第10章
クラスタリング

11T4066Y
松村拓真

本章の目的

クラスタリング…教師データなしで、入力データ間の類似度、比類似度を手掛かりに、データをいくつかのクラスタにグループ分けすること

比階層的、階層的な手法が存在

⇔混合分布で表現=混合正規分岐モデル

10.1 距離の公理

10.1.1 距離の公理

距離…データやクラスタ間の類似度を測る尺度

距離の公理

(1)非負性：$d(x, y) \geq 0$

(2)反射律：$d(x, y) = 0$となるのは $x = y$の場合のみ

(3)対称性: $d(x, y) = d(y, x)$

(4)三角不等式: $d(x, z) \leq d(x, y) + d(y, z)$

10.1　距離の公理

10.1.2　ミンコフスキー距離

d次元のデータ…N個、i番目のベクトル…$x_i = (x_{i1}, \ldots, x_{id})_T$

$$d(x_i, x_j) = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^a\right)^{1/b}$$

(1)$a = 1, b = 1$…市街地距離
(2)$a = 2, b = 2$…ユークリッド距離
(3)$a = 2, b = 1$…ユークリッド距離の2乗
(4) $a = b = \infty$…チェビシェフ距離

その他の代表的な尺度…キャンベラ尺度、方向余弦

10.2　非階層型クラスタリング(K-平均法)

- 目的…d次元のN個のデータをデータ間の距離を尺度に、
  あらかじめ定めたK個のクラスタに分割

k番目の代表ベクトルを支配するクラスタ…$\mathrm{M}(\mu_k)$

$\mathrm{M}(\mu_k)$に帰属するか否かを表す帰属変数…$q_{ik}(0 or 1)$

→K平均法の評価関数

$$J(q_{ik},\mu_k)=\sum_{i=1}^{N}\sum_{k=1}^{K}q_{ik}||x_i-\mu_k||^2$$

μ_kに関する最適化

$$\frac{\partial J(q_{ik},\mu_k)}{\partial \mu_k}=0\Rightarrow\mu_k=\frac{\sum_{i=1}^{N}q_{ik}x_i}{\sum_{i=1}^{N}q_{ik}}$$

10.2 非階層型クラスタリング(K-平均法)

- K-平均法のアルゴリズム

  初期化:N個のデータをランダムにK個のクラスタに振り分け、
  それぞれのクラスタの平均ベクトルを求め、μ_kとする。

  (1) q_{ik}に関する最適化

  (2) μ_kの最適化

  (3)(1)と(2)を収束するまで繰り返す。

K-メノイド法…代表ベクトルをデータベクトルに限った方法

(メノイド…クラスタ内のデータ点の内、その点以外のクラスタ内の他との非類似度の総和が最小になる点)

10.3 階層型クラスタリング(融合法)

- クラスタリングされていないN個のデータから、類似度の高い順に融合して大きなクラスタを作成し、N個のデータを一つのクラスタに統合する手法(融合の過程は樹形図で表現可能)

融合法のアルゴリズム

(1)$n = N$

(2)$n \times n$の距離行列を作成

(3)最も近距離の二つのデータを一つのクラスタにする

(4)$n = n - 1$

(5)$n > 1$であれば(2)へ、$n = 1$であれば終了

クラスタ間の類似度の定義は様々…

10.3 階層型クラスタリング(融合法)

10.3.1 単連結法

二つのクラスタA, B間で最も類似度の高いデータ間の距離をクラスタ間の距離にする

- クラスタに一つデータが追加されると、他のクラスタとの距離は小さくなるか、または変化しない
- 大きなクラスタができる傾向
- 同じ距離に二つのクラスタがある場合、どちらでも結果は同じ
- 近いデータが別なクラスに属する連鎖効果が起こる可能性

10.3 階層型クラスタリング(融合法)

10.3.2 超距離

二つのデータx_iとx_jが融合する直前の「クラスタ間の距離」

$$\tilde{d}(x_i, x_j)$$

超距離の性質

(1) $\tilde{d}(x_i, x_j) \leq d(x_i, x_j)$

(2) $\tilde{d}(x_i, x_j) \leq \tilde{d}(x_i, x_k) + \tilde{d}(x_k, x_j)$

(3) $\tilde{d}(x_i, x_j) \leq \max[\tilde{d}(x_i, x_k), \tilde{d}(x_k, x_j)]$

⇒(3)…超距離不等式(距離の三角不等式より強い条件)

10.3 階層型クラスタリング(融合法)

10.3.3 完全連結法

最遠隣距離(最も類似度の低いデータ間の距離)をクラスタ間の距離にする手法

- クラスタに一つデータが追加されると、他のクラスタとの距離は大きくなるか、または変化しない
- 同じようなサイズのクラスタができる傾向
- 連鎖効果は起こらない

10.3 階層型クラスタリング(融合法)

10.3.4 群平均法

二つのクラスのクラスタ内の全てのデータ対間の距離の平均でクラスタ間の距離を決定する方法

クラスタ間の距離

$$D(A,B) = \frac{1}{N_A N_B} \sum_{x \in A, x \in B} d(x,y)$$

クラスタA, Bを融合したクラスタCと他のクラスタXの距離

$$D(C,X) = \frac{N_A D(A,X)}{N_A + N_B} + \frac{N_B D(B,X)}{N_A + N_B}$$

10.3 階層型クラスタリング(融合法)

10.3.5 ウォード法

クラスタAとBの距離を融合したときのクラスタ内変動の増加分で定義し、距離の小さなクラスタから融合していく方法

⇒階層法の中で最も精度が高い

これらの手法の他に重心法、メディアン法が存在

10.4 確率モデルによるクラスタリング

- K-平均法、融合法→ハードクラスタリング

⇔混合分布モデル→確率的に決めるクラスタリング

10.4.1 混合正規分布モデル

$$\mathrm{p}(\mathrm{x}) = \sum_{k=1}^{K} \pi_k \, \mathrm{N}(\mathrm{x}|\mu_k, \sum\nolimits_K), 0 \le \pi_k \le 1, \sum_{k=1}^{K} \pi_k = 1$$

π_k…混合比

パラメータ…$\pi, \mu, \sum$

→全体で$K + dK + (d+1)dK/2$個

のパラメータを求める必要有

10.4 確率モデルによるクラスタリング

- **10.4.2 隠れ変数と事後確率**

  隠れ変数$(z = (z_1, \dots, z_K)^T)$

$$\sum_{k=1}^{K} Z_k = 1 \, , z = \underbrace{(0, \dots, 0,1,0, \dots 0)^T}_{\text{一つだけ}1}$$

  隠れ変数の事後確率

$$\gamma(z_k) \overset{def}{\cong} p(z_k = 1|x) = \frac{\pi_k \mathrm{N}(\mathrm{x}|\mu_k, \Sigma_K)}{\sum_{j=1}^{K} \pi_j \mathrm{N}(\mathrm{x}|\mu_j, \Sigma_j)}$$

10.4 確率モデルによるクラスタリング

10.4.3 完全データの対数尤度

観測データ…$X = (x_1, \ldots, x_N)$ $x_i = (x_{i1}, \ldots, x_{id})^T$

隠れ変数… $Z = (z_1, \ldots, z_N)$ $z_i = (z_{i1}, \ldots, z_{id})^T$

XとZを合わせた集合→完全データ

$$Y = (x_1, \ldots, x_N, z_1, \ldots, z_N) = (X, Z)$$

完全データの尤度を最大にするパラメータ

⇔混合正規分布のパラメータ

10.4 確率モデルによるクラスタリング

10.4.4 Q関数

隠れ変数に関する期待値

$$\mathrm{L} = \sum_{i=1}^{N}\sum_{k=1}^{K} E_{Z_{ik}}\{z_{ik}\} ln\pi_k$$

$$+\sum_{i=1}^{N}\sum_{k=1}^{K} E_{Z_{ik}}\{z_{ik}\}\left(-\frac{d}{2}\ln(2\pi)+\frac{1}{2}\ln\left|\sum\nolimits_{k}\right|^{-1}-\frac{1}{2}(x_i-\mu_k)^T\sum\nolimits_{K}^{-1}(x_i-\mu_k)\right)$$

⇒隠れ変数に関する期待値$E_{Z_{ik}}\{z_{ik}\}$ =隠れ変数の事後確率$\gamma(z_{ik})$

隠れ変数の期待値を事後確立で置き換えた関数⇒Q関数

10.4 確率モデルによるクラスタリング

10.4.5 EMアルゴリズム

Eステップ…隠れ変数の事後確率の導出

Mステップ…求めた確立をQ関数に代入し、
Q関数を最大にするパラメータを導出

EMアルゴリズム

(1)$\pi_k, \mu_k, \sum_K$を初期化

(2)Eステップ:現在のパラメータを用いた$\gamma(z_{ik})$の推定

(3)Mステップ:$\gamma(z_{ik})$を用いたパラメータの再推定
Q関数の各パラメータによる最大化

(4)完全データの対数尤度に変化があり、収束×→(2)へ
変化がなくなり、収束○→終了