L08 クラスター分析

樋口さぶろお

龍谷大学 先端理工学部 数理・情報科学課程

理論物理学特論 L08(2021-11-16 Tue)

最終更新: Time-stamp: "2021-11-17 Wed 08:48 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

L07-Q1

Quiz 解答:混合ガウス分布のナイーブベイズ

$$P(Y=0|X=1) = \frac{\frac{1}{(2\pi 2^2)^{1/2}}\mathrm{e}^{-\frac{(1-2)^2}{2\cdot 2^2}} \cdot \frac{3}{10}}{\frac{1}{(2\pi 2^2)^{1/2}}\mathrm{e}^{-\frac{(1-2)^2}{2\cdot 2^2}} \cdot \frac{3}{10} + \frac{1}{(2\pi (1/2)^2)^{1/2}}\mathrm{e}^{-\frac{(1-6)^2}{2\cdot (1/2)^2}} \cdot \frac{7}{10}}$$

L07-Q2

Quiz 解答:混合ガウス分布の線形判別関数

$z = \frac{2-6}{2} \cdot \frac{x-4}{2}.$

L07-Q3

Quiz 解答:2 次元混合ガウス分布の線形判別関数

$$z = {}^{\mathrm{t}}\left(\Sigma^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\right)\boldsymbol{x} = {}^{\mathrm{t}}\left(\begin{pmatrix} 4^{-1} & 0 \\ 0 & 6^{-1} \end{pmatrix}\begin{pmatrix} -2 \\ -6 \end{pmatrix}\right)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = -\frac{1}{2}x_1 - x_2.$$

ここまで来たよ

7 線形判別分析

8 クラスター分析
- 教師なし学習としてのクラスタリング
- 階層的クラスタリング, Ward 法
- 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, または指定なし
- 訓練データ $(x_{\mathrm{train}i})$ $(i = 1, \ldots, n)$ を, 近いものごとにクラスに分類.
  - 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

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

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

ここまで来たよ

7 線形判別分析

8 クラスター分析
- 教師なし学習としてのクラスタリング
- 階層的クラスタリング, Ward 法
- 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, Ward 法

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
- クラスタとクラスタの距離

に拡張

例: 平均同士の距離, 最短距離, 最長距離など. (あとから出てくる Ward 距離も参照)

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, dendro=木の)

- 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.

ここまで来たよ

7 線形判別分析

8 クラスター分析
- 教師なし学習としてのクラスタリング
- 階層的クラスタリング, Ward 法
- 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

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

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.

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).

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

Ward 法-トリッキーな距離の定義

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)

L08-Q1

Quiz(クラスター分析)

$n = 4$ の 2 変量データ A(1,1),B(3,1),C(4,5),D(6,7) を考える. クラスターの代表位置は平均 (重心) とする.

1. クラスタ間距離を, 最近接データ点間のマンハッタン距離, としたとき階層的クラスター分析を行い, デンドログラムを描こう.
2. クラスタ間距離を, 平方和の増分 (ウォードの方法で用いられる), としたとき, クラスター (A,B) と (C,D) の間の距離を計算しよう.

ここまで来たよ

7 線形判別分析

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

k-means, 非階層的クラスタリング

ユーザが最終的クラスタの個数 $k = C$ を指定する.

1. k 個の初期「クラスタの代表値 $\boldsymbol{x}_{\mathrm{C}k}$ 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, 代表値 $\boldsymbol{x}_{\mathrm{C}k}$ 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値 (mean) に更新する.

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

th-d09-cluster.ipynb

L08-Q2

Quiz(クラスター分析)

2 変量データ A(1,1),B(3,1),C(4,5),D(6,7) を考える.
データ点間の距離はマンハッタン距離, クラスターの代表位置は平均 (重心) とする.

1. k-means 法で $k = 2$ とする. 初期のクラスターの代表位置を, $(4, 4), (5, 7)$ とするとき, 最初の 2 回の繰り返しでのクラスターを求めよう.