第８講　クラスタリングと次元削減

【学習到達目標】

・クラスタリングの基本概念と代表的な手法を理解し、適切な場面での適用方法を説明できる。

・次元削減の目的と代表的な手法（主成分分析（PCA）など）を理解し、データの可視化や前処理に役立てられる。

・クラスタリングと次元削減の違いや関係性を理解し、実データ分析においてこれらの手法を適切に選択・適用できる。

１．教師なし学習

第７講で取り扱った回帰分析や分類などの教師あり学習は、データに対して「正解」の情報が与えられていた。一方、本講で取り扱うクラスタリングや次元削減は教師なし学習の一種であり、「正解」の情報が無いデータから何かしらの規則や特徴を見出そうとするものである。これらの手法はデータの構造理解や可視化、前処理などに広く用いられている。

２．クラスタリング

第７講で扱った分類モデルは、ラベルが付与されたデータを学習し、各データに適切なラベルに分類していた。一方で、クラスタリングにおいては、ラベルが付与されていないデータが与えられ、それらを似た特徴を持つグループ（クラスター）に分類する。こうした手法は、ラベルがついていないデータにおいて、傾向を把握する上で重要となる。具体的なアルゴリズムとしては、以下に紹介する k-means 法や階層的クラスタリングが挙げられる。

(1)k-means 法

k-means 法は以下の手順によって、データをk個のクラスターに分類する。

1. k個の「クラスター中心」を定める

2. 各データ点を最も中心までの距離が近いクラスターに割り当てる

3. 各クラスターに関して、割り当てられたデータ点の平均値を新たなクラスター中心に定める

4. 2~3 の手続きをクラスター中心の位置が収束するまで繰り返す

この手順を図示すると、以下の図 1 のようになる。なお、クラスター数kをあらかじめ決めておく必要がある点には注意が必要である。

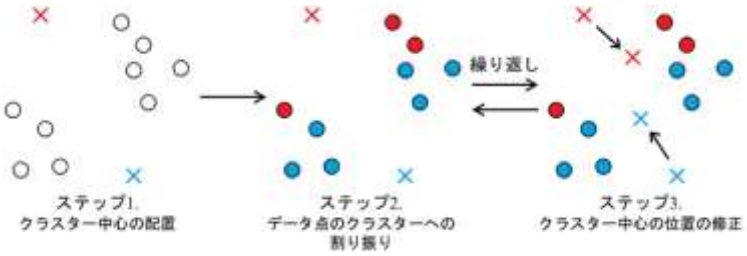

図 1　k-means 法

k-means 法においては、大抵の場合「近さ」の基準には通常のユークリッド距離を使う。

(2)階層的クラスタリング

階層的クラスタリングは以下の手続きによって、互いに「類似した」データから順にクラスターにまとめていく方法である。

1. 現時点で存在する全てのクラスターのうち、最も「類似度」が高い対を一つのクラスターに併合する

2. 新しいクラスターと他のクラスターの「類似度」を計算する

3. 1~2 の手続きを繰り返す

クラスターを統合していく過程で、図 2 のように木の枝分かれのような形のグラフが作られる。これをデンドログラム（樹形図）と呼ぶ。

上記の手続きはクラスターの数が 1 個になるまで続けることができるが、そこまでの過程で、データが任意の数のクラスターに分類された状態を経ているし、どの段階でどのクラスター同士が結合したのかもデンドログラムに記録される。よって、始めにクラスター数kを決めておく必要がある k-means 法とは異なり、**一連の操作が終了した後でも、何個のクラスターに分類するか指定できる**。

階層的クラスタリングにおいて、他のクラスターと結合していないデータ点は、「1 つの点によって構成されるクラスター」として扱う。

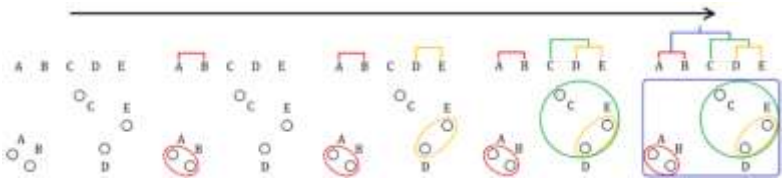

図 2　階層的クラスタリング

また、データ点間の「類似度」をどう定義するか、また、それが定義できたとして、データ点をまとめて作ったクラスター間の「類似度」の方をどのように定義するかに関しては、Ward 法や最近距離法、最遠距離法、群平均法など複数の方法がある。

なお、階層的クラスタリングにおいては、始めに全てのデータ点間の類似度を計算する必要がある上に、クラスターの統合を一回行うたびに、新しいクラスターと他のクラスターの類似度の計算もしなければいけない点には注意が必要である。この性質から、特にデータ点の数が多いデータセットの場合、k-mean 法よりも計算時間が長くなる傾向がある。よって、**データ点の数が多いものの、あらかじめクラスターの数がある程度分かっている場合は k-mean 法、データ点はそこまで多くないがクラスターの数が分からない、またはクラスターの結合の詳細な過程をデンドログラムとして可視化したい場合は階層的クラスタリング**、などのように適宜使い分ける必要がある。

(3)クラスタリングの実践例

図 3 に k-means 法および階層的クラスタリングの実行例を示す。使用したデータセットは互いに異なる 3 種類の 2 変量正規分布によって生成したデータ点の集合であり、(a)がクラスタリングをする前のデータ、(b)と(c)がそれぞれ k-means 法と階層的クラスタリングによって分類された状態を示す。

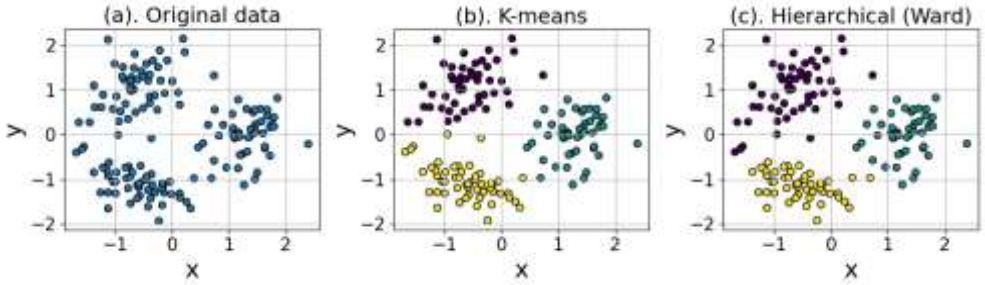

図 3　クラスタリングの実践例

3. 次元削減

次元削減は高次元のデータ（つまり、多数の変数を含むデータ）を、なるべく情報を失わずに低次元データに変換し、本質的な特徴を抽出する手法である。代表的な方法としては以下に挙げる主成分分析(PCA)が存在する。こうした手法は、高次元データを 2 次元や 3 次元に変換し、グラフとして**可視化する場合**や、他の機械学習手法によって分析する際の**前処理として変数の数を減らす場合**などに用いられる。

p個の変数をもつデータ点を、p次元空間の点とみなす考え方から、データが含む変数の数を次元と呼ぶ。例えば、各人の身長と体重が記載されたものは 2 次元データ、身長と体重に加えて座高も記載されているならば 3 次元データとなる。

(1)主成分分析(PCA)

主成分分析は以下の手順によってデータの特徴量を抽出する方法である。

1. データ点の集合の「ばらけ具合」(＝共分散)を計算する。

2. ばらけ具合が最も大きい方向$\boldsymbol{u}_1$、二番目に大きい方向$\boldsymbol{u}_2$、三番目に大きい方向$\boldsymbol{u}_3$、…を(互いに直交するように)求める。

3. 各データに関して、2.で求めた方向$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ に**沿って取った座標(これを主成分という)**を特徴量として抽出する。

つまり、元のデータセットにおいて、データ点が大きくばらけている方向のみに注目し、逆にばらけ具合が小さい方向に関しては無視することによってデータの次元を減らすのが主成分分析である。実際の分析時は、なるたけ情報を失いたくない場合ほど、抽出する主成分の数Mを大きめにする。

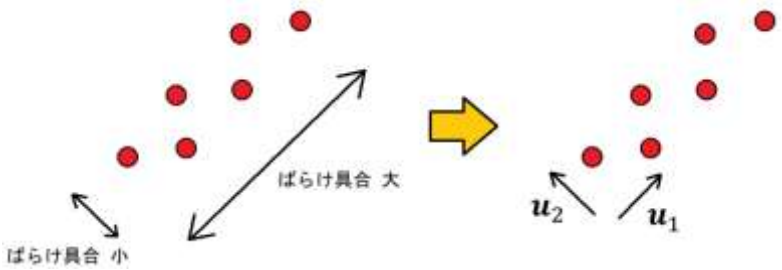

図 4　2 次元データにおける$\boldsymbol{u}_1, \boldsymbol{u}_2$のとり方

例えば、図 4 のような 2 次元データの場合、$\boldsymbol{u}_1$はデータ点が最もばらける方向に、$\boldsymbol{u}_2$はそれと直交する方向にとる。この際、$\boldsymbol{u}_1$に沿った主成分のみを抽出すれば、データが持つ情報を極力失わずに、データを 1 次元に削減できる

$\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$ を単位ベクトル（長さ 1 のベクトル）とすると、データ点$\boldsymbol{x}_n$の主成分は$\boldsymbol{x'}_n \cdot \boldsymbol{u}_i$ $(i = 1,2, \ldots)$と表せる。ただし$\boldsymbol{x'}_n$は$\boldsymbol{x}_n$を中心化した値である。

2 次元データにおいて 2 つの主成分を両方取り出すと、情報は失わないが次元を減らすこともできない。

(2)非線形な手法

主成分分析は直線的なデータ点のばらつきに基づいて次元を削減しているため、元のデータが曲線的に複雑に入り組んだパターンを持っている場合、そうした構造を捉えることができないという欠点を持っている。そうした問題に対処するため、t-SNE[4]や UMAP[5]といった非線形な手法も開発されている。

例として図 5(a)のような、データ点がロールケーキ状に分布している場合を考える。この場合、データ点が存在する 2 次元のシートが曲がりくねった状態で 3 次元空間に埋め込まれている。図 5(b)の主成分分析はそうした構造を捉え切れていない一方で、 (c)の t-SNE や(d)の UMAP はある程度捉えることができていることが分かる。

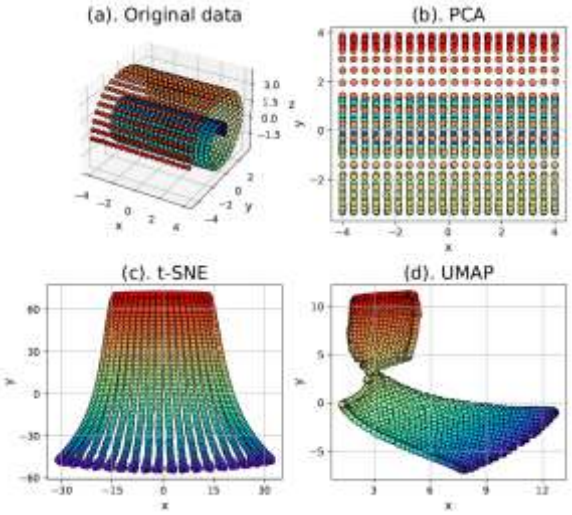

図 5　次元削減の実践例

4. 教師なし学習の注意点

冒頭で述べたように、クラスタリングも次元削減も教師なし学習の一種であり、データに「正解」の情報が無い（もしくは、仮にそうした情報があったとしても使用しない）状況で学習を進めるものである。そのため、どちらの手法においても、アルゴリズムの使用者が本来取り組みたかった課題において重要な情報を軽視し、逆にどうでもいい情報を重視してしまう危険性が存在する。

教師なし学習を使う限り、こうした問題はどれだけ洗練されたアルゴリズムを採用しても完全に無くすことはできない。よって、実際に分析を進める際はドメイン知識と照らし合わせつつ、分析結果が妥当かどうかを人間の目でチェックする必要がある。

例えばクラスタリングであれば本来の想定とは全く異なる基準で分類をしてしまう恐れがあるし、次元削減でも本来重要な情報を切り捨ててしまう危険性がある。

参考文献

[1] G. James, D. Witten, T. Hastie, R. Tibshirani 著、落海浩、首藤信通訳、「R による統計的学習入門」（朝倉書店、2018）

[2] C. M. Bishop 著、元田浩、栗田多喜夫、樋口知之、松本裕治、村田昇 監訳、「パターン認識と機械学習上・下」（丸善出版、2012）

[3] 日本統計学会編、「日本統計学会公式認定　統計検定準 1 級対応　統計学実践ワークブック」（学術図書、2020）

[4] L. van der Maaten, and G. Hinton, 2008. *Journal of Machine Learning Research*, **9**(Nov) (2008), 2579—2605.

[5] L. McInnes, J. Healy, and J. Melville, *arXiv preprint*, arXiv:1802.03426 (2018)

画像引用元

いらすとや　https://www.irasutoya.com/

課　題

1．クラスタリングの代表的な手法を 2 つ挙げ、それぞれの特徴と適用例について説明してください。

2．主成分分析（PCA）の基本的な仕組みと、その結果得られる主成分の意味について説明してください。さらに、PCA を用いる際の注意点も述べてください。

3．高次元データに対して次元削減を行う目的と、その際に考慮すべきポイントについて具体的に説明してください。