クレジット:
UTokyo Online Educationデータマイニング入門 2018 森　純一郎



東京大学
THE UNIVERSITY OF TOKYO

CC BY NC ND

# データマイニング入門 第7回

2018年度

# 学習目標

クラスタリングの概念について理解する

特徴量ベクトルと距離・類似度について理解する

階層的クラスタリングについて理解する
- 併合の方法
- デンドログラム

k-meansクラスタリングについて理解する
- コスト関数

クラスタリングの実装について理解する

# クラスタリング

データ間の類似度（距離）に基づきデータをグループ化する

データの内的結合と外的分離としてのクラスタ

- 同じクラスタのデータは類似しているが異なるクラスタのデータは類似していないような性質

データ全体の中でグループごとの共通の性質を知ることができる

- ネットワーク分析で学んだコミュニティ抽出はノードのクラスタリングと考えられる

クラスタリングは与えられたデータのみからクラスを推定する「教師なし学習」

# 特徴量ベクトル

データ分析では、まず対象データを“よく表す”特徴を抽出する必要がある

基本的には人が”よい”と判断した特徴量を用いる
- テキスト
  - 単語と重み付けによるによる特徴量ベクトル
- 画像
  - ピクセル情報による特徴量ベクトル
- 信号
  - 振幅、周波数情報による特徴量ベクトル

各特徴量は数値化（離散、連続）され、対象データをそれらを組みにしたベクトルで表現する
- TFIDFベクトル

特徴量ベクトルによって張られる空間を特徴空間と呼び、対象データはこの特徴空間内に特徴量を次元として位置付けられる
- テキストのベクトル空間モデル

データ間の距離は特徴空間における特徴量ベクトルの距離で定量化できる
- コサイン類似度

# クラスタリング

n次元特徴空間上にm個のデータ$x^1,x^2,\ldots,x^m$が分布しており、各データはk個の異なるクラスのいずれかに所属しているものとする

ただし、観測できるのは各データの特徴空間上での位置のみであり、その所属クラスは観測できないものとする（教師なし学習）

同一クラスに属するデータは互いに類似していると考えられ、それらは特徴空間上で互いに近接し、クラスごとにまとまったグループとして観測されるはずである

このグループを**クラスタ**と呼び、特徴空間でのデータの分布状況からこのクラスタを見出す処理を**クラスタリング**という

# 類似度と距離尺度

## 距離の公理

- 非負：距離は負にならない
- 距離が0ならばAとBは同じ点、 AとBは同じ点なら距離は0
- 対称：AからBの距離tとBからAの距離は等しい
- 三角不等式：AからCを経由したBへ距離は、AからBへの距離以上

ユークリッド距離 $\Sigma_{i=1}^{n}(x_i - y_i)^2$

マンハッタン距離 $\Sigma_{i=1}^{n}|x_i - y_i|$

マハラノビス距離 $\sqrt{(\vec{x} - \vec{y})^T \Sigma^{-1} (\vec{x} - \vec{y})}$

# COS類似度とユークリッド距離

cos類似度は2つのベクトルの要素同士の相関を計算し、それを2つのベクトルのユークリッド空間における長さで割ることでそれぞれのベクトルを正規化した**正規化相関係数**とみなせる

正規化された（長さが1）のベクトル同士のcos類似度は：

$$cos\theta = \vec{x} \cdot \vec{y}$$

一方、これらのベクトル間のユークリッド距離の2乗は：

$$(|\vec{x} - \vec{y}|)^2 = \Sigma_{i=1}^{n} (x_i - y_i)^2 = 2(1 - \vec{x} \cdot \vec{y})$$

正規化されたベクトルにおいてはcos類似度による近さはユークリッド距離による近さと等しい

# 凝縮型階層的クラスタリング

データ一つが個々のクラスタの状態から、順次クラスタを併合しクラスタの階層を生成する

- 各データをそれぞれ1つのクラスタとして初期化する
- クラスタの併合を繰り返す（クラスタの併合ごとに全体のクラスタ数は1ずつ減っていく）
  - 最も近いクラスタを1つのクラスタとして併合する
  - * データ間の距離尺度を事前に決めておく
  - * 2つのクラスタ間の併合方法を事前に決めておく
    - 最短距離法, 最長距離法, 群平均法, Ward法 など
- 最終的に全体のデータが1つのクラスタとなったら終了

データの階層構造が得られる、クラスタ数を事前に決める必要がない

一方、適切な距離尺度や距離計算（クラスタ併合）方法を選択する必要がある

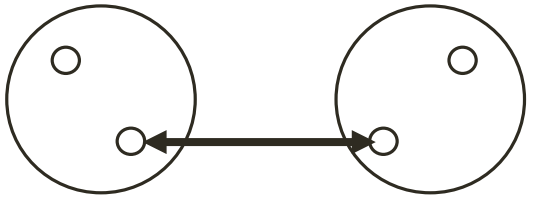
距離計算方法によっては計算コストがかかる $O(n^2)$

# 凝縮型階層的クラスタリング

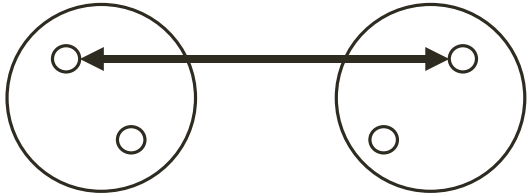
## 最短距離法

- クラスタ同士の要素間距離の最小値をクラスタ間の距離とする
  - クラスタの大きさが偏りやすい、外れ値に弱い



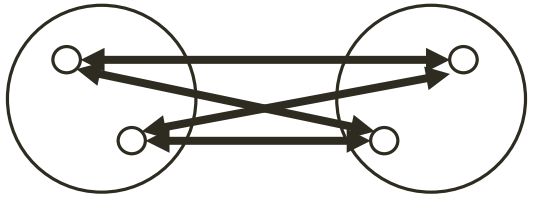
## 最長距離法

- クラスタ同士の要素間距離の最大値をクラスタ間の距離とする
  - クラスタの大きさが揃いやすいが外れ値に弱い



## 群平均法

- クラスタ同士の要素間距離の平均値をクラスタ間の距離とする
  - 外れ値に強く実用的



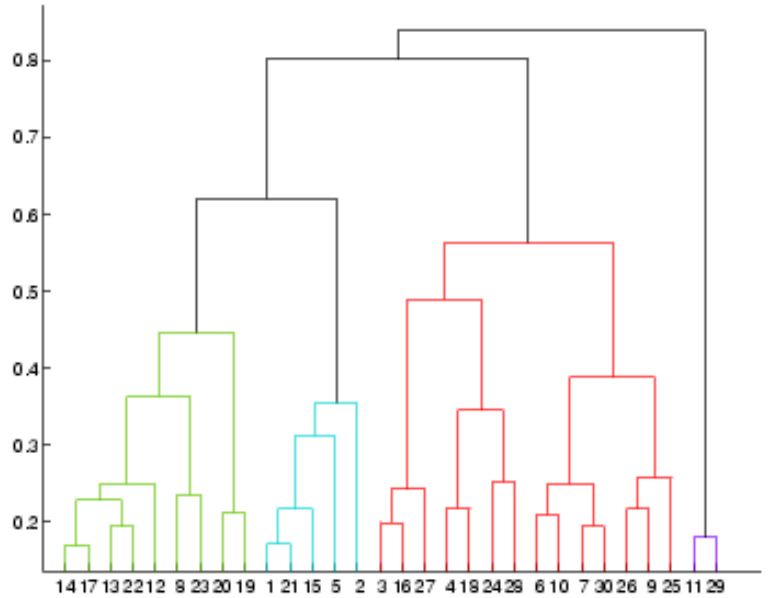
この他に各クラスタの中心同士の距離をクラスタ間の距離とする方法などもある

# 凝縮型階層的クラスタリング

デンドログラム

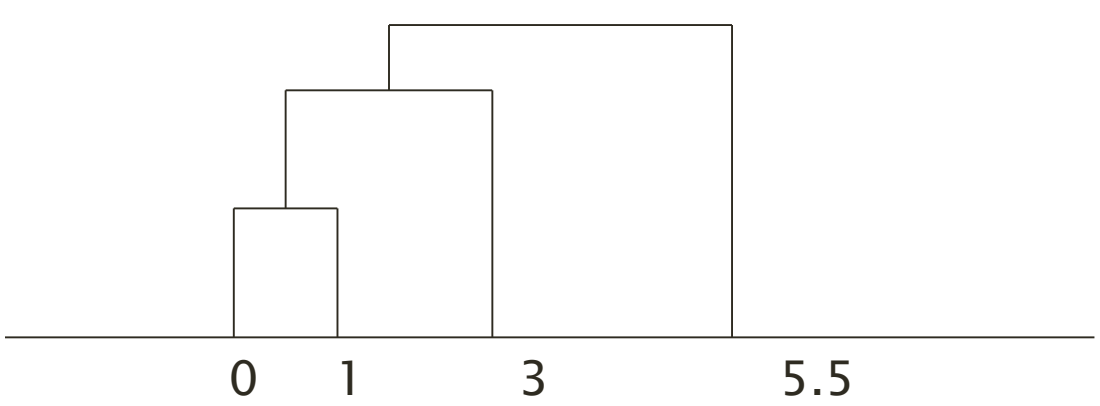
- クラスタリングの併合の過程を木として図示したもの
- 木の葉から併合までの高さは併合したクラスタ同士の類似度（距離）に対応
- 任意の距離水準でデンログラムの木を切断することで、その距離に応じたクラスタを決定可能



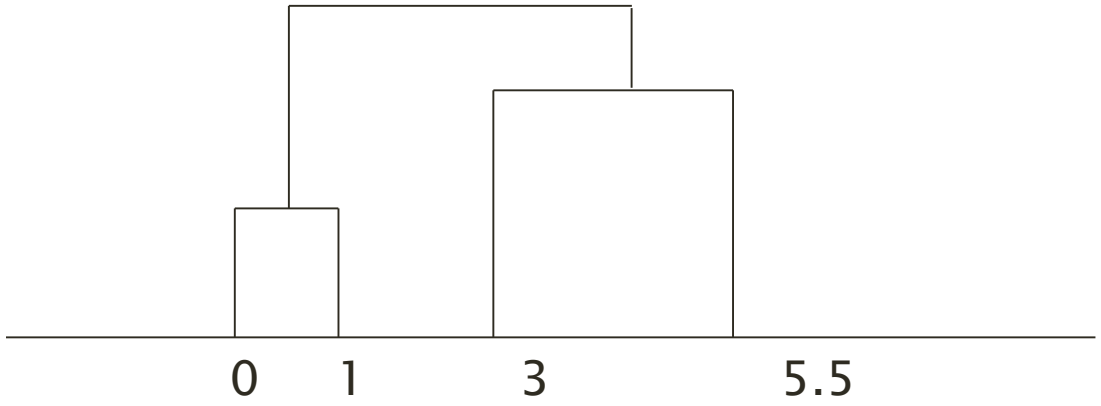

 

# 凝縮型階層的クラスタリング

1次元での階層化的クラスタリング




最短距離法の例



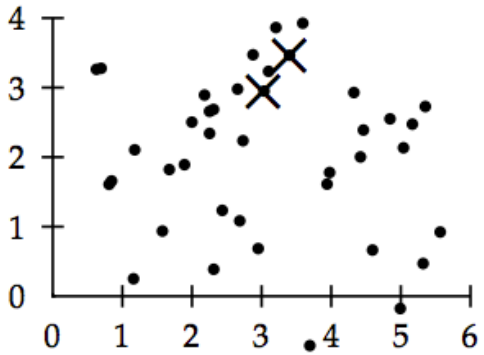

最長距離法の例

# K-MEANS法

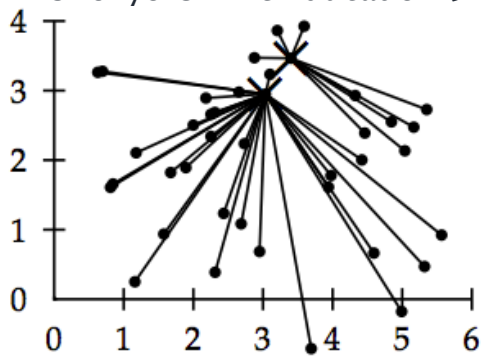
分割最適化に基づくクラスタリング

- クラスタの良さを表す関数を最適化する分割を求める
- クラスタの中心とそのクラスタに属するデータとの2乗ユークリッド距離の和が最小になるようなクラスタの分割を求める
  - クラスタ数とランダムな中心（クラスタ数と同数）を与える
    - クラスタの数だけデータからランダムに中心を選んでもよい
  - 以下を繰り返してコスト関数を小さくするようなクラスタの分割を見つける
    - データの中心への割り当て
      - 各データを最も近い中心に割り当て、クラスタ中心とする
    - 中心の更新
      - クラスタ中心に割り当てられたデータの平均でそのクラスタの中心として更新する
  - 終了の条件
    - 繰り返しが一定回数すぎた
    - 中心への割り当てが変化しない
    - コスト関数の値が閾値以下

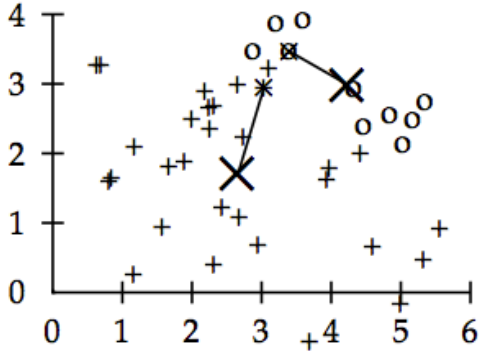
 



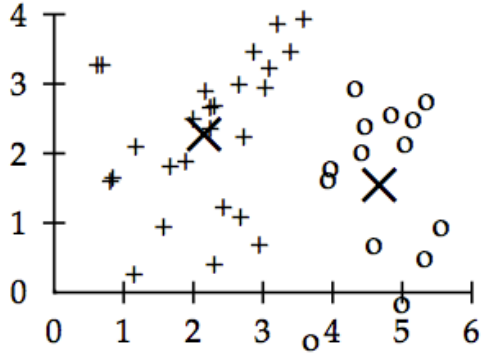

selection of seeds



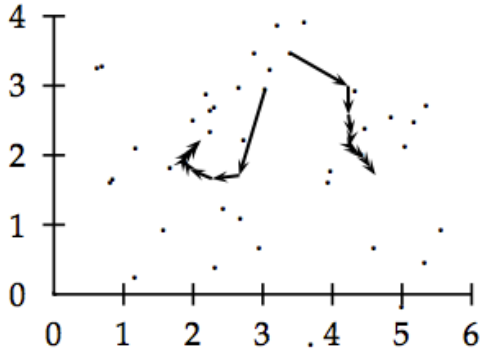

assignment of documents (iter. 1)




recomputation/movement of $\vec{\mu}$'s (iter. 1)




$\vec{\mu}$'s after convergence (iter. 9)

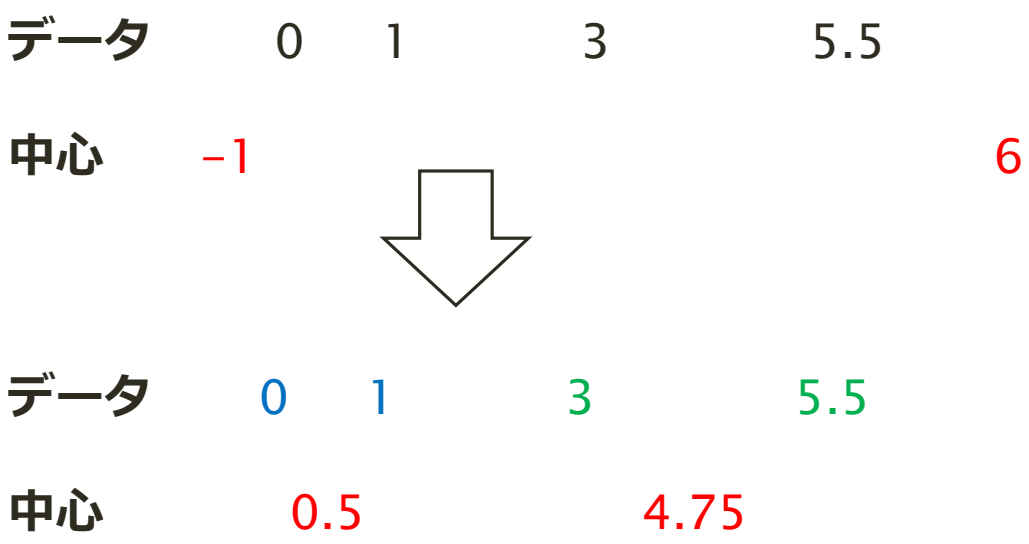



movement of $\vec{\mu}$'s in 9 iterations

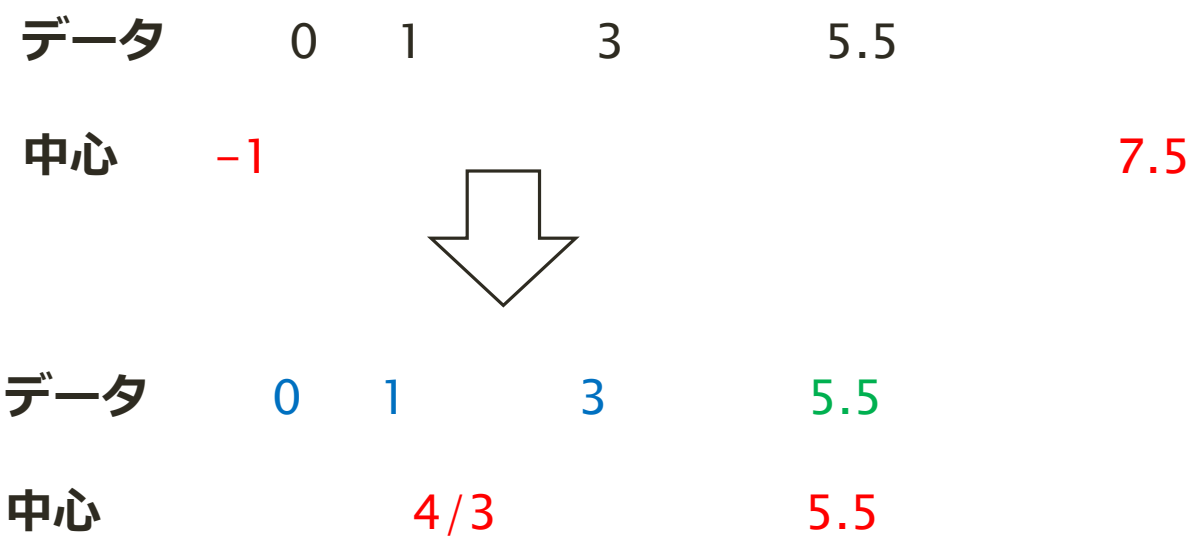
"Introduction to Information Retrieval",
Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze,
Cambridge University Press. 2008.
Fig.16.6: A K-means example for K=2 in $R^2$.
https://nlp.stanford.edu/IR-book/pdf/16flat.pdf
(ref. 8 Jan. 2019)

# K-MEANS法

例1




例2



# K-MEANS法

## K-means法のコスト関数（残差平方和）と最適化

$$J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K)=\Sigma_{i=1}^{m}\|x^{(i)}-\mu_{c^{(i)}}\|^2/m$$

*課題の実装では/mを省略

$x^{(i)}$: データ（$n$次元の特徴量ベクトル）

（データ数を$m$として、$i=1,\ldots,m$）

$\mu_k$: クラスタ中心（$n$次元のベクトル）

（クラスタ数を$K$として、$k=1,\ldots,K$）

$c^{(i)}$: データ$x^{(i)}$が所属するクラスタ（$1,\ldots,K$）

$\mu_{c^{(i)}}$: データ$x^{(i)}$が所属するクラスタの中心

K-means法では各繰り返しでコスト関数Jを
以下のように最適化していると考えられる

- 中心の割り当てではc(i)に関する最適化
  （Jを最小化するようなc(i)の推定）
- 中心の更新ではμkに関する最適化
  （Jを最小化するようなμkの推定）

繰り返しでJは単調に減少する

ユークリッド距離の2乗を用いるのはこの最適化を効率化するためとも考えられる

# K-MEANS法

K-means法は各データとそれが割り当てられるクラスタ中心の平方ユークリッド距離を最小化する最適化問題を解いている

K-means法で得られる解（クラスタリング結果）は局所最適解であり、クラスタリングの初期状態に大きく依存

大域的最適解を求めるには、複数の初期状態でクラスタリングを行い、その中から最も良い結果を選ぶ

- クラスタリングの良し悪しは、コスト関数の値で評価すればよい

# K-MEANS++

初期の中心を以下のように選ぶことによりK-means法におけるクラスタリングの局所最適解を回避する

- 任意の点を初期中心候補として最初に選ぶ
- その後はすでに選ばれた点からの平均距離が最も遠い点を別の初期中心候補として選ぶ
- 以上の初期中心候補選択をクラスタの数だけ繰り返し、最終的な初期中心を決定するこれを k 回繰り返す
  - 直感的にはお互いがなるべく離れている初期中心を選んでいる

この他にあらかじめデータに前処理として階層的クラスタリングを適用してなるべくお互いに離れているデータを初期中心として選ぶ方法もある

# クラスタ数の決定

## エルボー法

- クラスタ数を増やしていった時にコスト関数の値が最も減少した時（以降のコスト関数の減少がゆるやかになった時）のクラスタ数として選ぶ
  - より厳密には、クラスタ数を増やしていった時のクラスタ内のデータ間距離の減少を観察する
  - クラスタリング結果が大域的最適解であれば、クラスタ数を増やせばコスト関数は単調に減少する
  - この時、よいクラスタリング結果（各クラスタがよくまとまっており、またクラスタ同士は離れていれば）であれば、クラスタ内のデータは特徴空間において互いに距離が小さく、コスト関数は大きく減少する
- コスト関数がなだらかに減少し、急激に減少する”エルボー”がない時もある

著作権の都合によりここに挿入されていた画像を削除しました

“Tutorial: How to determine the optimal number of clusters for k-means clustering”, Tola Alade
Fig. Elbow Method for optimal k
https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f

## この他にAICなどの情報量を考慮したクラスタ数選択方法もある

# クラスタ数の決定

クラスタ数の決定は、クラスタリング結果をどのように使うかによるのでエルボー法などで決定されたクラスタ数を一概に最適なものとしてみなすことは難しい

目的に応じたクラスタリングの評価関数・尺度を定めることも必要

各データが所属するクラスタの正解ラベルがあれば

- 純度、正規化相互情報量、Randインデックス、F値

などでクラスタリングの結果を評価できる

- 教師あり学習の評価

# 復習：確率変数としてのデータ

データが生成される母集団は確率分布（確率密度関数）f(x)を持っている

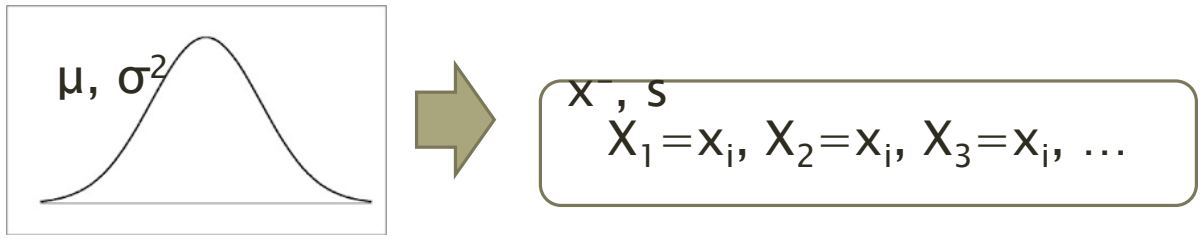
- f(x)は連続型でも離散型でもよい

母集団からデータ{$X_1$, $X_2$, …}をランダムに選ぶので、各データ$X_i$は母集団の確率分布f(x)に従う確率変数とみなせる

- データは同一の確率分布に従う独立な確率変数（独立同分布）

データから母集団の確率分布（パラメータ）が推測できれば、データの一般的な特徴を把握できる

- 母平均μ、母分散$\sigma^2$は代表的なパラメータ




母集団の確率分布 f(x)　　標本（データ）

# 復習：正規分布

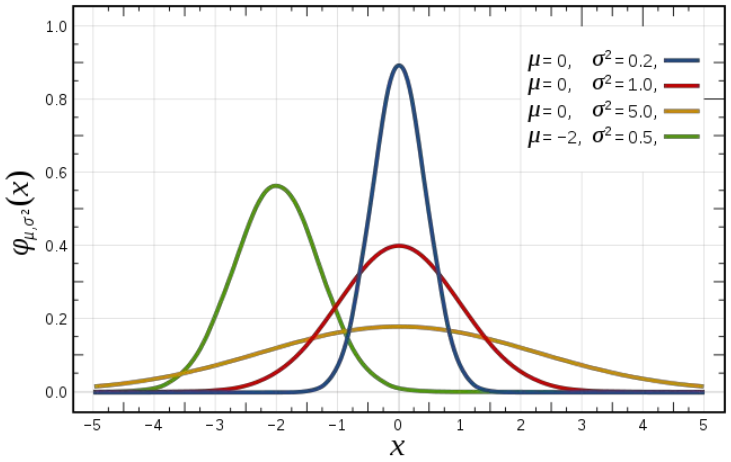
確率変数が正規分布に従うとき、その確率密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (x \in \mathbb{R})$$

正規分布に従う確率変数の平均（期待値）E(X)、分散V(X)は

- $E(X)=\mu$, $V(X)=\sigma^2$

正規分布は平均、分散の2つの母数で決まる




https://ja.wikipedia.org/wiki/%E6%AD%A3%E8%A6%8F%E5%88%86%E5%B8%83#/media/File:Normal_Distribution_PDF.svg

確率変数の平均（期待値）と分散

- 離散型
  - 平均
    - 確率変数のとりうる値とその確率の積

$$E(X) = \Sigma_i x_i P(X = x_i)$$

  - 分散

$$V(X) = \Sigma_i (x_i - E(X)) P(X = x_i)$$

- 連続型
  - 平均

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

  - 分散

$$V(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

# 復習：最尤法によるパラメータ推定

## 最尤法によるパラメータ推定

- 最尤原理
  - 現実の標本は確率最大のものが実現した
- 尤度関数
  - パラメータθの元で、観測したデータがどの程度起こりうるかを表す関数
    - 確率分布（確率密度関数）の積をパラメータθの関数とみなしたもの

$$L(\theta) = \Pi_{i=1}^{n} f(x_i|\theta)$$

  - 尤度関数を（パラメータがとる値の集合の空間で）最大にするものを観測したデータに対するパラメータの推定量とする
- 対数尤度
  - 尤度関数の対数をとって和の形にしたもの
  - 観測したデータについて対数尤度を最大にするパラメータの推定：最尤推定

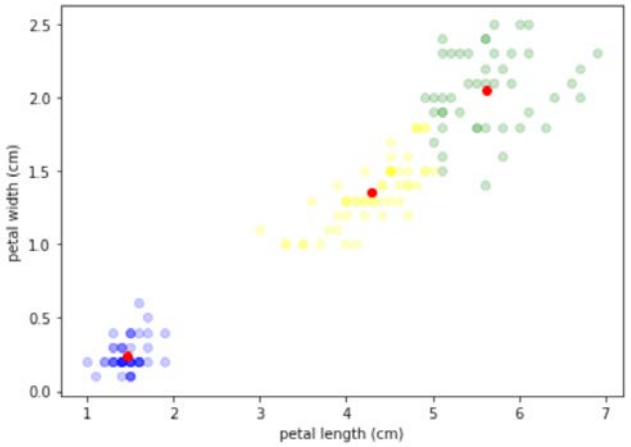
$$logL(\theta) = \Sigma_{i=1}^{n} f(x_i|\theta) \qquad \frac{\partial logL(\theta)}{\partial \theta} = 0$$

# 確率的なK-MEANS法

課題のk-means法の実装では、中心への割り当てのステップにおいて、データを最近接の中心に割り当てることでデータの所属クラスタを”clusters”として表した

“clusters”をデータがクラスタに所属するかするかしないかの1or0で表現すると、k-means法はデータがいずれかのクラスタに所属するハードクラスタリングとなっている

ここでデータが各クラスタに確率的に所属するソフトクラスタリングを考えてみる




$$\begin{matrix} & \text{clusters} \\ x^{(1)} & 1 \\ \cdots & \cdots \\ x^{(i)} & 3 \\ \cdots & \cdots \\ x^{(m)} & 2 \end{matrix} = \begin{matrix} 1 & 0 & 0 \\ \cdots & & \\ 0 & 0 & 1 \\ \cdots & & \\ 0 & 1 & 0 \end{matrix}$$

ハードクラスタリング

$$\begin{matrix} & w_1 & w_2 & w_3 \\ x^{(1)} & 0.8 & 0.1 & 0.1 \\ \cdots & \cdots & & \\ x^{(i)} & 0.2 & 0.1 & 0.7 \\ \cdots & \cdots & & \\ x^{(m)} & 0.2 & 0.7 & 0.1 \end{matrix}$$

ソフトクラスタリング

クラスタが$k$個あるとして、データ$x^{(i)}$が所属するクラスタを表す***潜在変数***$z^{(i)} = 1, 2, \ldots, k$を考える。

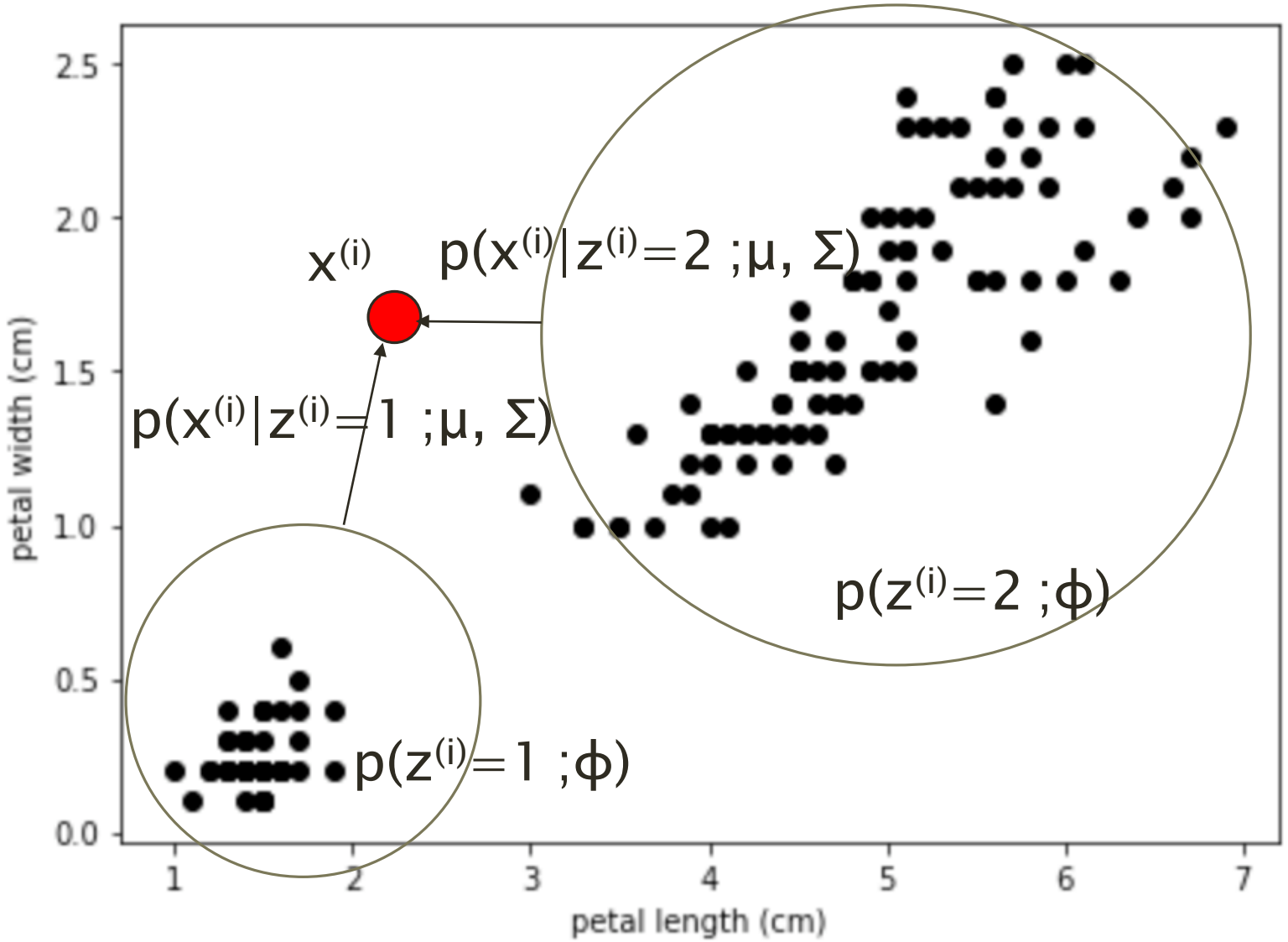
クラスタ$j$が選ばれる確率$p(z^{(i)} = j)$は多項分布に従い、多項分布のパラメータを$\phi_j (\Sigma_{j=1}^{k} \phi_j = 1)$とする。

クラスタ$j$からデータ$x^{(i)}$が生成される確率$p(x^{(i)}|z^{(i)} = j)$は正規分布に従い、クラスタ$j$に対する正規分布のパラメータを$\mu_j$, $\Sigma_j$とする（***混合ガウス分布***）。

クラスタ$z^{(i)}$を選びデータ$x^{(i)}$が生成される確率は$p(x^{(i)}|z^{(i)})p(z^{(i)})$なので、パラメータ$\phi, \mu, \Sigma$の元でデータ$x^{(i)}$が観測されうる対数尤度$L(\phi, \mu, \Sigma)$は以下のようになる。

$$L(\phi, \mu, \Sigma) = \Sigma_{i=1}^{m} \Sigma_{z^{(i)}=j}^{k} p(x^{(i)}|z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi)$$



# EMアルゴリズムとしてのK-MEANS法

潜在変数（クラスラベル）$z$が既知であれば、各$z = j$について尤度を最大にするパラメータを解析的に求められる。

$z$が未知の状態で観測されたデータ$x$について尤度を最大にするパラメータ$\phi, \mu, \Sigma$を推定するため、まずデータ$x^{(i)}$を観測した時、それがクラスタ$z^{(i)} = j$に所属する確率（$z^{(i)}$の事後確率）を計算する。

$$w_j^{(i)} = p(z^{(i)} = j | x^{(i)}; \phi, \mu, \Sigma) = \frac{p(x^{(i)} | z^{(i)}=j; \mu, \Sigma) p(z^{(i)}=j; \phi)}{\Sigma_{z^{(i)}=j}^{k} p(x^{(i)} | z^{(i)}; \mu, \Sigma) p(z^{(i)}; \phi)}$$

これは、データ$x^{(i)}$の各クラスタへの確率的な割り当てを推定しており、k-means法の中心への割り当てステップに対応している（***EMアルゴリズム***のE-step）

# EMアルゴリズムとしてのK-MEANS法

次に、各データの推定されたクラスタへの割り当てを元に、パラメータ（$\phi, \mu, \Sigma$）を更新する。

$$\phi_j = \frac{1}{m}\Sigma_{i=1}^{m} w_j^{(i)}$$

$$\mu_j = \frac{\Sigma_{i=1}^{m} w_j^{(i)} x^{(i)}}{\Sigma_{i=1}^{m} w_j^{(i)}}$$

$$\Sigma_j = \frac{\Sigma_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)((x^{(i)} - \mu_j)^T}{\Sigma_{i=1}^{m} w_j^{(i)}}$$

これは、仮に決められたデータのクラスタへの割り当てを元に各クラスタのパラメータを更新（最尤推定）しており、k-means法の中心の更新のステップに対応している（***EMアルゴリズム***のM-step）