

# クラスタリング (Clustering)

## クラスタリング (Clustering)

クラスタリングはデータにある関係性に基づいてグループ化する教師なし学習の 1 つ。グループ化されたデータの集合をクラスタと呼ぶ。関係性としては、ユークリッド距離やマンハッタン距離などの距離とコサイン類似度と Jaccard 係数などの類似度がよく使われる。コサイン類似度はデータの構造 (方向) が似ているほど上昇し、Jaccard 係数は共有する要素を持っているほど上昇する。距離によるクラスタリングでは、なるべく距離が近いデータを同じクラスタに遠いデータを異なるクラスタに所属させる。類似度によるクラスタリングでは、なるべく類似度が高いデータを同じクラスタに類似度が低いを異なるクラスタに所属させる。

ユークリッド距離

$$\begin{aligned} L_2(\mathbf{x}, \mathbf{y}) &= \|\mathbf{x} - \mathbf{y}\|_2 \\ &= \left( \sum_{i=1}^m (x_i - y_i)^2 \right)^{\frac{1}{2}} \end{aligned}$$

マンハッタン距離

$$\begin{aligned} L_1(\mathbf{x}, \mathbf{y}) &= \|\mathbf{x} - \mathbf{y}\|_1 \\ &= \sum_{i=1}^m |x_i - y_i| \end{aligned}$$

コサイン類似度

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}$$

Jaccard 係数

$$\begin{aligned} \cos(\mathbf{x}, \mathbf{y}) &= \frac{|X \cap Y|}{|X \cup Y|} \\ &= \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \mathbf{x}^\top \mathbf{y}} \end{aligned}$$

## 代表的な手法

クラスタリングの代表的な手法として階層的クラスタリング, k-means 法, k-medoids 法, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), スペクトラルクラスタリングが挙げられる。

## 階層的クラスタリング

階層的クラスタリングはクラスタを順次併合または分割しながらデータを階層的に統合または結合する手法である。デンドログラムを用いてクラスタの階層構造を視覚化できるため、クラスタの階層構造を理解したい場合やクラスタ数が事前に不明確な場合に用いられる。個々のデータをクラスタとし、近いクラスタを順次統合するボトムアップ型の手法を凝集型クラスタリングと呼ぶ。また、全体を1つのクラスタとし、分割を繰り返すトップダウン型の手法を分割型クラスタリングと呼ぶ。

課題はどの時点の分割がふさわしいかを判定するのはユーザが決める必要があることと、大規模なデータに対しては計算負荷が高いこと。

近いクラスタを測る指標として、以下にクラスタ間の距離の代表的なものを示す。

単リンク法 (最短距離法)

$$d(C_i, C_j) = \min\{d(u, v) | u \in C_i, v \in C_j\}$$

完全リンク法 (最長距離法)

$$d(C_i, C_j) = \max\{d(u, v) | u \in C_i, v \in C_j\}$$

グループ平均法

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{u \in C_i} \sum_{v \in C_j} d(u, v)$$

ワード法

$$\begin{aligned} d(C_i, C_j) &= q(C_i \cup C_j) - q(C_i) - q(C_j) \\ q(C_i) &= \sum_{u \in C_i} \|\mathbf{x}_i - \mu_i\|_2^2 \\ \mu_i &= \frac{1}{|C_i|} \sum_{u \in C_i} \mathbf{x}_u \end{aligned}$$

## k-means 法

k-means 法は各クラスタの重心を代表点とし、データをその代表点に最も近いクラスタに割り当てる作業を収束するまで繰り返す手法である。代表点とクラスタ内のデータ点の距離の二乗の総和を最小にする問題。単純なクラスタリング問題やクラスタ数が明確な場合、計算量を抑えたい場合に用いられる。

課題は事前にクラスタ数を知っている必要があることと、クラスタの形状は円(球)状であることを仮定しているので複雑なデータや、特定の方向に分散したデータを上手く分類できないこと。また、結果は初期値に依存し最適なクラスタを出力する保証はない。

## k-medoids 法

k-medoids 法は k-means 法の代表点のとり方を重心の座標ではなく、実際のデータ点を代表点とする手法。具体的にはその点以外のクラスタ内の点でまでの非類似度の総和が最小になる点 (medoids) を代表点とする。medoid とデータ点の距離の総和 (二乗の総和ではない) を最小化する問題。k-means 法よりも外れ値に強い。また座標がわからなくても各点の距離が分かればクラスタリングできるので、集合での距離にもとづいてクラスタリングをしたいときに用いられる。

課題は k-means 法よりも遅いこと。

## Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN はデータの密度を基準としてクラスタリングを行う手法である。手順はまず適当な点を選択し、その点と一定距離にある点の数を数える。その点の数がしきい値以上ならクラスタにし、未満ならノイズ点とする。この操作を繰り返す。すべての点に対してクラスタの割り振りやノイズの判断ができれば、DBSCAN のアルゴリズムは終了する。パラメータとして距離のしきい値最小のデータ点の数のしきい値 (クラスタを形成するために必要な最小点数) を設定する必要がある。密度に基づいたクラスタリングやクラスタの形状が不規則な場合に用いられる。クラスタの数を設定する必要がないこと、外れ値に強いことが利点である。また点同士の距離を判断基準とするためデータの分布より密度が優先されるので、データの分布に左右されない。

課題は大規模なデータに対しては計算コストが高くなる傾向があること、密度がデータ間で異なる場合に距離のしきい値を設定するのが難しく精度が悪いこと、どのようなパラメータにするかによって結果が大きく変わるため、パラメータの設定に精度が左右されやすいこと。また高次元データでは、距離尺度がスパースになるため密度を基準としたクラスタリングが不正確になる場合があるので、主成分分析 (PCA) などで次元圧縮を行い、対策をする必要がある。

## スペクトラルクラスタリング

スペクトラルクラスタリングは重み付き無向グラフで表現される関係データをクラスタリングする手法である。クラスタ間にまたがる重みである Cut を正規化した RatioCut を最小化する問題となる。グラフの隣接行列と次数行列からなるラプラシアン<sup>1</sup>の性質を利用することで、RatioCut 最小化問題の緩和問題は固有値問題に定式化可能。よってラプラシアンの下位  $k$  個の固有値に対する固有ベクトルを求めているとも解釈できる。また、固有ベクトルによりグラフ頂点を  $k$  次元のユークリッド空間に埋め込み、ユークリッド空間におけるクラスタリングアルゴリズムを実行しているとも解釈できる。スペクトラルクラスタリングでは連結性に注目してクラスタリングを行うため、k-means 法などではうまくクラスタリングできなかった円 (球) 状でないデータをうまくクラスタリングできる。

課題はデータをグラフに変換する必要があること。変換方法としては、データ点同士がすべて結ばれる全結合法と、距離が  $\epsilon$  以下のデータ点同士だけを結ぶ  $\epsilon$  近傍法と、各データ点について、距離が近い上位  $k$  個のデータ点と結ぶ  $k$  近傍法などがある。全結合法の利点はデータの全体的な関係性を反映するため比較的安定した結果が得られること。しかし課題は大規模データセットでは計算コストが高くなること。 $\epsilon$  近傍法の利点はスパースな (疎な) グラフを生成できるため、大規模データでも扱いやすいこと。またデータの局所的な構造に焦点を当てることができる。しかし課題はパラメータ  $\epsilon$  の選択が難しいこと。値が小さすぎると孤立点が多くなり、逆に大きすぎると全結合法に近くなる。 $k$  近傍法の利点は、 $\epsilon$  近傍法と同じくスパースな (疎な) グラフを生成できるため、大規模データでも扱いやすいこと。さらに  $\epsilon$  近傍法よりもパラメータ選択が簡単である。課題も  $\epsilon$  近傍法と同じく、パラメータ  $\epsilon$  の選択が難しいこと。値が小さすぎると孤立点が多くなり、逆に大きすぎると全結合法に近くなる。

## クラスタリングの評価指標

クラスタリングの評価指標として、最適なクラスタ数を決定するための方法であるエルボー法、各データ点が適切なクラスタに属しているかを評価するシルエットスコア、クラスタ間の分離と内部のコンパクトさを評価するダブス・ボルディン指数が挙げられる。

### エルボー法

エルボー法はクラスタ内誤差平方和 (SEE) 値が小さいほど歪みのない (クラスタリングがうまくいっている) 良いモデルという考え方から最適なクラスタ数の検討を付ける手法。クラスタごとの SSE 値をプロットした図を参照し、SSE 値が”ヒジ”のようにガクンと曲がった点が最適なクラスタ数とみなす、つまり SSE 値の低下が最大のクラスタ数を探す。しかし、現実のデータを使ってエルボー図を書いた場合きれいなヒジが現れることは少ない。他の評価指標と

併用することが推奨される。

## シルエットスコア

シルエットスコアはクラスタ内の凝集度とクラスタ間の分離度を用いて計算されるクラスタリングの効果を測る指標の 1 つである。クラスタ間の距離が離れているほど、クラスタ内の密度が高いほど良いモデルという考え方からクラスタリングの効果を数値化する。凝集度は各データ点に対して、その点が属するクラスタ内の他の点との平均距離。これはデータ点とそのクラスタにどれだけフィットしているかを示す。凝集度が小さいほど、そのデータ点はクラスタ内でより密接している。分離度は各データ点に対して、最も近い別のクラスタとの平均距離。この距離は、異なるクラスタとの分離を示す。分離度が大きいほど、そのデータ点は他のクラスタから遠く離れている。シルエットスコアは  $-1$  から  $1$  の範囲で、 $1$  に近いほどクラスタリングの品質が高い。  $0$  に近いほどクラスタの境界が曖昧であることを示し、負の値はクラスタリングが不適切である可能性がある。

課題は高次元データでは計算コストが高くなる場合がある。またシルエットスコアは球状や均一な密度のクラスタを前提としているため、複雑な形状や異なる密度のクラスタには適用が難しい場合があるので、DBSCAN やスペクトラルクラスタリングなど、形状や密度に強い手法を併用する。

シルエットスコア  $a$  は凝集度、  $b$  は分離度を表す。

$$s = \frac{b - a}{\max(a, b)}$$

## ダブース・ボルディン指数

ダブース・ボルディン指数は最もコンパクトでよく分離されたクラスタを生成したクラスタリングを特定するための指標の 1 つである。各クラスタについて、クラスタサイズの合計をクラスタ間距離で割った比率が最大となる代替クラスタを見つけ、その値をデータ中の全クラスタについて平均化したものである。  $k = 2$  が理想的なクラスタサイズになりがち。

ダブース・ボルディン指数

$$D_i = \max_{i \neq j} \left( \frac{|C_i| + |C_j|}{d(C_i, C_j)} \right)$$
$$DB = \frac{1}{n} \sum_{i=1}^n D_i$$