, d_{dd}, d_{de}\} = \min\{1,2,0,1\} = 0, \end{aligned}$$

$$D = \begin{array}{c} \\ ab \\ cd \\ de \end{array} \begin{array}{c} \begin{array}{ccc} ab & cd & de \end{array} \\ \begin{pmatrix} 0 & 2 & 2 \\ 2 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} \end{array}$$

这三类之间的最小距离为 $d_{(cd)(de)}=0$, 合并它们得到两类：(cde) 和 (ab) , 合并时的单连结距离为0 (这一步合并不需要在树图 Height 上标出) 。最后，它们之间的单连结距离 $d_{(ab),(cde)}=d_{bc}=d_{be}=2$，合并成一类，得到树图（下左图）





(2) 完全连结. $d_{ab}=1, d_{cd}=d_{de}=1$ 最小，分别合并 a,b 和 c,d, d,e，得到三类 $(ab),(cd),(de)$ ，它们之间的完全连结距离:

$$\begin{aligned} d_{(ab)(cd)} &= \max\{d_{ac}, d_{ad}, d_{bc}, d_{bd}\} = \max\{3,4,2,3\} = 4, \\ d_{(ab)(de)} &= \max\{d_{ad}, d_{ae}, d_{bd}, d_{be}\} = \max\{4,3,3,2\} = 4, \\ d_{(cd)(de)} &= \max\{d_{cd}, d_{ce}, d_{dd}, d_{de}\} = \max\{1,2,0,1\} = 2, \end{aligned}$$

$$D = \begin{array}{c} \\ ab \\ cd \\ de \end{array} \begin{array}{c} \begin{array}{ccc} ab & cd & de \end{array} \\ \begin{pmatrix} 0 & 4 & 4 \\ 4 & 0 & 2 \\ 4 & 2 & 0 \end{pmatrix} \end{array}$$

这三类之间的最小距离为 $d_{(cd)(de)} = 2$, 合并它们得到两类：(cde) 和 (ab) , 合并时的（完全连结）距离为 2。最后，(cde) 和 (ab) 之间的完全连结距离 $d_{(ab),(cde)} = d_{ad} = 4$，合并成一类，得到树图（上右图）”

比较单连结和完全连结：单连结解法中，由于 $d_{cd} = d_{de} = 1$, 将 c,d,e 合并成了一类，但这不是很合理，因为 $d_{ce} = 2$, c,d,e 三个点并不是两两等距的。完全连结更合理一些，它先把 c,d 合并成一类，然后与 e 合并，当需要划分为三类时，三类为 $(a,b),(c,d),(e)$。

3. 对于上题的距离矩阵，应用 K-中心方法（K-medoid）将 a-e 聚集为 $K = 2$ 类，假设初始指定两类的中心 (medoids) 各为 a 和 b，写出迭代过程。

   解：假设初始中心为 a,b, 其它点与中心的距离如下

$$D = \begin{array}{c} \\ a \\ b \end{array} \begin{array}{c} \begin{array}{ccc} c & d & e \end{array} \\ \begin{pmatrix} 3 & 4 & 3 \\ 2 & 3 & 2 \end{pmatrix} \end{array}$$

   每个点划分到距其更近的中心所代表的类中，所以 c,d,e 都划分到 b 所在的类，因此得到两类: $(a),(bcde)$, 其中第一类只有一个元素 a，中心为 a; 第二类 $(bcde)$ 中各个点与其它类内点距离之和分别为 $d_b = d_{bc} + d_{bd} + d_{be} = 2 + 3 + 2 = 7$, 类似地 $d_c = d_d = d_e = 5$, 所以 c,d,e 都可作为该类 $(bcde)$ 的中心，下面分别考虑这三种可能。

   - 先取 c 为类 $(bcde)$ 的中心，其它点与中心 a,c 的距离

$$D = \begin{array}{c} \\ a \\ c \end{array} \begin{array}{c} \begin{array}{ccc} b & d & e \end{array} \\ \begin{pmatrix} 1 & 4 & 3 \\ 2 & 1 & 2 \end{pmatrix} \end{array}$$

     因此 b 划分到 a 类，d,e 划分到 c 为中心的类。因此得到两类 $(ab),(cde)$, 第一类 a,b 都可为中心，第二类 d 和其他类内点距离之和 $d_d = d_{dc} + d_{de} = 1 + 1 = 2$, 最小，第二类的中心更新为 d, 计算其它点到 a,d 的距离，或到 b,d 的距离，划分不再改变。因此最终两类为 $(ab),(cde)$, 两类中心分别为 a/b, d.

   - 若取 d 为第二类中心，其它点与中心 a,d 的距离

$$D = \begin{array}{c} \\ a \\ d \end{array} \begin{array}{c} \begin{array}{ccc} b & c & e \end{array} \\ \begin{pmatrix} 1 & 3 & 3 \\ 3 & 1 & 1 \end{pmatrix} \end{array}$$

     我们划分 b 到中心为 a 的类，划分 c,e 到中心为 d 的类，即划分两类为 $(a,b),(cde)$, 由前面的结果, 这是最终划分。

   - 若取 e 为第二类中心，其它点与中心 a,e 的距离

$$D = \begin{array}{c} \\ a \\ e \end{array} \begin{array}{c} \begin{array}{ccc} b & c & d \end{array} \\ \begin{pmatrix} 1 & 3 & 4 \\ 2 & 2 & 1 \end{pmatrix} \end{array}$$

故划分 b 到中心为 a 的类，划分 c,d 到 e 为中心的类，得到两类 (a,b), (c,d,e)。这是最终划分。

综上，最终聚类结果为 $(a,b),(c,d,e)$ 两类，两类的中心分别为 a/b 和 d。