クラスター分析 - 非階層的方法

数理科学続論J

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

村田 昇

2019.12.13

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**

2.1

クラスター分析の復習

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
- 個体間の類似度・距離(非類似度)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

3.2

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

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
   - 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
2. 形成されているクラスターの間の距離を計算
3. 最も近い2つを統合し新たなクラスターを作成
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

3.4

非階層的クラスタリング

4.1

非階層的方法

- 対象とするデータ: p 次元変数
  $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top$
- 観測データ: n 個の個体
  $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \ (i = 1, 2, \ldots, n)$
- 推定する関係式: 対応 C (個体 i が属するクラスター番号 $C(i)$)
- 非階層的クラスタリング:
  - 対応 C の **全体の良さ** を評価する損失関数を設定
  - 観測データ $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ の最適な対応関係 $C(i)$ を決定

k-平均法の損失関数

- クラスターの個数 k を指定
- 2つの個体 i, i' の **近さ=損失** を距離の二乗で評価

$$\| \boldsymbol{x}_i - \boldsymbol{x}_{i'} \|^2 = \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

- 損失関数 $W(C)$: クラスター内の平均の近さを評価

k-平均法の性質

- クラスター l に属する個体の平均:

$$\bar{\boldsymbol{x}}_l = \frac{1}{n_l} \sum_{i:C(i)=l} \boldsymbol{x}_i$$

- 損失関数 $W(C)$ の等価な表現:

$$W(C) = 2 \sum_{l=1}^{k} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_l\|^2$$

4.4

クラスター対応の最適化

- 最適化: 損失関数 $W(C)$ を最小とする C を決定
- 貪欲な C の探索:
  - 原理的には全ての値を計算すればよい
  - 可能な C の数: k^n 通り (有限個のパターン)
  - サンプル数 n が小さくない限り実時間での実行は不可能
- 近似的な C の探索:
  - いくつかのアルゴリズムが提案されている
  - 基本的な考え方: **Lloyd-Forgyのアルゴリズム**

$$\bar{\boldsymbol{x}}_l = \arg\min_{\mu} \sum_{i:C(i)=l} \|\boldsymbol{x}_i - \boldsymbol{\mu}\|^2$$

(標本平均と変動の平方和の性質を利用)

4.5

Lloyd-Forgyのアルゴリズム

1. クラスター中心の初期値 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ を与える
2. 各データの所属クラスター番号 $C(i)$ を求める

$$C(i) = \arg\min_l \|\boldsymbol{x}_i - \boldsymbol{\mu}_l\|$$

3. 各クラスター中心 $\boldsymbol{\mu}_l$ $(l = 1, 2, \ldots, k)$ を更新する

$$\boldsymbol{\mu}_l = \frac{1}{n_l} \sum_{i:C(i)=l} \boldsymbol{x}_i$$

(n_l は $C(i) = l$ となるデータの総数)

4. 中心が変化しなくなるまで 2,3 を繰り返す

4.6

Lloyd-Forgyのアルゴリズムの性質

- 結果は確率的で初期値 $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ に依存
- アルゴリズムの成否は確率的
  (最適解が得られない場合もある)
- 一般には複数の初期値をランダムに試して損失を最小とする解を採用

4.7

R: 関数 `kmeans( )`

- k-平均法を実行するための標準的な関数
  - クラスターの数 k はオプション `centers` で指定
  - オプション `algorithm` で最適化アルゴリズムを指定
    (既定値は Hartigan-Wong アルゴリズム)
  - オプション `nstart` で初期値の候補の数を指定
- 結果は変数のスケールにも依存
  - 例えば測定値の単位により異なる
  - 必要ならば主成分分析の場合と同様に実行前にデータを標準化する

演習: 非階層的クラスタリング

- 12-kmeans.r を確認してみよう

4.9

クラスター構造の評価指標

凝集係数 (agglomerative coefficient)

- 階層的方法の評価
- データ $\boldsymbol{x}_i$ と最初に統合されたクラスター C の距離:

$$d_i = D(\boldsymbol{x}_i, C)$$

- 最後に統合された2つのクラスター C', C'' の距離:

$$D = D(C', C'')$$

- **凝集係数** AC:

凝集係数の性質

- 定義より $0 \leq AC \leq 1$
- 1に近いほどクラスター構造が明瞭
- banner plot の面積比
  (banner plot: l_i をデータ毎に並べた棒グラフ)

5.3

シルエット係数 (silhouette coefficient)

- 非階層的方法の評価 (階層的方法でも利用可)
- C^1, C^2: $\boldsymbol{x}_i$ を含む，および一番近いクラスター
- C^1 と $\boldsymbol{x}_i$ の距離: $d_i^1 = D(\boldsymbol{x}_i, C^1 \setminus \boldsymbol{x}_i)$
- C^2 と $\boldsymbol{x}_i$ の距離: $d_i^2 = D(\boldsymbol{x}_i, C^2)$
- **シルエット係数** S_i:

$$S_i = \frac{d_i^2 - d_i^1}{\max(d_i^1, d_i^2)}$$

5.4

シルエット係数の性質

- 定義より $-1 \leq S_i \leq 1$
- 1に近いほど適切なクラスタリング
- 全体の良さを評価するには S_i の平均を用いる

5.5

演習: クラスター分析の評価

- 12-eval.r を確認してみよう

5.6