第4章　階層的手法

4.1　アルゴリズムと数値例
4.2　デンドログラム
4.3Rによる解析

５月１２日
鈴木直也

4.1　アルゴリズムと数値例

以下の図で示された5つのデータを階層的手法で
クラスタリングします。

データ	座標
データ1	（2，1）
データ2	（1，2）
データ3	（2，4）
データ4	（4，4）
データ5	（4．5，2）

階層的手法では、まず各データが自身からなるクラスタと考え、
　データ数分のクラスタを作ります。

クラスタ間距離と併合

次にクラスタ間での距離を測り、最も近い
クラスタ同士を併合します。
ここではクラスタAとクラスタBの距離D（A，B）
を以下の式で与えられるものとする。

$$D(A,B)=\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}d(a_i,b_j)$$

ただし、$A=\{a_1,a_2,a_3,\cdots,a_m\}, B=\{b_1,b_2,b_3,\cdots,b_n\}$
であり

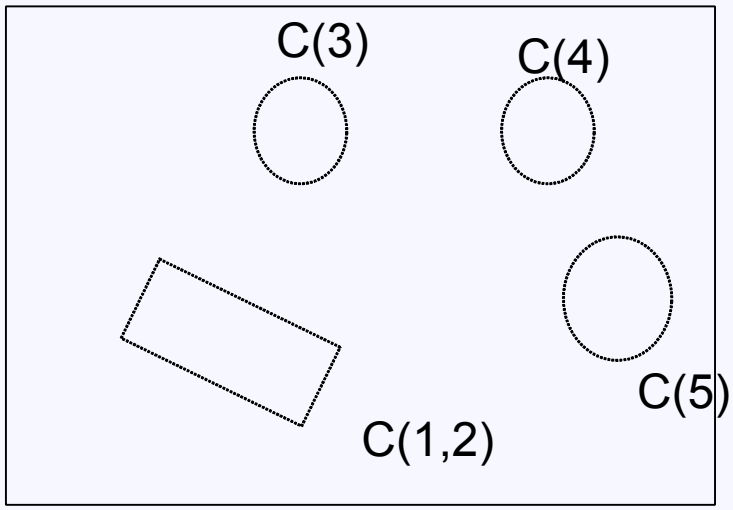
$d(a_i,b_j)$　はデータ a_iと b_j の距離とする。

距離行列

距離を測ったら、距離行列を作成します。

	C(1)	C(2)	C(3)	C(4)
C(2)	1.4142			
C(3)	3.0000	2.2361		
C(4)	3.6056	3.6056	2.0000	
C(5)	2.6926	3.5000	3.2016	2.0616

ここで最も近いクラスタ同士を併合します

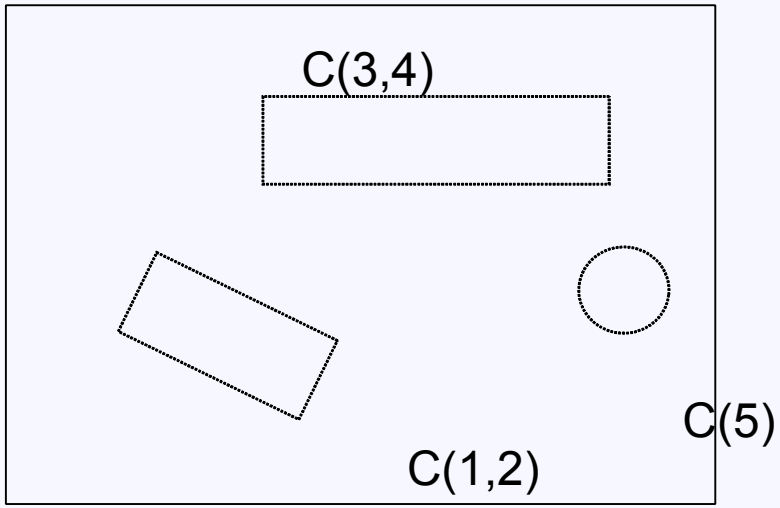


距離行列（2）

- 再びクラスタ間の距離を求め、最も近いクラスタどうしを併合します。

	C(1,2)	C(3)	C(4)
C(3)	2.6180		
C(4)	3.6056	2.0000	
C(5)	3.0963	3.2016	2.0616

ここではクラスタC(3)とクラスタC(4)が併合され、クラスタC(3,4)が作成されます。

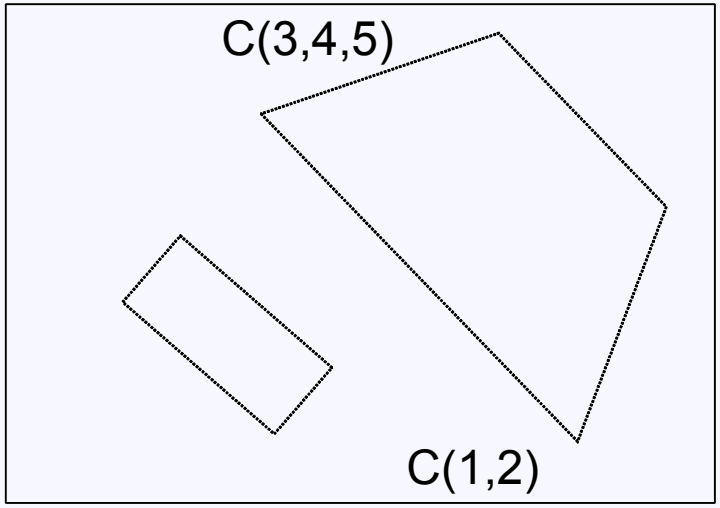


距離行列（3）

さらに距離行列を求めると

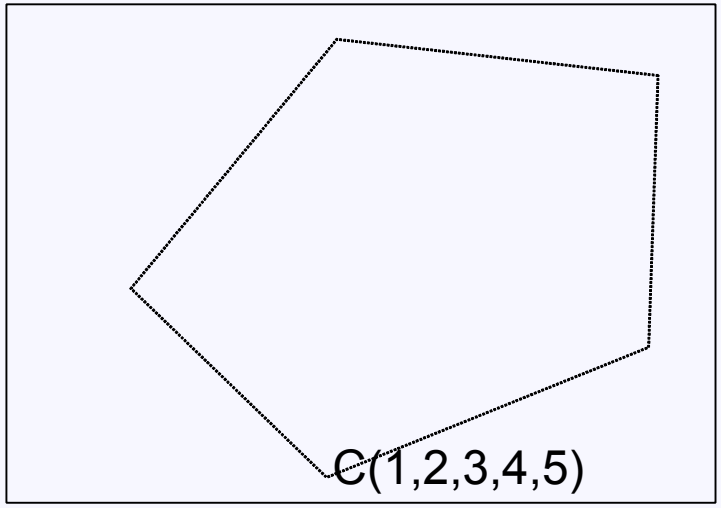
	C(1,2)	C(3,4)
C(3,4)	3.1118	
C(5)	3.0963	2.6316

よってクラスタC(3,4,5)ができます。



距離行列（４）

最後に2つのクラスタC(1,2)とC(3,4,5)を
併合して終了です。



このクラスタリング手法は群平均法と呼ばれて
います。

階層的手法のアルゴリズム

階層的手法では距離行列が更新されて行くことで処理が進んでいきます。距離行列の更新箇所は新たに作られたクラスタと既存のクラスタとの距離だけです。

このため階層的手法のアルゴリズムはクラスタAとクラスタBからクラスタCができるとき、その他のクラスタXとクラスタCとの距離を現在の距離行列から求めることに対応します。

単連結法

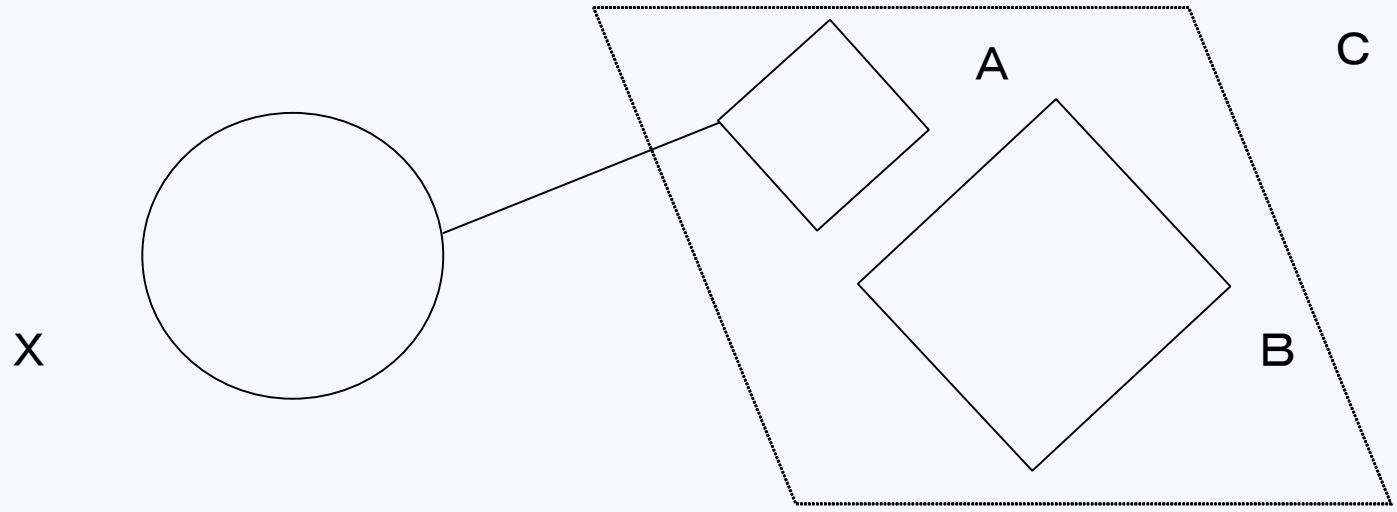
2つのクラスタにおけるデータ間の距離が最小となるようなデータ対を選ぶ手法

式で表すと以下のとおりである。

$$D(A,B) = \min_{a_i \in A, b_j \in B} d(a_i, b_j)$$

また、距離行列の更新は以下のとおりである。

$$D(C,X) = \min\{D(A,X), D(B,X)\}$$



完全連結法

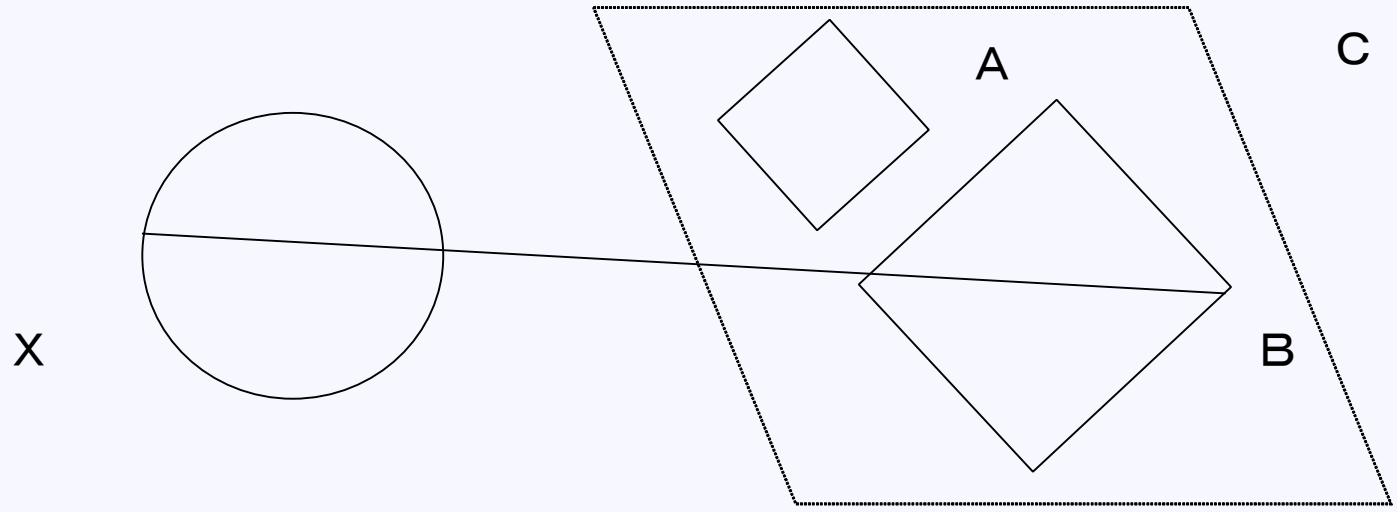
単連結法とは逆に2つのクラスタ間で、距離が最大となるデータ対を選ぶ手法

式で表すと以下のとおりである。

$$D(A,B) = \max_{a_i \in A, b_j \in B} d(a_i, b_j)$$

また、距離行列の更新は以下のとおりである。

$$D(C,X) = \max\{D(A,X), D(B,X)\}$$



ウォード法

クラスタAとクラスタBを併合したときに、クラスタ内の平方和の増加分が最小となるように併合する手法

式であらわすと以下のとおりである。

$$D(A,B) = E(A \cup B) - E(A) - E(B)$$

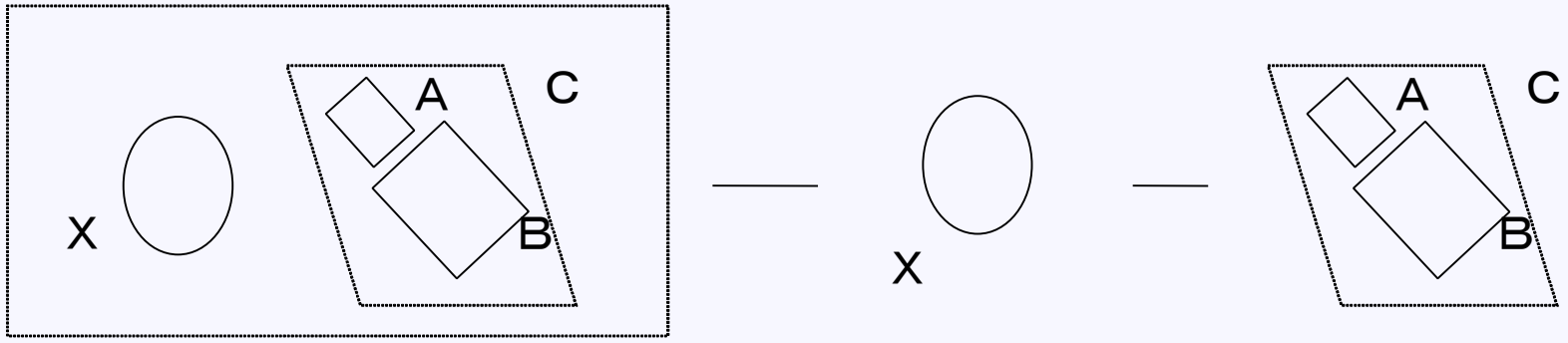
ここでE（X）はクラスタXの平方和を意味し、クラスタX内の各データxに対して、クラスタXの重心$center(X)$との距離の平方の和で、以下の式で定義されます。

$$E(X) = \sum_{x \in X} d(x, center(X))^2$$

ウォード法（続き）

距離行列の更新は以下のとおりです。

$$D(C,X)=\frac{|A|+|X|}{|A|+|B|+|X|}D(A,X)+\frac{|B|+|X|}{|A|+|B|+|X|}D(B,X)$$
$$-\frac{|X|}{|A|+|B|+|X|}D(A,B)$$



平方和　　平方和　　平方和

重心法

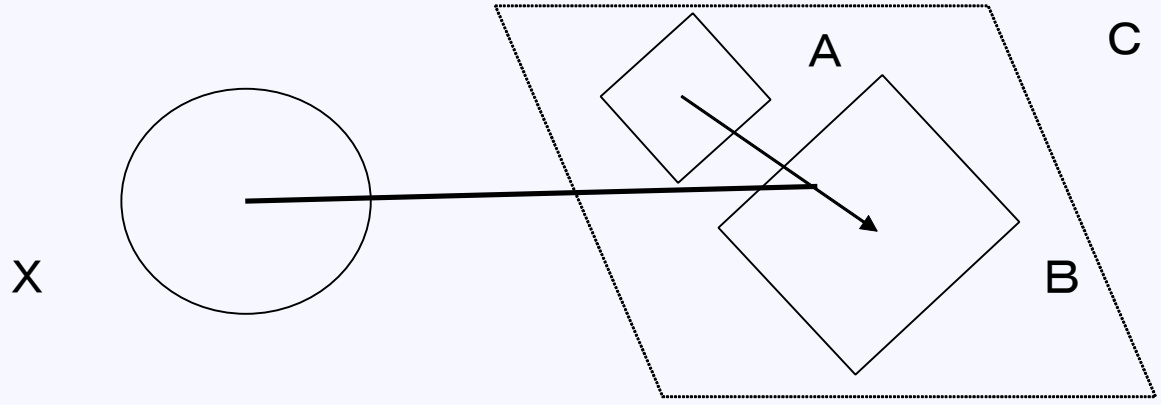
クラスタAとクラスタBの距離をクラスタAの重心とクラスタBの重心の距離の2乗で定義される手法

以下の式で表される。

$$D(A,B) = \left\| center(A) - center(B) \right\|^2$$

また、距離行列の更新式は以下のとおりである。

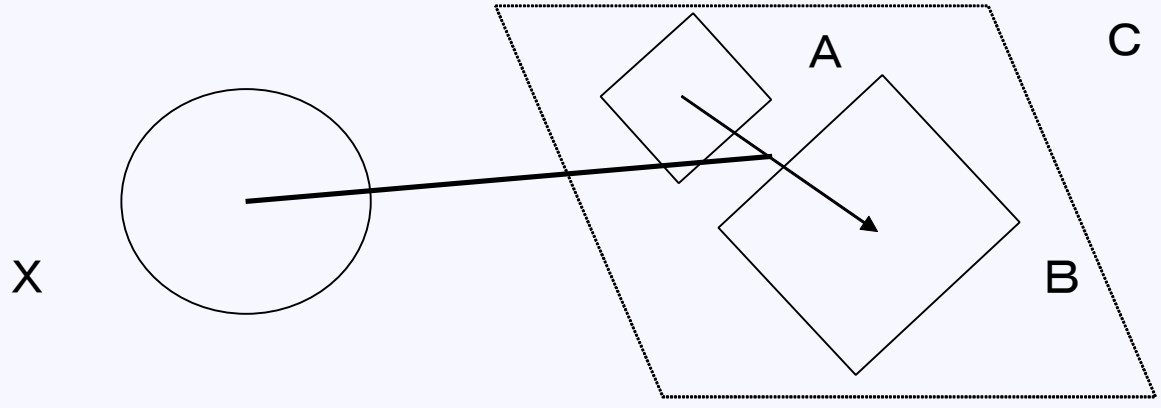
$$D(C,X) = \frac{|A|}{|A|+|B|} D(A,X) + \frac{|B|}{|A|+|B|} D(B,X) - \frac{|A||B|}{(|A|+|B|)^2} D(A,B)$$



メディアン法

重心法と似ているが、重心のとり方が異なる。
クラスタAとクラスタBの重心の中点をとる。
また、距離行列の更新式は以下のとおりである。

$$D(C,X)=\frac{1}{2}D(A,X)+\frac{1}{2}D(B,X)-\frac{1}{4}D(A,B)$$

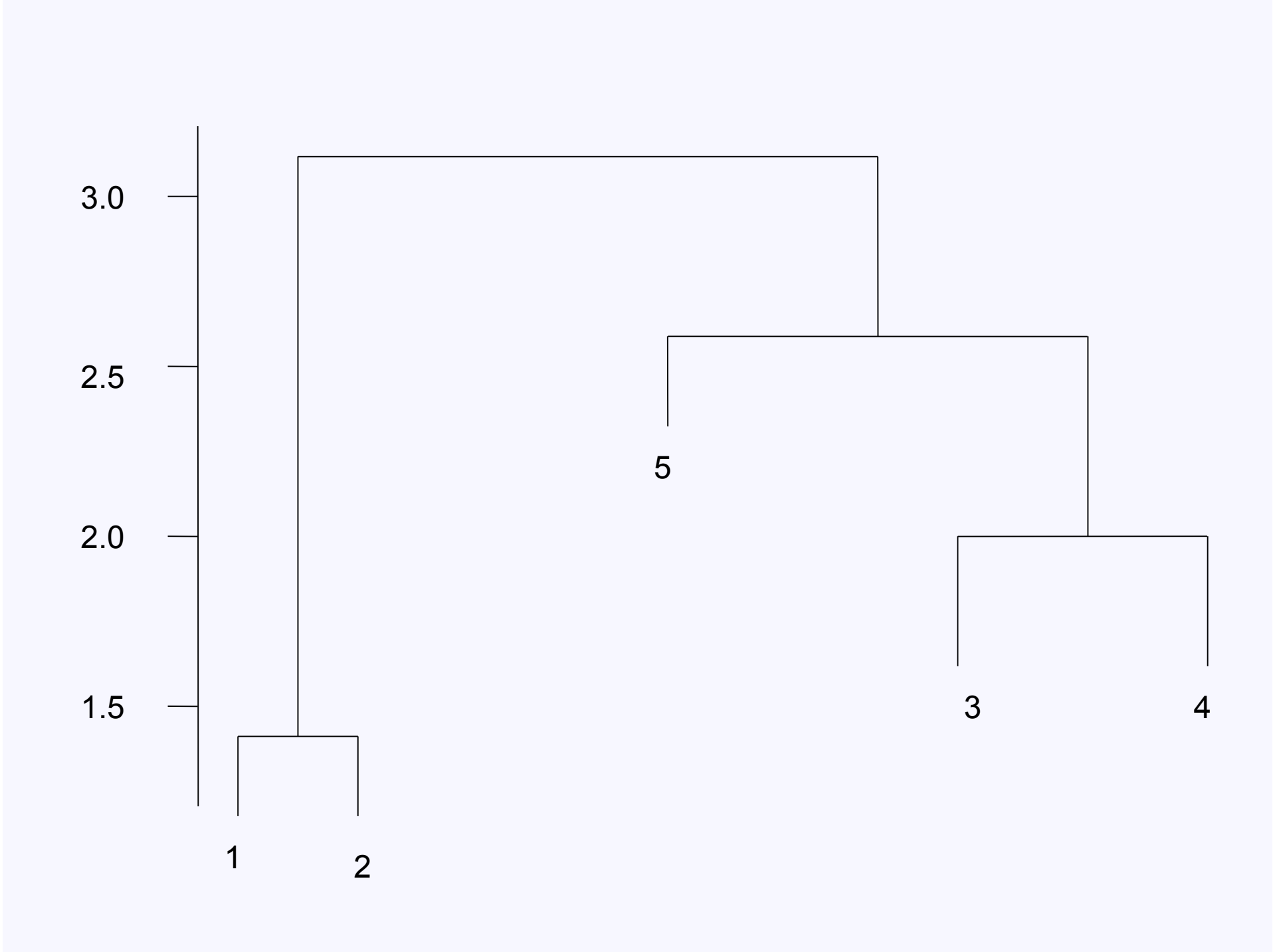


4.2デンドログラム

階層的手法においてクラスタが併合されていく様子をデンドログラムというグラフによって表すことができる。

前章の例を基に、デンドログラムを作成してみます。

- 距離1.4142の C(1) と C(2) を併合して C(1,2)を作成
- 距離2.0000の C(3) と C(4) を併合して C(3,4)を作成
- 距離2.6316の C(3,4) と C(5) を併合して C(3,4,5)を作成
- 距離3.1066の C(1,2) と C(3,4,5) を併合して C(1,2,3,4,5)を作成


3.0
2.5
2.0
1.5
1
2
5
3
4

4.3 Rによる解析

Rではさまざまな階層的クラスタリングが、
関数hclust()により提供されています。

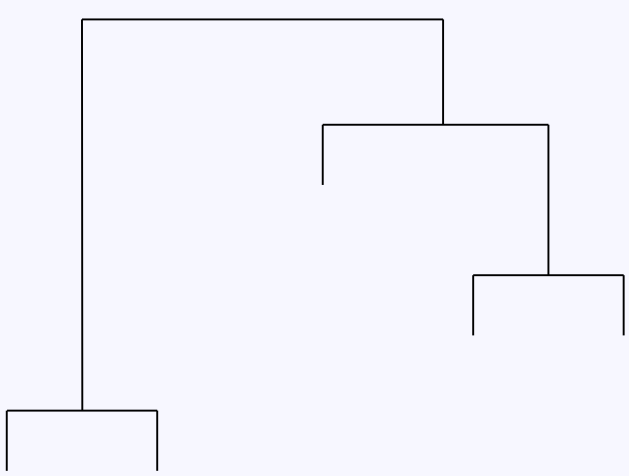
$$hclust(d, method = "complete", members = NULL)$$

第1引数は距離行列で、第2引数でクラスタリング手法を指定します。第3引数のmembersはデンドログラムの途中からクラスタリングを行いたいような特殊な場合なので通常は指定する必要はありません。

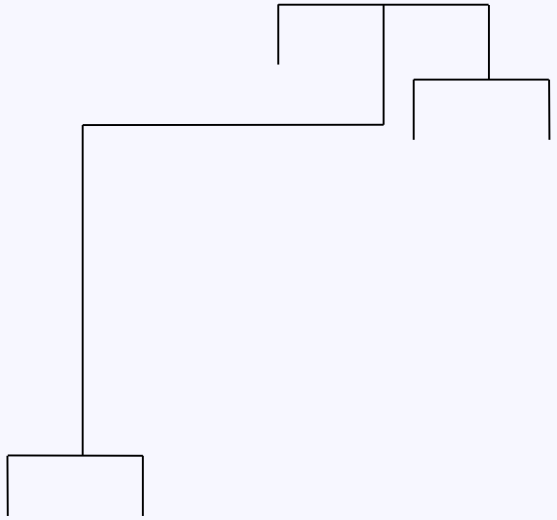
コマンド

```
> a <-matrix(c(2,1,1,2,2,4,4,4,4.5,2),nrow=5,byrow=TRUE))
> d=dist(a)
> hc <- hclust(d,method="average")
> plot(hc)
> hc <- hclust(d,method="centroid")
> plot(hc)
```

Rによる群平均法と重心法でのクラスタリング結果



群平均法



重心法

終わり