クラスター分析 - 階層的方法

数理科学続論J

(Press `?` for help, `n` and `p` for next and previous slide)

村田 昇

2018.12.06

講義の予定

- **第1日: クラスター分析と階層的クラスタリング**
- 第2日: 非階層的クラスタリング

クラスター分析

3.1

クラスター分析とは

- 個体の間に隠れている **集まり=クラスター** を発見する方法
- 個体間の距離・類似度を定義:
  - 同じクラスターに属する個体どうしは近い性質をもつ
  - 異なるクラスターに属する個体どうしは異なる性質をもつ
- さらなるデータ解析やデータの可視化に利用
- 教師なし学習の代表的な手法の一つ

クラスター分析の考え方

- 階層的方法:
  - データ点およびクラスターの間に **距離** を定義
  - 距離に基づいてグループ化:
    - 近いものから順にクラスターを **凝集**
    - 近いもの同士が残るようにクラスターを **分割**
- 非階層的方法:
  - クラスターの数を事前に指定
  - クラスターの **集まりの良さ** を評価する損失関数を定義
  - 損失関数を最小化するようにクラスターを形成

階層的クラスタリング

4.1

凝集的方法の手続き

1. データ・クラスター間の距離を定義する
   - データ点とデータ点の距離
   - クラスターとクラスターの距離
   - (データ点とクラスターの距離はデータ1点をクラスターと考える)
2. データ点および形成されているクラスターを対象にそれぞれの間の距離を求める
3. 最も近い2つを統合し新たなクラスターを作成する (データ点とデータ点，データ点とクラスター，クラスターとクラスタのいずれの場合もあり得る)
4. クラスター数が目的の数になるまで2,3の手続きを繰り返す

4.2

データ間の距離

- データ: 変数の値を成分としてもつベクトル

$$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top, \boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp})^\top \in \mathbb{R}^p$$

- 距離: $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$
- 代表的なデータ間の距離:
  - ユークリッド距離 (Euclidean distance)
  - ミンコフスキー距離 (Minkowski distance)
  - マンハッタン距離 (Manhattan distance)

ユークリッド距離

- 最も一般的な距離
- 各成分の差の2乗和の平方根 (2ノルム)

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

ミンコフスキー距離

- ユークリッド距離を q 乗に一般化した距離
- 各成分の差の q 乗和の q 乗根(q ノルム)

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left\{ |x_{i1} - x_{j1}|^q + \cdots + |x_{ip} - x_{jp}|^q \right\}^{1/q}$$

4.5

マンハッタン距離

- $q = 1$ のミンコフスキー距離
- 格子状に引かれた路に沿って移動するときの距離

$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) = |x_{i1} - x_{j1}| + \cdots + |x_{ip} - x_{jp}|$$

4.6

クラスター間の距離

- データ点同士の距離をどのように使うかで定義
  - データ点の距離から陽に定義する方法
  - クラスターを統合したときに成り立つクラスター間の距離の関係を用いて再帰的に定義する方法
- クラスター: いくつかのデータ点からなる集合

$$C_a = \{\boldsymbol{x}_i | i \in \Lambda_a\}, \quad C_b = \left\{\boldsymbol{x}_j | j \in \Lambda_b\right\}$$

- 2つのクラスター間の距離: $D(C_a, C_b)$
- 代表的なクラスター間の距離
  - 最短距離法 (単連結法; single linkage method)
  - 最長距離法 (完全連結法; complete linkage method)
  - 群平均法 (average linkage method)

4.7

最短距離法

- 最も近い対象間の距離を用いる方法:

$$D(C_a, C_b) = \min_{\boldsymbol{x}_i \in C_a,\ \boldsymbol{x}_j \in C_b} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- 統合前後のクラスター間の関係:

$$D(C_a + C_b, C_c) = \min\left\{ D(C_a, C_c), D(C_b, C_c) \right\}$$

4.8

最長距離法

- 最も遠い対象間の距離を用いる方法:

$$D(C_a, C_b) = \max_{\boldsymbol{x}_i \in C_a,\ \boldsymbol{x}_j \in C_b} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- 統合前後のクラスター間の関係:

$$D(C_a + C_b, C_c) = \max\left\{D(C_a, C_c), D(C_b, C_c)\right\}$$

4.9

群平均法

- 全ての対象間の平均距離を用いる方法:

$$D(C_a, C_b) = \frac{1}{|C_a||C_b|} \sum_{\boldsymbol{x}_i \in C_a,\ \boldsymbol{x}_j \in C_b} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  ただし $|C_a|, |C_b|$ はクラスター内の要素の数を表す
- 統合前後のクラスター間の関係:

$$D(C_a + C_b, C_c) = \frac{|C_a| D(C_a, C_c) + |C_b| D(C_b, C_c)}{|C_a| + |C_b|}$$

距離計算に関する注意

- データの性質に応じて距離は適宜使い分ける
  - データ間の距離の選択
  - クラスター間の距離の選択
- 変数の正規化は必要に応じて行う
  - 物理的な意味合いを積極的に利用する場合はそのまま
  - 単位の取り方などによる分析の不確定性を避ける場合は平均0，分散1に正規化
- データの性質を鑑みて適切に前処理

演習: 階層的クラスタリング

- 11-hclust.r を確認してみよう