【データサイエンス基礎講座】:2014年12月度

データサイエンス基礎講座（超初級・実践編）
2014年11月26日～12月17日＜全5回＞演習用資料

R入門

インストールから活用まで

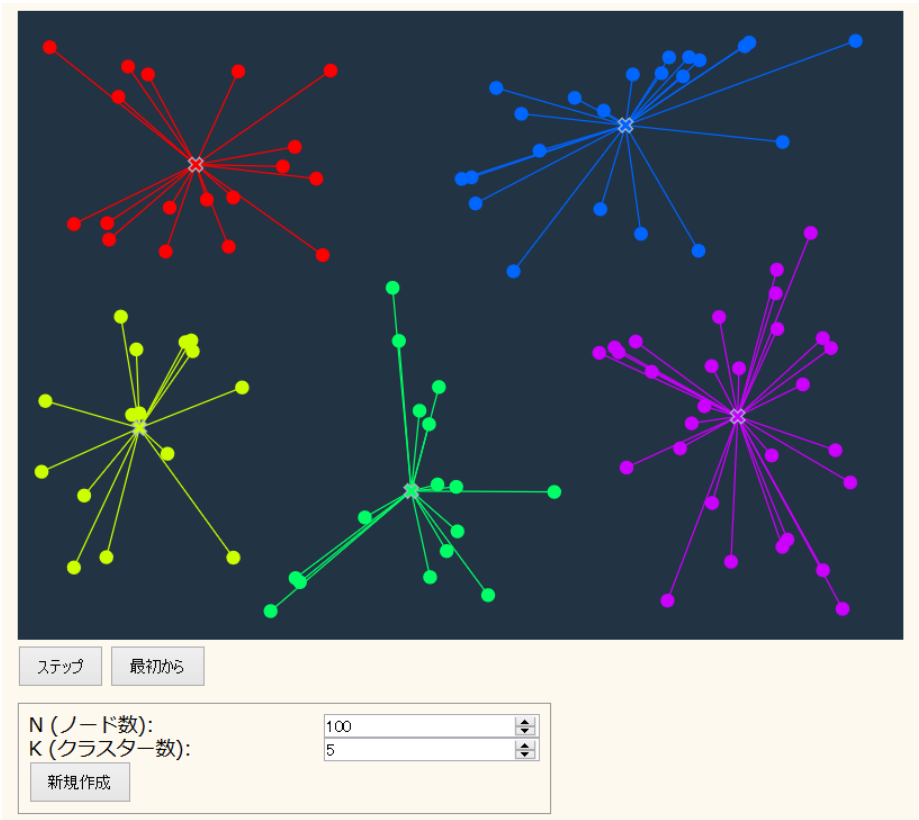
・主催:株式会社インプレス
・企画/製作:フューチャーブリッジパートナーズ株式会社

第5限目

- k平均法
- アンサンブル学習

k平均法

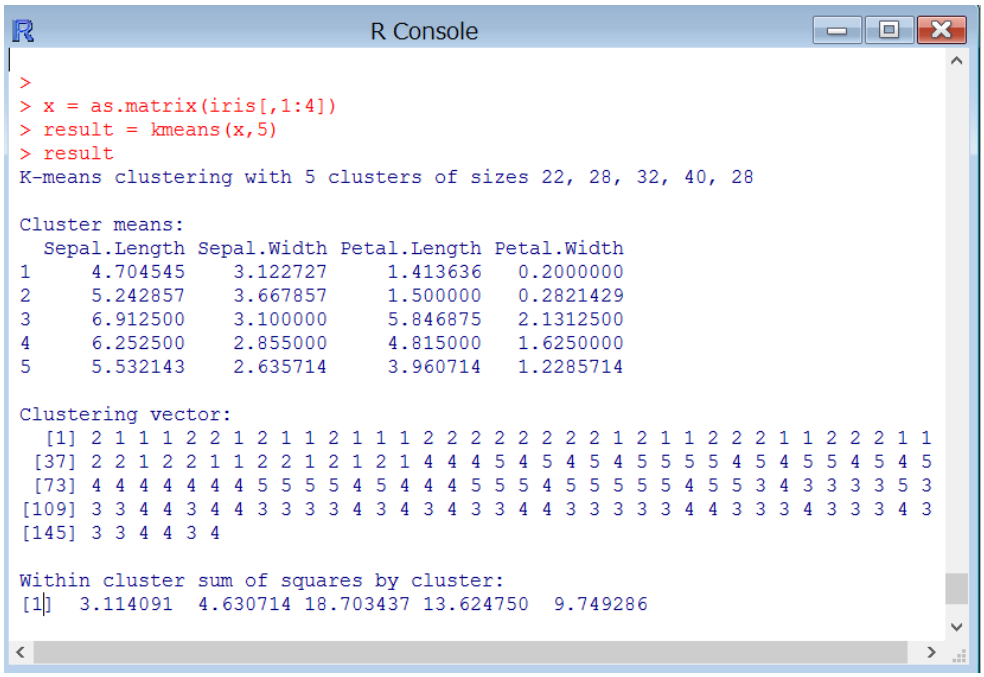
- k平均法とは？
  - 非階層クラスタリング
  - クラスタの数をkとしてクラスタをする方法
  - クラスタリング方法
    - 初期化：データをランダムにk個に分類し、クラスタの重心を求める
    - クラスタの決定：あるデータに対し、クラスタの重心の中で最も近いクラスタがデータの属するクラスタとする。
    - クラスタの中心の再計算：新しく属したクラスタについて重心を再計算し、収束するまで続ける。
    - ビジュアル化
    - http://tech.nitoyon.com/ja/blog/2013/11/07/k-means/



K平均法

- RによるK平均法
  - kmeans(x, k, nstart=5, iter.max=10, algorithm=c(“Hartigan-Wong”)
    - x → データセット
    - k → クラスタの数
    - nstart -> 初期値に試すデータの数
    - iter.max → 計算回数上限
    - algorithm → 計算アルゴリズム

- データのセット
  - x = as.matrix(iris[,1:4])
  - result = kmeans(x,5)
  - result
  - plot(x, col=result$cluster)



k平均法

- クラスタの数(k)をどのように決めるか？
  - カーネル主成分分析
    - 第4次限で取り上げたカーネル法を用いて、主成分分析
    - 成分をプロットし、それを目視して、クラスタ数を決める
  - ギャップ統計量
    - ギャップ統計量の最大値を、クラスタ数とする
    - 元データ(L_k)と同じ範囲から構成される乱数データ(L'_k)を用意する
    - ギャップ統計量(G_k)は、$\log(L'_k)/\log(L_k)$の最大値

ギャップ統計量Gk

$$G_k = \log \frac{L'_k}{L_k} = \log L'_k - \log L_k$$

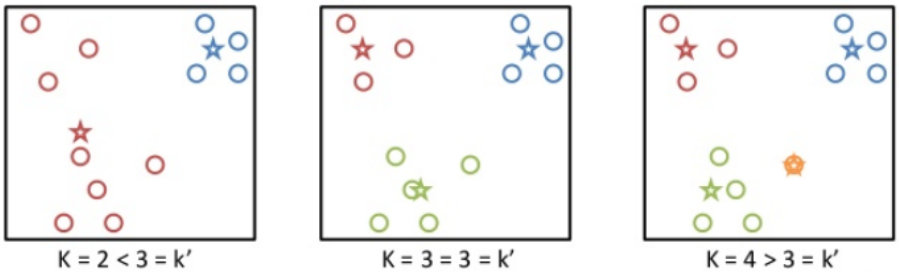
元のデータ

$$L_k = \sum_{i=1}^{n} \min_{l=1,\cdots,k} \| x_i - c_l \|$$

同一範囲内の一様乱数

$$L'_k = \sum_{i=1}^{n} \min_{l=1,\cdots,k} \| x_i - c_l \|$$

正しいクラスタ数K' = 3の場合、



K = 2 < 3 = k'　　K = 3 = 3 = k'　　K = 4 > 3 = k'

本来異なるクラスタを同一のクラスタから別のクラスタとして分けると、評価関数は大きく下がる！

同一のクラスタのデータをさらに分割しても評価関数は大きく下がらない。

k平均法

- ギャップ統計量
  - 標準でパッケージがインストールできないので、手動でRファイルを読み込み
  - 次のファイルを保存
    - http://dev.ecstaff.net/kmeansGap.R
    - http://dev.ecstaff.net/gapStat.R
  - 保存したRファイルを有効にする
    - ファイル　→　ディレクトリの変更　（Rファイルを保存したディレクトリ）
    - source(“gapStat.R”)
    - source(“kmenasGap.R”)
    - x = iris[,1:4]
    - kmeansGap(x)