情報管理

第９回：クラスタリング
教師なしクラスタリング(2)

今回の講義内容

前回は教師なしクラスタリングの方法としてK-means法と混合正規分布について解説しました。

今回は教師なしクラスタリングの方法として，「階層クラスタリング」という方法について解説します。

階層クラスタリングによる教師なしクラスタリング

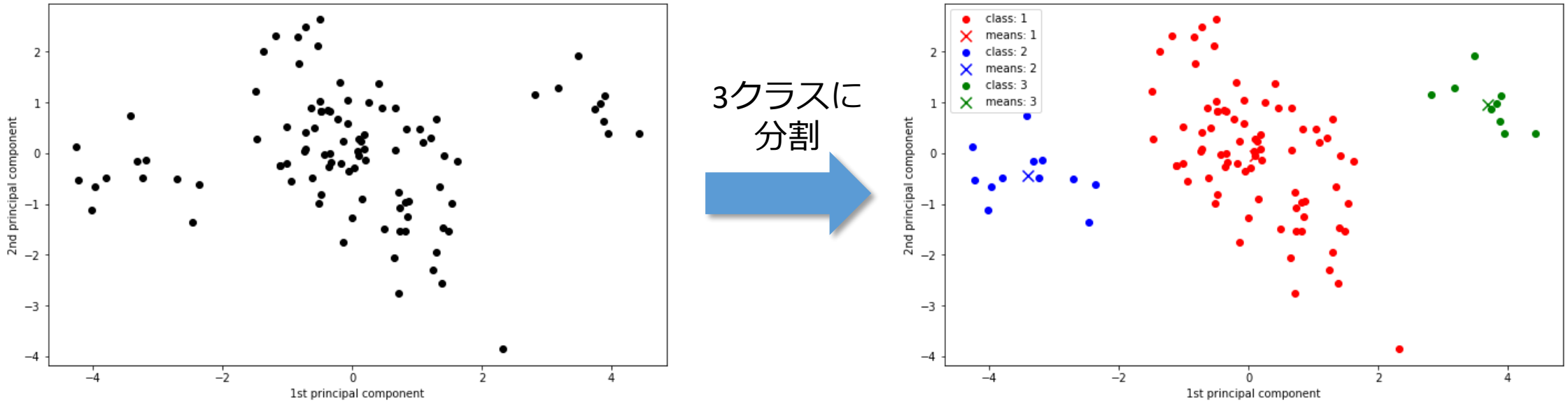
- 最短距離法
- 最長距離法
- ウォード法

教師なしクラスタリングの解説は今回までとなります。

階層クラスタリング

階層クラスタリングの目的

K-meansや混合正規分布は，あらかじめ分割するクラスの数(K)を決め，
それに従って教師無しクラスタリングをしていました。



この方法では，クラスの分割数を変えてクラスタリングしたい場合，その都度処理を最初からやり直す必要があります。

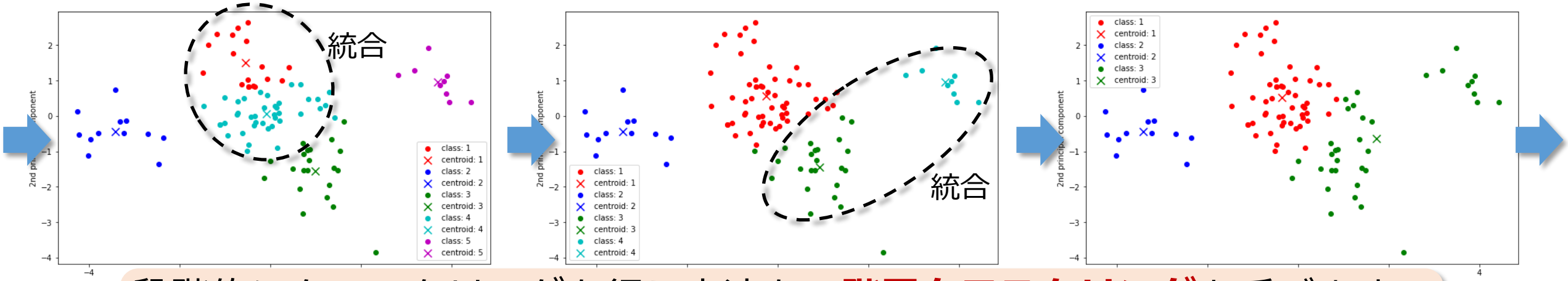
階層クラスタリングの目的

段階的にクラス数を変化させて，一つ前の結果を流用しながらクラスタリングできれば，毎回クラスタリングをやり直す必要がありません。

例：クラス数5の結果に対して，似ている２クラスを統合 → クラス数4の結果が得られる。

同様にしてクラス数4の結果を使ってクラス数3の結果を得る。

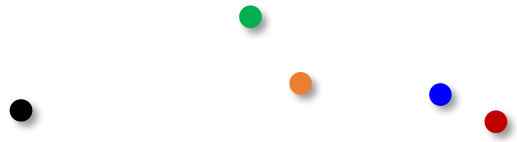
クラス数: 5　　クラス数: 4　　クラス数: 3



段階的にクラスタリングを行う方法を，**階層クラスタリング**と呼びます。
K-meansや混合正規分布は「**非階層クラスタリング**」と呼ばれます。

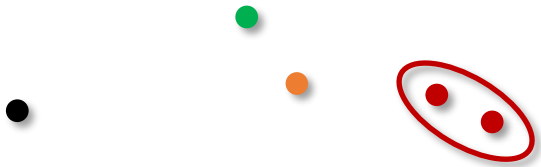
階層クラスタリングの基本アルゴリズム

初期状態は，データ一つ一つをクラスと扱います。（クラス数=データ数）



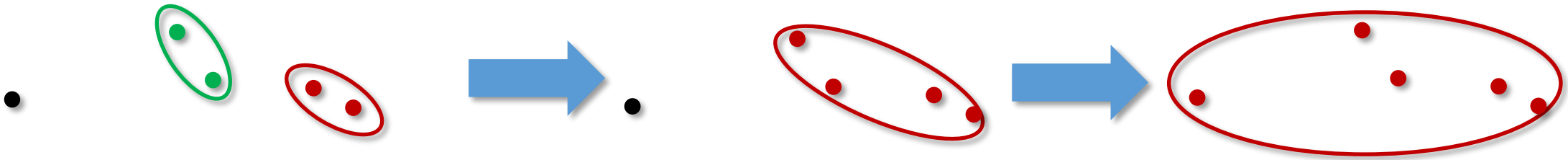
（それぞれのデータがクラスに相当）

全てのクラス同士で距離を測り，最も近いクラス同士を統合します。
これにより，クラス数が一つ減ります。



一つのクラスに統合

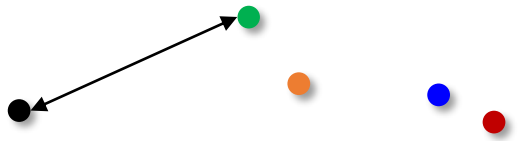
以上の処理を，クラス数が１，つまり全てのクラスが統合されるまで繰り返します。



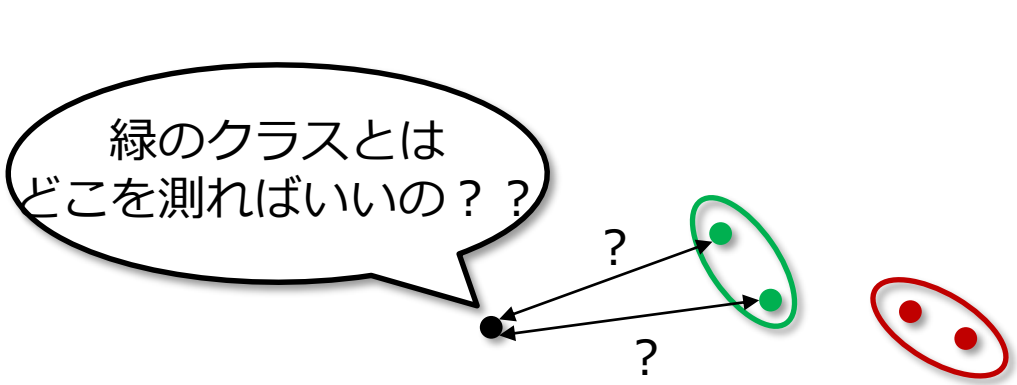
クラス数：3　　クラス数：2　　クラス数：1

クラス同士の距離はどうやって測る？

初期状態は，クラス＝サンプルなので，サンプル同士の距離を測って最小のペアを見つければよいです。



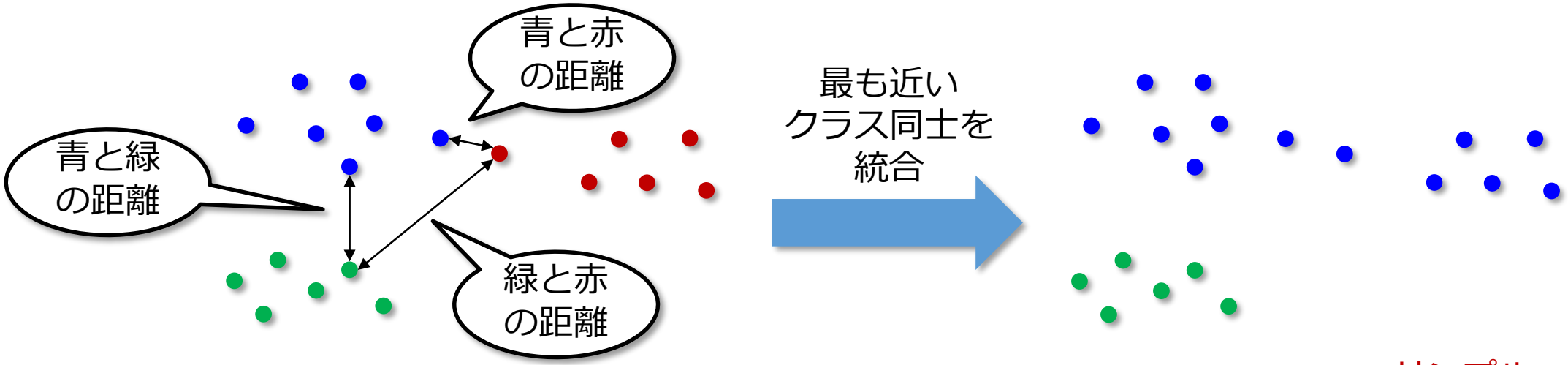
統合されたクラスと距離を測る場合，どこを測れば良いでしょうか？？



クラス間の距離の測り方によって，それぞれ異なるアルゴリズムになります。

最短距離法

クラス同士の距離を測る際，最も近いサンプル点同士の距離を採用します。



クラスc_1とクラスc_2の最短距離は以下で定義されます。

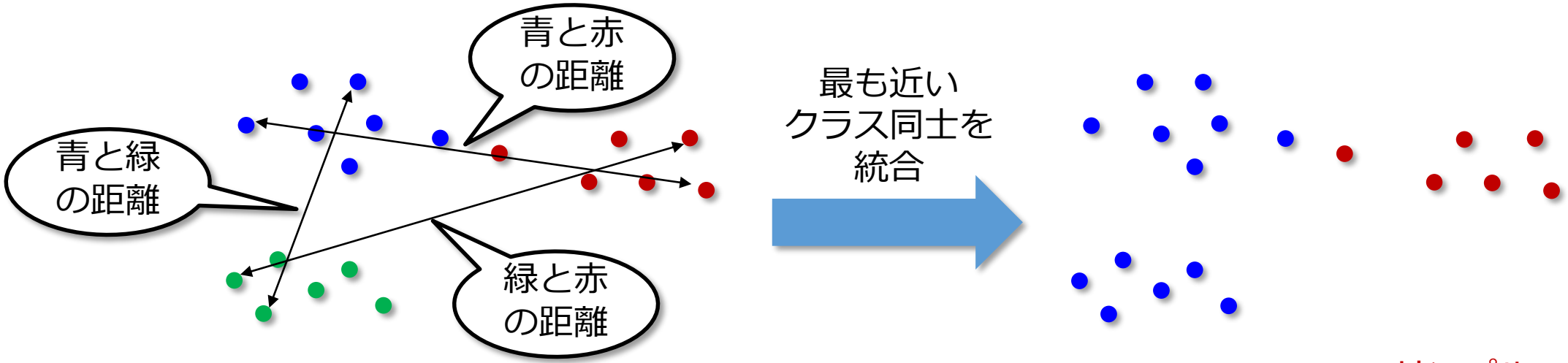
$$shortest\_distance(c_1, c_2) = \min_{x_1 \in c_1, x_2 \in c_2} d(x_1, x_2)$$

サンプルx_1と
サンプルx_2の
ユークリッド距離

クラスc_1に属するサンプルと
クラスc_2に属するサンプルで
最も距離が短くなるサンプルの組み合わせから算出

最長距離法

クラス同士の距離を測る際，最も遠いサンプル点同士の距離を採用します。



クラスc_1とクラスc_2の最短距離は以下で定義されます。

$$longest\_distance(c_1, c_2) = \max_{x_1 \in c_1, x_2 \in c_2} d(x_1, x_2)$$

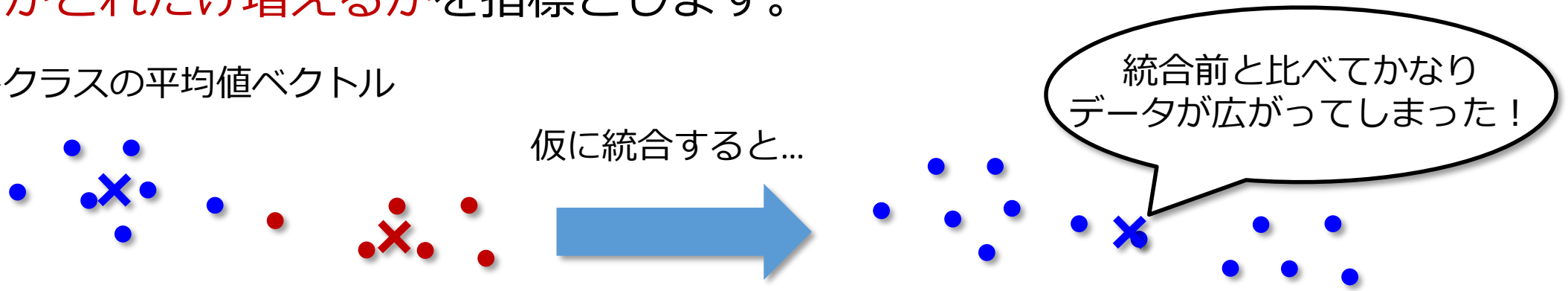
サンプルx_1と
サンプルx_2の
ユークリッド距離

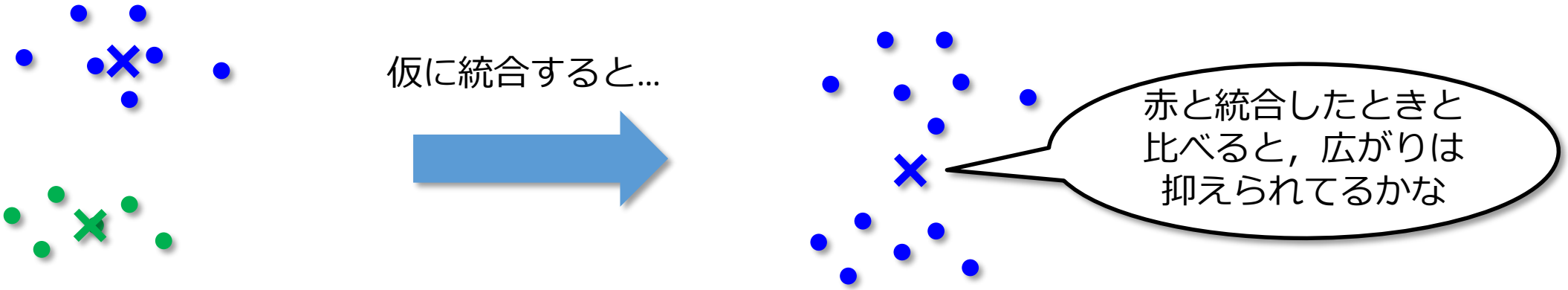
クラスc_1に属するサンプルと
クラスc_2に属するサンプルで
最も距離が長くなるサンプルの組み合わせから算出

ウォード法

正確には距離ではありません。仮に統合した場合，クラス内のデータの広がりがどれだけ増えるかを指標とします。

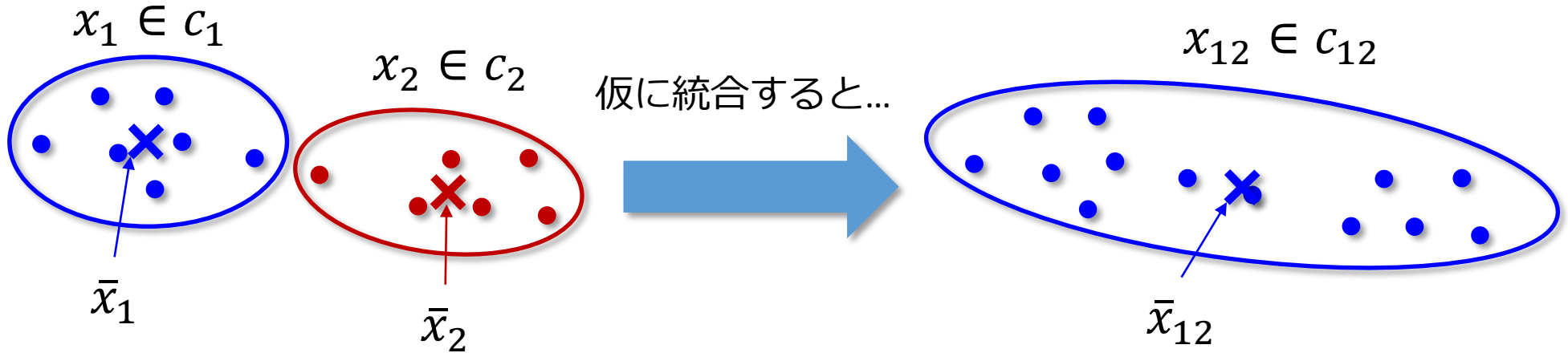
✖：各クラスの平均値ベクトル





ウォード法

クラスc_1とクラスc_2のウォード法による距離は以下で定義されます。



$$ward\_distance(c_1, c_2) = \sum_{x_{12} \in c_{12}} |x_{12} - \bar{x}_{12}|^2 - \sum_{x_1 \in c_1} |x_1 - \bar{x}_1|^2 - \sum_{x_2 \in c_2} |x_2 - \bar{x}_2|^2$$

統合前と後のクラスの広がりの差分

統合したクラスc_{12}内のデータと，平均値ベクトルとのユークリッド距離の二乗の総和（統合後の広がり）

クラスc_1内のデータと，平均値ベクトルとのユークリッド距離の二乗の総和（クラスc_1の広がり）

クラスc_2内のデータと，平均値ベクトルとのユークリッド距離の二乗の総和（クラスc_2の広がり）

階層クラスタリングのアルゴリズム・まとめ

初期状態では，各データのサンプルそれぞれがクラスとなる。

全てのクラスの組み合わせで，クラス同士の距離を測る。
距離の測り方は最短距離，最長距離，ウォード法距離などがある。

最も距離が小さかったクラス同士を統合する。

以上を，クラス数が 1 になるまで繰り返す。

09_01_hierarchical_clustering.ipynb を動かして，最短距離法の動作を確認しましょう。

おわりに

今回は，教師なしクラスタリングの方法として，階層クラスタリングを解説しました。

クラスを段階的に統合していく方法は，「凝集型クラスタリング」とも呼ばれます。

階層クラスタリングは非階層クラスタリングと比べて計算量が大きいですが，結果を見ながら分割数を決められることが利点です。

次回以降は，教師ありクラスタリングについて解説していきます。

レポート課題

第９回ファイル一式に含まれる，”report09.ipynb” に従って，’todoufuken.csv’のデータに対して最短距離法とウォード法を適用し，結果の違いについて考察しなさい。

小問1: 最短距離法を実施して挙動を確認せよ。
（分割数を固定しても良いし，分割数ごとにアニメーションを作成しても良い）

小問2: ウォード法によるクラスタリングを実装し，挙動を確認せよ。
分割数=6のときの結果を図示せよ。
（追加でアニメーションを作成しても良い）

小問3: 最短距離法とウォード法の結果の違いについて考察せよ。

レポート提出期限：7/5(火) AM10:30，ipynbファイルをhtmlファイルに変換して提出