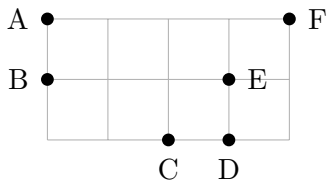
# HW 9　参考解答

**1.** 考虑下图平面方格上的 5 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.** 注：图中实际有 $A$–$F$ 共 $6$ 个点。

由图读出各点坐标：$A=(1,3)$，$B=(1,2)$，$C=(3,1)$，$D=(4,1)$，$E=(4,2)$，$F=(5,3)$。

对任意两点 $(x_1,y_1)$ 与 $(x_2,y_2)$，Manhattan 距离为 $d=|x_1-x_2|+|y_1-y_2|$。计算得距离矩阵

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

单连结法的类间距离定义为 $d(u,v)=\min_{i\in u,\,j\in v} d(i,j)$。

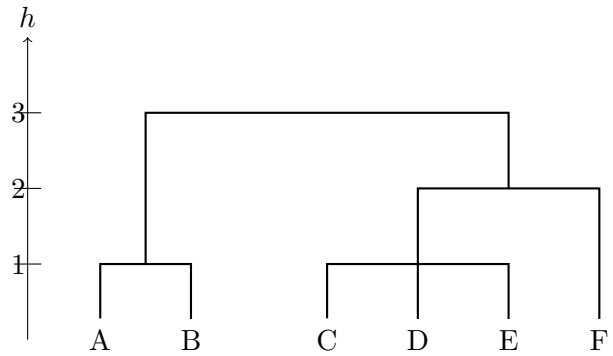
$h=1$：最小距离为 $1$，出现在 $d(A,B)=d(C,D)=d(D,E)=1$。合并 $\{A,B\}$；由于 $d(C,D)=1$ 且 $d(D,E)=1$，在单连结下发生链式合并，得 $\{C,D,E\}$；$F$ 独立。更新类间距离：

$$\begin{aligned} d\big(\{A,B\},\{C,D,E\}\big)&=\min(4,5,4,3,4,3)=3,\\ d\big(\{A,B\},F\big)&=\min(4,5)=4,\\ d\big(\{C,D,E\},F\big)&=\min(4,3,2)=2. \end{aligned}$$

$h=2$：合并 $\{C,D,E\}$ 与 $F$，得 $\{C,D,E,F\}$。此时 $d\big(\{A,B\},\{C,D,E,F\}\big)=\min(3,4)=3$。

$h=3$：合并全部，得 $\{A,B,C,D,E,F\}$。

树图如下：



**2.** 假设物件 $a$–$e$ 的距离矩阵如下：

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

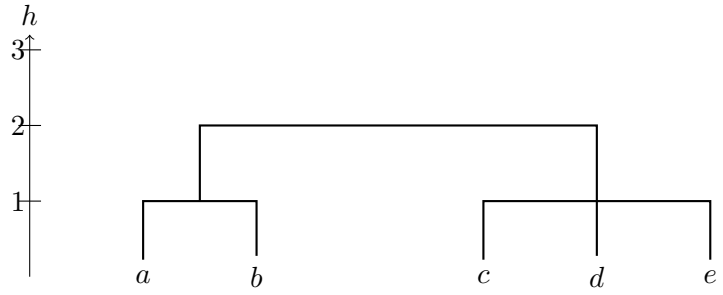
试应用单连结以及完全连结的层次聚集聚类方法进行聚类，画出树图，比较两种方法得到的结果。

**解.**

**(i) 单连结法** 类间距离 $d(u,v) = \min_{i\in u,\, j\in v} d(i,j)$。

$h = 1$：$d(a,b) = d(c,d) = d(d,e) = 1$。合并 $\{a,b\}$；$c, d, e$ 链式合并为 $\{c,d,e\}$。

$h = 2$：$d\big(\{a,b\},\{c,d,e\}\big) = \min(3,4,3,2,3,2) = 2$，合并为 $\{a,b,c,d,e\}$。

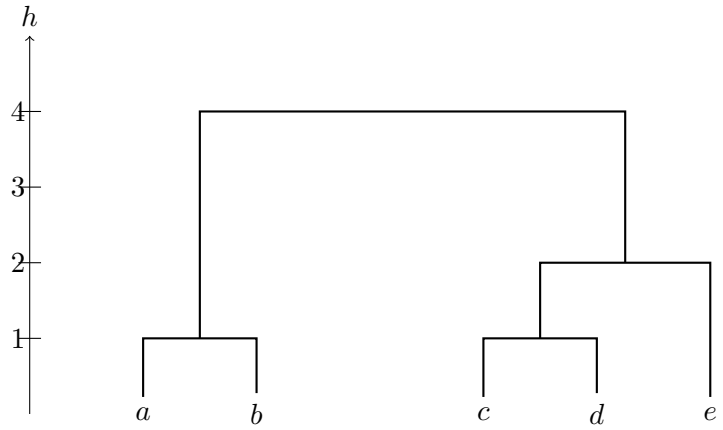



**(ii) 完全连结法** 类间距离 $d(u,v) = \max_{i\in u,\, j\in v} d(i,j)$。

$h = 1$：$d(a,b) = 1$，合并 $\{a,b\}$；$d(c,d) = 1$，合并 $\{c,d\}$。虽然 $d(d,e) = 1$，但 $d\big(e,\{c,d\}\big) = \max(2,1) = 2 \neq 1$，故 $e$ 暂不并入。

$h = 2$：$d\big(e,\{c,d\}\big) = \max(2,1) = 2$，合并得 $\{c,d,e\}$。

$h = 4$：$d\big(\{a,b\},\{c,d,e\}\big) = \max(3,4,3,2,3,2) = 4$，合并为 $\{a,b,c,d,e\}$。




**(iii) 比较** 单连结法在高度 2 即完成全部合并，完全连结法则延迟到高度 4。单连结只看两类之间最近的一对点，容易产生链式效应，使合并过快；完全连结看最远的一对点，倾向于产生紧凑、分离度高的类，因此类间合并推迟到更大的距离。

**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.** K-medoid 算法选取实际样本点作为类中心，目标是最小化总代价 $TC=\sum_i d(x_i, m_{c(i)})$，其中 $m_{c(i)}$ 为 $x_i$ 所属类的中心。

**迭代 1** $M=\{a,b\}$。

分配：$c,d,e$ 均离 $b$ 更近（$d(c,b)=2<3$，$d(d,b)=3<4$，$d(e,b)=2<3$），故 $C_1=\{a\}$，$C_2=\{b,c,d,e\}$，$TC=0+(0+2+3+2)=7$。

更新中心：$C_1$ 仅含 $a$，不变。$C_2$ 中各点作为中心的局部代价：$b$: 7，$c$: 5，$d$: 5，$e$: 5。三者并列，按字母序选 $c$。新中心 $M=\{a,c\}$。

**迭代 2** $M=\{a,c\}$。

分配：$d(b,a)=1<2=d(b,c)$，归 $a$；$d(d,c)=1<4=d(d,a)$，$d(e,c)=2<3=d(e,a)$，均归 $c$。故 $C_1=\{a,b\}$，$C_2=\{c,d,e\}$，$TC=(0+1)+(0+1+2)=4$。

更新中心：$C_1$ 中 $a,b$ 代价均为 1，保持 $a$。$C_2$ 局部代价：$c$: 3，$d$: 2，$e$: 3，选 $d$。新中心 $M=\{a,d\}$。

**迭代 3** $M=\{a,d\}$。

分配：$d(b,a)=1<3=d(b,d)$，归 $a$；$d(c,d)=1<3=d(c,a)$，$d(e,d)=1<3=d(e,a)$，均归 $d$。故 $C_1=\{a,b\}$，$C_2=\{c,d,e\}$，$TC=(0+1)+(1+0+1)=3$。

更新中心：$C_1$ 中 $a,b$ 代价均为 1，保持 $a$；$C_2$ 中 $d$ 的代价 2 仍最小，保持 $d$。中心不变，划分不变，算法收敛。

**结论：**最终聚为 $\{a,b\}$（中心 $a$）与 $\{c,d,e\}$（中心 $d$），总代价 $TC=3$。