クラスタリング

橋本 玄基

2017/06/07

1　概要

クラスタリングは教師なし学習の一つであり，データ群を外的要因無く，アルゴリズムに従って分類する手法である．近年流行している深層学習は，教師付き学習であり．ラベルの貼られた大量のデータを学習して未来のデータを予測するために用いられる．対して教師なし学習は，正しいラベルが不明であるデータに用いられ，妥当であろうラベル分けを行いデータ解析の補助のために用いられる．例として，顧客のグループ化による広告メールの差別化等があげられる．橋本，清野らによるプログラミング教育のためのログデータ解析についてもログデータはラベルの無いデータであるため教師なし学習であるクラスタリングを用いて解析を行っている．

2　種類

2.1　階層クラスタリング

階層クラスタリングは性質が近いデータをまとめていき，小さな分類を重ね大きな分類を作成する手法である．最短距離法，ウォード法などがあげられる．どの時点の階層を見るかで任意のクラスタに分類でき，樹形図等からデータの関係を非常に細かく見ることができるが，計算量が膨大かつモデルが複雑であるため，大量データの解析には向いていない．

2.2　非階層クラスタリング

非階層クラスタリングは分布や境界値などを更新することで，分類をする手法である．k-means 法，混合正規分布モデル，凸クラスタリング，ノンパラメトリックベイズモデルなどがあげられる．

3　距離

データから作られた特徴ベクトル間がどれだけ離れているかを表す指標となる距離にはいくつか種類がある．本節ではそれを紹介する．

3.1 定義

本節でしようする記号の定義を以下に示す．

$$
\begin{aligned}
\boldsymbol{x}(\boldsymbol{y}) &:= \text{ベクトル } (\boldsymbol{x} = x_1, \cdots, x_n) \\
x_i(y_i) &:= \text{ベクトル } \boldsymbol{x} \text{ の } i \text{ 次元の要素} \\
n &:= \text{ベクトルの次元数} \\
N &:= \text{データの数} \\
X &:= \text{データの集合．} (|X| = N, X = \boldsymbol{x}_1, \cdots, \boldsymbol{x}_N) \\
d(\boldsymbol{x}, \boldsymbol{y}) &:= \text{ベクトル } \boldsymbol{x}, \boldsymbol{y} \text{ 間の距離}
\end{aligned}
$$

3.2 ユークリッド距離 (Euclidean distance)

もっとも一般的な距離であり，中心から円形に離れると距離が大きくなる．一般化した式は

$$
d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(x_1, y_1)^2 + \cdots + (x_n, y_n)^2} (n \text{ 次元}) \tag{1}
$$

となる．距離の大小さえわかれば良い問題においては，平方根を取り除いた，2 乗ユークリッド距離が用いられることもある．

3.3 標準化ユークリッド距離

魚の体長 [cm] と重量 [g] など，単位の違うデータがベクトルとなっているデータは多く存在する．それらのベクトルの距離をユークリッド距離で図ろうとすると体長が 10[cm] 違うことと体重が 10[g] 違うことが同じ意味合いとなってしまう．そのため，データを分散で正規化したもの同士の距離を測ることを考える．これを標準化ユークリッド距離という．各次元の分散が $\sigma_1, \cdots, \sigma_n$ で与えられた時，式は以下のとおりである．

$$
d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\left(\frac{x_1 - y_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{x_n - y_n}{\sigma_n}\right)^2} \tag{2}
$$

距離計算の前に前処理としてデータ全体を分散で割る正規化の作業を行うほうが一般的かもしれない．

3.4 マンハッタン距離

ユークリッド距離は 2 乗することにより，値を正にしているがマンハッタン距離は絶対値計算により正とする．そのため，円形ではなく中心からひし形状に距離が遠くなっていく．式は以下のとおりである．

$$
d(\boldsymbol{x}, \boldsymbol{y}) = |x_1 - y_1| + \cdots + |x_n - y_n| \tag{3}
$$

マンハッタンは碁盤上に街ができており，ある地点同士を移動するにはどの道を通っても同じブロック数を通過することが由来である．ユークリッド距離に対し，いくつもの軸で差があることを遠いと強く評価する．

3.5 チェビシェフ距離

チェビシェフ距離は斜めに離れた距離と x 軸 y 軸方向に離れた距離を区別しない．チェスのキングや将棋の王がある地点に移動するのにかかる手数が距離になるとイメージするとわかりやすいだろう．最も差がある

次元についてのみ注目した距離といえる．中心から正方形状に距離が遠くなっていく．

$$d(\boldsymbol{x}, \boldsymbol{y}) = \max_i(|x_i - y_i|) \tag{4}$$

3.6 ミンコフスキー距離

ユークリッド，マンハッタン，チェビシェフ距離を一般化したものをミンコフスキーといい下式で表す．$a = b = 1$ のときマンハッタン距離，$a = b = 2$ のよきユークリッド距離，$a = b = \infty$ のときチェビシェフ距離となる．a は次元の違いによる重み，b はデータ間の大きな差に与える重みを変える．

$$d(\boldsymbol{x}, \boldsymbol{y}) = \left(\sum_{i=1}^{n} (|x_i - y_i|)^a \right)^{1/b} \tag{5}$$

3.7 マハラノビス距離

データには相関があるものも多く存在する．例えば身長と体重などである．このようなデータにユークリッド距離を適用すると，平均的な人より身長と体重が大きい同じ体格で体が大きい人，身長は小さいのに体重が大きい肥満体型の人，身長は高いが体重が少ない痩せ型の人が同じ距離に位置してしまう．そこで，相関が強い方向の距離を短く考えるようにしたのがマハラノビス距離である．以下のように定義される．

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x} - \boldsymbol{y})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{y})} \tag{6}$$

3.8 Dynamic Time Warping(DTW)

時系列データに適用する距離も紹介しておこう．DTW は時系列データの位相ズレを解消する距離尺度である．アルゴリズムを以下に示す．DTW はベクトルの次元数が異なっても距離計算が可能であるためベクトル $\boldsymbol{y}$ の次元数を m と置く．

1. $m \times n$ 行列 $\boldsymbol{D}$ を，$\boldsymbol{D}_{1,1} = 0$，それ以外の要素を ∞ と定義する．
2. 全ての $i(= 2, \cdots, n)$，$j(= 2, \cdots, m)$ について以下の式で $\boldsymbol{D}_{i,j}$ を求める．

$$\boldsymbol{D}_{i,j} = \sqrt{(x_i, y_j)^2} + \min(\boldsymbol{D}_{i-1,j}, \boldsymbol{D}_{i,j-1}, \boldsymbol{D}_{i-1,j-1}) \tag{7}$$

3. $\boldsymbol{D}_{n,m}$ を距離とする．

4 階層クラスタリング

先述のとおり，クラスタをまとめて最終的に 1 つのクラスタとするのが，階層クラスタリングである．手法の違いとしてはクラスタ間の距離の計算方法にある．代表として，4 つの手法を紹介する．

4.1 定義

本節で仕様する記号を以下に定義する．

$$
\begin{aligned}
P, Q &:= \text{ある時点で存在するクラスタ} \\
d(P, Q) &:= \text{クラスタ間の距離} \\
\boldsymbol{p_i}, \boldsymbol{q_i} &:= \text{クラスタ } P(Q) \text{ に属するデータベクトル} \\
d(\boldsymbol{p}, \boldsymbol{q}) &:= \text{あるデータベクトル間の距離} \\
E(P) &:= \text{あるクラスタ } P \text{ の重心と全データの距離との 2 乗和} \\
&= \sum_{i=0}^{|P|} (\boldsymbol{p}_i - \boldsymbol{\mu}_P)^2 \\
\boldsymbol{\mu}_P &:= \text{あるクラスタ } P \text{ のセントロイド (重心ベクトル)}
\end{aligned}
$$

4.2 最短距離法

最短距離法はクラスタ同士で最も近いデータ間の距離を採用する．高速ではあるが，クラスタに 1 つずつデータがまとまっていくチェイン現象が起きやすく．分類性能は低い．

$$
d(P, Q) = \min(d(\boldsymbol{p}_i, \boldsymbol{q}_j)) \tag{8}
$$

4.3 最長距離法

最長距離法はクラスタ同士で最も遠いデータ間の距離を採用する．最短距離法よりも分類性能は高いが，外れ値の影響が大きくなってしまう．

$$
d(P, Q) = \max(d(\boldsymbol{p}_i, \boldsymbol{q}_j)) \tag{9}
$$

4.4 群平均法

群平均法はクラスタ同士の全データ間の距離の平均を採用する．そのため計算量は前述の二手法よりも大きくなる．しかし，外れ値の影響やチェイン現象のおきやすさは改善される．

$$
d(P, Q) = \frac{\sum_{i=0}^{|P|} \sum_{j=0}^{|Q|} d(\boldsymbol{p}_i, \boldsymbol{q}_j)}{|P| + |Q|} \tag{10}
$$

4.5 Ward 法

距離を測るクラスタの重心とのデータの距離の 2 乗和 $E(P), E(Q)$ と 2 つのクラスタをあわせた重心との全データの距離の 2 乗和 $E(P \cup Q)$ を用いて距離を下式のように計算する．

$$
d(P, Q) = E(P \cup Q) - (E(P) + E(Q)) \tag{11}
$$

群平均法よりさらに計算量が多いが，分類性能はとても高い．

5 k-means 法 (k 平均法)

5.1 概要

k-means 法は，予めクラスタ数を決定した上で，クラスタリングを行う．2 つの手順，クラスタ毎の中心の更新と最も近いクラスタ中心に所属クラスを更新を 2 つの手順を繰り返してクラスタを確定する．k-means 法はアルゴリズムが単純であり，高速である．

5.2 アルゴリズム

クラスタ数が k, それぞれのクラスタに属するデータ集合を $C_1, C_2, \cdots, C_k$, クラスタ中心を $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k$ とする．このとき観測データ集合 $X = \{\boldsymbol{x_i} \mid i = 1, 2, \cdots, n\}$ であるとするとアルゴリズムは以下のようになる．

1. X の中からデータをランダムに k 個選びそれぞれを $\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_k$ とする．
2. 全データ $\boldsymbol{x}_i (i = 1, 2, \cdots, n)$ を最も近いクラスタ中心に対応するクラスタ集合に所属させる．
3. 各クラスタ中心を式 12 の通りに更新する．

$$\boldsymbol{c}_i = \frac{1}{|C_i|} \sum_{\boldsymbol{x} \in C_i}^{j} \boldsymbol{x} \tag{12}$$

4. クラスタ集合に変化がなくなるまで手順 (2)，(3) を繰り返す．

5.3 k-means++

5.2 節より分かる通り，k-means 法は初期値にランダムな値が用いられるため，収束する結果も必ず良くなるとは限らない．また，同じ結果になった際も初期値によては繰り返し回数がとても多くなってしまう．この問題をいくらか改善するために k-means++ 法が提案された．アルゴリズムは以下のとおりである．

1. クラスタ中心 $\boldsymbol{c}_1$ を X の中からランダムに 1 つ選ぶ．
2. $D(\boldsymbol{x})(\boldsymbol{x} \in X)m$ を求める．$D(\boldsymbol{x})$ は $\boldsymbol{x}$ と決定された最も近いクラスタ中心の最短距離を表す．
3. 式 13 の確率分布に従って次のクラスタ中心 $\boldsymbol{c}_i$ を決定する．

$$\frac{D(\boldsymbol{x})^2}{\sum_{\boldsymbol{x} \in X} D(\boldsymbol{x})^2} \tag{13}$$

4. k-means 法の手順 (2) から (4) を行う．

5.4 k-means 法の弱点

k-means 法はクラスタ数が既知でなければ使用できない．また大きな弱点として，クラスタ毎のデータ数はほぼ等しくなるため，分布数が大きく異なるクラスを正しく分類することはできない．

6 混合正規分布モデル (GMM)

6.1 概要

混合正規分布モデルは, あるデータ群はいくつかの正規分布によって発生したと考え, それぞれの正規分布のパラメータ (平均と分散) を求める数理モデルである. 各正規分布をクラスタとみなし，クラスタリングの手段として用いられることになる. 混合正規分布モデルによるクラスタリングの特徴として,1 つのクラスタに属するのてはなく, どのクラスタであるかの確率が出力されるソフトなクラスタリングである点があげられる.

6.2 アルゴリズム

クラスタ数が k のとき，クラスタを ω_i，事前確率を π_i，正規分布の平均を $\boldsymbol{\mu}_i$，分散（2 次元以上の場合は分散共分散行列）を $\boldsymbol{\Sigma}_i$ とおく $(i = 1, 2, \cdots, k)$．また，$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)^t$，$\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \cdots, \boldsymbol{\theta}_k, \pi_1, \cdots, \pi_k)^t$ と定義する．そして，n 個のデータを $\boldsymbol{x}_j (k = 1, 2, \cdots, n)$ とおいた時，アルゴリズムは以下のようになる．なお，$P(x)$ は確率関数，$p(x)$ は確率密度関数である．

1. $\pi_i, \boldsymbol{\theta}_i$ の初期値を与える．
2. 各 $\boldsymbol{x}_j$ に対して $P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta})$ を求める．

$$P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta}) = \frac{\pi_i \cdot p(\boldsymbol{x}_j|\omega_i; \boldsymbol{\theta}_i)}{\sum_{l=1}^{k} \pi_l \cdot p(\boldsymbol{x}_j|\omega_l; \boldsymbol{\theta}_l)} \tag{14}$$

3. 次式によって $\hat{\pi}_i$,$\hat{\boldsymbol{\mu}}_i$,$\hat{\boldsymbol{\Sigma}}_i$ を求めた後，$\pi_i = \hat{\pi}_i$,$\boldsymbol{\mu}_i = \hat{\boldsymbol{\mu}}_i$,$\boldsymbol{\Sigma}_i = \hat{\boldsymbol{\Sigma}}_i$ としてパラメータを更新する．

$$\hat{\pi}_i = \frac{1}{n} \sum_{j=1}^{n} P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta}) \tag{15}$$

$$\hat{\boldsymbol{\mu}}_i = \frac{\sum_{j=1}^{n} P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta}) \boldsymbol{x}_j}{\sum_{j=1}^{n} P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta})} \tag{16}$$

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{j=1}^{n} P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta})(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)^t}{\sum_{j=1}^{n} P(\omega_i|\boldsymbol{x}_j; \boldsymbol{\theta})} \tag{17}$$

4. 対数尤度 $\log p(\boldsymbol{x}; \boldsymbol{\theta})$ を求め増分が定めた閾値以下なら終了し，そうでなければ (2) に戻る．

6.3 混合正規分布モデルの弱点

k-means 法と同様に，クラスタ数が既知である必要がある．しかし，情報量基準を用いてモデル選択を行うことが可能である．しかしそのためには各クラス多数で計算を繰り返すため，計算量がとても多くなる．