クレジット:
UTokyo Online Education 数理手法Ⅲ 2018 寒野善博

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

最小木問題と階層的クラスタリング

1. 最小木問題

連結で閉路をもたないグラフのことを，木とよぶ．また，連結な無向グラフが与えられたとき，その部分グラフのうち元のグラフの頂点をすべてもつ木のことを全域木とよぶ．最小木問題とは，重み付き無向グラフ $G = (V, E, c)$ が与えられたとき，G の全域木のうち重みが最小のものを求める問題である．以下では，簡単のため，G の辺の重みはすべて互いに異なるものとする．

最小木問題は，クラスカル (Kruskal) のアルゴリズム（アルゴリズム 1）やプリム (Prim) のアルゴリズムを用いて解くことができる．

アルゴリズム 1（クラスカルのアルゴリズム）

1: G の辺を重みが小さい順に並べたものを $e_1, e_2, \ldots, e_{|E|}$ とおく．
2: $T \leftarrow \emptyset$, $k \leftarrow 1$.
3: **while** $|T| < |V| - 1$ **do**
4:　　**if** $T \cup \{e_k\}$ が閉路をもたない **then**
5:　　$T \leftarrow T \cup \{e_k\}$.
6:　　**end if**
7:　　$k \leftarrow k + 1$.
8: **end while**

2. 最大間隔 k-クラスタリング

いくつかのデータ点が与えられたとき，それらを似たものどうしのいくつかのグループに分類することをクラスタリングとよび，個々のグループのことをクラスターとよぶ．

たとえば，図 2 は，図 1 の 10 個の点からなるデータを 4 個のクラスターに分けた例である．ただし，データ点は平面上にあり，似た点どうしは近くにあって，似ていない点どうしは遠くにあるものとしている．このように，2 つのデータ点の似ていない度合いを測る尺度を，クラスタリングでは距離とよぶ．したがって，この距離は，物理的な距離とは異なるものであって構わない．

似たデータ点を同じクラスターに入れるということは，クラスター間の間隔をなるべく大きくするということでもある．データ点を $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m \in \mathbb{R}^n$ で表し，これらをクラスター $C_1, \ldots, C_k$ に分類することを考える．点 $\boldsymbol{p}_l$ と点 $\boldsymbol{p}_h$ との距離を $d(\boldsymbol{p}_l, \boldsymbol{p}_h)$ で表し，クラスター C_i とクラスター C_j との間隔を

$$D(C_i, C_j) = \min\{d(\boldsymbol{p}_l, \boldsymbol{p}_h) \mid \boldsymbol{p}_l \in C_i,\ \boldsymbol{p}_h \in C_j\}$$

で定義する．つまり，C_i に属する点と C_j に属する点で，最も近いものの間の距離を，C_i と C_j の間隔とする．そして，間隔 $D(C_1, C_2), D(C_1, C_3), \ldots, D(C_{k-1}, C_k)$ のうち最小のものが最大になるようにクラスター $C_1, \ldots, C_k$ を定めたい．このようなクラスタリングを，最大間隔 k-クラスタリングとよぶ [*1]．

[*1] クラスター間の距離やクラスターの大きさには多様な定め方があり，最大間隔 k-クラスタリング以外にもさまざまなクラスタリング手法が実際に用いられている．

例として，図 1 の 10 個の点からなるデータを考える．ただし，データ点間の距離は平面上の物理的な距離（Euclid 距離）であるものとする．図 2 は，$k = 4$ としたときの最大間隔 k-クラスタリングを示している．

図 1 の例についてさらに考察すると，$k = 10$ は各データ点自身がそれぞれ 1 つのクラスターを成している場合に対応している．次に，$k = 9, 8, \ldots$ とクラスター数を減らしていくと，最終的に $k = 1$ はすべてのデータ点が同じクラスターに属する場合に対応する．このようにクラスター数を一つずつ減らしていったときの最大間隔 k-クラスタリングの結果は，図 3 のような樹形図（デンドログラム）にまとめることができる．この図は，たとえば $k = 9$ のときは，点 6 と点 8 が同じクラスターに属し，それ以外の点それぞれが一つのクラスターを成していることを意味している．そして，横軸は点

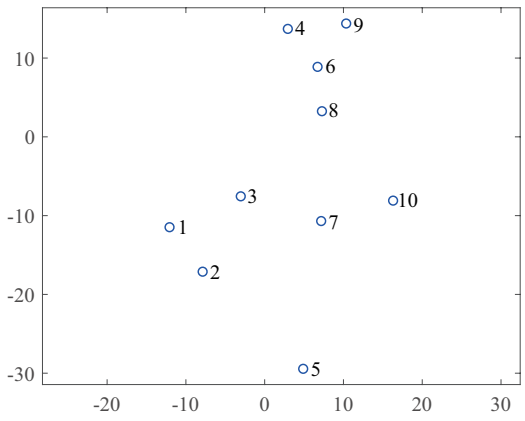

図 1: 平面上のデータ

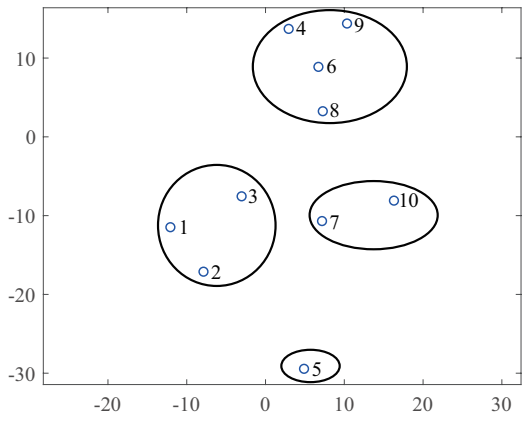

図 2: クラスタリングの例．図 1 のデータを 4 個のクラスターに分類した場合

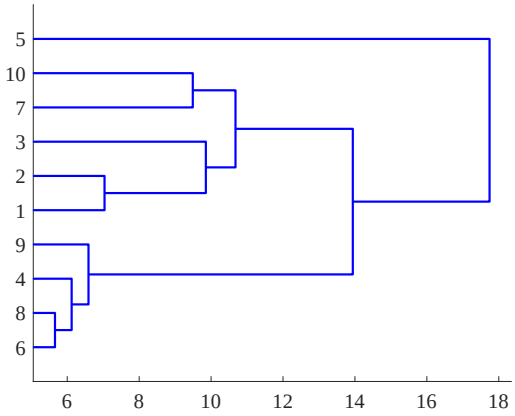

図 3: 樹形図（デンドログラム）

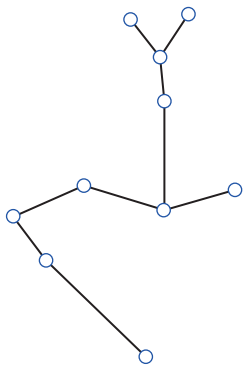

図 4: 最小木

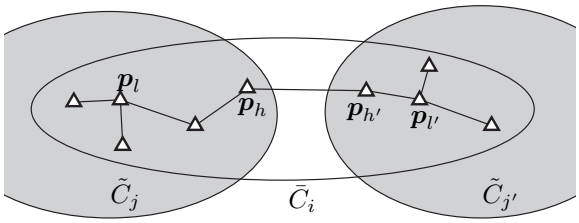

図 5: 命題 1 の証明で用いる図

6 と点 8 の間の距離を表している．また，たとえば $k = 8$ のときは，点 4, 点 6, 点 8 が同じクラスターに属し，それ以外の 7 個の点それぞれが一つのクラスターを成している．このように，いくつかのクラスターをまとめて新たなクラスターとするような包含関係をもつクラスターの分析法を，階層的クラスタリングとよぶ．特に，クラスター間の間隔を前述の $D(C_i, C_j)$ のように定めた階層的クラスタリングのことは，単リンク法とよばれている．

単リンク法は，次に述べるように，最小木（図 4）と密接な関係がある．

命題 1. データ点 $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m$ を頂点とみなし，任意の 2 頂点の間に辺を張り，辺の重みはデータ点間の距離として得られるグラフ G を考える．G にクラスカルのアルゴリズムを適用して，辺が $m - k$ 本だけ得られた状態で停止する．このときに得られているグラフを $\bar{G}$ とおくと，$\bar{G}$ の連結成分は最大間隔 k-クラスタリングである．

> **証明**. $\bar{G}$ の連結成分を $\bar{C}_1, \ldots, \bar{C}_k$ で表す．クラスカルのアルゴリズムで次に加えられる辺の重みを d^* とおくと，d^* は $\bar{C}_1, \ldots, \bar{C}_k$ のうち最も近いクラスター間の間隔に等しい．そこで，$\bar{C}_1, \ldots, \bar{C}_k$ とは異なるクラスタリング $\tilde{C}_1, \ldots, \tilde{C}_k$ を考えると，そのうち最も小さいクラスター間の間隔が d^* 以下であることを示せばよい．
>
> いま，クラスタリング $\bar{C}_1, \ldots, \bar{C}_k$ とクラスタリング $\tilde{C}_1, \ldots, \tilde{C}_k$ とは異なることから，前者では同じクラスターに属する 2 点で後者では別のクラスターに属するものが存在する．その 2 点を $\boldsymbol{p}_l, \boldsymbol{p}_{l'}$ とおき，これらは $\bar{C}_i$ に属するとする．また，$\boldsymbol{p}_l$ は $\tilde{C}_j$ に属し，$\boldsymbol{p}_{l'}$ は $\tilde{C}_{j'}$ に属するとする（図 5）．いま，$\boldsymbol{p}_l$ と $\boldsymbol{p}_{l'}$ は $\bar{G}$ の連結成分 $\bar{C}_i$ に属する頂点であるから，クラスカルのアルゴリズムを停止した時点で $\boldsymbol{p}_l$ から $\boldsymbol{p}_{l'}$ への道が得られており，その道に属する辺の重みはすべて d^* 以下である．また，その道の辺のうち，片方の端点は $\tilde{C}_j$ に属しもう一方の端点は $\tilde{C}_{j'}$ に属するものが存在する（図 5 では，$\boldsymbol{p}_h$ と $\boldsymbol{p}_{h'}$ を結ぶ辺がこれにあたる）．この辺の長さも d^* 以下であるから，クラスター $\tilde{C}_j$ とクラスター $\tilde{C}_{j'}$ との間隔も d^* 以下である． □

図 4 は，図 1 のデータに対応するグラフの最小木である．この最小木の辺を短い順につないでいってみれば，図 3 の樹形図が得られることが確認できる．

（以上）