L08 クラスタ – 分析・クラスタ分析

樋口さぶろお　https://hig3.net

龍谷大学 先端理工学部 数理・情報科学課程

多変量解析☆演習　L08(2021-11-18 Thu)

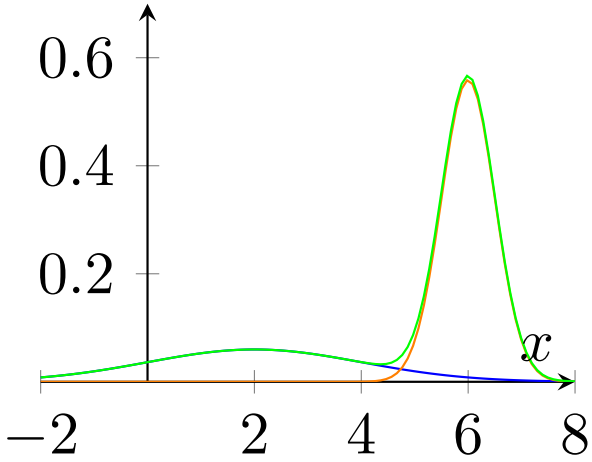
最終更新: Time-stamp: "2021-11-19 Fri 12:26 JST hig"

今日の目標

- ロジスティック回帰, 判別分析とクラスタ分析 (クラスタリング) の違いを説明できる
- 非/階層的クラスタ分析の違いを説明できる
- 主なクラスタリングアルゴリズムを説明できる
- scikit-learn でクラスタ分析できる

L06-Q1

Quiz 解答:混合ガウス分布の確率密度関数



L06-Q2

Quiz 解答:混合ガウス分布の確率

1. $f(4) = \frac{3}{10}\frac{1}{\sqrt{2\pi\cdot 2^2}}\mathrm{e}^{-\frac{(4-2)^2}{2\cdot 2^2}} + \frac{7}{10}\frac{1}{\sqrt{2\pi\cdot(1/2)^2}}\mathrm{e}^{-\frac{(4-6)^2}{2\cdot(1/2)^2}}$.
2. $P(X \leq 4) = \frac{3}{10}(\frac{1}{2} + I(\frac{4-2}{2})) + \frac{7}{10}(\frac{1}{2} + I(\frac{4-6}{1/2}))$.

ここまで来たよ

6 混合ガウス分布

8 クラスタ – 分析・クラスタ分析
- 教師なし学習としてのクラスター分析
- 階層的クラスタリング
- 平方和の分解
- 非階層的クラスタリング, k-means

クラスター分析

ロジスティック回帰, ナイーブベイズ, 線形判別分析 分類. 教師あり学習 (supervised learning) $y_{\mathrm{train}i}$ は教師データ

- モデルを仮定. 特に, $y = 0, 1$ の 2 クラス. (とにかくクラスの個数は固定).
- 訓練データ $(x_{\mathrm{train}i}, y_{\mathrm{train}i})$ $(i = 1, \ldots, n)$ で学習 (=パラメタを推定).
- モデルとパラメタを使ってテストデータ x_{test} に対して, y_{test} を予測.

クラスター分析, クラスタリング 分類. 教師なし学習 (unsupervised learning)

- クラスの個数は指定, または後出しで決める. 自動的には決まらない.
- 訓練データ $(x_{\mathrm{train}i})$ $(i = 1, \ldots, n)$ を, 「近い」点をまとめてクラスター (データ点の集合) に類別.
  - 一発の式はなく, 反復法的な感じで.

ニュートン法 数値計算法

- テストデータ x_{test} がさらに与えられても分類できる.

データ空間での距離

データ点 $\boldsymbol{x}_i \in \mathbb{R}^p$.
x_{ik} $(i = 1, \ldots, n, k = 1, \ldots, p)$. i 番目のデータ (行) の k 番目のコラムの値.
近い $\Leftrightarrow$ 距離が小さい

距離空間, 位相

データ点 $\boldsymbol{x}, \boldsymbol{x}'$ の間の距離には様々なものが考えられる. 例.

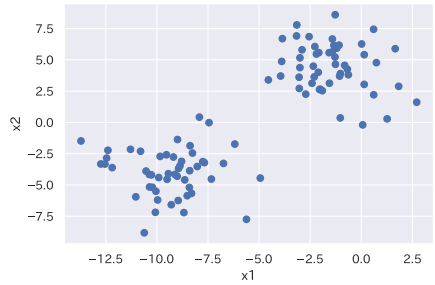
$$d(\boldsymbol{x}, \boldsymbol{x}') = \left(\sum_{k=1}^{p} |x_k - x'_k|^r \right)^{1/r}$$

どの単位を使うかで, 距離の大小の順は変わる. すべての k を平等に扱うには, 無難には, あらかじめデータを標準化しておく.

マハラノビスの距離

$$\tilde{x}_{ik} = \frac{x_{ik} - \overline{x}_{\cdot k}}{\sigma_{\cdot k}}$$

$$\tilde{x}_{ik} = \frac{x_{ik} - x_{\min k}}{x_{\max k} - x_{\min k}} \in [0, 1]$$



あとで, データ点とクラスター (データ点の集合) との距離も考える

ここまで来たよ

6 混合ガウス分布

8 クラスタ – 分析・クラスタ分析
- 教師なし学習としてのクラスター分析
- 階層的クラスタリング
- 平方和の分解
- 非階層的クラスタリング, k-means

階層的クラスタリング

永田棟方 多変量解析法入門 §12

準備

データ点とデータ点の距離を,

- データ点とクラスタ (データ点の集合) の距離
- クラスターとクラスタの距離 (クラスター間距離)

に拡張
様々な定義があるので, 適切なもの, 問題に指定されたものを使う

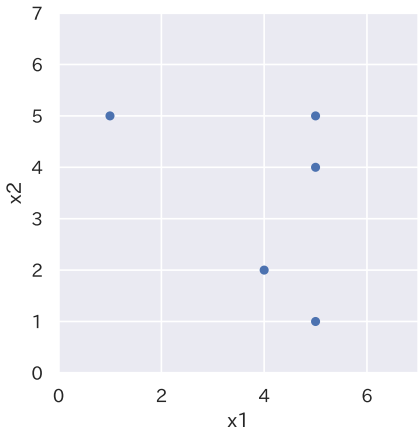
- クラスター内の平均 (重心) 間の距離. クラスターの代表を平均としてその間の距離
- データ点ペアの最短距離
- データ点ペアの最長距離

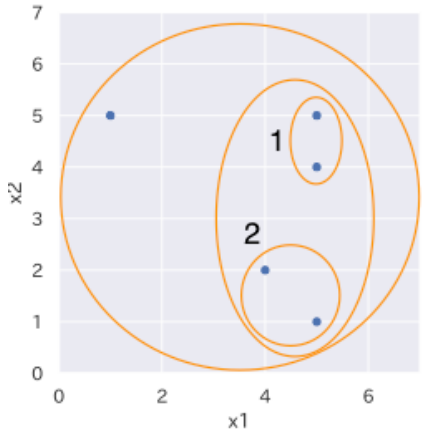
階層的クラスタリングのアルゴリズム

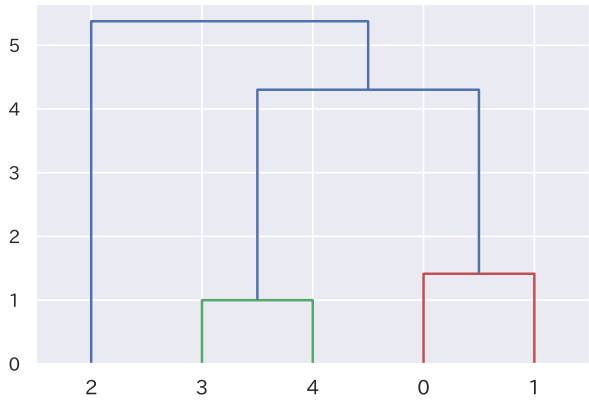
当初の $n_{\mathrm{C}} = n$ クラスターを, まとめて減らしていく.

1. すべてのデータ点を, 1 点のみからなるクラスターとみなす. $n_{\mathrm{C}} = n$.
2. 最短距離が指定の距離により大きい間, 次を行う
   1. $n_{\mathrm{C}} \times n_{\mathrm{C}}/2$ 個のクラスター間距離を計算する.
   2. 最短距離にある 2 クラスターを合併して 1 個のクラスターとみなす. n_{C} を 1 減らす.

合併の過程を描いたもの: デンドログラム (dendrogram, 樹形図)







- クラスタの個数は自動的には決まらない.
- 指定の距離=0 として 1 クラスタになるまで行い, デンドログラムを眺めて, 望みのクラスタ個数になるように指定の距離を後出しで決めることはできる.
- クラスタの分かれ方は, 「望みのもの」になるとは限らない. 各クラスタ内のデータ点の個数が同程度になる保証はない. 鎖現象.

階層的クラスタリングが得意/苦手なデータ

階層的クラスタリングの結果が, 人間の直観と同じとは限らない.
`https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html`

Example

階層的クラスタリングの結果が, 人間の直観と異なるような訓練データ x_{train} の例を考えよう.

L08-Q1

Quiz(階層的クラスター分析)

次の 2 変量データを考える.

	x_1	x_2
A	10	2
B	10	4
C	2	10
D	10	8
E	10	10

距離をユークリッド距離, クラスターの代表点は平均 (重心) とする.
階層的クラスター分析を行い, デンドログラムを描こう.

ここまで来たよ

6 混合ガウス分布

8 クラスタ – 分析・クラスタ分析
- 教師なし学習としてのクラスター分析
- 階層的クラスタリング
- 平方和の分解
- 非階層的クラスタリング, k-means

平方和の分解とクラスタの良さの評価と Ward の距離

本質は $p = 1$ 次元で見えるのでその表現で.
x_{ik}: k 番目のクラスターに属する i 番目のデータ点.
$i = 1, \ldots, n_k . k = 1, \ldots, C . \sum_{k=1}^{C} n_k = n.$
偏差平方和

$$
\begin{aligned}
S =& \sum_{i,k} [x_{ik} - \overline{x}_{..}]^2 \\
=& \sum_{i,k} [(x_{ik} - \overline{x}_{.k}) + (\overline{x}_{.k} - \overline{x}_{..})]^2 \\
\stackrel{!}{=}& \sum_{k} \sum_{i} (x_{ik} - \overline{x}_{.k})^2 + \sum_{k} n_k (\overline{x}_{.k} - \overline{x}_{..})^2 \\
=& S_{\mathrm{W}} + S_{\mathrm{B}}
\end{aligned}
$$

群=クラスター
$x_{..} = \frac{1}{n} \sum_{i,k} x_{ik}$ 全体平均
$x_{.k} = \frac{1}{n_k} \sum_{i} x_{ik}$ 群内平均
S_{W}: 群内平方和 Within
S_{B}: 群間平方和 Between

‘よい’ クラスタリングとは, S_{B} に対して S_{W} が小さいことでは?
実は, 線形判別分析の Fisher の線形判別関数は比 $S_{\mathrm{B}}/S_{\mathrm{W}}$ を最大化するものになっていた.
Ward 法=トリッキーな距離の定義での階層的…クラスター間距離=(そのクラスターを合併したときの群内平方和 S_{W} の増分)
CH 基準 (カリンスキ-ハラバシュ基準) クラスタの個数 C が異なる場合にも対応させたもの (回帰の自由度調整済決定係数みたいなアイデア).

$$
\mathrm{CH}_C = \frac{(n - C) S_{\mathrm{B}}}{(C - 1) S_{\mathrm{W}}}
$$

ここまで来たよ

6 混合ガウス分布

8 クラスタ – 分析・クラスタ分析
- 教師なし学習としてのクラスター分析
- 階層的クラスタリング
- 平方和の分解
- 非階層的クラスタリング, k-means

非階層的クラスタリング, k-means

ユーザが最終的クラスタの個数 $k = n_{\mathrm{C}}$ を指定する.

1. k 個の初期「クラスタの代表点 $\boldsymbol{x}_{\mathrm{C}k}$ を定める (バリエーションあり).
2. 収束するまで (すべてのデータ点の属するクラスタが変わらなくなるまで), 以下を繰り返す.
   1. n 個のデータ点に対して, 代表点 $\boldsymbol{x}_{\mathrm{C}k}$ が最も近いクラスタに属させる.
   2. $\boldsymbol{x}_{\mathrm{C}k}$ を, 属するデータ点の平均 (mean) に更新する.

結果は初期代表値に依存する.
どの $k = n_{\mathrm{C}}$ がいいかを知るには, さまざまな $k = n_{\mathrm{C}}$ で試行して, 群内平方和などの指標で比較する.

L08-Q2

Quiz(非階層的クラスター分析 (k-means))

次の 2 変量データを考える.

	x_1	x_2
A	1	1
B	1	3
C	5	1
D	5	3

データ点間の距離はユークリッド距離, クラスターの代表位置は平均 (重心) とする.
k-means 法で $k = 2$ とする. 初期のクラスターの代表点を, $(1, 0), (3, 3)$ とする. 最初に得られるクラスターとその代表点, 更新後に得られるクラスターとその代表点, を求めよう.