統計学 Lecture 12 クラスタリング

国立大学法人
東京海洋大学
Tokyo University of Marine Science and Technology

機械学習法

機械学習とは？ データから関係性や特徴を学ぶこと.

- 教師あり学習 (Supervised learning):
  - 入力xと出力yのようにデータがペアになっている
  - この入出力データから関係性を見出す，あるいは入力xからyを予測する為に学習することが目的
  - yが離散値のとき，判別または分類とよぶ
  - yが連続値のとき，回帰分析とよぶ

- 教師なし学習 (Unsupervised learning):
  - 入力xのみがデータとして観測される
  - データの次元縮小やクラスタリングなどを通して，データの特徴を取り出すことが目的

- 強化学習 (Reinforcement learning):
  - 利得が最大になるように行動を逐次的に選択するルールを学習することが目的
  - 動的最適法，マルコフ決定過程，深層強化学習などが例にあたる

機械学習のタイプ

教師あり学習

環境変数　⇔　資源密度　(関係を知りたい)

教師なし学習

形態測定　(似ているグループを見つけたい)

強化学習(教師なし学習)

環境変数　⇔　CPUE

(環境変数とCPUEの情報を基に儲けが最大になるように漁場を逐次的に選択するルールを模索．行動の最適化)

機械学習のタイプ

教師あり学習

環境変数　⇔　資源密度　（関係を知りたい）

トレーニングデータ (x_i, y_i) を用いて関係を学習

トレーニングデータ以外の新たな入力に対して
予測能力（汎化能力）のある方法（写像，矢印）を見つけたい

機械学習のタイプ

教師なし学習

形態測定　(似ているグループを見つけたい)

入力
×

入力の特徴について学習する

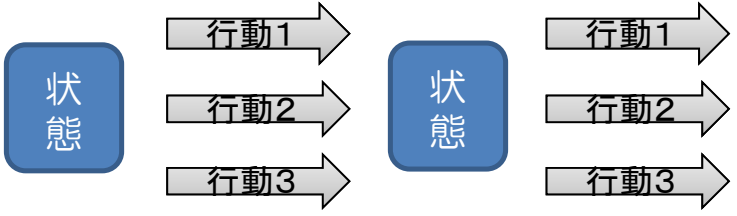
- 要約（次元の削減，縮小）
- 類型化，クラスタリング

機械学習のタイプ

強化学習(教師なし学習)

環境変数　⇔　CPUE

(環境変数とCPUEの情報を基に儲けが最大になるように漁場を逐次的に選択するルールを模索．行動の最適化)



マルコフ的に状態に依存して行動を行う．
行動によって（正解ではなく報酬が決まり），
　将来の報酬が大きくなるような行動を模索したい

機械学習(教師あり，教師なし) の分類

教師あり学習

正則化アプローチ

- 回帰分析
- Lasso, Elastic net
- サポートベクターマシン

アンサンブル学習

- Bagging
- Boosting
- ランダムフォレスト

- ニューラルネットワーク
- 深層学習
- ・・・

教師なし学習

分類（離散的）

- クラスタリング
- Mixture model
- ・・・

次元縮小（連続的）

- 主成分分析
- 多次元尺度法

教師なし学習 (unsupervised learning)

教師なし学習 (unsupervised learning)

Principal Component Analysis (PCA) 主成分分析:

- データの次元縮小，可視化，事前処理などの目的で利用

Multi-dimension scaling (MDS) 多次元尺度法:

- 次元縮小によりデータを2次元平面あるいは3次元空間で可視化する方法

Clustering クラスタリング:

- データの類似度からグルーピングを行い，データの特徴を取り出す方法)
  - ここではクラスタリング法について紹介します.

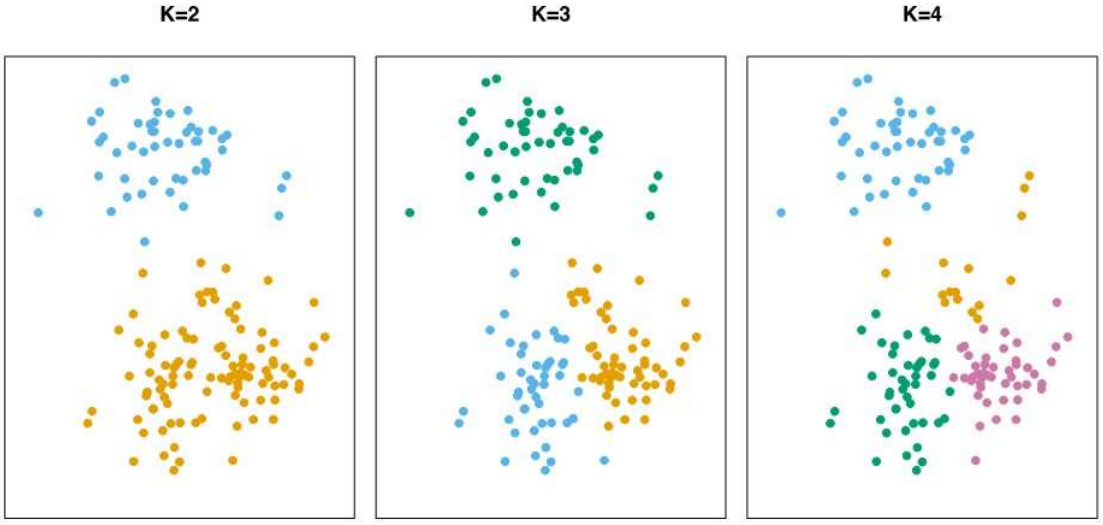
Clustering クラスター法

K-means法
階層的クラスタリング

k-means method (k平均法)

Concept:

- K個のクラスターにデータを分割したい)
- Kの値は事前にきめている
- 各データはK個のクラスターのどこか1つに含まれる
- アイデアとしては，クラスター内の違いがなるべく小さくなるような均質クラスターを作りたい



Algorithm

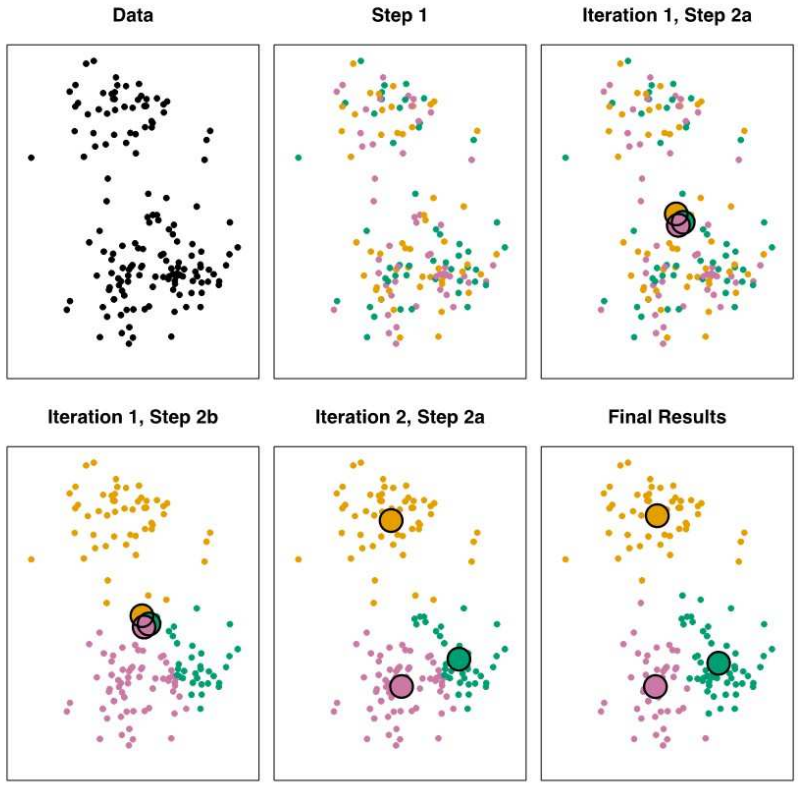
For finding a local minimum:

1) n個のデータをランダムに1からKのグループに割り当てる

2) 上記の仮のクラスター毎に，データの中心を計算する

3) ステップ2でK個の中心ができ，n個のデータを一番近い中心のクラスターに割り当てなおす

4) 中心が変動が小さくなり安定するまで，この2と3を繰り返す



Example data (Iris data アヤメデータ)

The following 3 difference species 次の3種のアヤメデータ (50本ずつ)

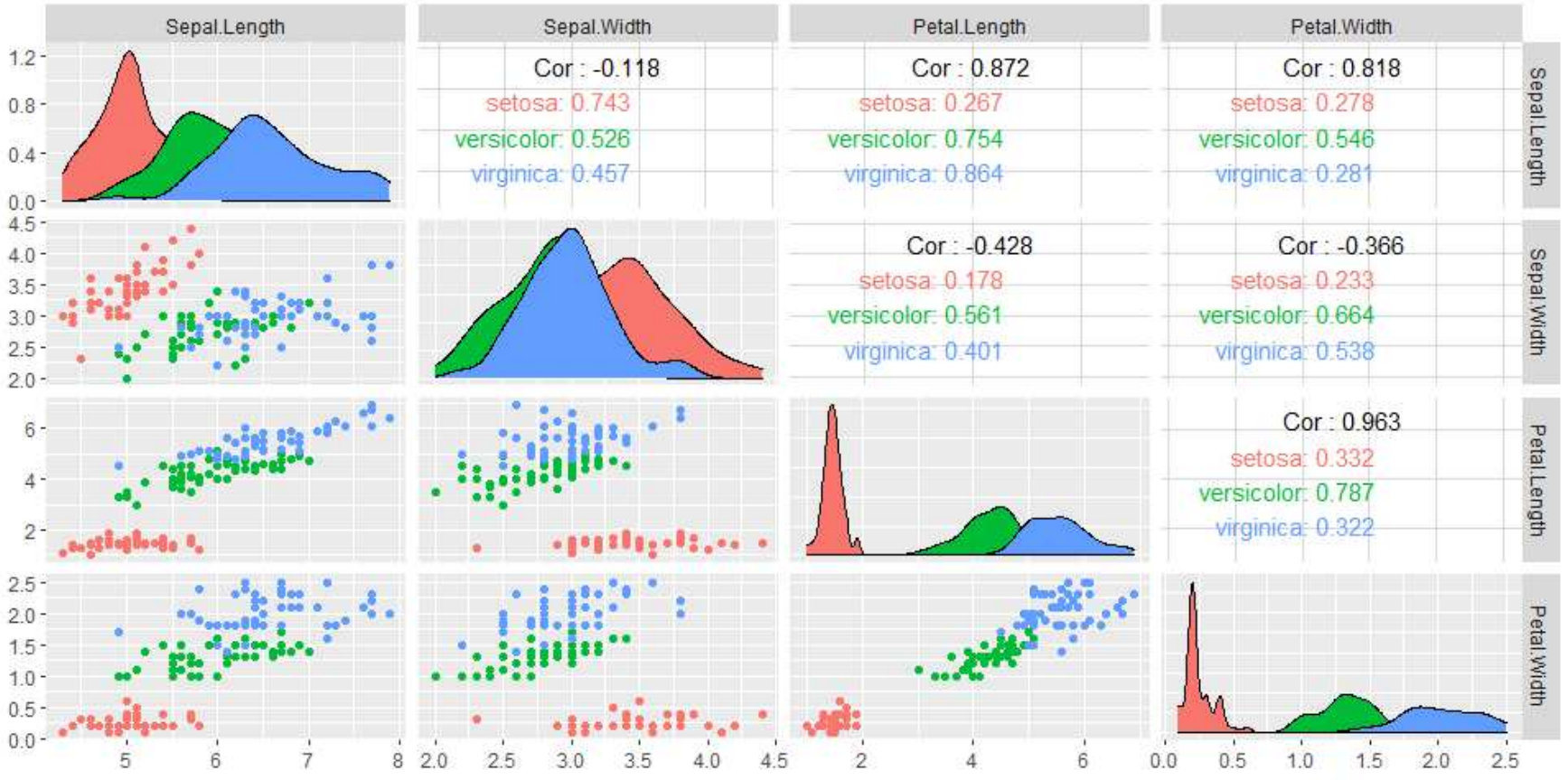
- Setosa
- Versicolor
- Virginica

Data (4 measurements 4次元の測定値, 3*50=150 individuals 150本の花)

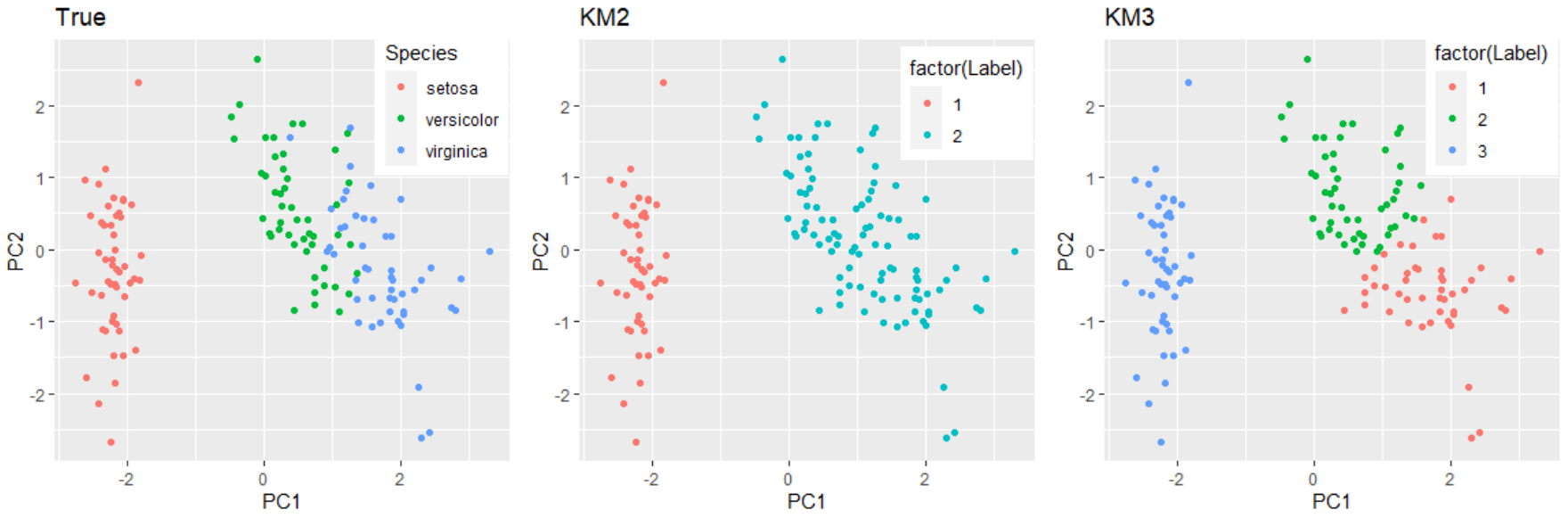
- Length and width of “Sepal” (がく片の長さと幅)
- Length and width of “Petal” (花弁の長さと幅)

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
149	6.2	3.4	5.4	2.3	virginica
150	5.9	3.0	5.1	1.8	virginica

Data visualization データ可視化

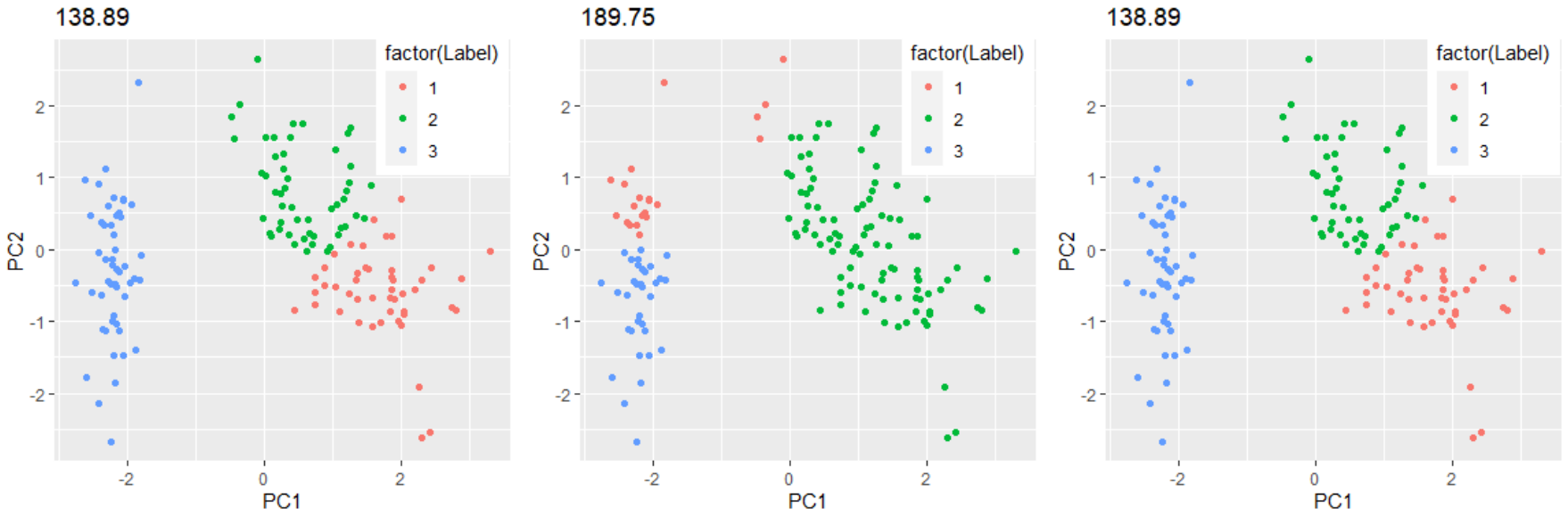


Clustering: K-means method with K=2 & 3 (10 repeats)



A total of 10 initial assignments were used. (10通りの初期割り当てで10回計算，最良の結果を選択)

Clustering: K-means method with K= 3 (only one initial assignment)

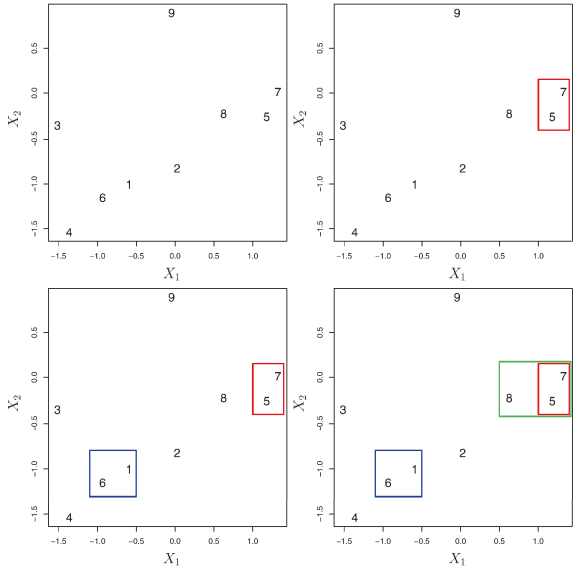


もし初期割り当てを1度しか行わないと，アリゴリズムが変な分割を提示してしまうこともあるので，前シートのように何度か繰り返すことが重要

Hierarchical clustering

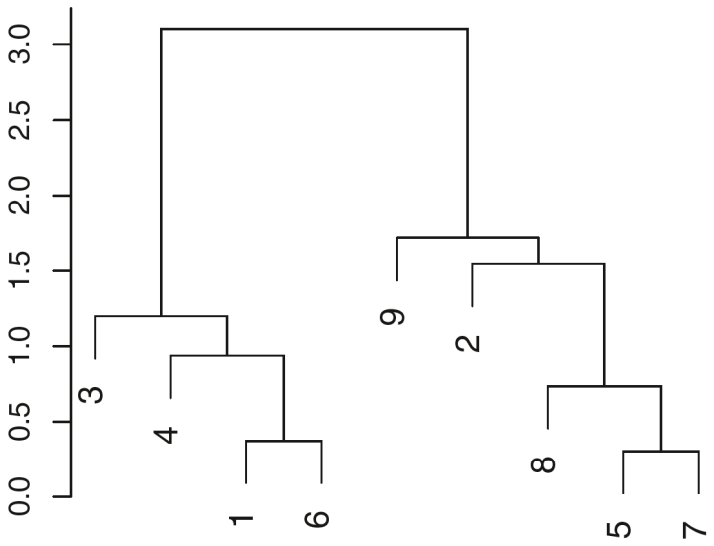
Concept:

- Bottom-up clustering
- 近いものを順番にグループ化していく



Sequence of grouping (from ISLR)

Hierarchical clustering



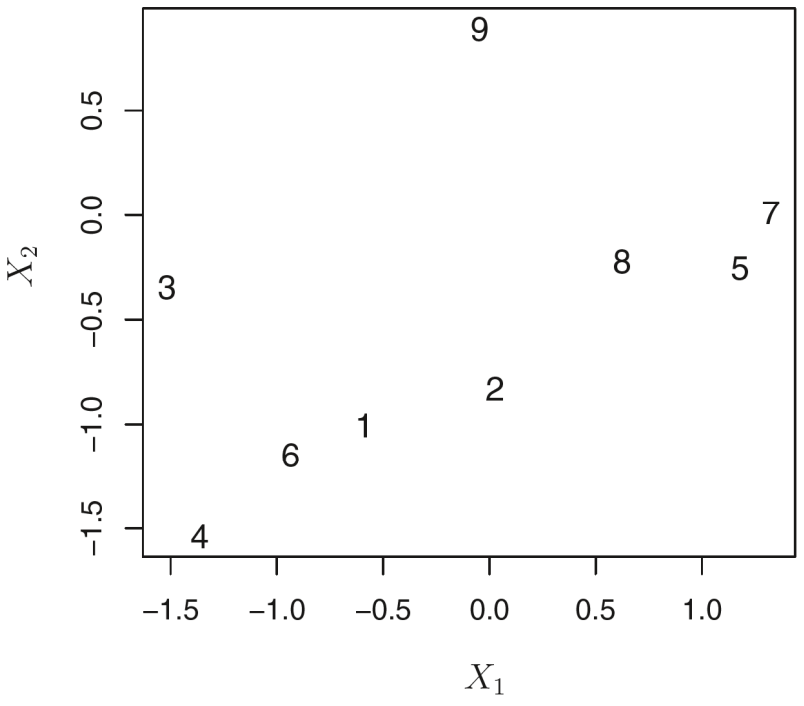
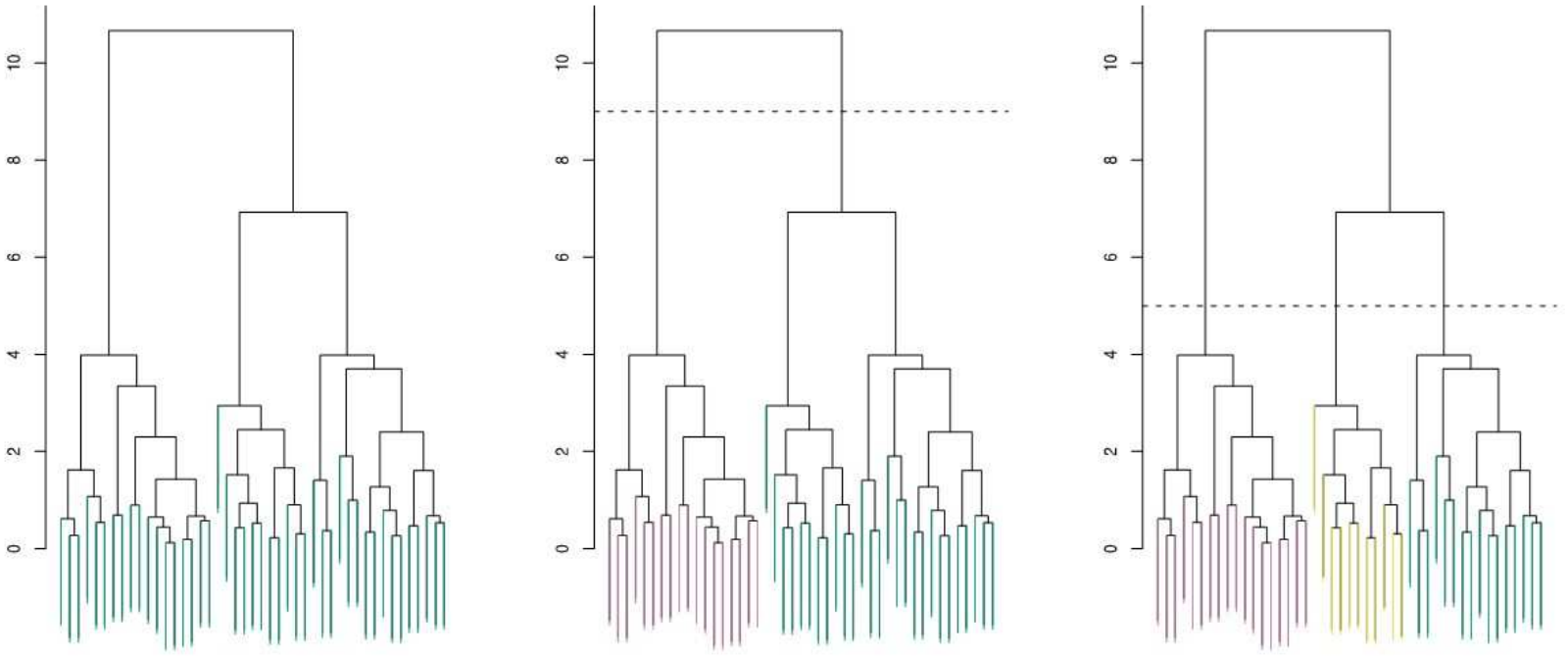


Image of dendrogram after hierarchical clustering (from ISLR)

Options

- Choice of similarity metric (距離の定義の仕方)
  - ユークリッド距離
  - 相関
- Choice of linkage （複数の点からなるクラスター間の距離の定義)
  - single (最短)
  - complete (最長)
  - average (平均)
  - centroid (中心)
  - …
- Choice of K （いくつのクラスターに分けるか？)

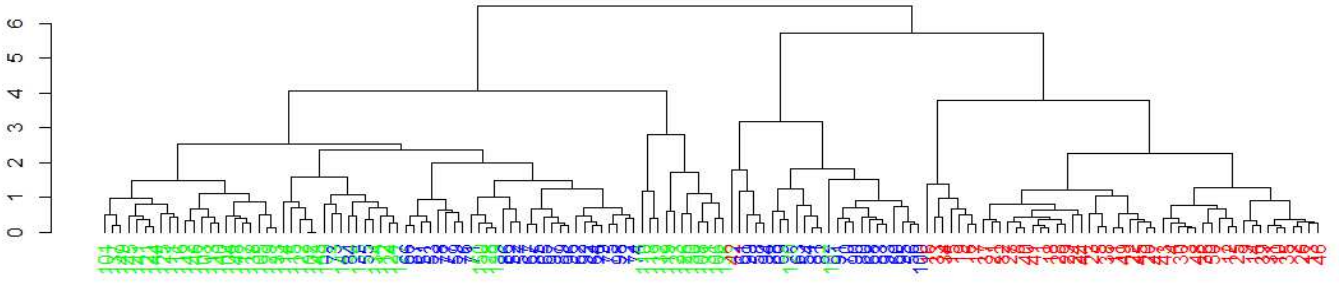
Outcomes as dendrogram (系統樹)



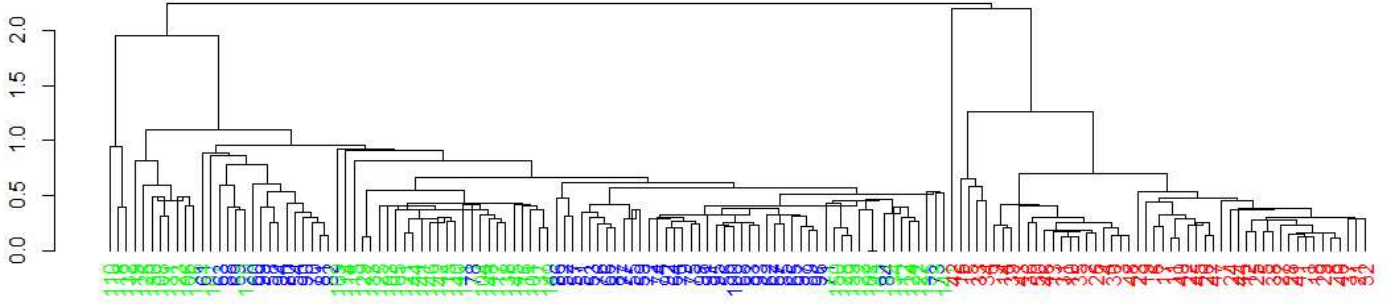
Grouping after hierarchical clustering (from ISLR)

Iris again: Hierarchical clustering with “average” linkage

“complete”



“centroid”



Iris again: Comparison of results