クラスタリング

クラスタリングとは

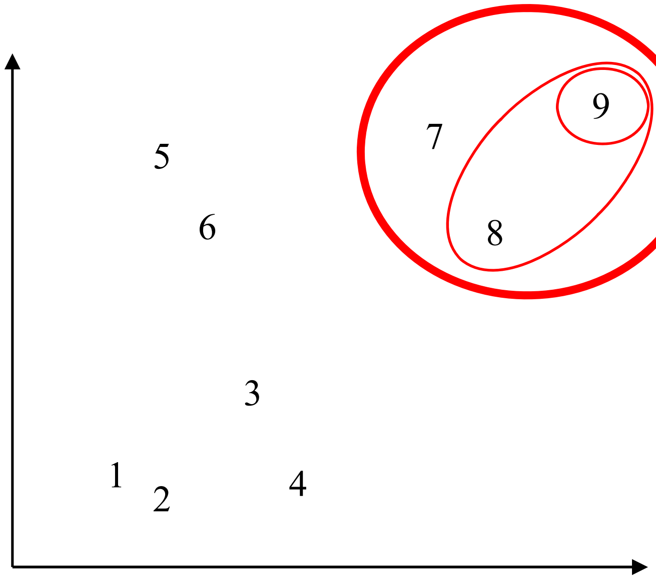
- データ間に距離を定義し、距離が近いデータ同士をグループ（クラスター）にまとめる作業

- 塩基配列のクラスタリング
  - 何度も同じ遺伝子の配列が部分的に読まれデータベースに登録される
  - マッチ率等により距離を定義
  - 例： UniGene（NIH/NCBI） EST配列をクラスタリング
    ヒト 5,112,666配列 ⇒ 53,032 クラスター

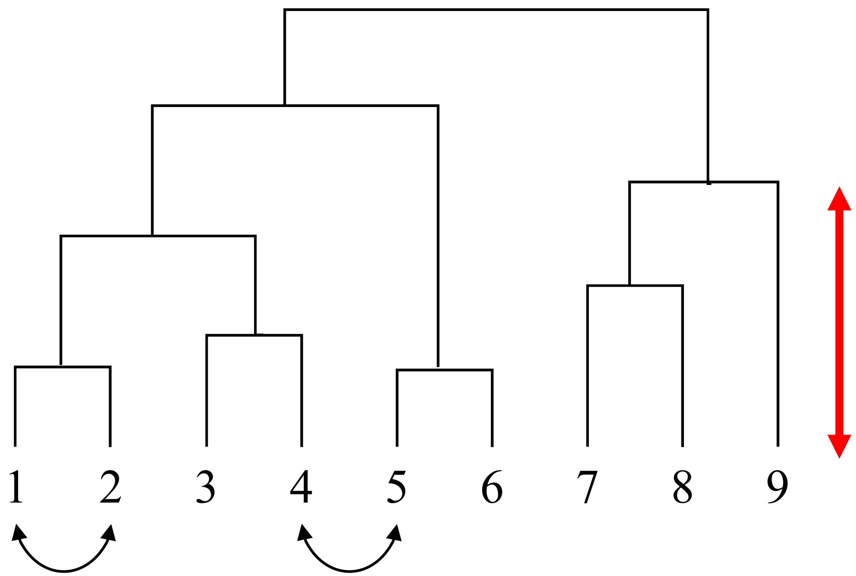
クラスタリングのアルゴリズム

- 階層的クラスタリング
  - ボトムアップ型
  - トップダウン型
- *k*-クラスタリング（*k*個のグループへ分類）
  - k-means 法
  - ゴンザレスの最遠点選択法
- 高次元空間の点を2、３次元へ埋め込みグループを視覚化する方法

階層的クラスタリング



Dendrogram



高さが
類似度を
表現

類似　　類似していない

各ノードで部分木の左右を交換しても構わない

階層的クラスタリング　ー　ボトムアップ型

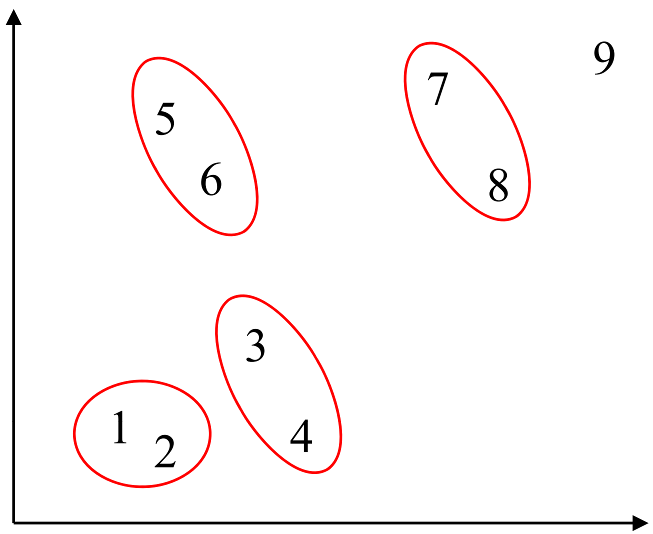
近いクラスター同士を融合するプロセスを繰り返す
クラスター（点の集合）C_i, C_j 間の距離に結果が依存

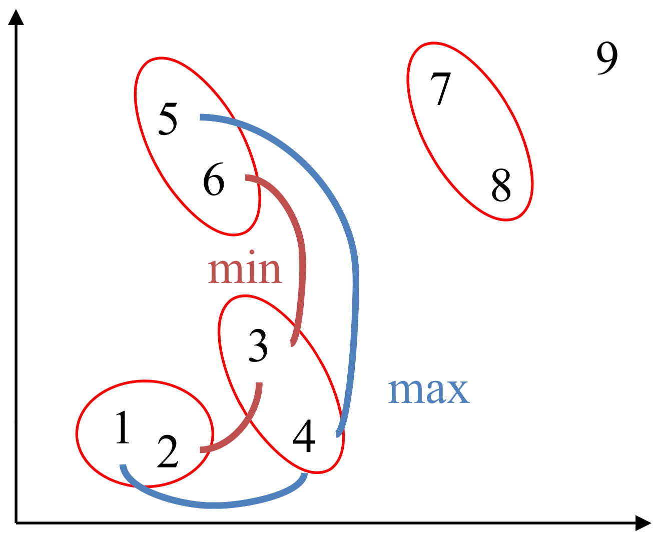
距離の例：

$d(\vec{x},\vec{y}) : \vec{x},\vec{y}$　間のユークリッド距離

$$D_{\min}(C_i, C_j) = \min\{d(\vec{x},\vec{y}) \mid \vec{x}\in C_i, \vec{y}\in C_j\}$$

$$D_{\max}(C_i, C_j) = \max\{d(\vec{x},\vec{y}) \mid \vec{x}\in C_i, \vec{y}\in C_j\}$$





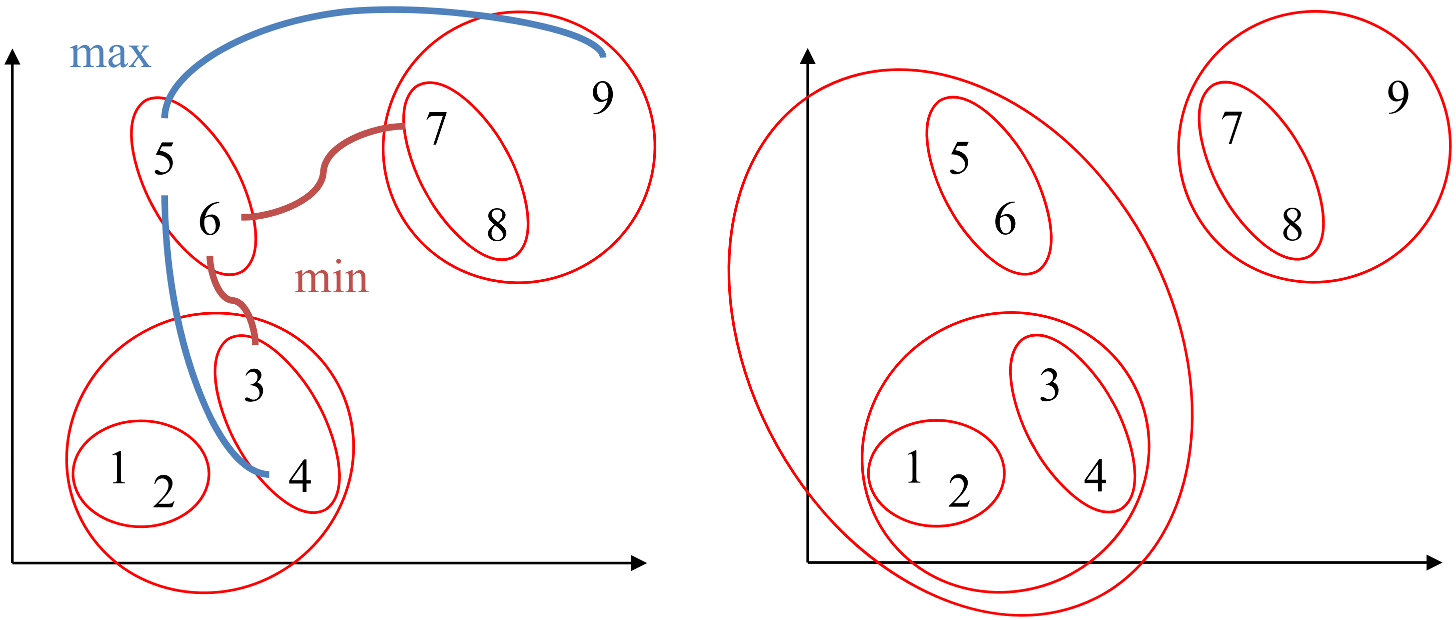
階層的クラスタリング　ー　ボトムアップ型

近いクラスター同士を融合するプロセスを繰り返す
クラスター（点の集合）C_i, C_j 間の距離に結果が依存

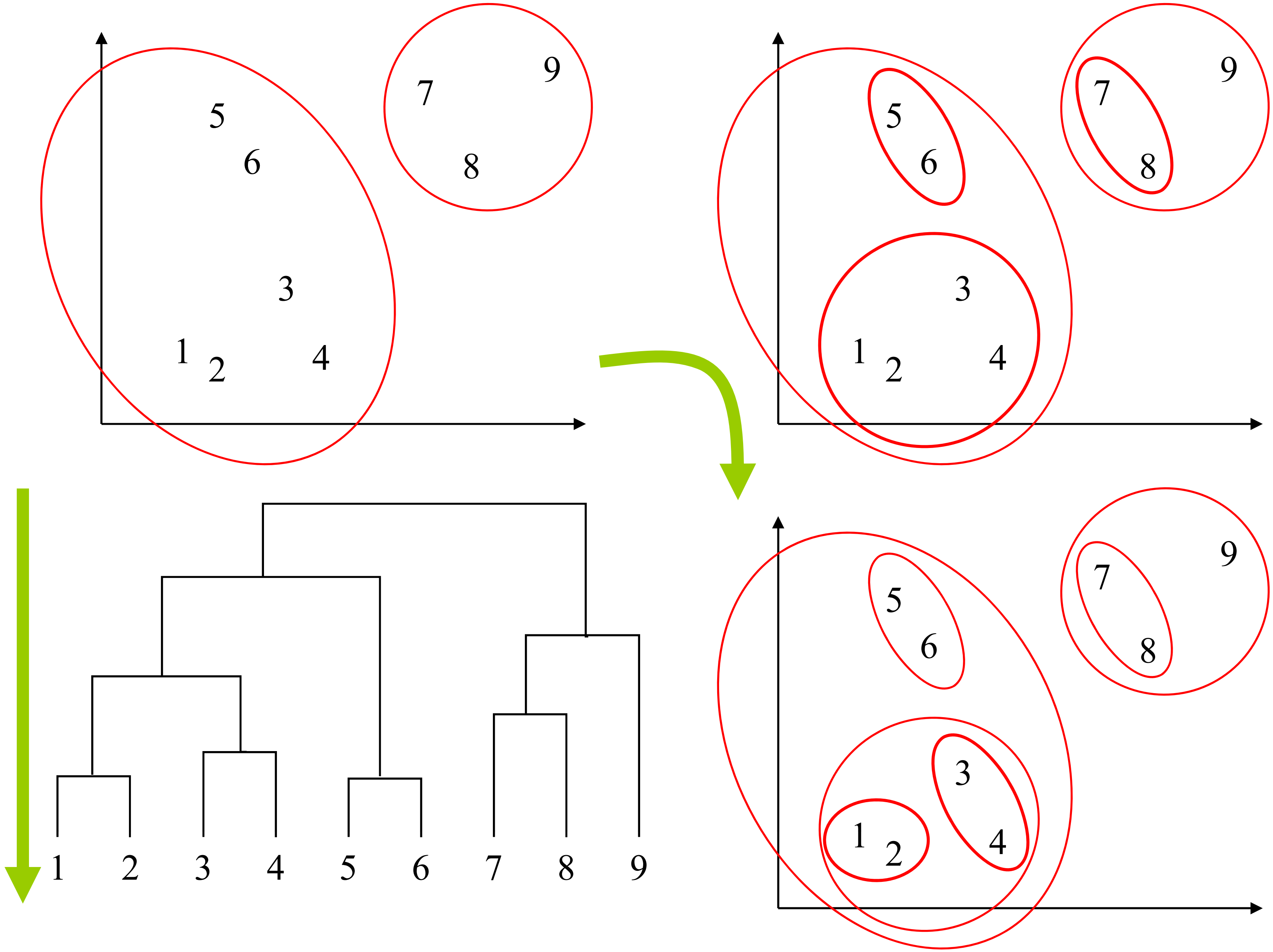
距離の例: $d(\vec{x},\vec{y}) : \vec{x},\vec{y}$　間のユークリッド距離

$$D_{\min}(C_i, C_j) = \min\{d(\vec{x},\vec{y}) \mid \vec{x} \in C_i, \vec{y} \in C_j\}$$

$$D_{\max}(C_i, C_j) = \max\{d(\vec{x},\vec{y}) \mid \vec{x} \in C_i, \vec{y} \in C_j\}$$

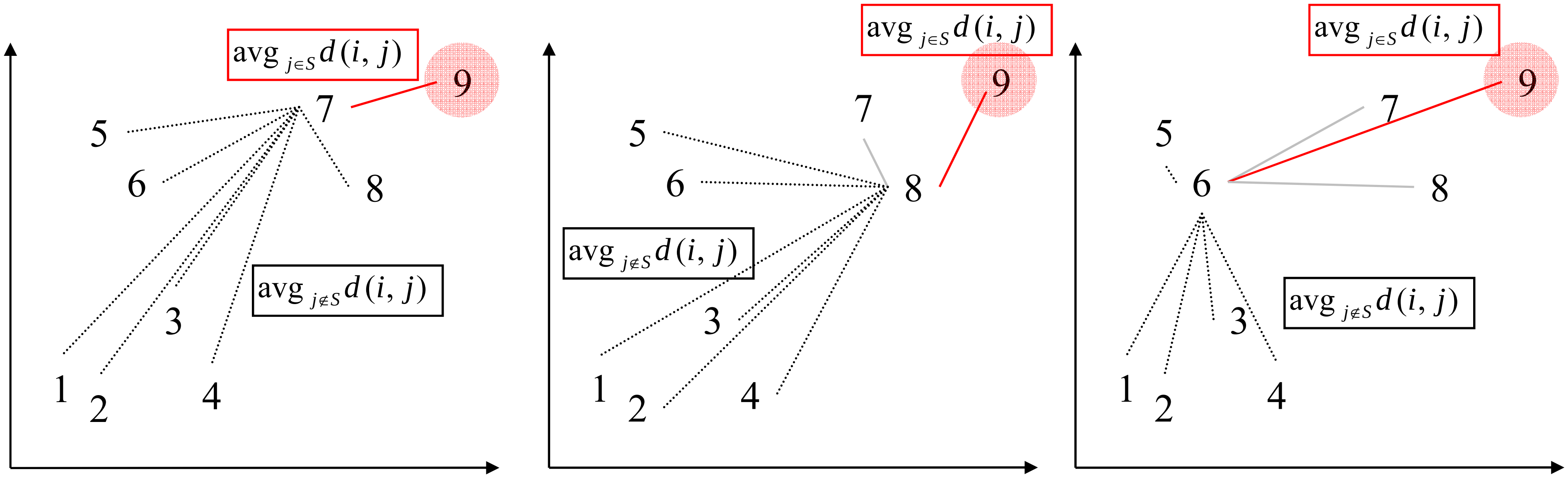


階層的クラスタリング：トップダウン型



階層的クラスタリング：トップダウン分割型の例

S-plus で使われている diana 法　　L. Kaufman, P. Rousseeuw. "*Finding Groups in Data– An Introduction to Cluster Analysis.*" Wiley Series in Probability and Mathematical Sciences, 1990.



S={9}, i=7　　S={9}, i=8　　S={9}, i=6

V は点全体の集合　　　S はクラスターの候補

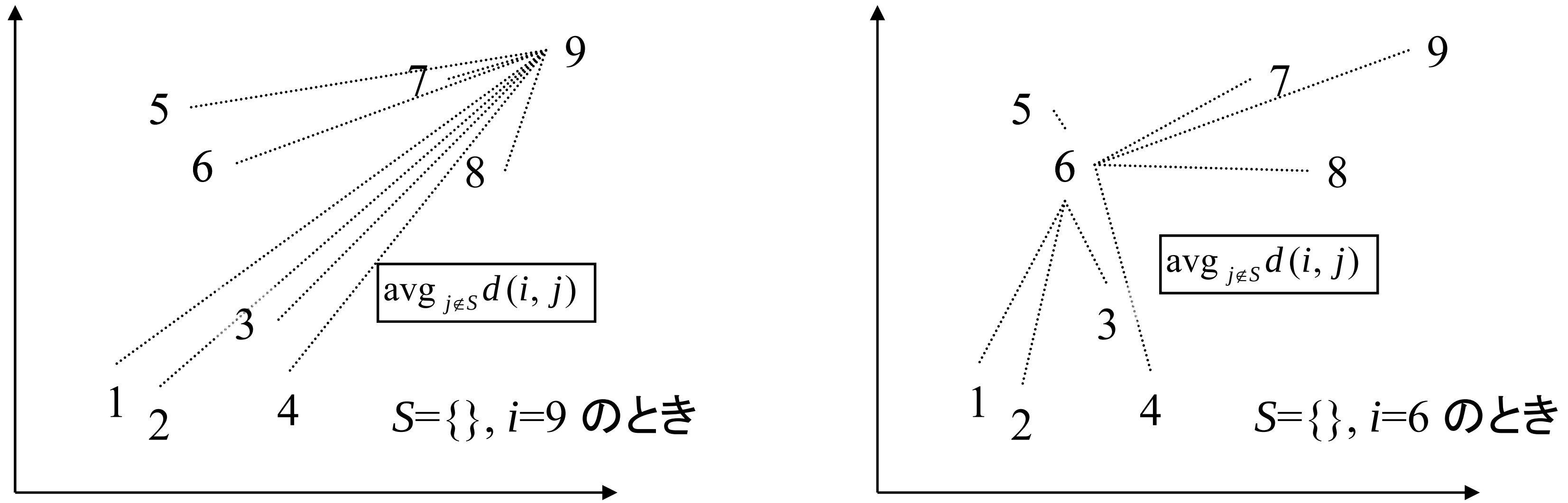
S に追加すべき点は存在　するか？

avg を距離の平均値とする　とき、$i \in V - S$ について

$V(i,S) \equiv \mathrm{avg}_{j\notin S}\, d(i,j) - \mathrm{avg}_{j\in S}\, d(i,j)$

$V(i,S)$ を最大にする　i は「S に最も近く　$V-S$から遠い」と解釈

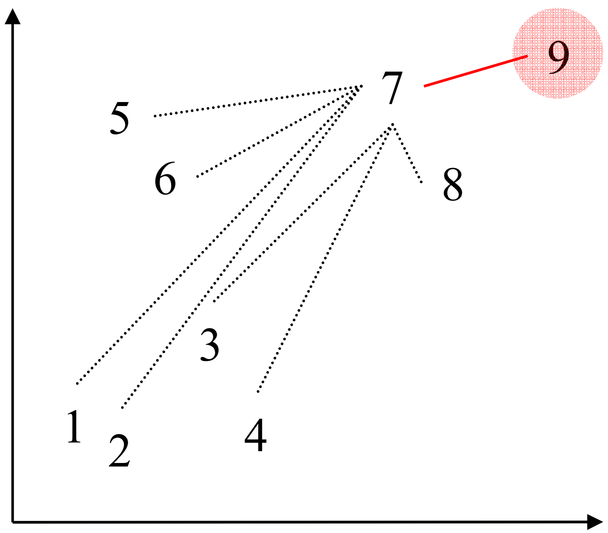
階層的クラスタリング: diana法
初期化 S= {} のとき　他の点から一番離れている点を選らぶ



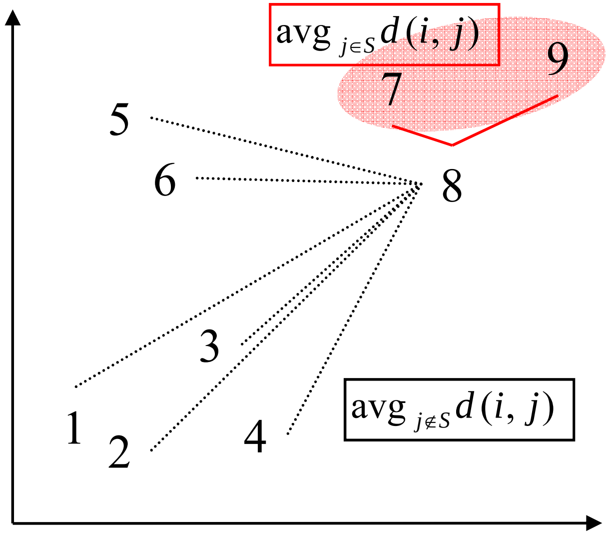
$$V(i,S) \equiv \mathrm{avg}_{j \notin S} d(i, j) - \mathrm{avg}_{j \in S} d(i, j)$$

$V(i,S)$ を最大にする i は「S に最も近く $V-S$ から遠い」と解釈

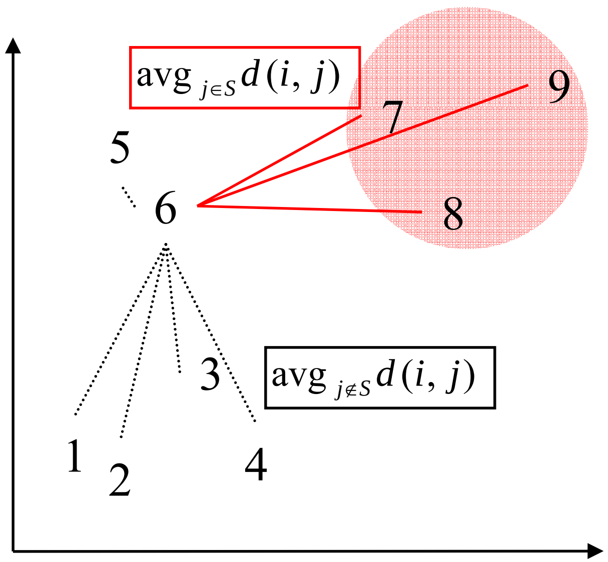
階層的クラスタリング：トップダウン分割型



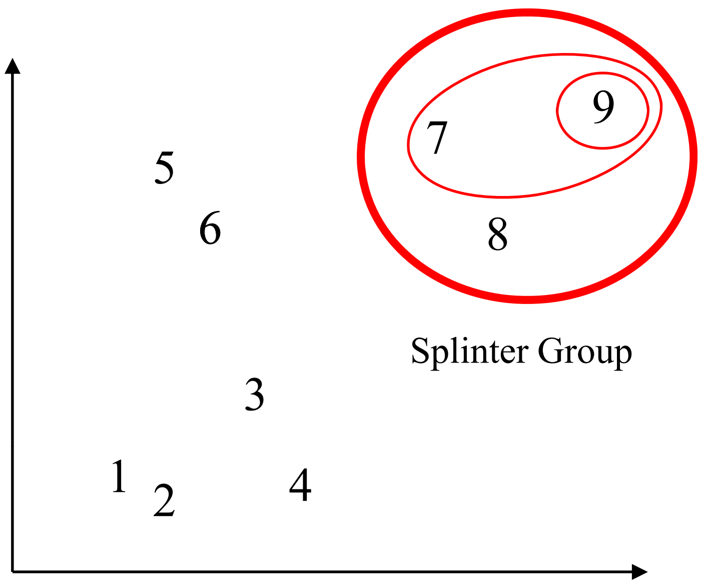
S={9}, i=7



S={7,9}, i=8



S={7,8,9}, i=6



初期化ステップ： S を空集合 {} に初期化

繰り返しステップ：

$V(i,S)$ を最大にする $i \in V - S$ を h とする。

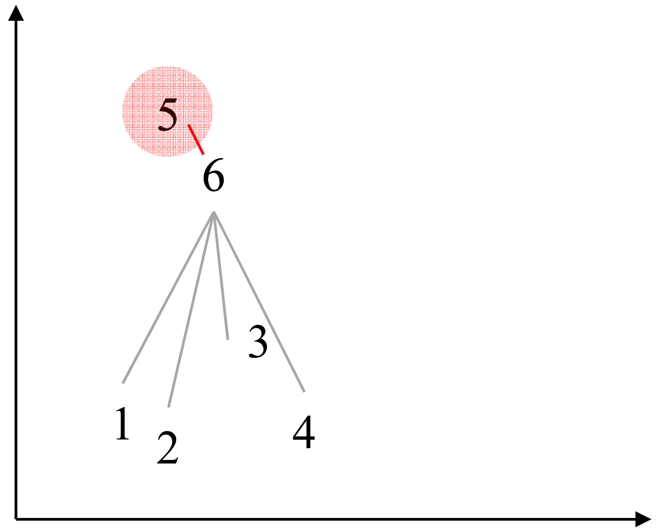
終了判定：

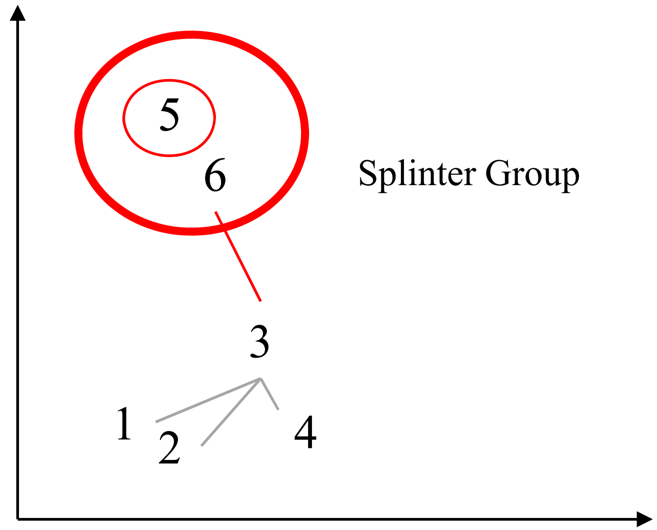
$V(h,S) > 0$ ならば h は S に近いので S に追加し、
他にも S に近い要素があるか調べるため、
繰り返しステップを実行。

$V(h,S) \le 0$ ならば h は S に近くなく、追加せず終了。
S を「Splinter Group」と呼び、V から分離。

以上のステップを繰り返し Splinter Group を分離。

階層的クラスタリング：トップダウン分割型





初期化ステップ：　S を空集合 $\{\}$ に初期化

繰り返しステップ：

　$V(i,S)$ を最大にする $i \in V - S$ を h とする。

終了判定：

　$V(h,S) > 0$ ならば h は S に近いので S に追加し、

　他にも S に近い要素があるか調べるため、

　繰り返しステップを実行。

　$V(h,S) \leq 0$ ならば h は S に近くなく、追加せず終了。

　S を「Splinter Group」と呼び、V から分離。

以上のステップを繰り返し Splinter Group を分離。

問題

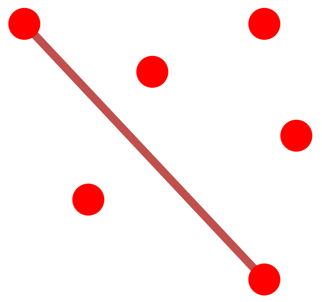
クラスタリングの結果に自由度があり、複数の妥当な候補がありうる例を考えよ

クラスターの評価

クラスター S_i の評価：　直径

$$diameter(S_i) = \max\left\{ \left\|\vec{x}_1 - \vec{x}_2\right\| \mid \vec{x}_1, \vec{x}_2 \in S_i \right\}$$

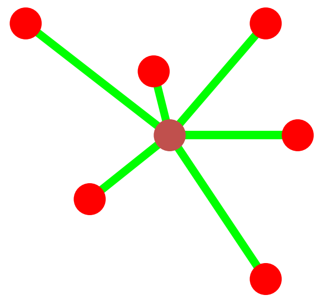
$$\left\| (x_1, \cdots, x_d) \right\| = \left(\sum_{i=1,\cdots,d} x_i^2 \right)^{1/2}$$



S_i の評価：重心からの距離の分散

$$c(S_i) = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \vec{x}$$

$$\mathrm{var}(S_i) = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \left\|\vec{x} - c(S_i)\right\|^2$$



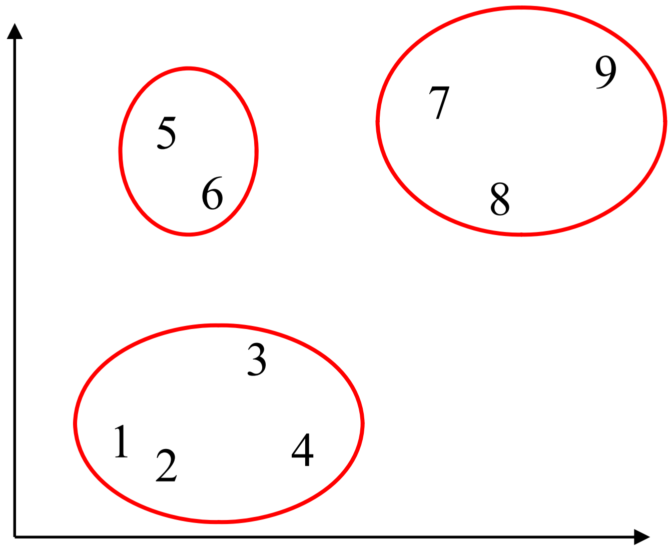
問題

直径が同一で、重心からの距離の分散が大きく異なるクラスターの例（S_1 と S_2）を考えよ

$$\mathrm{diameter}(S_1) = \mathrm{diameter}(S_2),\ \mathrm{var}(S_1) >> \mathrm{var}(S_2)$$

k - クラスタリング

d 次元ユークリッド空間 R^d 内の点集合 S を、S を覆い $\left(S = S_1 \cup S_2 \cup \cdots \cup S_k\right)$ かつ、互いに交わらない k 個の部分集合（クラスターと呼ぶ）$S_1, S_2, \cdots, S_k$ に分解すること。



3-クラスタリング

k - クラスタリング
誤差二乗平均によるクラスターの評価

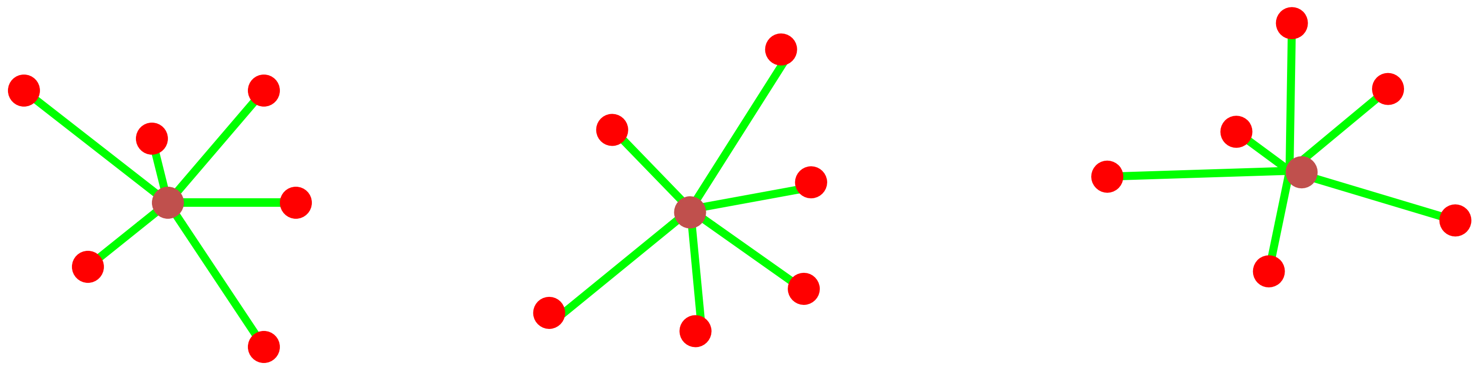
S の k –クラスタリングを $\{S_1, \ldots, S_k\}$

- 各クラスターの重心は　$c(S_i) = \frac{1}{|S_i|} \sum_{\vec{y} \in S_i} \vec{y}$

- Sの各点 $\vec{x}$と、$\vec{x}$が属するクラスターの重心間の距離の分散

$$\frac{1}{|S|} \sum_{i=\{1,\ldots,k\}} \sum_{\vec{x} \in S_i} \left\| \vec{x} - c(S_i) \right\|^2$$

を*mean squared error* (誤差二乗平均) と呼ぶ



k - クラスタリング

誤差二乗平均を最小化する k-クラスタリングを計算する
問題はNP困難(現実的な時間で解けない)

できるだけ小さくすると言われているアルゴリズムとして
k-means 法がある

k-means 法の様々な変形が広く使われている

$k-means$法

クラスターの代表点の集合をT、
Tの点$\vec{y}$を代表点とするクラスターを$S_{\vec{y}}$と表現.

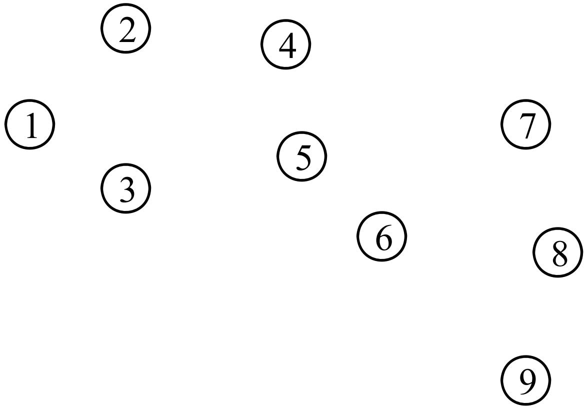
1. （初期化）Sからk個の点を選択し、Tの初期集合とする.

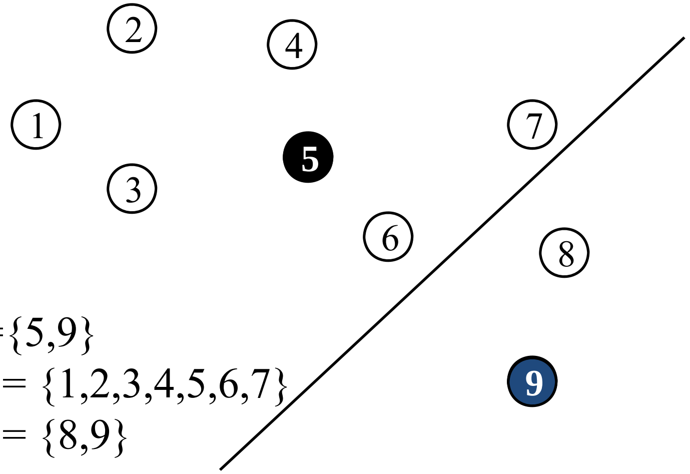
2. （再クラスタリング）各代表点$\vec{y}\in T$について$S_{\vec{y}}$を空集合にリセット.
Sの各点$\vec{x}$に最も近いTの点$\vec{y}$ $\left(\|\vec{x}-\vec{y}\| = \min_{\vec{z}\in T}\|\vec{x}-\vec{z}\|\right)$を計算し、$\vec{x}$を$S_{\vec{y}}$に追加.

3. （代表点を再計算）$S_{\vec{y}}$に登録された点全体の重心は代表点$\vec{y}$から
ずれている可能性がある.各代表点$\vec{y}\in T$を重心$c\left(S_{\vec{y}}\right)=\frac{1}{\left|S_{\vec{y}}\right|}\sum_{\vec{u}\in S_{\vec{y}}}\vec{u}$に更新.

4. 誤差二乗平均が改善しなくなるまで、ステップ２と３を繰返す.

k-means 法による 2-クラスタリング

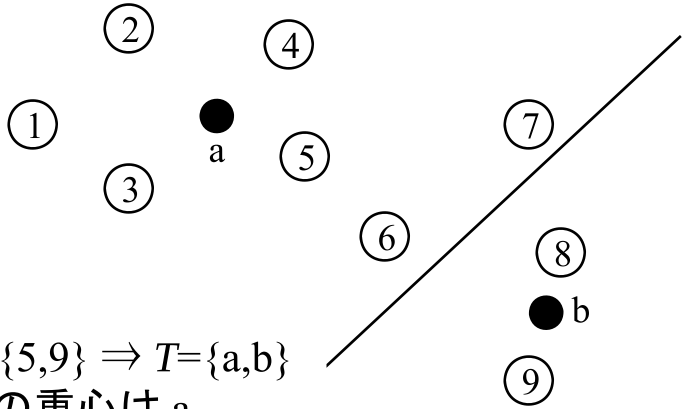




$T=\{5,9\}$
$S_5=\{1,2,3,4,5,6,7\}$
$S_9=\{8,9\}$

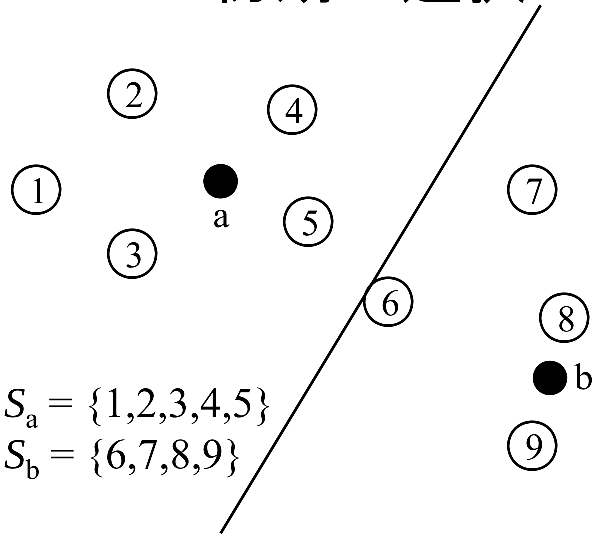
初期の選択

左下へ



$T=\{5,9\} \Rightarrow T=\{a,b\}$
S_5 の重心は a
S_9 の重心は b

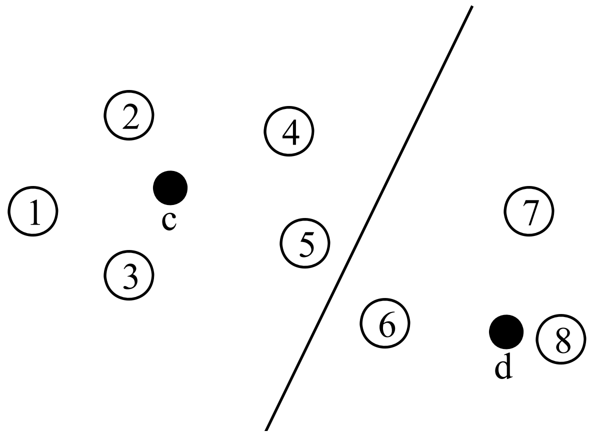
重心を再計算



$S_a=\{1,2,3,4,5\}$
$S_b=\{6,7,8,9\}$

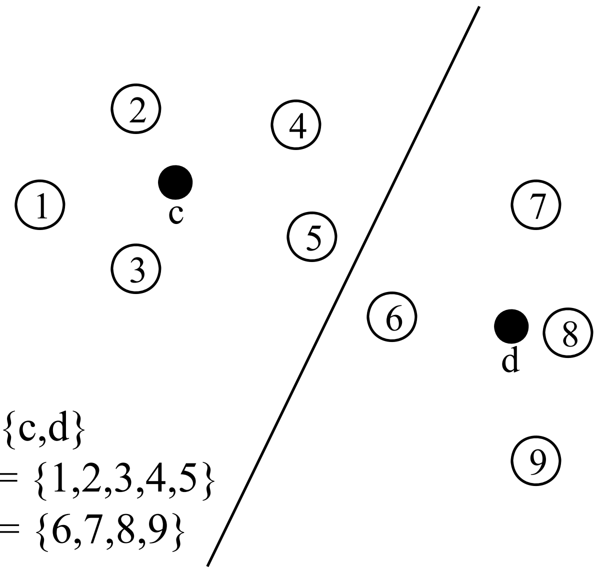
再クラスタリング

次ページ



$T=\{a,b\} \Rightarrow T=\{c,d\}$
S_a の重心は c
S_b の重心は d

重心を再計算



$T=\{c,d\}$
$S_c = \{1,2,3,4,5\}$
$S_d = \{6,7,8,9\}$

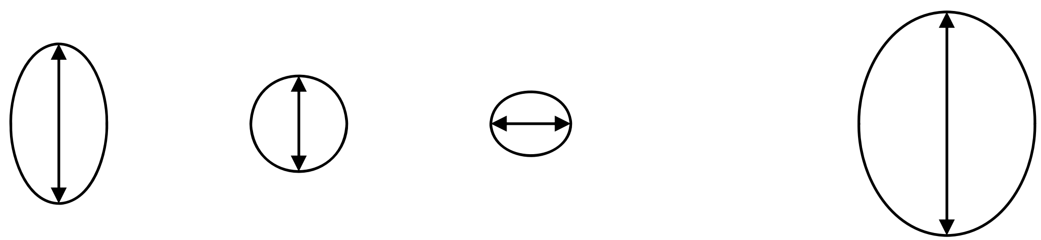
再クラスタリング

誤差二乗平均は収束
クラスターに変化なし

k-クラスタリング　クラスターの評価に直径を使う場合

S の $k-$クラスタリング $C=\{S_1, S_2, \cdots, S_k\}$ に対して

$$q(C)=\max\{\,diameter(S_i)\mid i=1,\cdots,k\,\}$$ と定義



$q(C)$ を最小化する $k-$クラスタリング C を
効率的に計算できるか？

与えられた B に対して $q(C)\le B$ となる $k-$クラスタリングが
存在するか否かを決定する問題はＮＰ完全

近似的解法

$opt = \min\left\{ q(C) \mid C \text{ は } S \text{ の } k\text{−クラスタリング} \right\}$ とおく

$q(C) \leq 2 \bullet opt$ となる k−クラスタリング C を生成する

アルゴリズムが存在する

Gonzalez's farthest point heuristics

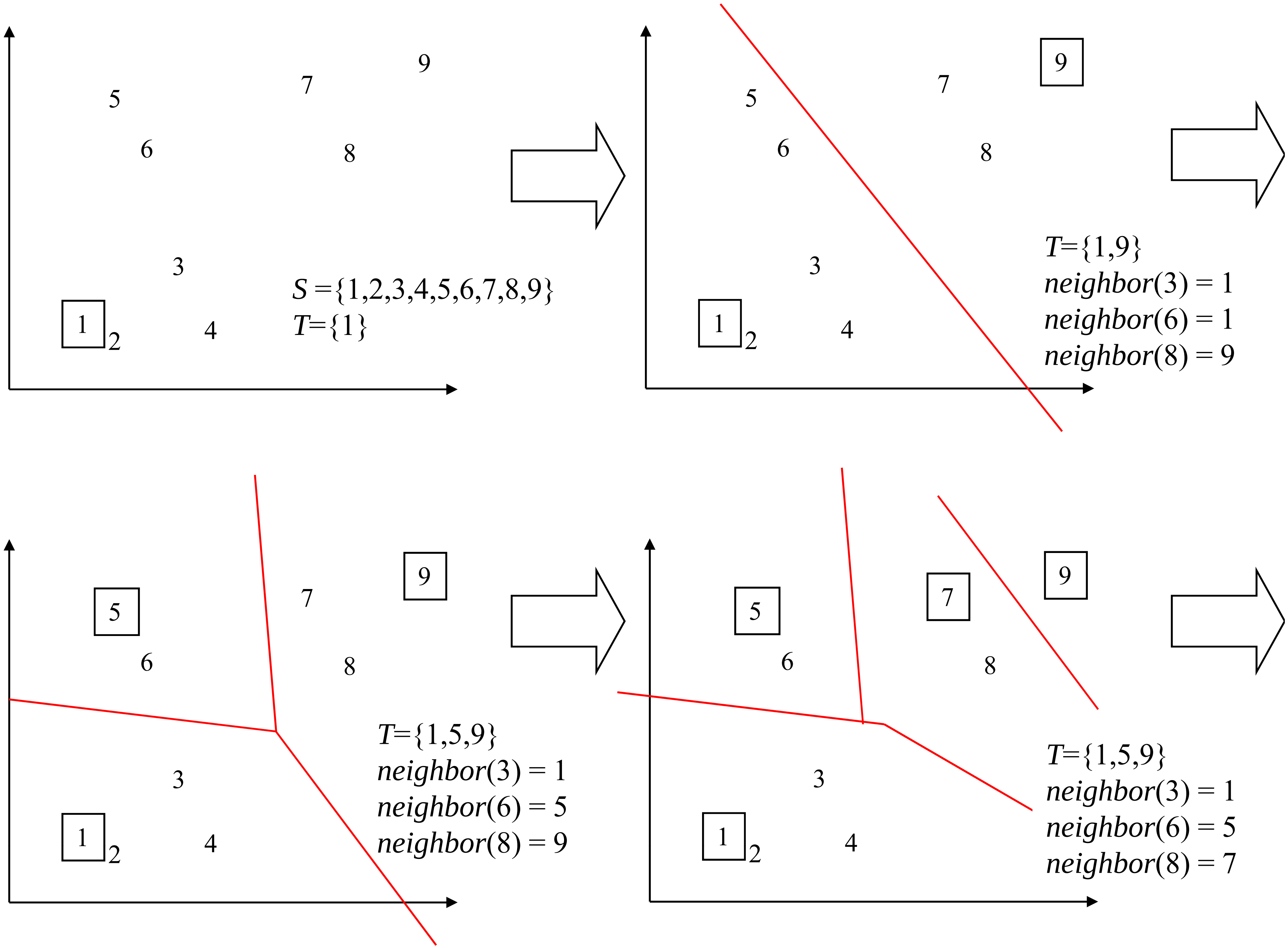
- T を S のクラスターの代表点の集合とし、空集合で初期化
- S から 1 点 $\vec{c}_1$ を選択し T に追加
- 各 $j = 2, \ldots, k$ について以下のステップを実行

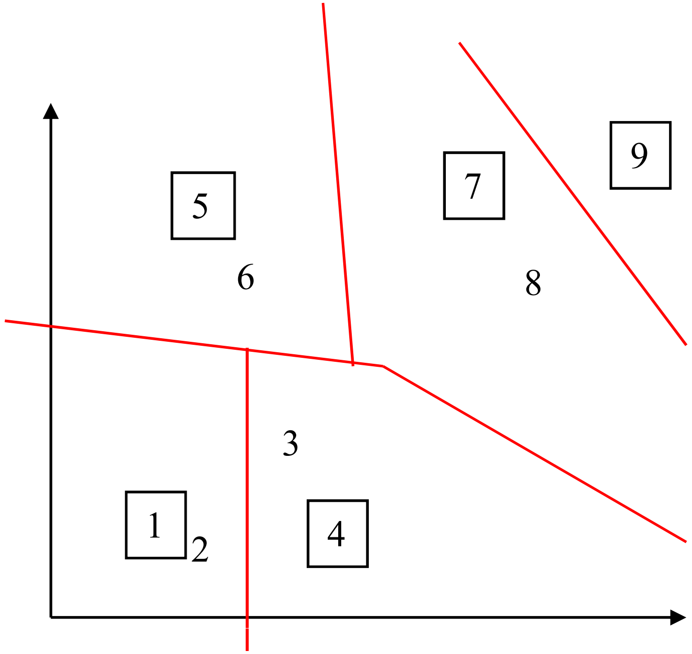
1. $\vec{x} \in S - T$ に最も近い T の点を $neighbor(\vec{x})$ と記述

$\vec{x}$ は $neighbor(\vec{x})$ を代表点とするクラスターに属すると定義

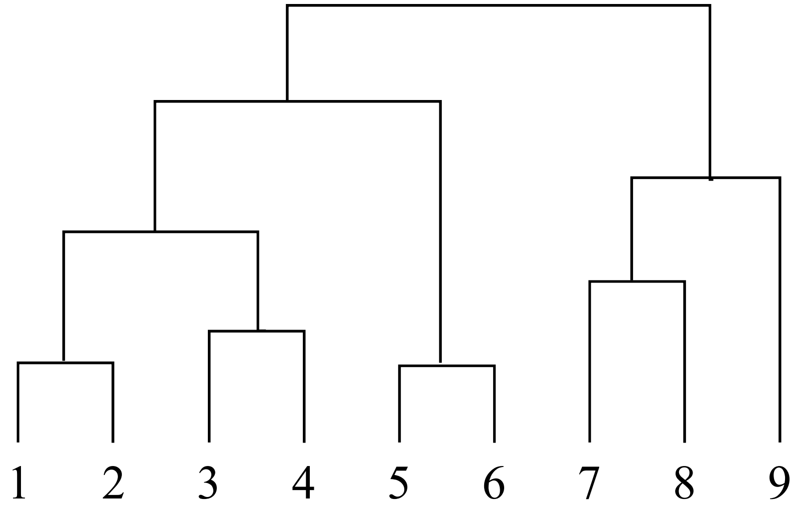
2. 属するクラスターの代表点との距離が最大の点 $\vec{c}_j \in S - T$ (farthest point) を T に追加

$$\left\| \vec{c}_j - neighbor(\vec{c}_j) \right\| = \max\left\{ \left\| \vec{x} - neighbor(\vec{x}) \right\| \mid \vec{x} \in S - T \right\}$$


S ={1,2,3,4,5,6,7,8,9}
T={1}
T={1,9}
neighbor(3) = 1
neighbor(6) = 1
neighbor(8) = 9
T={1,5,9}
neighbor(3) = 1
neighbor(6) = 5
neighbor(8) = 9
T={1,5,9}
neighbor(3) = 1
neighbor(6) = 5
neighbor(8) = 7



T={1,5,9}
neighbor(3) = 4
neighbor(6) = 5
neighbor(8) = 7



問題

Gonzalez's farthest point heuristics　において

$q(C) = 2 \bullet opt$ となる $k-$クラスタリング C

が生成される例をつくれ

点間の距離をできるだけ保存して高次元を低次元に埋め込みクラスターを視覚化する

Multi-dimensional Scaling

Latent Semantic Indexing

Self-Organizing Maps (SOM)

Multi-dimensional Scaling

高次元における2点 i,j 間の距離 $d_{i,j}$

点 i を低次元への写像した結果 $f(i)$

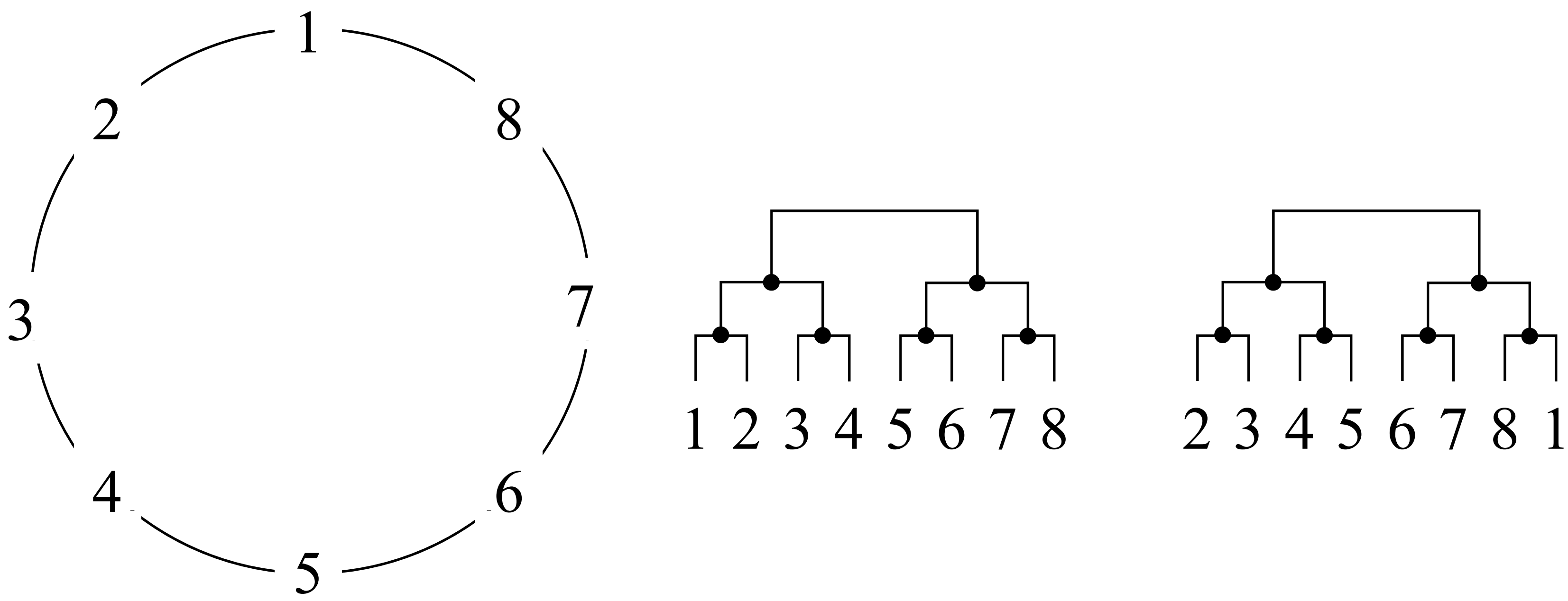
点間の距離をできるだけ保つ写像 f が望ましい

最小化したい指標 $$\frac{\sum_{i,j}(d_{f(i),f(j)} - d_{i,j})^2}{\sum_{i,j} {d_{i,j}}^2}$$

解答例

クラスタリングの結果に自由度があり、複数の
候補がありうる例を考えよ

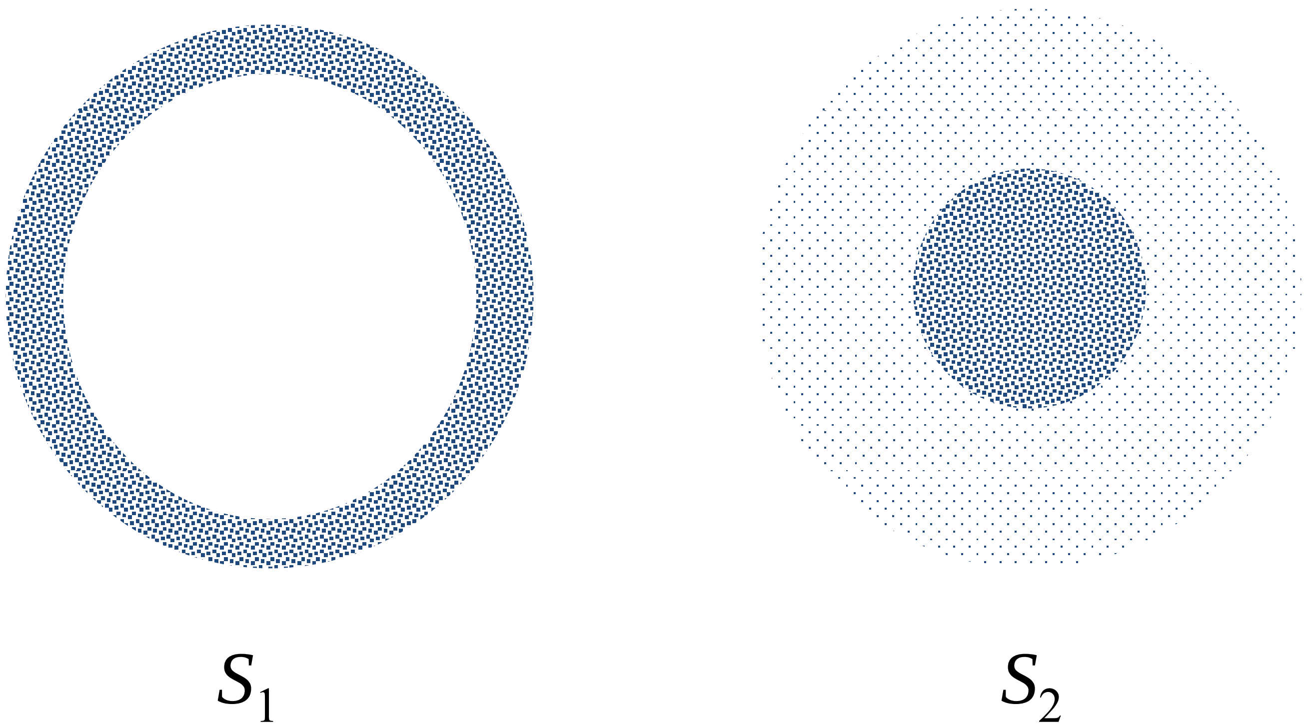
円周に等間隔に並んだ点列のクラスタリング



解答例

直径が同一で、重心からの距離の分散が大きく異なるクラスターの例を考えよ

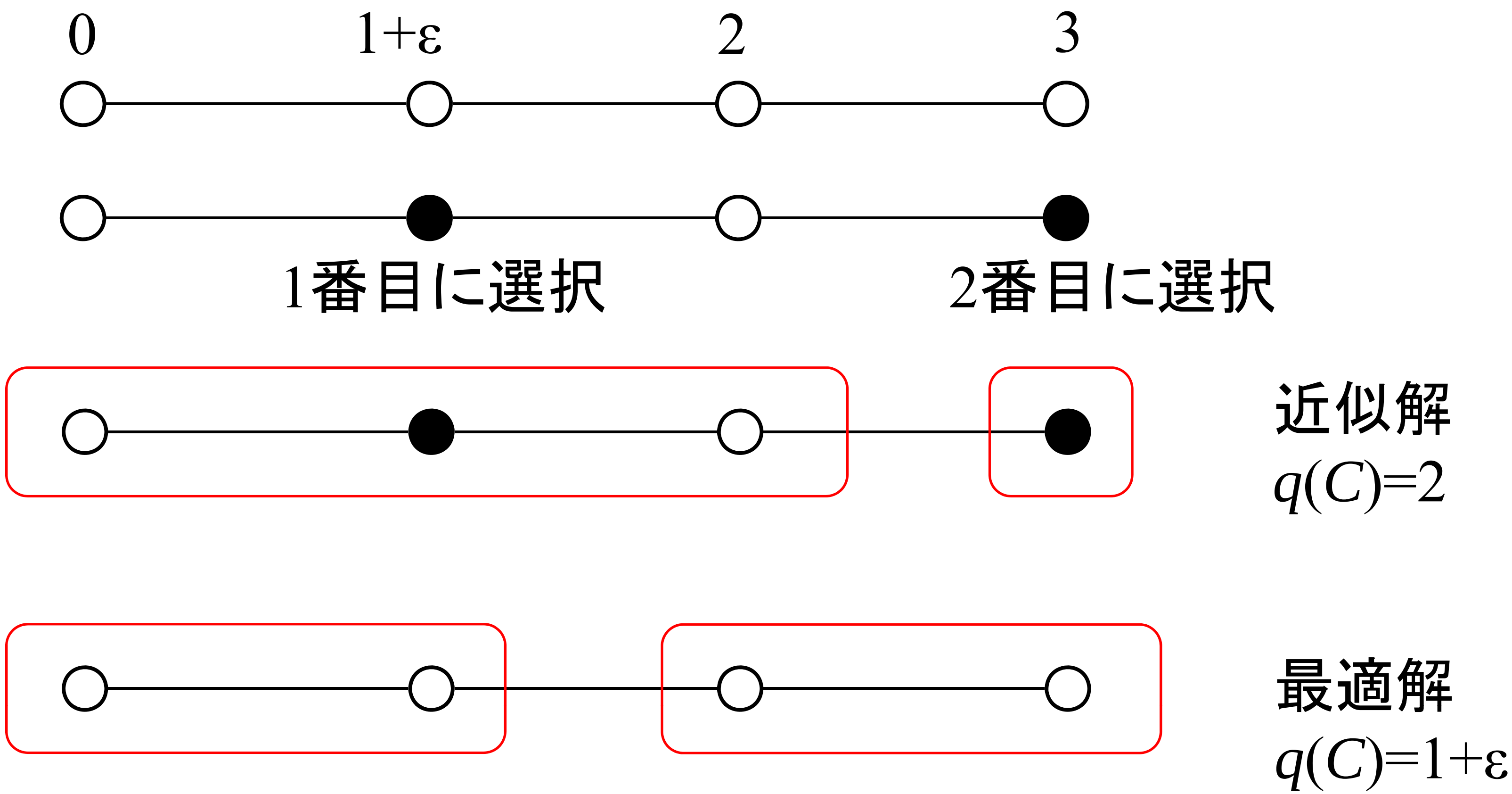
$$\mathrm{diameter}(S_1) = \mathrm{diameter}(S_2),\ \mathrm{var}(S_1) >> \mathrm{var}(S_2)$$



解答例

直線上に置かれた4点の 2-クラスタリング

ε>0 は限りなく0に近い数



付録　Gonzalez’s farthest point heuristics の証明

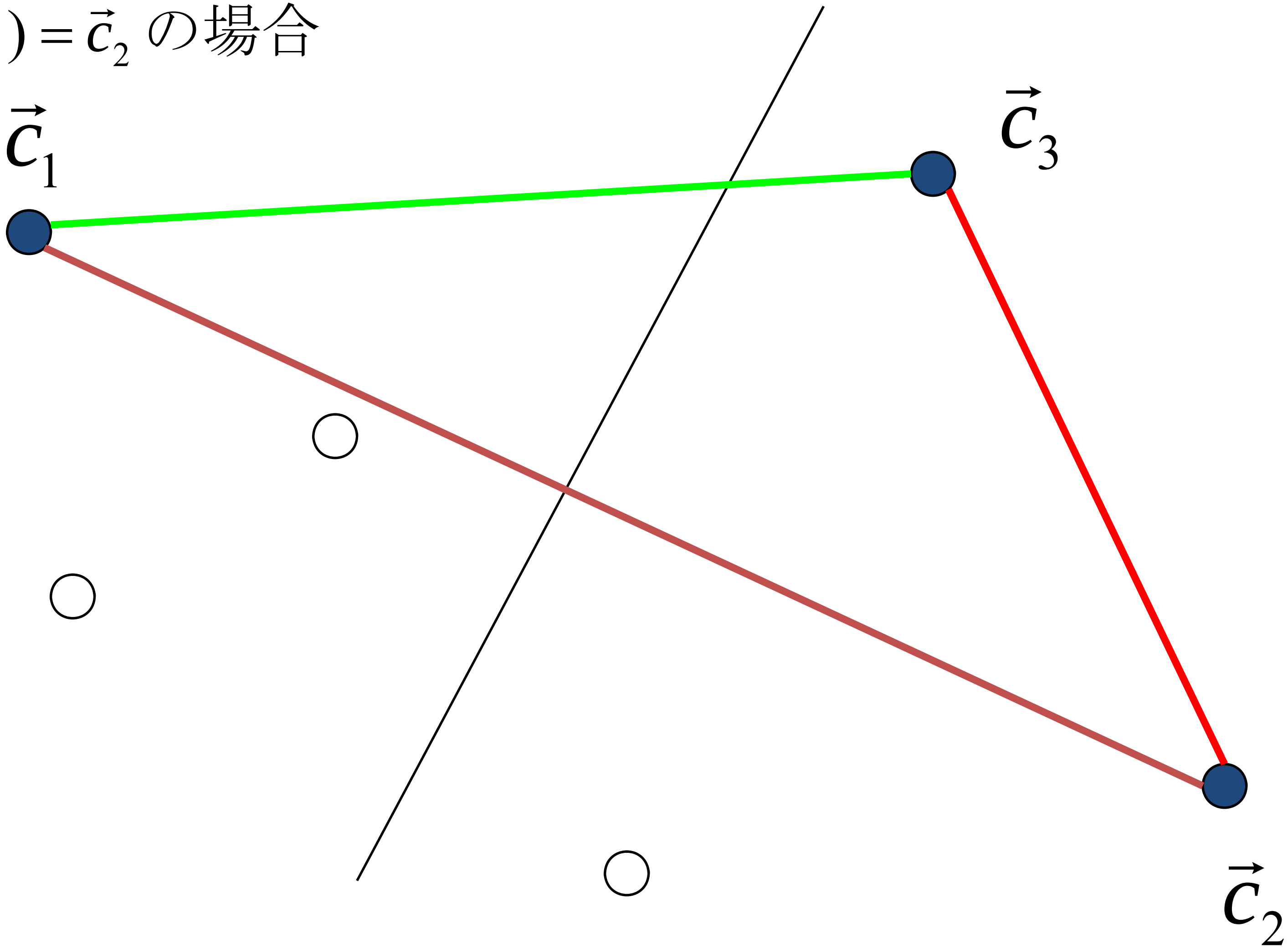
定理　Gonzalez’s farthest point heuristicsが生成するk－クラスタリングをC_G
$q(C)$が最小のk－クラスタリングをC_{opt}とすれば　$q(C_G) \le 2 \cdot q(C_{opt})$

補題　$T = \{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{j-1}\}$のとき、
$T \cup \{\vec{c}_j\}$の任意の２点は
$\|neighbor(\vec{c}_j) - \vec{c}_j\|$以上離れている

つまり$1 \le h < i \le j$について

$$\|neighbor(\vec{c}_j) - \vec{c}_j\| \le \|\vec{c}_h - \vec{c}_i\|$$

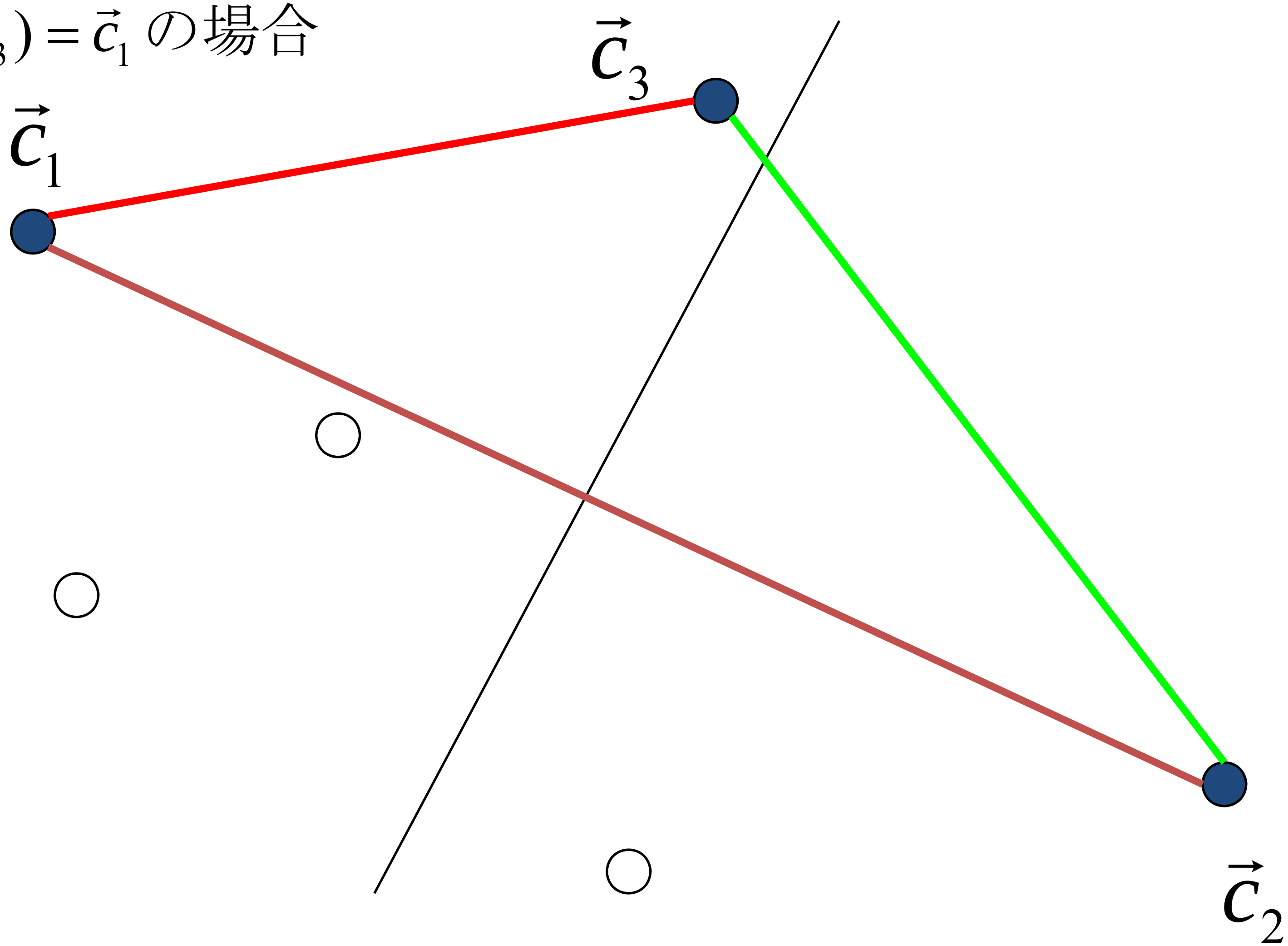
$neighbor(\vec{c}_3) = \vec{c}_2$ の場合



$neighbor(\vec{c}_3) = \vec{c}_2$なので　$\|\vec{c}_2 - \vec{c}_3\| \leq \|\vec{c}_1 - \vec{c}_3\|$

$\vec{c}_2$ が $\vec{c}_3$ より先に選択されたので　$\|\vec{c}_1 - \vec{c}_3\| \leq \|\vec{c}_1 - \vec{c}_2\|$

$neighbor(\vec{c}_3) = \vec{c}_1$ の場合



$neighbor(\vec{c}_3) = \vec{c}_1$ なので $\|\vec{c}_1 - \vec{c}_3\| \le \|\vec{c}_2 - \vec{c}_3\|$

$\vec{c}_2$ が $\vec{c}_3$ より先に選択されたので $\|\vec{c}_1 - \vec{c}_3\| \le \|\vec{c}_1 - \vec{c}_2\|$

補題　$T = \{\vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{j-1}\}$のとき、$1 \le h < i \le j$について

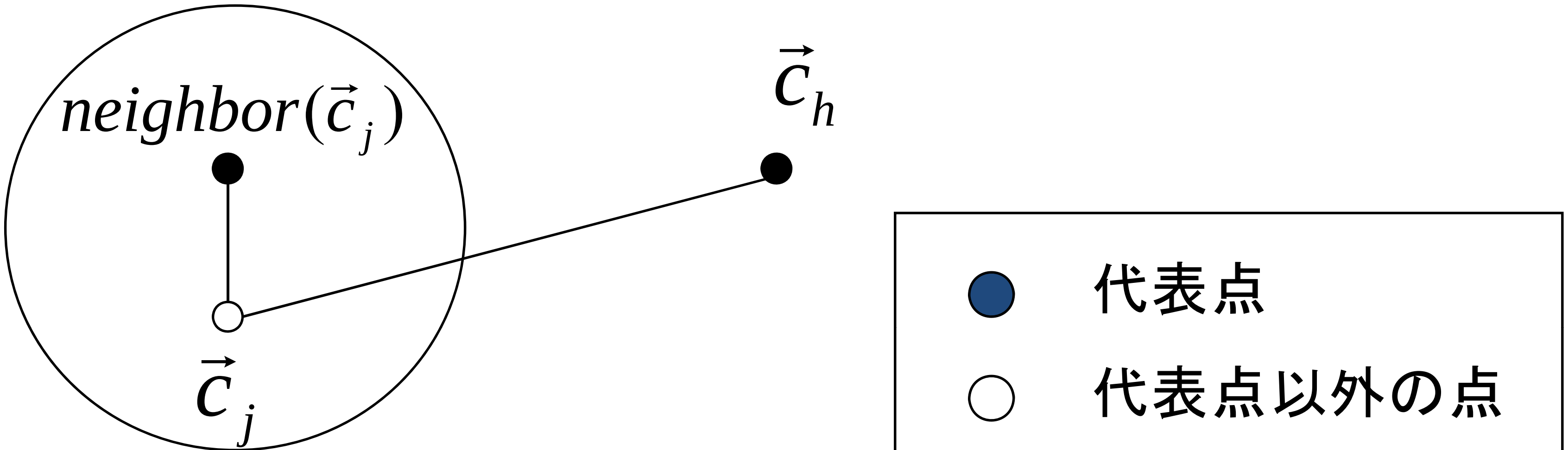
$$\|neighbor(\vec{c}_j) - \vec{c}_j\| \le \|\vec{c}_h - \vec{c}_i\|$$

つまり、$T \cup \{\vec{c}_j\}$の任意の２点間距離は$\|neighbor(\vec{c}_j) - \vec{c}_j\|$以上

jに関する帰納法.　　一般の場合を証明

- $i = j$　のとき、Tの中で最も$\vec{c}_j$に近いのは$neighbor(\vec{c}_j)$なので

$$\|neighbor(\vec{c}_j) - \vec{c}_j\| \le \|\vec{c}_h - \vec{c}_j\|$$

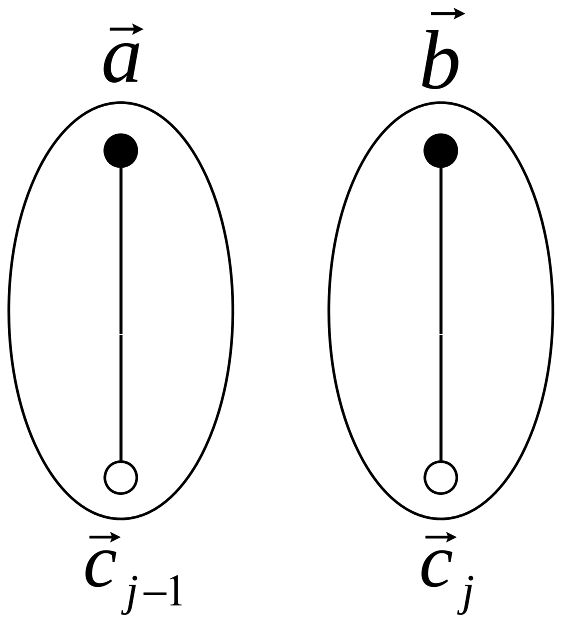


$i < j$　のとき：

- $j-2$個の代表点を選択した一つ前のステップの状態、

つまり$T = \left\{ \vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{j-2} \right\}$を考える。

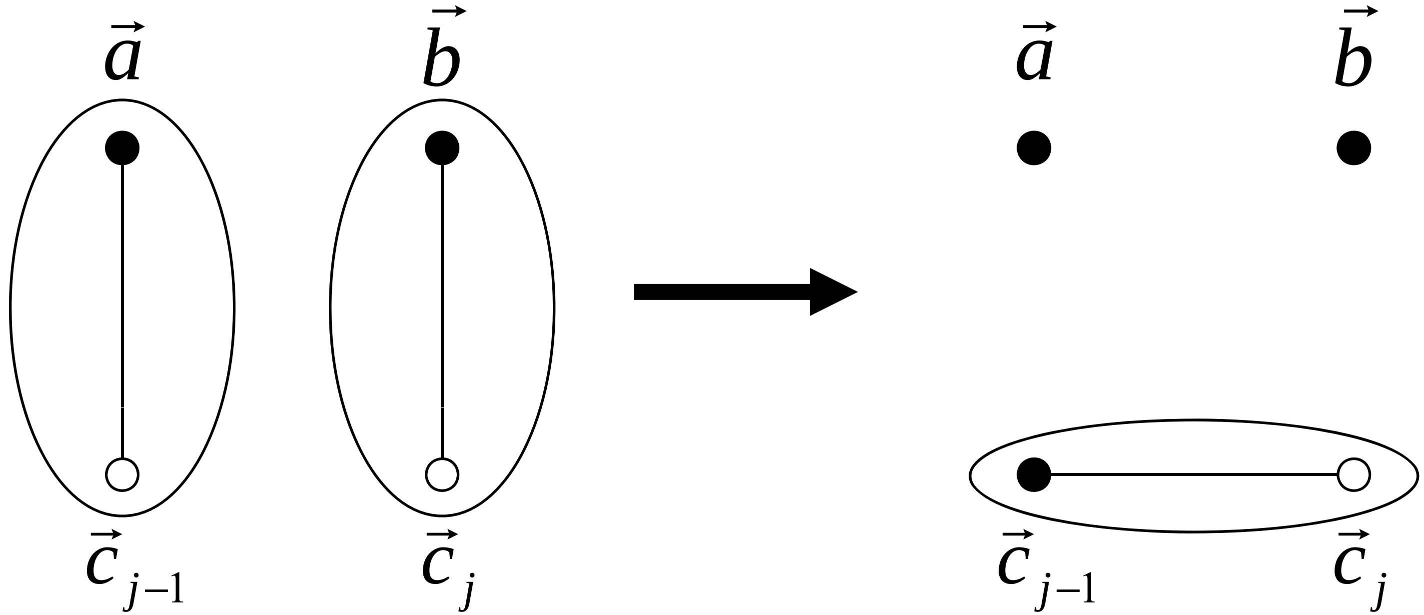
この時点での$\vec{c}_{j-1}$の属するクラスターの代表点 $neighbor(\vec{c}_{j-1})$を$\vec{a}$とすれば、帰納法の仮定より　$\left\| \vec{a} - \vec{c}_{j-1} \right\| \le \left\| \vec{c}_h - \vec{c}_i \right\|$



この時点で$\vec{c}_j$の属する
クラスターの代表点を$\vec{b}$

- $T = \left\{ \vec{c}_1, \vec{c}_2, \ldots, \vec{c}_{\underline{\underline{j-1}}} \right\}$ のとき、つまり $j-1$ 個代表点を選択後に
  $\left\| neighbor(\vec{c}_j) - \vec{c}_j \right\| \le \left\| \vec{a} - \vec{c}_{j-1} \right\|$ となることを示せば十分

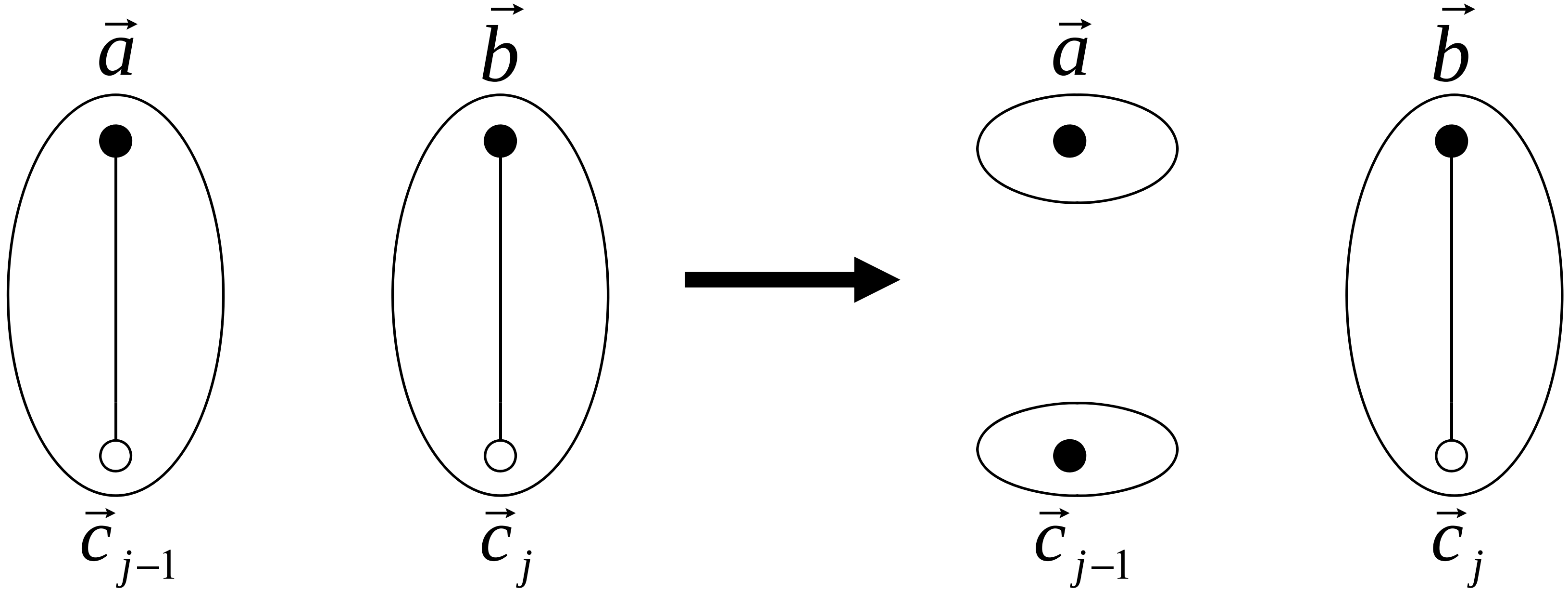
$neighbor(\vec{c}_j) = \vec{c}_{j-1}$ のとき



$\vec{c}_{j-1}$ が代表点として選択されたので
$\left\| \vec{b} - \vec{c}_j \right\| \le \left\| \vec{a} - \vec{c}_{j-1} \right\|$

$\vec{c}_j$ がより近い代表点 $\vec{c}_{j-1}$ のクラスターに移動したので
$\left\| neighbor(\vec{c}_j) - \vec{c}_j \right\| \le \left\| \vec{b} - \vec{c}_j \right\|$

$neighbor(\vec{c}_j) = \vec{b}$のとき、つまりクラスターを移動しないとき



$\vec{c}_{j-1}$が代表点として選択されたので　$\|neighbor(\vec{c}_j) - \vec{c}_j\| \leq \|\vec{a} - \vec{c}_{j-1}\|$

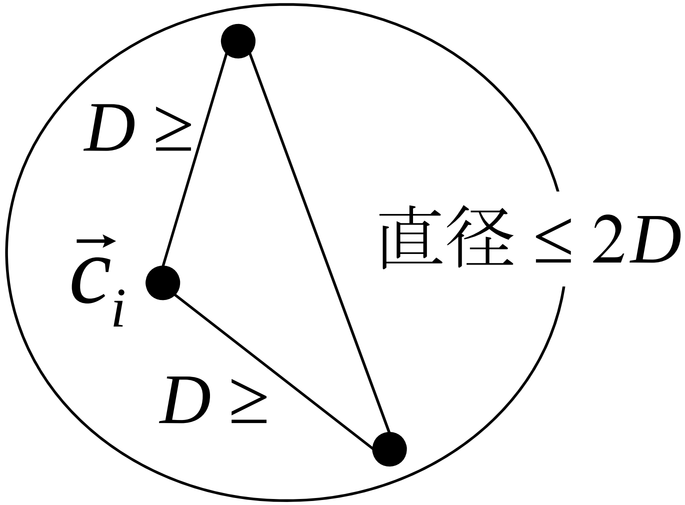
定理　Gonzalez's farthest point heuristicsが生成する$k-$クラスタリングをC_G
$q(C)$が最小の$k-$クラスタリングをC_{opt}とすれば　$q(C_G) \le 2 \cdot q(C_{opt})$

- Gonzalezの方法で選んだk個の代表点を$T = \{\vec{c}_1, \ldots, \vec{c}_k\}$
- ステップ２をもう一度実行して得られる代表点を$\vec{c}_{k+1}$，

  $D = \|\vec{c}_{k+1} - neighbor(\vec{c}_{k+1})\|$　とおく

- 各クラスターの直径は$2D$以下.　よって $\underline{q(C_G) \le 2D}$

ステップ2での代表点の選び方から
各クラスターの任意の元と
代表点との距離は最大でも D



- $k+1$個の代表点$T \cup \{\vec{c}_{k+1}\}$の任意の２点間の距離はD以上（補題より）

- 最適なクラスタリングC_{opt}のk個のクラスターのどれかは$k+1$個の代表点$T \cup \{\vec{c}_{k+1}\}$のうち２点を含むので　$\underline{D \le q(C_{opt})}$

- $q(C_G) \le 2D$より、$q(C_G) \le 2 \cdot q(C_{opt})$