クラスタリングとは

Copyright © 2024 by INIAD

クラスタリングとは？

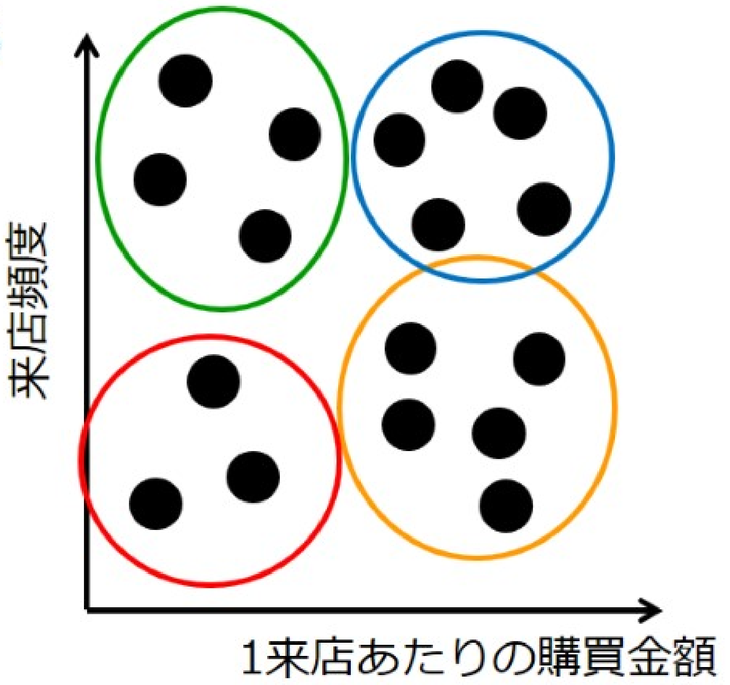
- データの散布図を描画すると、いくつかのかたまりに分かれることがあります
  - 例として、あるショップの顧客データが以下のようになっていたとします
  - 各グループ (クラスタ) の顧客は、購買行動が似ています。たとえば…

◆ **緑**：1来店あたりの購買金額は低いが、頻繁に来店している
→ 安定顧客

◆ **青**：頻繁に来店し、1来店あたりの購買金額も高い
→ 優良顧客

◆ **橙**：来店頻度は低いが、1来店あたりの購買金額が高い
→ 新規優良顧客

◆ **赤**：来店頻度が低く、1来店あたりの購買金額も低い
→ 非優良顧客



Copyright © 2024 by INIAD

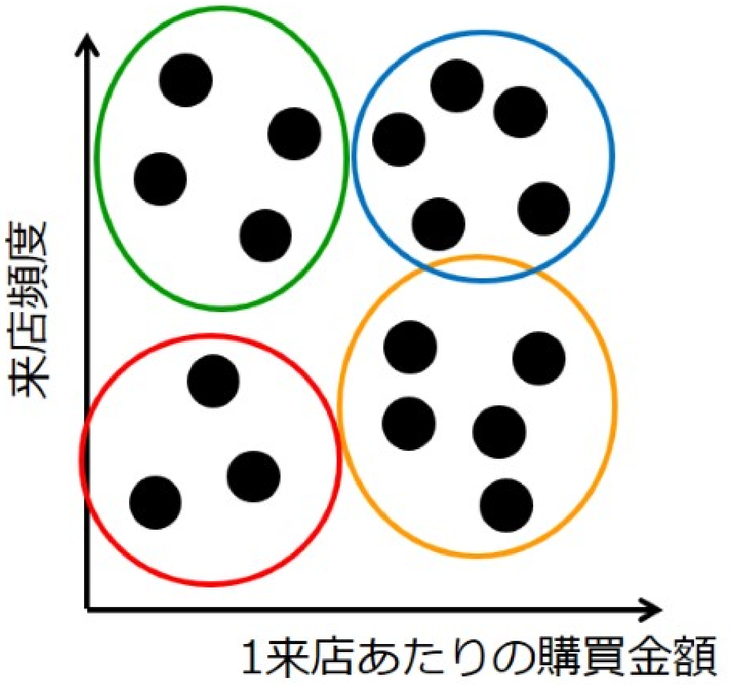
クラスタリングとは？

各クラスタの顧客に対する最適なプロモーション戦略はそれぞれ異なるはずです。

⬇

どんなクラスタがあり、各顧客がどのクラスタに分類されるかを判断するのは重要です。

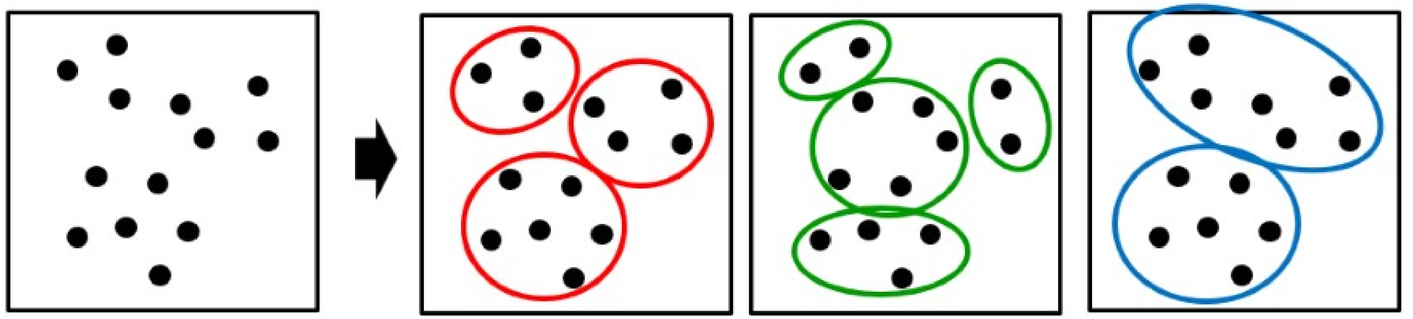
この例のように、クラスタを定義し、各データをクラスタに分類することを「クラスタリング」と呼びます。



Copyright © 2024 by INIAD

クラスタリングは自明ではない

- 「上手なクラスタリング」の要件
  - 均質性：同じクラスタ内のメンバーは互いに似ている。
  - 分離性：異なるクラスタに属するメンバーは互いに異なる。
- クラスタリングの難しさ
  - クラスタの個数は事前にはわからないことが多い。
  - メンバーの数は膨大であることが多い。
  - どのクラスタに分類するか迷うメンバーも多い。



どれが「一番よいクラスタリング」かは自明ではない! さまざまなアルゴリズムが提案されていて、目的によって使い分けることが多い。

Copyright © 2024 by INIAD

クラスタリングのアルゴリズム

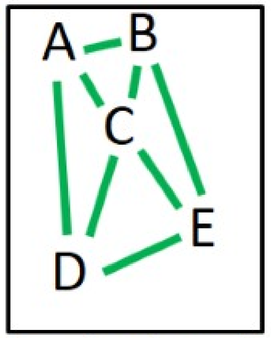
- クラスタリングのためのさまざまなアルゴリズムが提案されています。
- 階層型クラスタリング
  - 似たもの同士を次々とくっつけて、順にクラスタを大きくしていく。
  - 2つに分けるとしたら？ 3つに分けるとしたら… に答えてくれる。
  - クラスタ間の距離 (どのくらい「似ていない」か) の定義によって、最短距離法、最長距離法、重心法、群平均法、Ward (ウォード) 法など、さまざまなアルゴリズムがある。
- 非階層型クラスタリング
  - あらかじめクラスタの数を決めておき、その数にうまく分けるためのクラスタ位置・クラスタ境界を、反復を重ねて次第に改良していく。
  - k平均 (k-means) 法などが提案されている。

Copyright © 2024 by INIAD

重心法のアルゴリズム

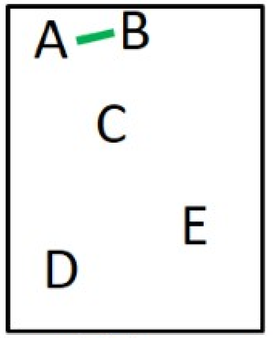
- 階層型クラスタリングのひとつです。
- もっとも近いもの同士を次々とペアにしていきます。

0



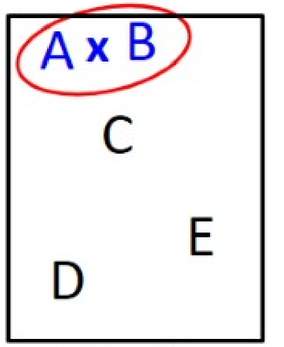
あらかじめ全ペア間距離を
求めておく

1



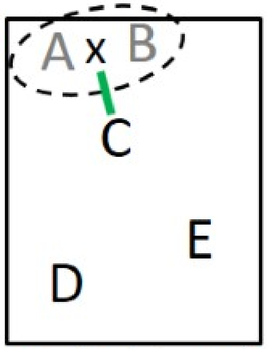
A-Bがもっとも近い

2



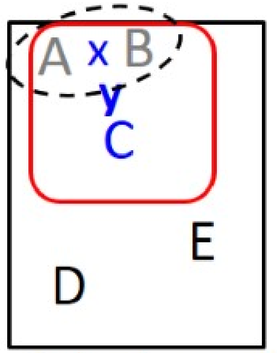
AとBをくっつける
A,Bの平均「x」を求める

3



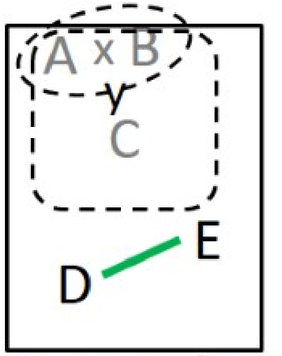
x,C,D,E の全ペア間距離の
中で x-Cがもっとも近い

4



x-Cをくっつける
A,B,Cの平均「y」を求める

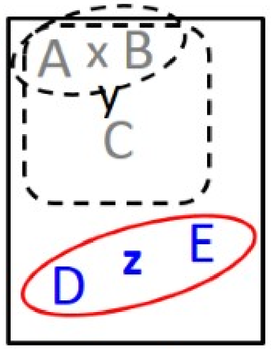
5



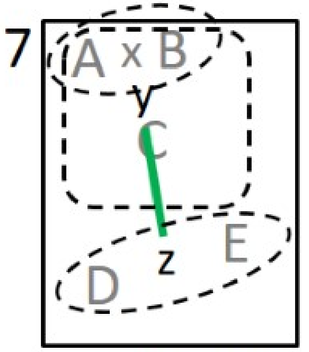
y,D,E の全ペア間距離の
中でD-Eがもっとも近い

Copyright © 2024 by INIAD

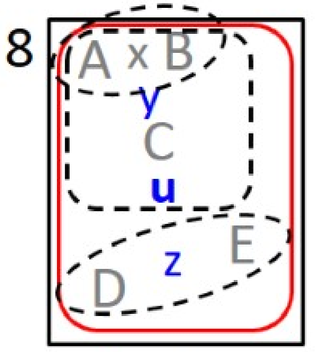
重心法のアルゴリズム



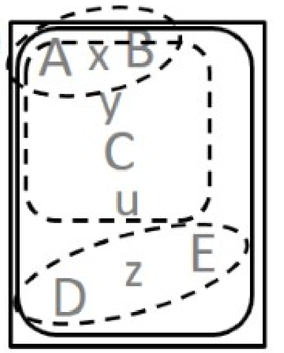
D-Eをくっつける
D,Eの平均「z」を求める



y-zが最後に残った

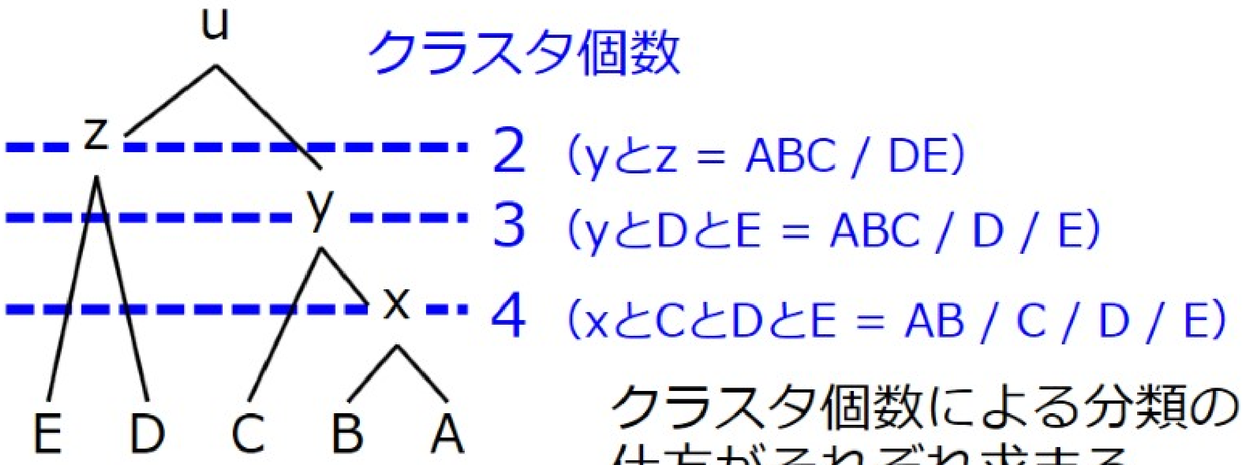


y-zをくっつける
A～Eの平均
「u」を求める



データがなく
なったので終了

- ペアのすべての組み合わせの間の距離を求めなければならないので、計算量大。
- 階層的クラスタリングの各段階を樹形図 (デンドログラム) に表現することができる。



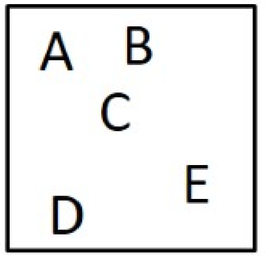
クラスタ個数による分類の
仕方がそれぞれ求まる

Copyright © 2024 by INIAD

k平均 (k-means) 法のアルゴリズム

- 非階層型クラスタリングの代表的なアルゴリズムです。
- あらかじめクラスタ数を決めておき、各メンバーを「一番近いクラスタ」に分類します。

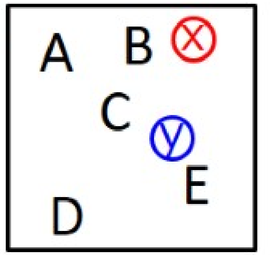
この5つのデータを
クラスタリングします



クラスタ数 k=2とします

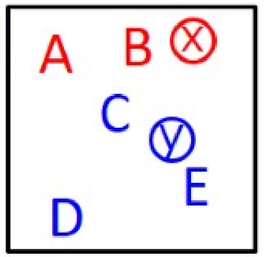
まず、各クラスタの
「中心」となる x, y を
任意の位置に置きます

1



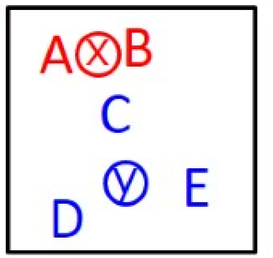
置いたクラスタ中心 x, y と
データA～Eの間の距離を求める

2



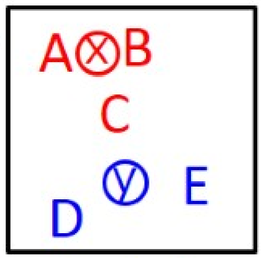
各データをもっとも近いクラスタ中心に割り当てる

3



各中心に割り当てられたデータの座標平均を、新たな中心位置とする

4

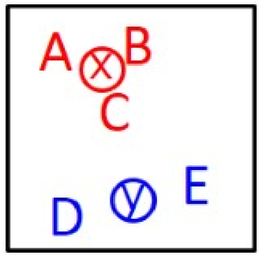


各データをもっとも近いクラスタ中心に割り当て直す

Copyright © 2024 by INIAD

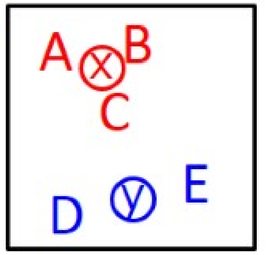
k平均 (k-means) 法のアルゴリズム

5



各中心に割り当てられたデータの座標平均を、新たな中心位置とする

6



各中心への割り当てが変わらないので終了

- 簡単のため2次元で説明しましたが、実際にはもっと多次元 (高次元) のデータを扱うことが多いです。
- クラスタ数はあらかじめ与えなければなりませんが、非常に高速に計算できるため、ビッグデータの解析によく用いられています。
- 一般に、最初の中心位置によって結果が変わる可能性があるため、異なる中心位置を初期値にして複数回クラスタリングを行い、結果がどれくらい安定しているかを調べることが多いです。

Copyright © 2024 by INIAD