ビジネスに役立つデータ分析（入門編）

瓜生真也（徳島大学デザイン型AI教育研究センター）

諸注意

資料置き場: https://github.com/uribo/cue2022aw_r104

投影するプレゼンテーション、ソースコードを置いています
（来週分は来週更新）

❶ Codeボタンをクリック

❷ Download ZIP をクリック

Rコードと実行環境

```r
# 気温とアイスの相関係数を求める
cor(df_icecream_temperature$temperature_average_c,
    df_icecream_temperature$value)
#> [1] 0.9144466
```

Rでの実行コマンドであることを示します

https://mybinder.org/v2/gh/uribo/cue2022aw_r104/main?urlpath=rstudio

から必要なパッケージ、データ、ソースコードを含んだRStudioが起動します

第四週:
データ分析入門(2)

瓜生真也（徳島大学デザイン型AI教育研究センター）

講座の内容

第一週

第二週

第三週

第四週

データサイエンス入門 (2)

クラスタリング

　K平均法

主成分分析

第五週

教師なし学習

得られている変数からデータを区分する

予測には興味がない、データの背後にある構造を理解しようとする

結果に対して主観的な判断が求められる（客観的に判断しにくい）

次元削減、異常検出、推薦のために利用される

教師あり学習（教師付き学習）

回帰、分類モデル

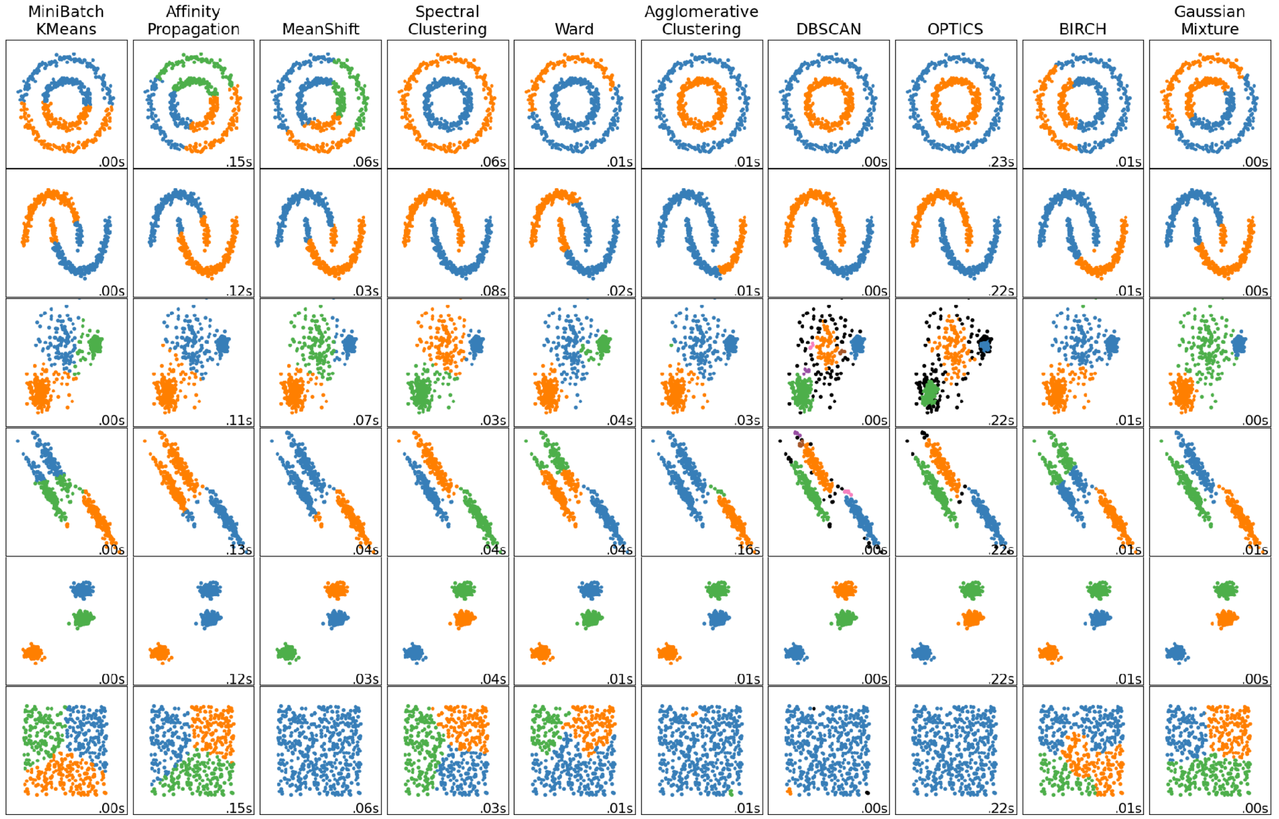
クラスタリング

クラスタリング

データ間の類似度・距離をもとに、データを未知のグループ（クラスター）に分割する

さまざまなアルゴリズム

- K平均法
- 階層クラスタリング
- DBSCAN

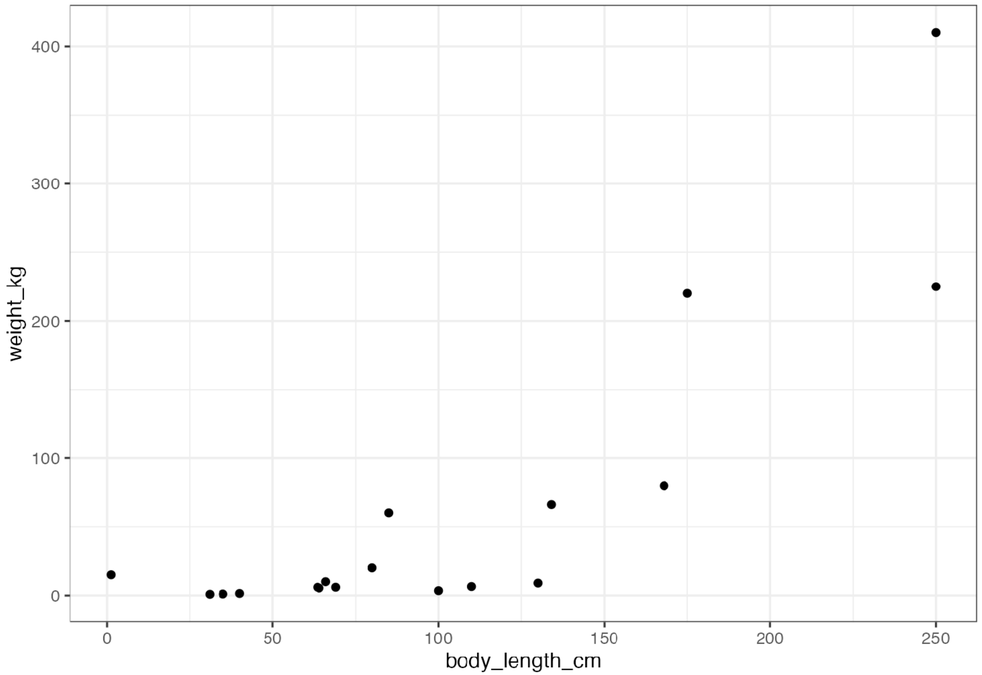


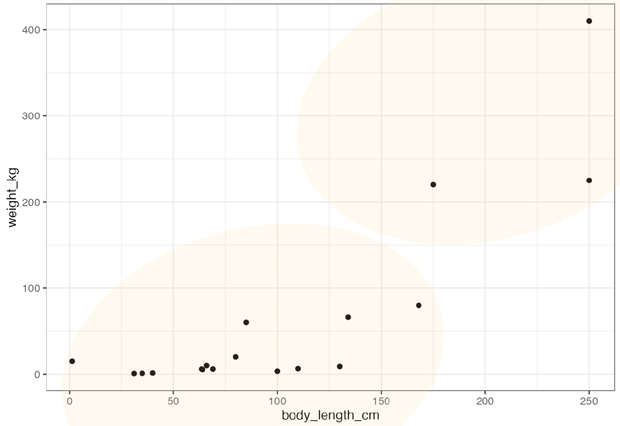
https://scikit-learn.org/stable/modules/clustering.html#k-means

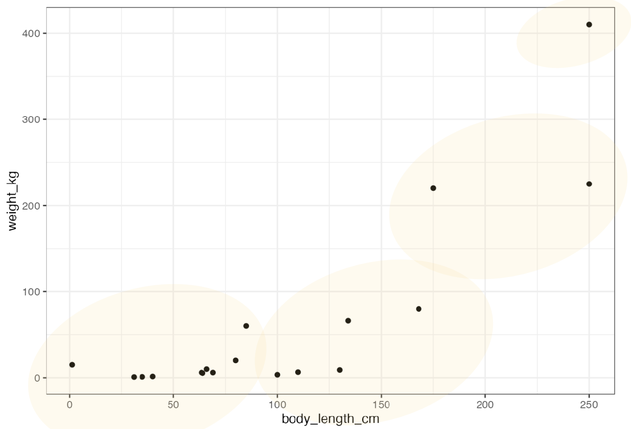
K平均法

クラスタの数 k をあらかじめ決めておき、クラスタ内の平均からの距離の二乗和が最小となるよう、入力をk個に分類する

動物の体長と体重データをいくつかのグループに分けるとしたら？







K平均法の手順

❶クラスタの数 k を決める
❷ランダムに各データをクラスタに割り振る
❸-1各クラスターの重心（平均値）を求め、各データからの距離（ユークリッド距離など）を求める
❸-2 各データを最も近い重心に対応するクラスタに振り分け直す

手順❷のランダムな割り振りのために、K平均法の結果は実行のたびにわずかに異なることがある

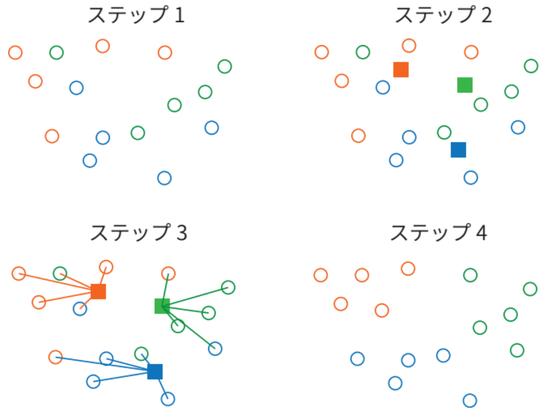


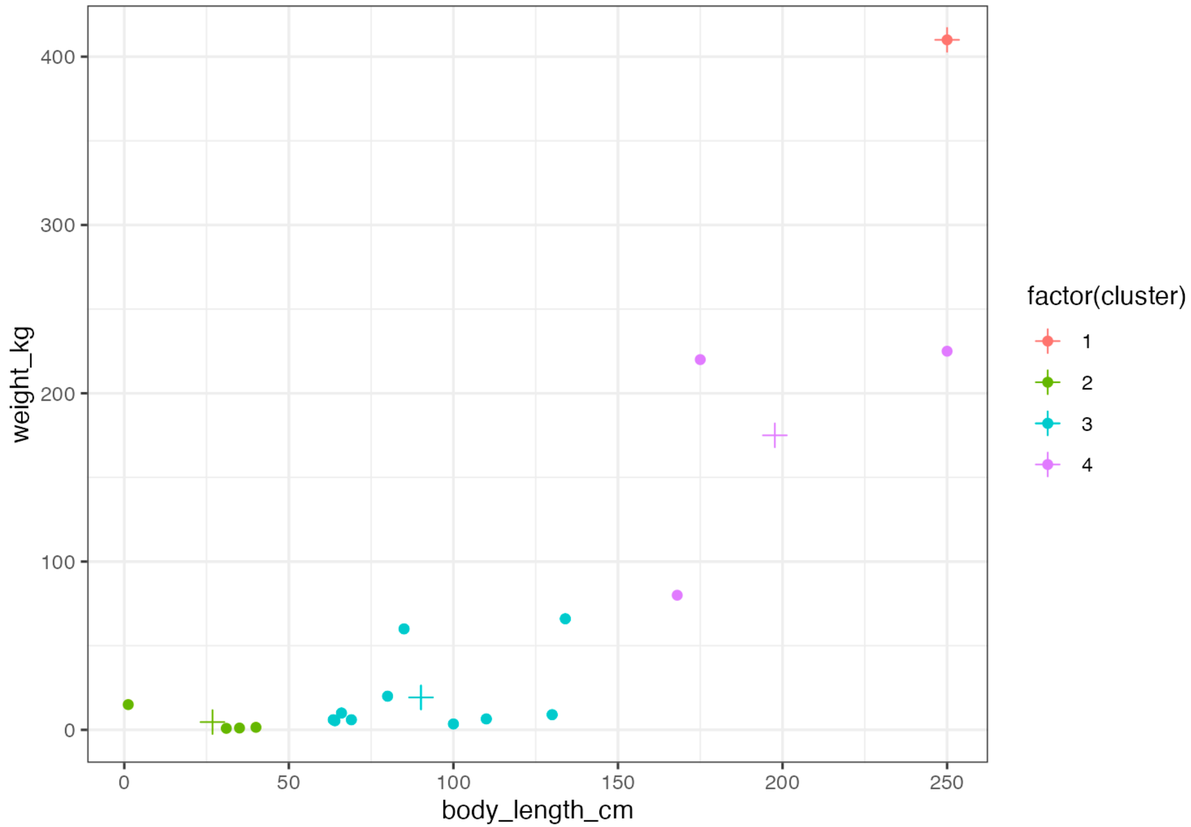
図 4.12 *k*-平均法の手順．まず，それぞれのサンプルにランダムにクラスタを割り当てて（ステップ 1），次に四角で示されるクラスタ平均を求めます（ステップ 2）．そして，クラスタ平均から距離の近いサンプルを再度，クラスタに割り当てなおします（ステップ 3，4）．この手順を繰り返しながら，サンプルのクラスタ割り当てを調整していきます．

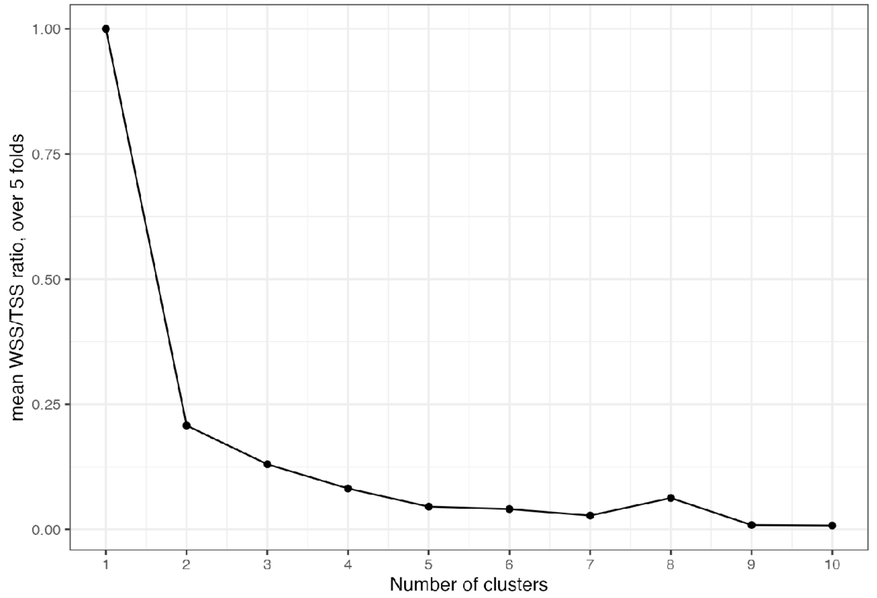
藤原幸一 (2022)より引用

Kの数をどうやって決める？

クラスタ内でのバラツキの和を最小にすることが最適化

k = 4の場合の動物データのクラスタリング結果





主成分分析

主成分分析 (Principal Component Analysis: PCA)

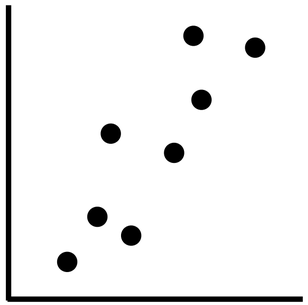
多次元（多変量）のデータがもつ情報を可能な限り維持した状態で低次元空間に縮約する

多次元を2次元、3次元に縮約　→可視化による関係の把握、解釈の促進

→相関、回帰による傾向の把握

線形射影の一種。変数の絞り込み（次元削減）の手法としても利用される

2次元のデータの関係



散布図等で示すことが可能

多次元のデータの関係

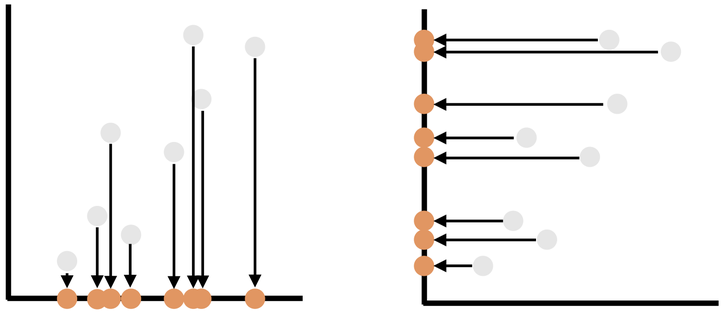
？

可視化、解釈が困難

次元の縮約と情報の損失

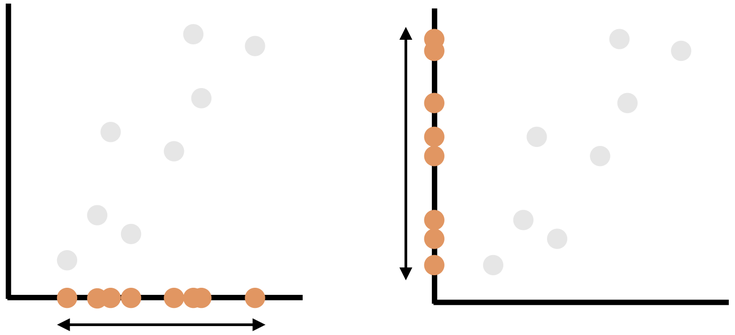
例）2次元のデータを1次元に縮約する

体重、身長... 2つの変数を1つの変数で表現する … 体の大きさ、BMI



縦軸と横軸上にデータを縮約する ● から ● へ

→縦軸と横軸それぞれで情報の損失が起こる



データの分散を元データの情報量として考える

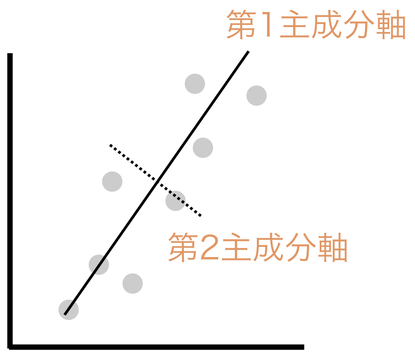
バラツキが大きい… 情報量が多い

主成分分析の目的

元のデータの情報損失ができるだけ小さくなるような（分散が最大となる）軸を探す

分散共分散行列または**相関行列**より固有値を求める

→データから変数間の相関関係を抽出



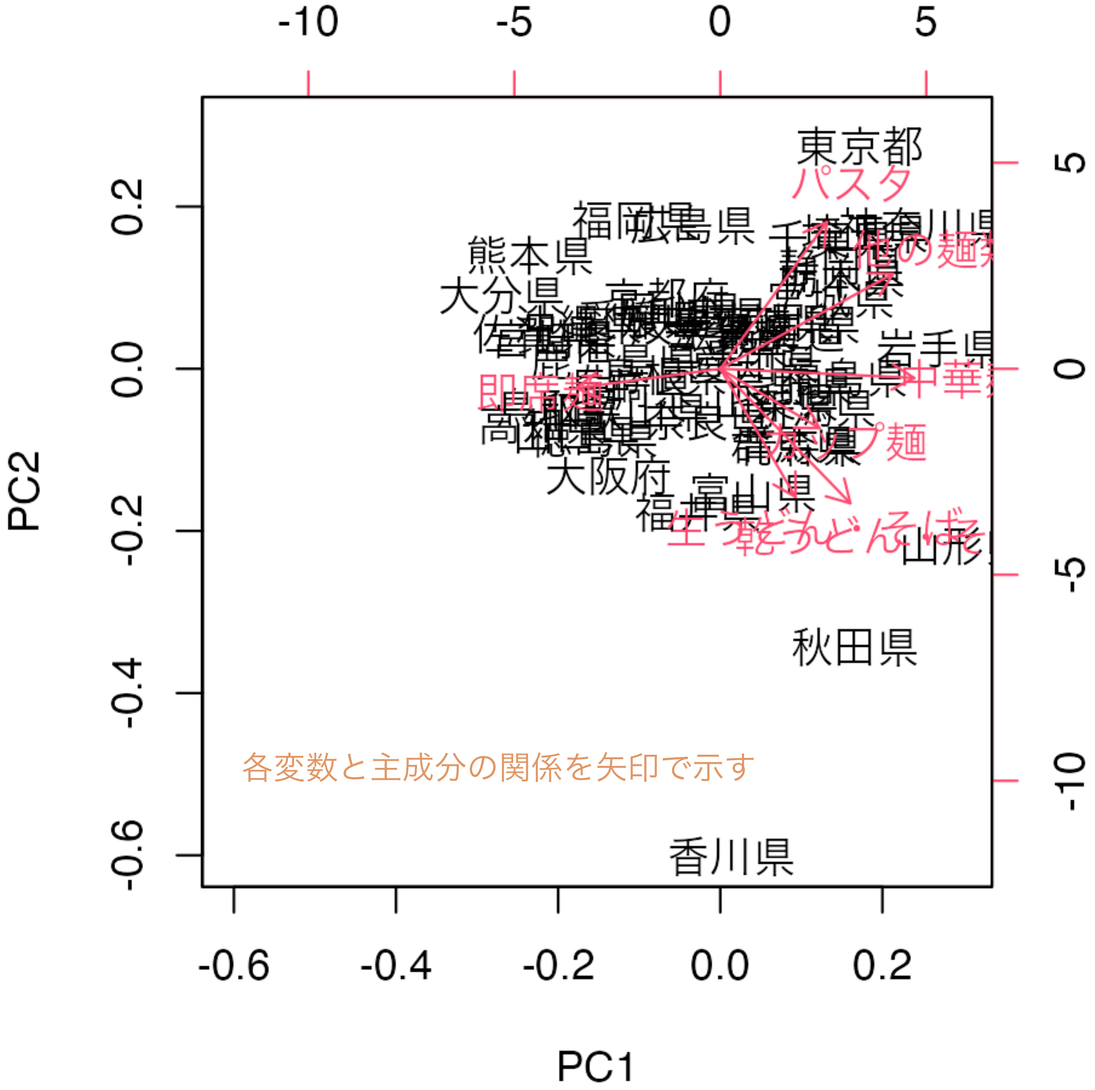
第1主成分 … データの分散が最も大きくなるような方向（軸）
データの「特徴」を最も表現する

第2主成分 … 第1主成分と直交する軸の中で、
軸上に射影したデータの分散が最も大きくなる軸
第1主成分では表現しきれなかった「特徴」を表現

変数の標準化を行わない場合、変数の単位に依存して結果の解釈が異なる

主成分分析の結果の解釈

主成分空間（第一主成分、第二主成分を両軸にとる）に主成分得点をプロット



参考文献・URL

G.James, D.Witten, T.Hastie, R.Tibshirani 著 落海浩、首藤信通 訳 (2018).
Rによる統計的学習入門 (朝倉書店) ISBN: 978-4-254-12224-4

藤原幸一 (2022).
スモールデータ解析と機械学習 (オーム社) ISBN: 978-4-274-22778-3