

AI（人工知能）概論【Ⅱ】（Ⅱ）

～ 教員のための実践的データサイエンス入門 ～

第8講 「クラスタリングと次元削減」

小松尚登(滋賀大学データサイエンス・
AIイノベーション研究推進センター)

第8講「クラスタリングと次元削減」

【目 的】

第8講では、データサイエンスにおいて重要な前処理・分析手法であるクラスタリングと次元削減の概念と、それぞれの代表的な手法について取り扱います。これらは双方とも教師なし学習の一種であり、事前にラベル付けされた情報がなくても、データの構造やパターンを把握するのに役立ちます。

【学修到達目標】

- ① 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。
- ② 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。
- ③ 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。

目次

1. 導入

2. クラスタリング

- K-means法
- 階層的クラスタリング

3. 次元削減

- 主成分分析(PCA)
- 主成分分析の欠点と発展的な次元削減手法

4. 教師なし学習の注意点

1. 導入

機械学習の種類

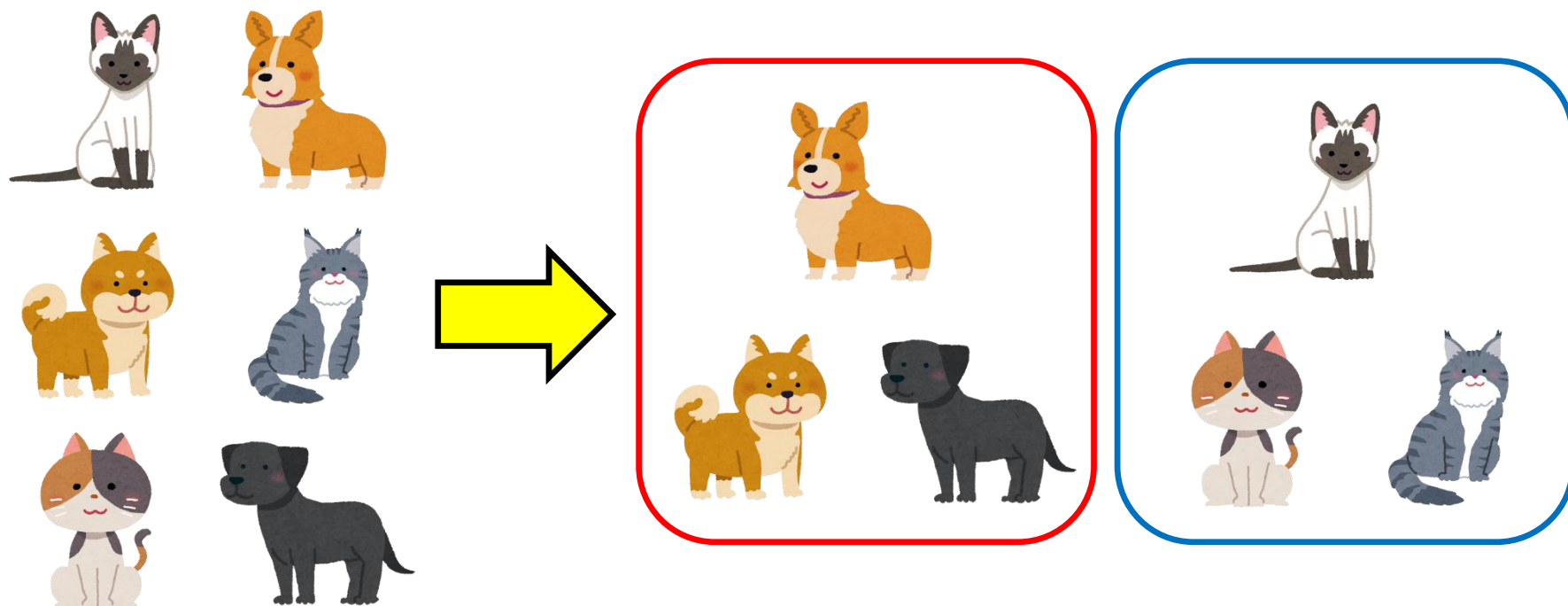
- 教師あり学習：「正解」のデータが与えられている
例)回帰分析、分類
- 教師なし学習：データに「正解」か否かの情報が無い
例)クラスタリング、次元削減
←今回はこれらについて説明
- 強化学習：「行動」に対して「報酬」が与えられた経験を学習

2. クラスタリング

データを似た特徴を持つグループ（クラスター）に分類

✓ 分類問題のように、データにラベルが付いている訳ではない

※画像引用元：いらすとや <https://www.irasutoya.com/>



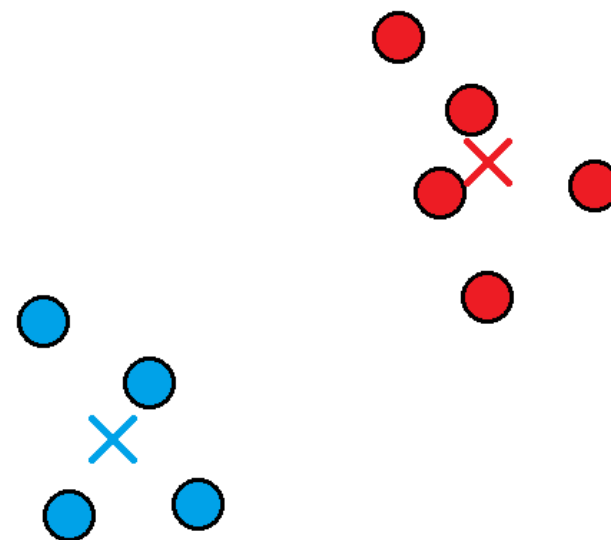
2. クラスタリング

k-means(k-平均)法：データを k 個のクラスターに分類する際によく用いられる方法

(手順)

1. k 個の「クラスター中心」(図の×印)を定める
2. 各データ点を最も中心に近いクラスターに割り当てる
3. 割り当てられたデータ点の平均値を新たな中心に定める
4. 2~3の手続きを繰り返す

※ k の値は予め決めておく必要あり。



2. クラスタリング

階層的クラスタリング：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。

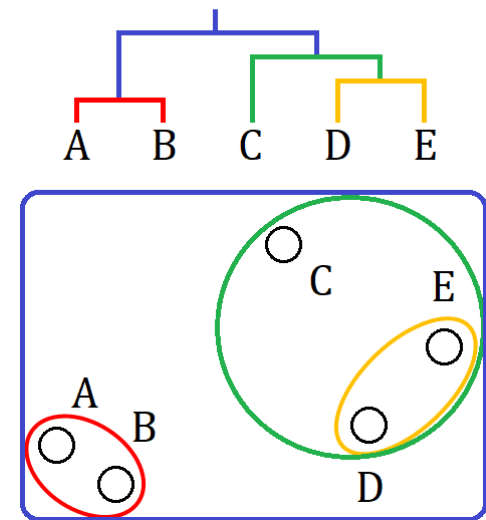
デンドログラム（樹形図）：階層的クラスタリングの際に得られる、木の枝分かかれのような形のグラフ

(手順)

- 1.N個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融合する
- 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

※「類似度」の定義は場合による

※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。



2. クラスタリング

階層的クラスタリングの各段階で、データ点が任意の数のクラスターに分類された状態が現れる。

- クラスタ数 k を事前に指定しなければならないk-means法とは異なり、好きな数のクラスターに分類することが可能

クラスタ数

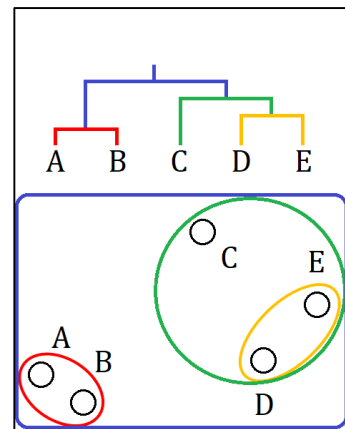
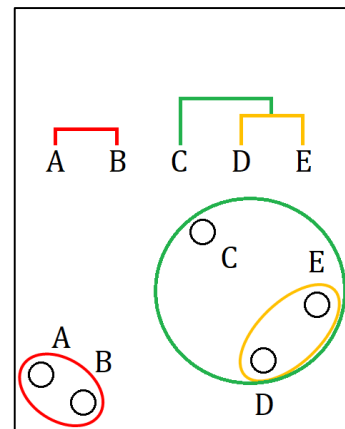
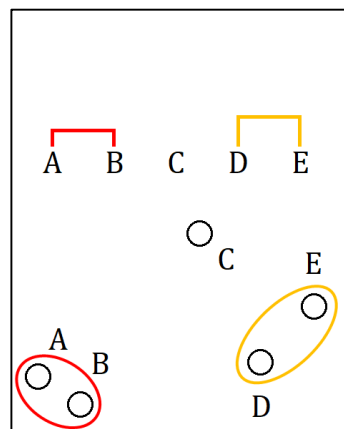
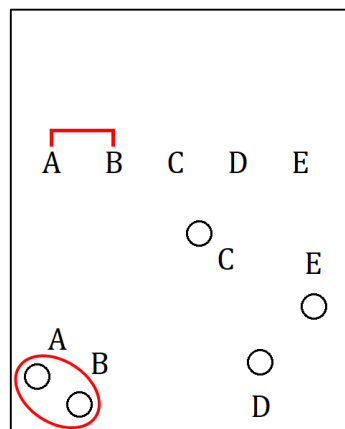
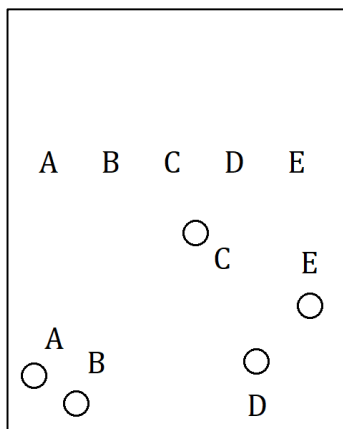
5

4

3

2

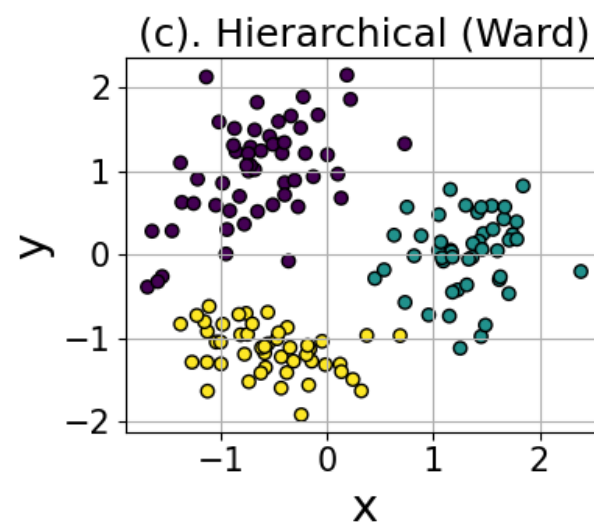
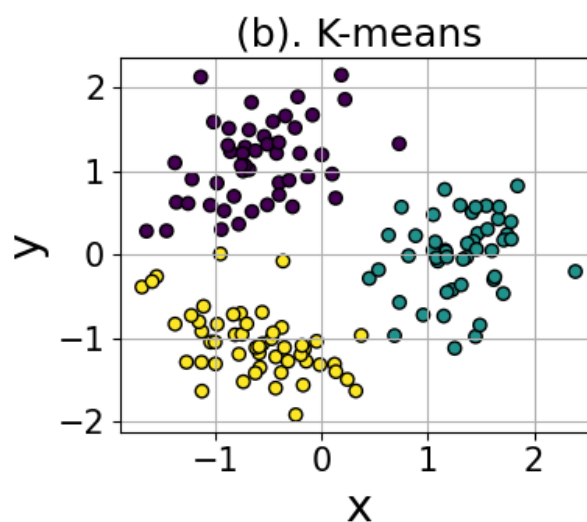
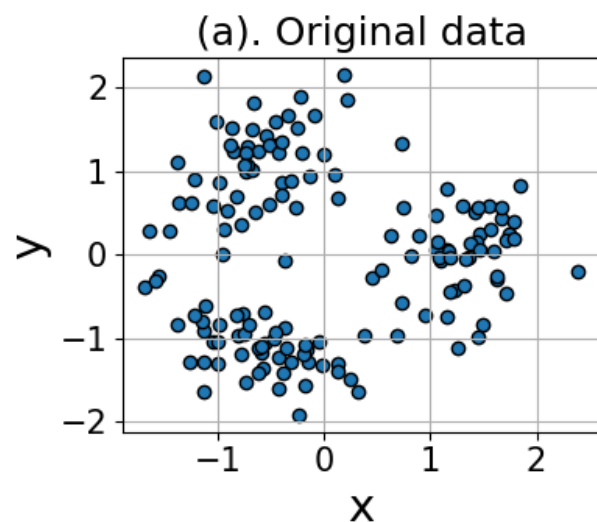
1



2. クラスタリング

実践例：2変数のダミーデータ

- (a): 元のデータ
- (b): k-means法(クラスター数 k を $k = 3$ と指定)
- (c): 階層的クラスタリング(クラスター数 k が $k = 3$ になるまで操作を実行)



2. クラスタリング

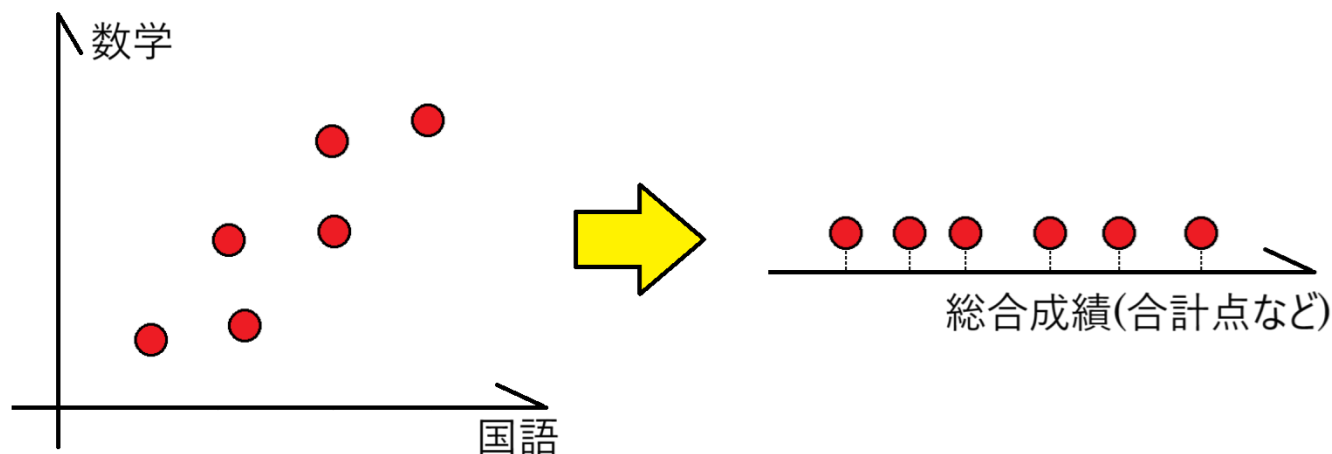
- 各手法のメリット、デメリット
 - k-means法：始めにクラスター数 k を指定する必要があるが、データ点が多い場合の計算コストは低め。
 - 階層的クラスタリング：始めにクラスター数を指定する必要が無いし、クラスター結合の過程をデンドログラムとして可視化できるが、データ点が多くなると計算コストが高くなる。
- 状況に応じて、適切な手法を選ぶ必要がある。

3. 次元削減

高次元データを低次元データに(なるべく情報を失わないように)変換し、本質的な特徴を抽出する

※データが含む変数の数を次元と呼ぶ。これは、 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によるものである。

※データの次元を減らす目的としては、「他の機械学習手法による分析のための前処理」、「4変数以上のデータを2変数や3変数に圧縮し、グラフとして可視化するため」などが挙げられる。



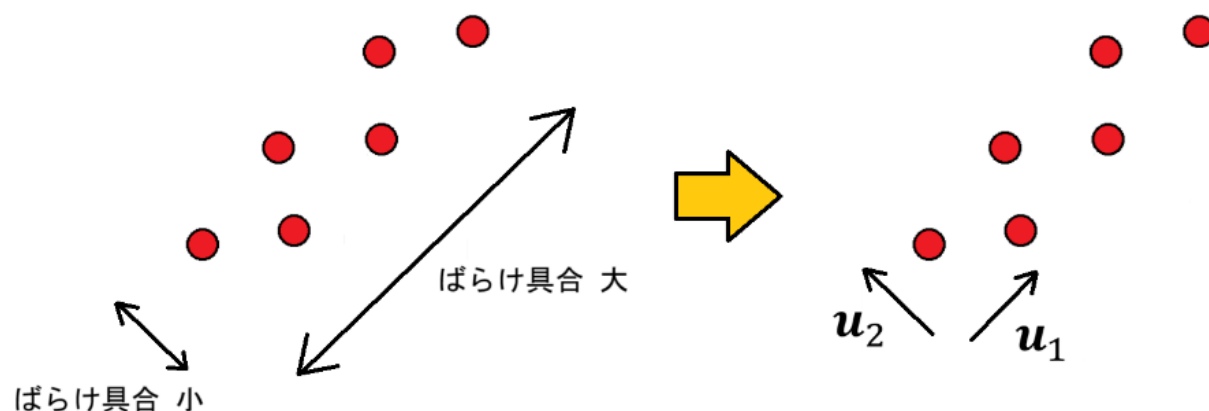
3. 次元削減

主成分分析：以下の手順で特徴量を抽出する。

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 $\mathbf{u}_1$ 、二番目に大きい方向 $\mathbf{u}_2$ 、三番目に大きい方向 $\mathbf{u}_3$ 、... 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 $\mathbf{u}_1, \mathbf{u}_2, \dots$ 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。

例) 2次元データ： $\mathbf{u}_1$ に対応する主成分が特徴量となる。

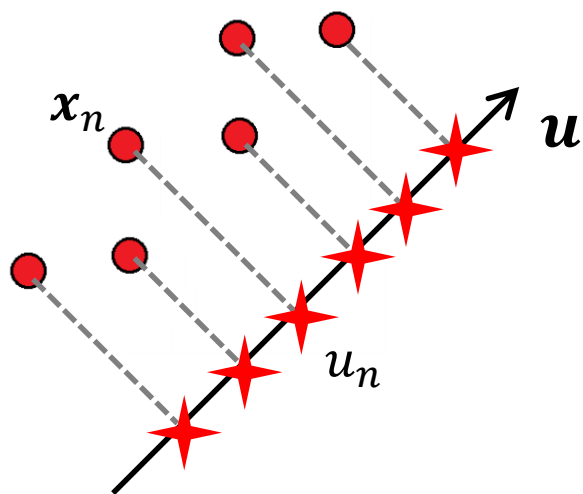
(※ 2次元データの場合、 $\mathbf{u}_1, \mathbf{u}_2$ 両方に関して座標を取ると、データの次元が減らないことに注意。)



3. 次元削減

Q. 「ばらけ具合が大きい方向」とはということか？

A. 以下のように、ある方向 $\mathbf{u}$ に沿って新しい座標軸をとることを考える。この際、図のようにデータ点 $\mathbf{x}_n$ から座標軸に向けて垂線を下ろすと、その垂線の足（図の星印）の座標は $u_n = \mathbf{x}_n \cdot \mathbf{u}$ と与えられる。「ばらけ具合が大きい方向」とは、この u_n の分散が大きくなるような方向を指す。



3. 次元削減

手順2.で現れる $\mathbf{u}_1, \mathbf{u}_2 \dots$ を単位ベクトル（長さを1のベクトル）とした場合、データ点 $\mathbf{x}_n$ の主成分は $\mathbf{x}_n$ 自身を中心化した値

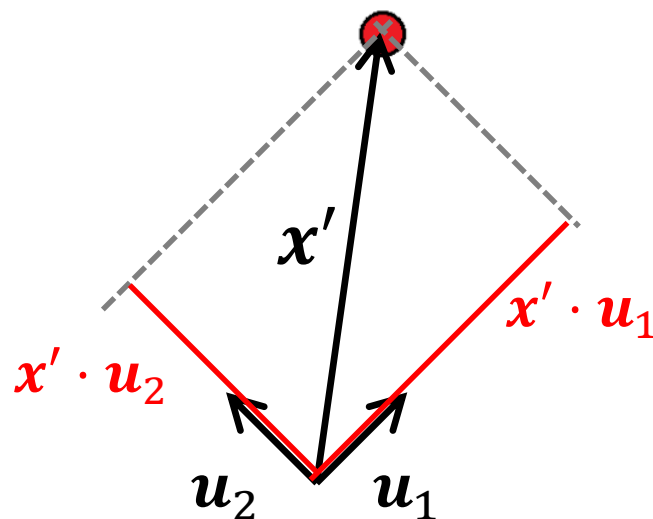
$$\mathbf{x}'_n = \mathbf{x}_n - \bar{\mathbf{x}},$$

where $\bar{\mathbf{x}} = \sum_{n=1}^N \mathbf{x}_n / N$

とそれらのベクトルとの内積

$$\mathbf{x}'_n \cdot \mathbf{u}_1, \mathbf{x}'_n \cdot \mathbf{u}_2, \dots$$

によって与えられる。

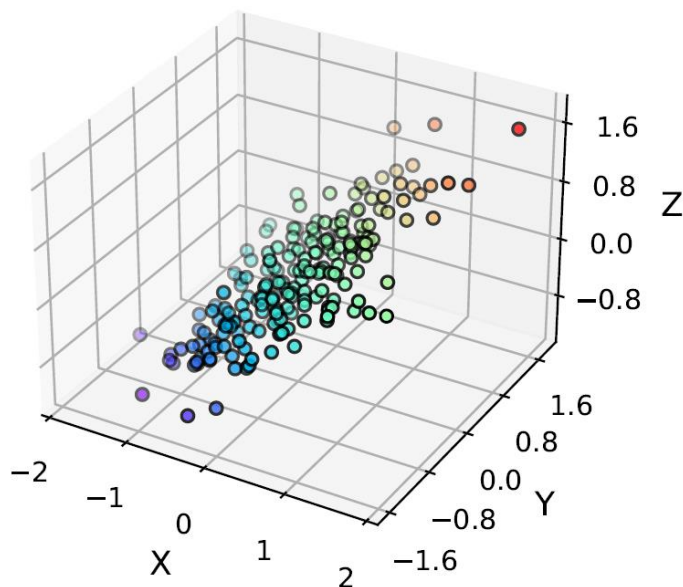


3. 次元削減

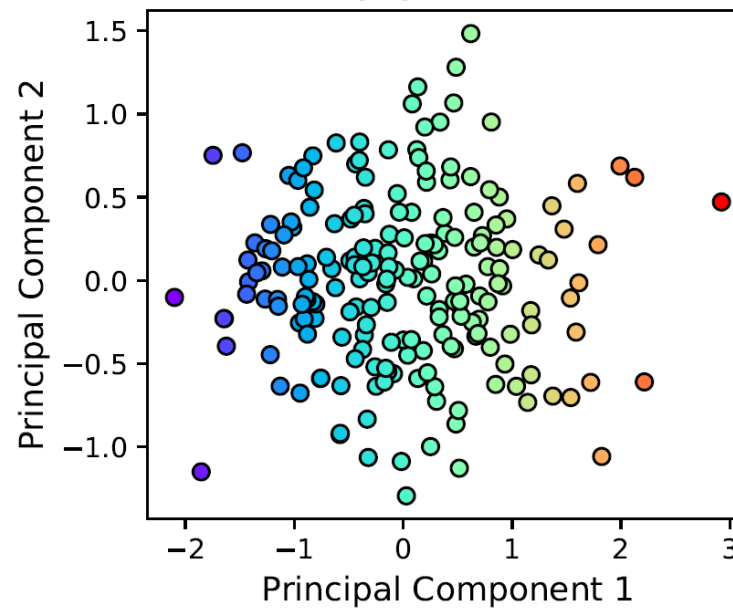
実践例：3変数のダミーデータ

- (a): 元のデータ
- (b): 主成分分析

(a). Original data

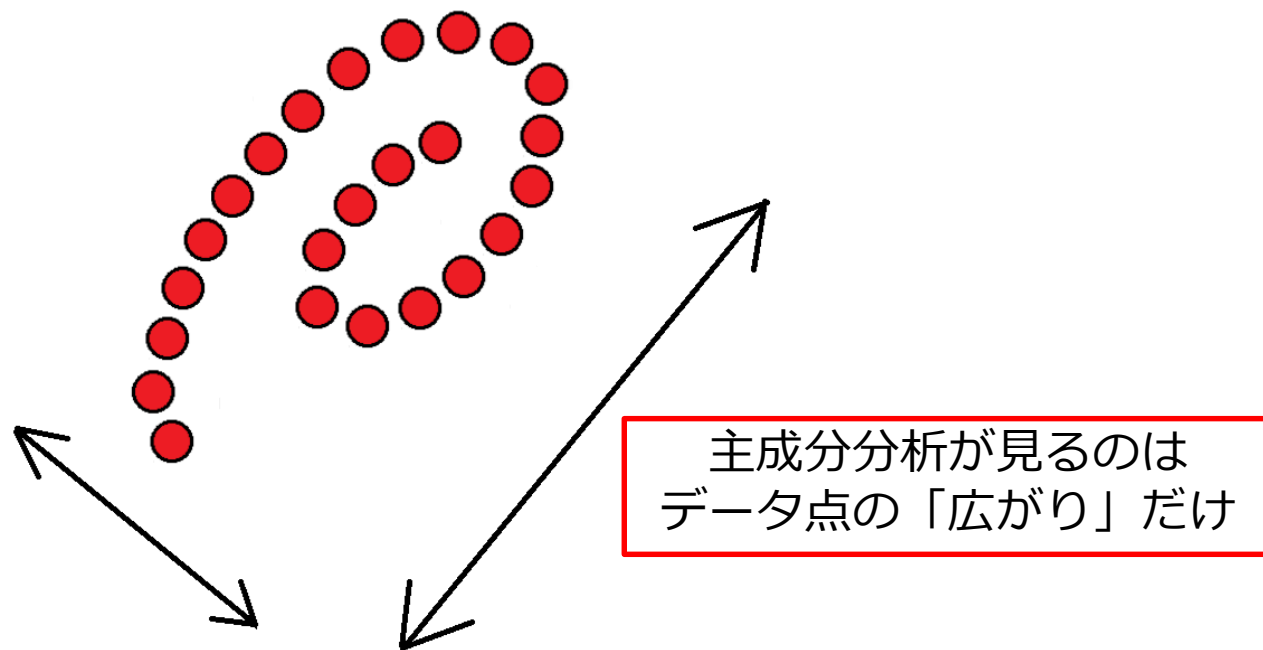


(b). PCA



3. 次元削減

- 主成分分析はデータ点の「広がり」だけを基準に次元削減をする。
- 図のように、データ点の分布に曲線的に複雑に入り組んだパターンが埋め込まれていたとしても、そうした構造を捉えることはできない

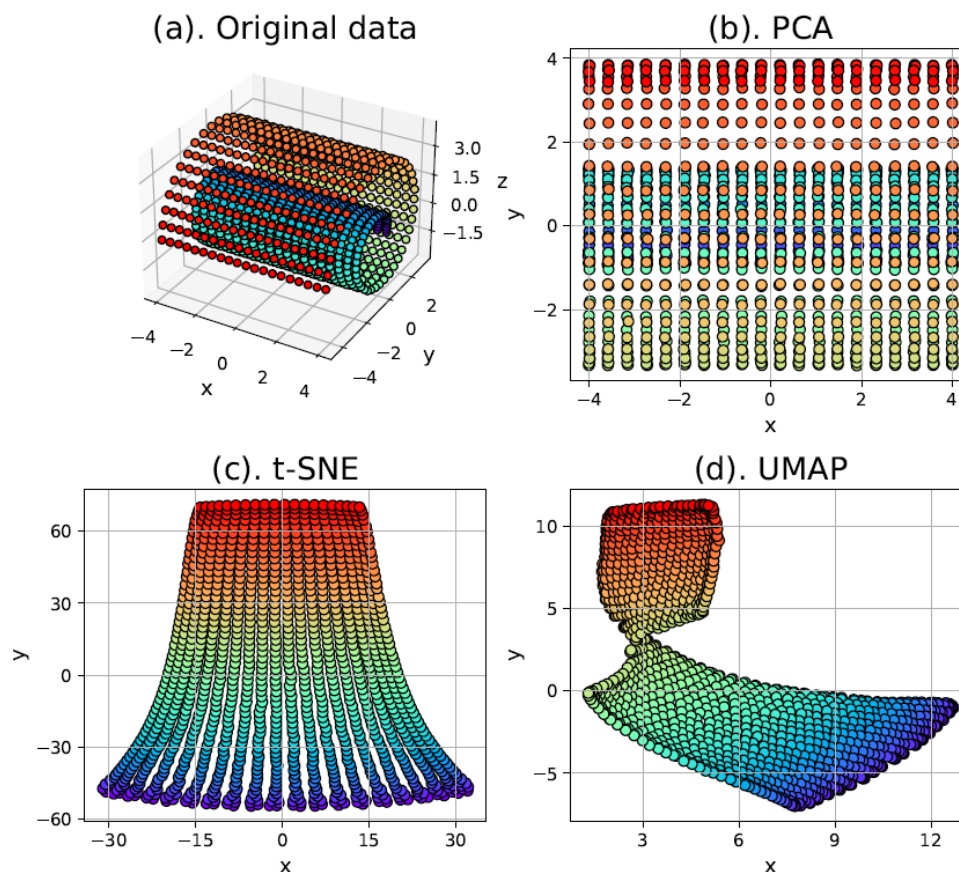


- 非線形な構造を捉えられる次元削減手法：**t-SNE**や**UMAP**などが存在

3. 次元削減

実践例：データ点がロールケーキ状に配置された三次元ダミーデータ

- (a): 元のデータ (b): 主成分分析 (c): t-SNE (d): UMAP

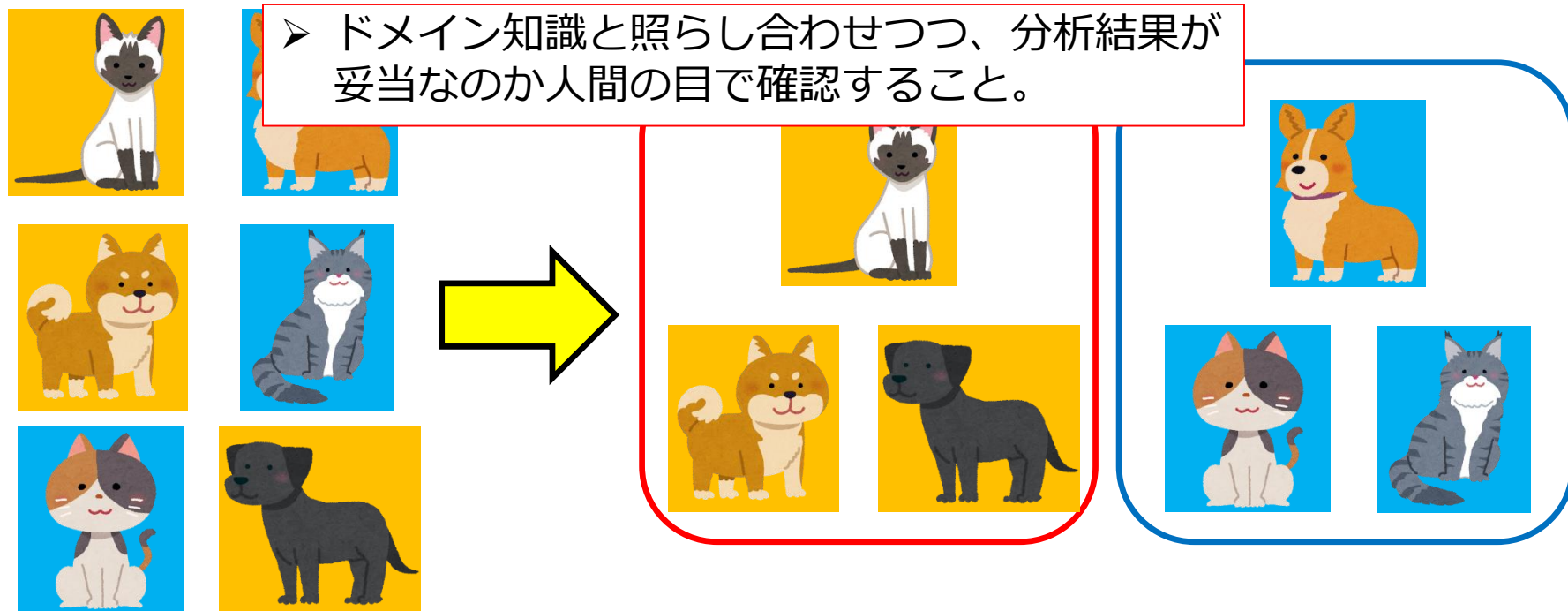


4. 教師なし学習の注意点

教師なし学習は、何が「正解」なのかにに関する情報が与えられていない。

- 人間側が本来組みたかった課題にとって重要な情報を見落したり、逆にどうでもいい情報に焦点を当てたりしてしまう危険性も

例：ラベルがついていない犬と猫の画像を分類して欲しかったのに…



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。