第10章 クラスタリング

12T4069L
佐鳥 恭太郎

類似度

- 類似度
  - データ（クラスタ）同士がどれだけ似ているかの尺度
  - 基本、距離で測る
    - xとyの距離$= d(x, y)$
  - 大きいほど似ている

  ⇔非類似度：大きいほど似ていない

基本的な距離

- データ:d次元のデータがN個
  - N個のデータ : $X = \{x_1, \dots, x_i, \dots, x_N\}$
  - i番目のデータ : $x_i = (x_{i1}, \dots, x_{id})^T$
- ユークリッド距離
  - $d(x_i, x_j) = \sqrt{\sum_{k=1}^{d} |x_{ik} - x_{jk}|^2}$
- ミンコフスキー距離
  - $d(x_i, x_j) = \left(\sum_{k=1}^{d} |x_{ik} - x_{jk}|^a\right)^{1/b}$

ミンコフスキー距離の派生

- a=1,b=1
  - 市街地（マンハッタン）距離
  - 碁盤の目状の経路を通るときの距離
- a=2,b=2
  - ユークリッド距離
  - 直線上の距離
- a=b=∞
  - チェビシェフ距離
  - $d(x_i, x_j) = \lim_{a\to\infty}\left(\sum_{k=1}^{d}\left|x_{ik} - x_{jk}\right|^a\right)^{1/a} = \max_k\left|x_{ik} - x_{jk}\right|$
  - 碁盤のマスいくつ離れているか（斜めと縦横同じ距離）
- aの増加=>特徴間の差に大きな重み
- bの増加=>差分累乗和に対する重みが小さくなる

尺度

- キャンベラ尺度
  - データを正規化する
  - $d(x_i, x_j) = \sum_{k=1}^{d} \frac{|x_{ik} - x_{jk}|}{|x_{ik}| + |x_{jk}|}$
- 方向余弦
  - ベクトル間の$\cos\theta$を用いる
  - $d(x_i, x_j) = \frac{\sum_{k=1}^{d} x_{ik} x_{jk}}{\sqrt{\left(\sum_{k=1}^{d} x_{ik}^2\right)\left(\sum_{k=1}^{d} x_{jk}^2\right)}}$

K-平均法

- K-平均法
  - 非階層型クラスタリング
  - d次元のN個のデータ$D = \{x_1, \dots, x_N\}, x_i \in R^d$を、あらかじめ定めたK個のクラスタに分類すること
  - 各クラスタの代表ベクトル(重心など)の集合M
    - $M = \{\mu_1, \dots, \mu_K\}$
    - K番目の代表ベクトルが支配するクラスタを$M(\mu_k)$
  - 帰属変数q_{ik}
    - $q_{ik} = \begin{cases} 1\ (x_i \in M(\mu_k)\text{の場合}) \\ 0\ (\text{それ以外の場合}) \end{cases}$

K-平均法の評価関数

- 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の最適化
  - $\frac{\partial J(q_{ik}, \mu_k)}{\partial \mu_k} = 2 \sum_{i=1}^{N} q_{ik}(x_i - \mu_k) = 0$

$$\Rightarrow \mu_k = \frac{\sum_{i=1}^{N} q_{ik} x_i}{\sum_{i=1}^{N} q_{ik}}$$

K-平均法アルゴリズム

- 初期化
  - N個のデータをランダムにK個のクラスタに振る。それぞれのクラスタの平均ベクトルを求め,それぞれ $\mu_k(k = 1, \dots, K)$とする。

①q_{ik}の最適化:

- $\mathrm{q_{ik}} = \begin{cases} 1\ (k = \arg\min_j \left\| x_i - \mu_j \right\|^2 \text{ の場合}) \\ 0\ (\text{それ以外の場合}) \end{cases}$

②μ_kの最適化:

- $\mu_\mathrm{k} = \frac{\sum_{i=1}^{N} q_{ik} x_i}{\sum_{i=1}^{N} q_{ik}}$

③収束するまで①,②を繰り返す。

K-メドイド法

- K-メドイド法
  - K-平均法の代表ベクトルをデータベクトルに限ったもの
  - 代表ベクトルの決定
    - $\mu_k = x_i$
    - $i = \arg\min_{x_j \in M(\mu_k)} \sum_{y \in \{M(\mu_k) - x_j\}} d(x_j, y)$
    - $d(x_j, y)$は非類似度の尺度であれば、距離でなくてもいい
  - K-平均法とは違い、距離の１乗で誤差が評価されるので外れ値の影響が少ない

融合法

- 融合法
  - 階層型クラスタリング
  - 未クラスタリングデータN個を類似度の高い順に2個ずつ融合させて、最後には一つのクラスタに統合する方法
  - 融合過程は樹状図という木の形で表せる
  - クラスタ間の類似度はデータ間の類似度を測る尺度と同じものを用いる

単連結法

- 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とする
  - $\mathrm{D}(A, B) = \min_{x \in A, y \in B} d(x, y)$
  - 性質
    - 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、又は変化しない
    - クラスタAとBが融合されてCとなるとき、他のクラスタXtの距離は
      $$\mathrm{D}(\mathrm{C}, \mathrm{X}) = \min[D(A, X), D(B, X)]$$
    - 大きなクラスタができる傾向がある
    - 同じ距離の二つのクラスがある場合、どちらを選んでも結果は同じ
    - 近いデータが別なクラスタに属する連鎖効果が現れる場合がある

超距離

- 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$\bar{d}(x_i, x_j)$のこと
  - 性質
    - $\bar{d}(x_i, x_j) \leq d(x_i, x_j)$
    - $\bar{d}(x_i, x_j) \leq \bar{d}(x_i, x_k) + \bar{d}(x_k, x_j)$
    - $\bar{d}(x_i, x_j) \leq \max[\bar{d}(x_i, x_k), \bar{d}(x_k, x_j)]$

完全連結法

- 完全連結法
  - クラスタ間で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
  - $D(A,B) = \max_{x \in A, y \in B} d(x,y)$
  - 性質
    - 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、又は変化しない
    - クラスタAとBが融合されてCとなったとき、他のクラスタXとの距離は

      $$\mathrm{D}(C,X) = \max[D(A,X), D(B,X)]$$

    - 大きなクラスタになりにくく、同じようなサイズのクラスができる傾向がある
    - 連鎖効果は現れない

群平均法

- 群平均法
  - 二つ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める
  - $\mathrm{D}(A,B) = \frac{1}{N_A N_B} \sum_{x \in A, y \in B} d(x,y)$
    - N_A, N_B:クラスタA,Bのデータ数
  - クラスタAとBが融合されてCとなったとき、他のクラスタXとの距離は

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

ウォード法

- ウォード法
  - クラスタAとBの距離を、それらを融合した時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
  - $D(A,B) = \sum_{x\in A,B} d(x,\mu_{AB})^2 - (\sum_{x\in A} d(x,\mu_A)^2 + \sum_{x\in B} d(x,\mu_B)^2) = S_{AB} - (S_A + S_B)$
    - $d(x,y)$:ユークリッド距離
    - μ_X:クラスタの平均ベクトル
    - S_X:平均からの距離の２乗和(変動)
  - 階層法の中で最も精度が高い

確率モデルによるクラスタリング

- ハードクラスタリング
  - 一つのデータは一つのクラスタにのみ分類される
  - K-平均法,融合法など

⇔ どのクラスタに属するかを確率的に決める

混合正規分布モデル

- k番目のクラスタを表すd次元正規分布関数
  - $N(x|\mu_k, \Sigma_{\mathrm{k}}) = \frac{1}{(2\pi)^{1/d}|\Sigma_{\mathrm{k}}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \Sigma_{\mathrm{k}}^{-1}(x-\mu_k)\right)$
- 全体の分布
  - $p(x) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_{\mathrm{k}})$
  - $0 \leq \pi_k \leq 1, \sum_{k=1}^{K} \pi_{\mathrm{k}} = 1$
- パラメータ
  - π_{k}:混合比
  - μ_{k}: d次元の平均ベクトル
  - Σ_{k}: $d \times d$の共分散行列

  } 推定しなければならない

  - $\pi = (\pi_1, \ldots, \pi_K), \mu = (\mu_1, \ldots, \mu_K), \Sigma = (\Sigma_1, \ldots, \Sigma_{\mathrm{K}})$

隠れ変数

- 隠れ変数
  - あるデータが、K個中のどのクラスタに属するかを表現する
  - モデルのパラメータを推定するために導入する
  - $z = (z_1, \ldots, z_K)^T$
  - 属していれば$z_k = 1$, 属していなければ$z_k = 0$
    - $z = (0, \ldots, 0, 1, 0, \ldots 0)^T$： どこか一つが1となる

隠れ変数の事後確率

- 変数xと隠れ変数zの同時確率
  - $p(x,z) = p(z)p(x|z)$
- 隠れ変数の分布
  - $p(z_k = 1) = \pi_k$より
  - $p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$
- 観測データの隠れ変数による条件付き分布
  - $p(x|z_k = 1) = N(x|\mu_k, \Sigma_\mathrm{k})$より
  - $p(x|z) = \prod_{k=1}^{K} [N(x|\mu_k, \Sigma_\mathrm{k})]^{z_k}$
- 全体の確率
  - $p(x) = \sum_{k=1}^{K} p(z)p(x|z) = \sum_{k=1}^{K} \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} [N(x|\mu_k, \Sigma_\mathrm{k})]^{z_k} = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_\mathrm{k}$
  - $z_k = 1, z_{j \neq k} = 0$
- 隠れ変数の事後確率
  - $\gamma(z_k) \stackrel{\text{def}}{=} p(z_k = 1|x) = \frac{p(z_k=1)p(x|z_k=1)}{p(x)} = \frac{\pi_k N(x|\mu_k,\Sigma_\mathrm{k})}{\sum_{j=1}^{K} \pi_j N(x|\mu_j,\Sigma_\mathrm{j})}$

完全データ

- 観測データ
  - $X = (x_1, \ldots, x_N), x_i = (x_{i1}, \ldots, x_{id})^T$
- 隠れ変数
  - $Z = (z_1, \ldots, z_N), z_i = (z_{i1}, \ldots, z_{iK})^T$
- 完全データ
  - $Y = (x_1, \ldots, x_N, z_1, \ldots., z_N) = (X, Z)$
  - 観測データと隠れ変数を合わせた集合

完全データの尤度

- 混合正規分布のパラメータは、完全データの尤度を最大にするパラメータで求める
- 完全データの尤度
  - $p(Y|\pi,\mu,\Sigma) = p(Z|\pi,\mu,\Sigma)p(X|Z,\pi,\mu,\Sigma) = \prod_{i=1}^{N}\prod_{k=1}^{K}[\pi_k N(x_i|\mu_k,\Sigma_k)]^{z_{ik}}$
- 最尤推定値を求めるために対数尤度関数にする
  - $\bar{L} = \ln p(Y|\pi,\mu,\Sigma) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\ln\pi_k + \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\ln N(x_i|\mu_k,\Sigma_k) =$
    $\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\ln\pi_k +$
    $\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(-\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma_k|^{-1} - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)$
- ただし、隠れ変数があるため、最尤推定値を直接求められない
  - EMアルゴリズムで最尤推定値を求める

Q関数

- EMアルゴリズムを用いるときに使用する
- 隠れ変数に関する期待値
  - $L = E_z\{\bar{L}\} =$
    $\sum_{i=1}^{N}\sum_{k=1}^{K} E_{z_{ik}}\{z_{ik}\}\ln \pi_k +$
    $\sum_{i=1}^{N}\sum_{k=1}^{K} E_{z_{ik}}\{z_{ik}\}\left(-\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma_{\mathrm{k}}|^{-1} - \frac{1}{2}(x_i-\mu_k)^T\Sigma_{\mathrm{k}}^{-1}(x_i-\mu_k)\right)$
  - $E_{z_{ik}}\{z_{ik}\} =$
    $\sum_{z_{ik}=\{0,1\}} z_{ik}p(z_{ik}|x_i,\pi_k,\mu_k,\Sigma_{\mathrm{k}}) = 1\times p(z_{ik}=1|x_i) = \gamma(z_{ik})$
  - 隠れ変数の期待値を事後確率で置き換える
- Q関数
  - Q=
    $\sum_{i=1}^{N}\sum_{k=1}^{K} \gamma(z_{ik})\ln \pi_k +$
    $\sum_{i=1}^{N}\sum_{k=1}^{K} \gamma(z_{ik})\left(-\frac{d}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma_{\mathrm{k}}|^{-1} - \frac{1}{2}(x_i-\mu_k)^T\Sigma_{\mathrm{k}}^{-1}(x_i-\mu_k)\right)$

EMアルゴリズム

- EMアルゴリズム
  - 隠れ変数がある場合に、確率モデルのパラメータの最尤推定値を求める優れた手法
  - 収束値は初期値に依存する
    - 何度か試行して良いものを選ぶ

EMアルゴリズム

① π_k, μ_k, Σ_kを初期化する

② Eステップ:現在のパラメータでの$\gamma(z_{ik})$の推定

③ Mステップ:推定した$\gamma(z_{ik})$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

- $N_k = \sum_{i=1}^{N} \gamma(z_{ik})$: k番目のクラスタに属するデータ数の推定値
- $\mu_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma(z_{ik}) x_i$
- $\Sigma_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik})(x_i - \mu_k)(x_i - \mu_k)^T}{N_k}$
- $\pi_k = \frac{N_k}{N}$

④ 完全データの対数尤度に変化がある => ②

変化なし => 収束したので終了する