機械学習 第１１回 モデル推定

立命館大学 情報理工学部

福森 隆寛

Beyond Borders

講義スケジュール

（第1～4回、第14回）　（第5～13回、第15回）

■ 担当教員：村上 陽平先生・福森 隆寛

1	機械学習とは、機械学習の分類		9	深層学習
2	機械学習の基本的な手順		10	アンサンブル学習
3	識別（1）		11	**モデル推定**
4	識別（2）		12	パターンマイニング
5	識別（3）		13	系列データの識別
6	回帰		14	強化学習
7	サポートベクトルマシン		15	半教師あり学習
8	ニューラルネットワーク			

■ 担当教員：叶 昕辰先生（第16回の講義を担当）

今回の講義内容

- 取り扱う問題の定義
- モデル推定
- クラスタリング
  - 階層クラスタリング
  - 分割最適化クラスタリング
    - k-meansアルゴリズム
- 異常検出
- 確率密度推定
  - EMアルゴリズム
- 演習問題

取り扱う問題の定義：教師なし･モデル推定

- **数値形式**の特徴ベクトルを入力して、その特徴ベクトルが生じるもとになったクラスを推定するモデルを考える

※ 教師なし学習の問題での学習データは、以下で構成される

入力データの特徴ベクトル（数値形式） ← $\{\boldsymbol{x}_i\},\ \ i = 1,2,\dots,N$ → 学習データの総数

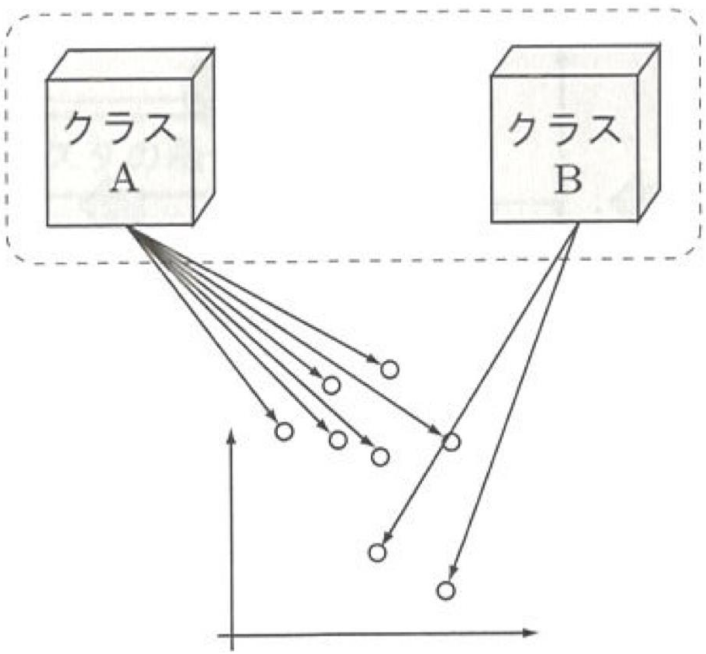
※ 正解情報は与えられていない

- 機械学習
  - 教師あり学習
  - 中間的学習
  - 教師なし学習
    - **モデル推定**
    - パターンマイニング

モデル推定

□ モデル推定

- 特徴ベクトルの要素が数値である場合に、
  その特徴ベクトルが生じるもとになったクラスを推定すること



モデル推定

特徴ベクトルの要素が数値ならば
特徴ベクトルをd次元空間上の
点として考えることができる

モデル推定
特徴空間上にあるデータのまとまりを
見つける問題

クラスタリング

□ クラスタリング

- 与えられたデータをまとまりに分ける操作
- 正確には、分類対象の集合を、**内的**(ないてき)**結合**と**外的**(がいてき)**分離**が達成されるような部分集合に分割すること
  - **内的結合**：一つのまとまりとして認められるデータは相互(そうご)の距離をなるべく近くする
  - **外的分離**：異なったまとまり間の距離は、なるべく遠くする

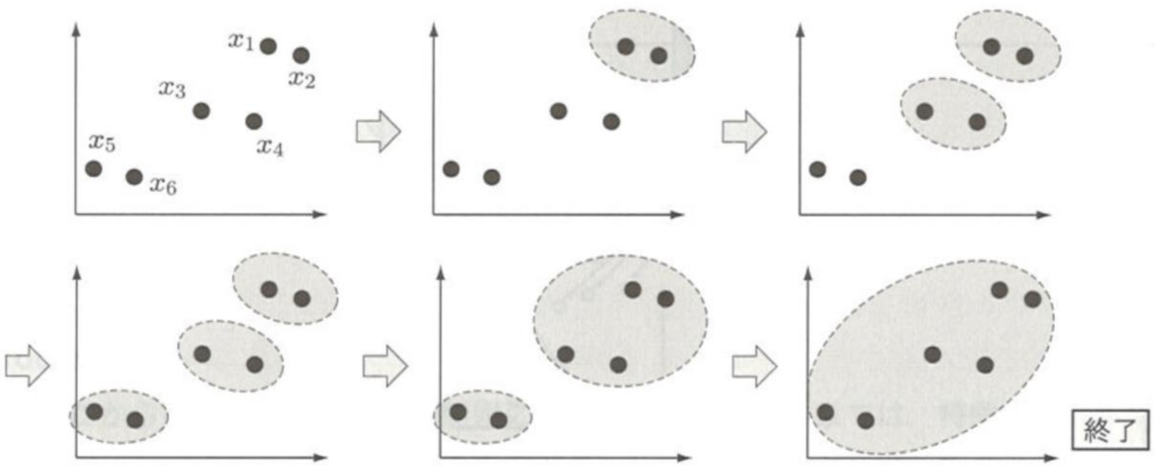
- クラスタリング手法の分類
  - 階層的手法：個々のデータをボトムアップ的にまとめる
  - 分割最適化手法：全体のデータをトップダウン的に分割する

クラスタリング：階層的クラスタリング

□ 階層的クラスタリング

- ■ 代表的な階層的手法
- ■ 近くのデータをまとめて小さいクラスを作り、
  - その小さいクラスタの近くのデータを取り込む
  - 小さいクラスタ同士をまとめて、少し大きめの新しいクラスを作る

のいずれかの手順を繰り返す



階層的クラスタリング

クラスタリング：階層的クラスタリング

□ 階層的クラスタリングのアルゴリズム

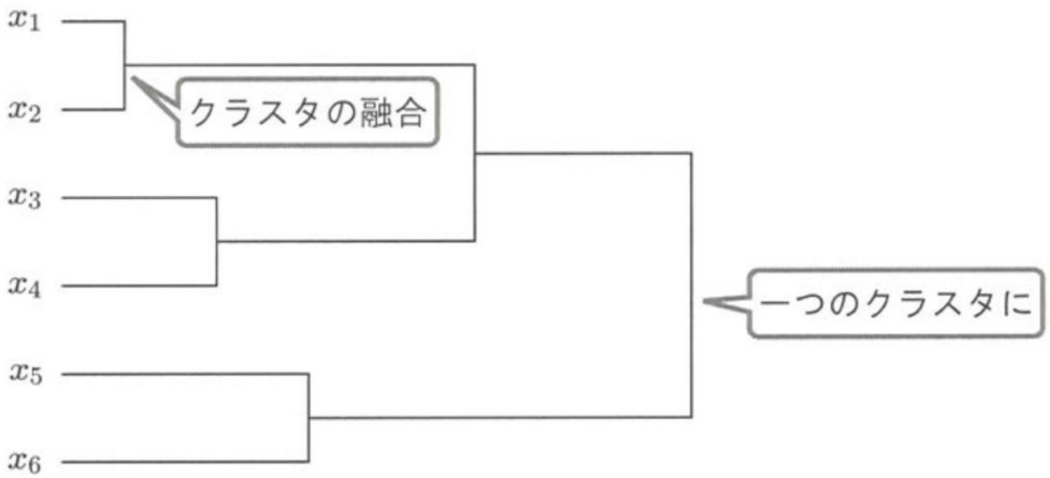
入力：正解なしデータ D
出力：クラスタリング結果の木構造

/* 学習データそれぞれをクラスタの要素としたクラスタ集合Cを作成 */
$C \leftarrow \{c_1, c_2, \ldots, c_N\}$
while $|C| > 1$ **do**
　/* 最も似ているクラスタ対（つい）$\{c_m, c_n\}$を見つける */
　$(c_m, c_n) \leftarrow \underset{c_i, c_j}{\operatorname{argmax}} \operatorname{sim}(c_i, c_j)$
　$\{c_m, c_n\}$を融合
end while

※ $\operatorname{sim}(c_i, c_j)$：クラスタ間の類似度を計算する関数

クラスタリング：階層的クラスタリング

- クラスタを融合(ゆうごう)する操作を木構造で記録
  - 全データNから始まって、1回の操作でクラスタが一つずつ減っていき、最後(さいご)は一つになる
  - この処理を途中(とちゅう)でやめることで、任意のクラスタ数からなるクラスタリング結果が得られる



階層的クラスタリング過程の木構造による表現

クラスタリング：階層的クラスタリング

◻ クラスタ間の類似度計算

- **単連結(たんれんけつ)法**：最も近いデータ対の距離を類似度とする
  - クラスタが一方向に伸(の)びやすくなる傾向
- **完全連結法**：最も遠いデータ対の距離を類似度とする
  - クラスタが一方向に伸びるのを避ける傾向
- **重心法**：クラスタの重心間の距離を類似度とする
  - クラスタが単連結と完全連結の間を取ったように伸びる
- **Ward法**
  - クラスタ融合(ゆうごう)の前後(ぜんご)にて「クラスタ内のデータと、クラスタ中心との距離の二乗和」を求め、融合後から融合前を引いたものが類似度
  - 階層的クラスタリングでよく用いられる類似度基準である

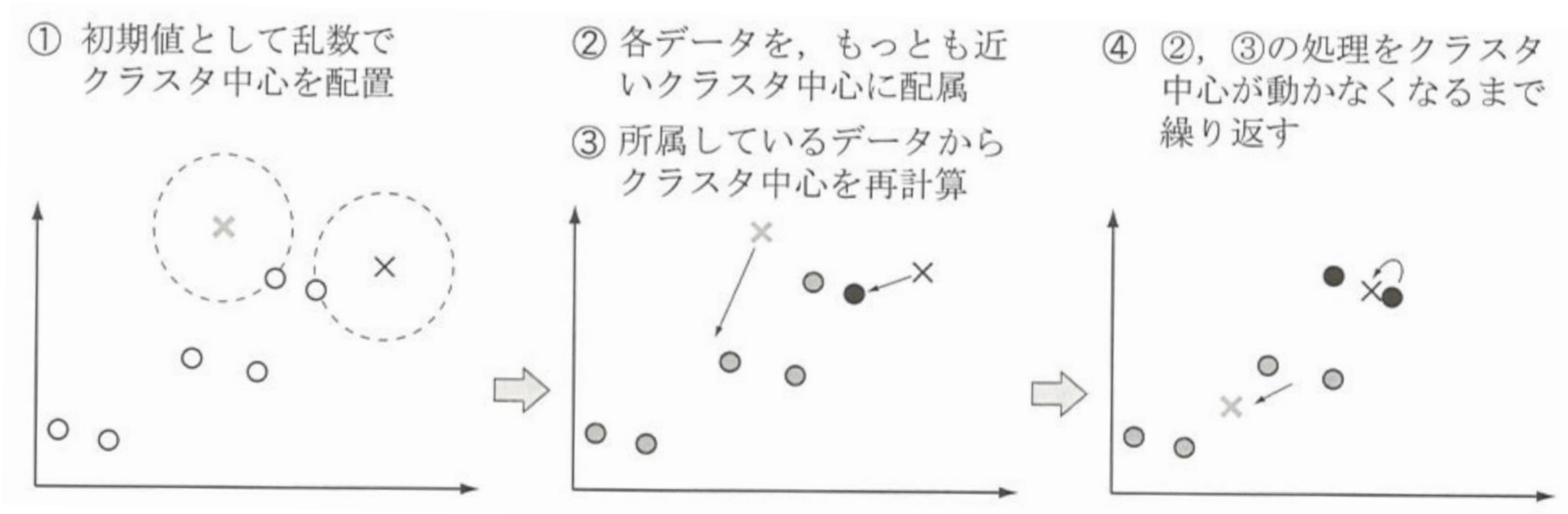
クラスタリング：分割最適化クラスタリング

□ 分割最適化クラスタリング

- 全体的な視点(してん)でまとまりの良いクラスタを求める方法
  - 階層的クラスタリングは、全体的な視点から見ると、いびつなクラスタを形成することがある
- データ分割の良さを評価する関数を定め、その評価関数の値を最適化することを目的とする
- データ数Nが大きいと、全ての分割を評価するのは不可能
  - データを分割する場合の数は、Nに対して指数的
  - 例えば、2つのクラスタに分ける場合： 2^N
- 一般的に探索によって準最適解を求めることを考える
  - その代表例が **k-meansクラスタリング**

クラスタリング：分割最適化クラスタリング

- k-meansクラスタリング（k-平均法）の手順
  1. 事前にクラスタ数kを与えて
各クラスタの平均ベクトルを乱数（らんすう）で生成
  2. 各データを最も近い平均ベクトルをもつクラスタに所属させる
  3. 各クラスタの平均ベクトルを再計算
  4. 全クラスタの平均ベクトルが動かなくなるまで繰り返す



クラスタリング：分割最適化クラスタリング

- k-meansクラスタリング
  - データ分割の良さの評価関数を
    「各データと所属するクラスタ中心との距離の総和」
    と定義すると…　　（クラスタの平均ベクトル）
    - クラスタ中心の位置更新によって評価関数の値が増えない
    - 再計算で中心ベクトルが動けば、あるデータが、より近くのクラスタに所属替えしたということによって距離の総和が減る
  - この方法によって、局所的最適解に辿り着くことができる
    - あくまで「局所的」なので、実際は異なった初期値で複数回学習を行い、評価値の最も良いものを結果として採用する

クラスタリング：分割最適化クラスタリング

k-meansアルゴリズム

入力：正解なしデータ D
出力：クラスタ中心 $\boldsymbol{u}_j\ (j = 1, \ldots, k)$

入力空間上にk個の点をランダムに設定し、それらをクラスタ中心$\boldsymbol{u}_j$とする
repeat
　　for all $\boldsymbol{x}_i \in D$ **do**
　　　　各クラスタ中心$\boldsymbol{u}_j$との距離を計算し、最も近いクラスタに割り当てる
　　end for
　　/* 各クラスタについて、以下の式で中心の位置を更新 */

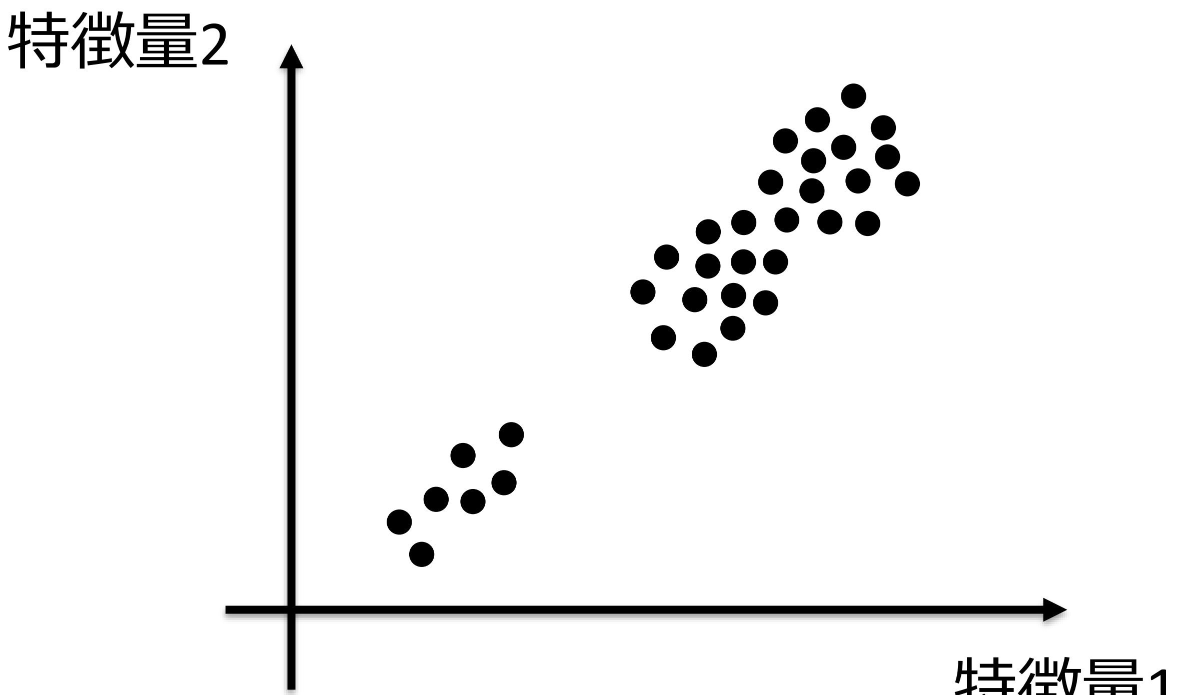
$$\boldsymbol{u}_j \leftarrow \frac{1}{N_j} \sum_{k=1}^{N_j} \boldsymbol{x}_k$$

（※ N_j ：クラスタjのデータ数）

until クラスタ中心$\boldsymbol{u}_j$が変化しない
return $\boldsymbol{u}_j\ (j = 1, \ldots, k)$

演習問題11-1（10分間）

- 以下の教師なしデータに対して、分割数が2の場合と3の場合のそれぞれで分割最適化クラスタリングを適用したとき、どのような結果が期待されるか？



クラスタリング：分割最適化クラスタリング

- k-meansアルゴリズムの問題点
  - 事前にクラスタ数kを決めないといけない
    - kが少ない：共通する性質を持たないクラスタができる
    - kが多い：各クラスタが個々のデータにマッチして
クラスタリングの目的に合わない結果となる

- **X-meansアルゴリズム**
  - クラスタ数を適応的に決定する方法
  - 最初は2分割から始まって、
得られたクラスタに対して分割が不適当と判断されるまで
k-means法によるクラスタリングを繰り返す

クラスタリング：分割最適化クラスタリング

- X-meansアルゴリズムにおける分割の判断基準
  - BIC (Bayesian Information Criterion)
    - BICが小さいと、「得られたクラスタリング結果がデータをよく説明している」かつ「詳細(しょうさい)になりすぎていない」ことを表現

$$\mathrm{BIC} = -2\log L + q\log N$$

- $\log L$：モデルの対数尤度
  - 各クラスタの正規分布を所属するデータから最尤推定し その分布から各データが出現する確率の対数値を得て それを全データについて足し合わせたもの
- q：モデルのパラメータ数（クラスタ数に比例(ひれい)）
- N：データ数

BIC以外にも、AIC (Akaike Information Criterion) のようなモデルの対数尤度とパラメータの複雑さのバランスをとった式なども用いられる

異常検出

□ 異常検出

- ■ 教師なし学習の実用（じつよう）的な応用例
- ■ 入力$\{\boldsymbol{x}_i\}$に含まれる異常値を、教師なし信号で見つける

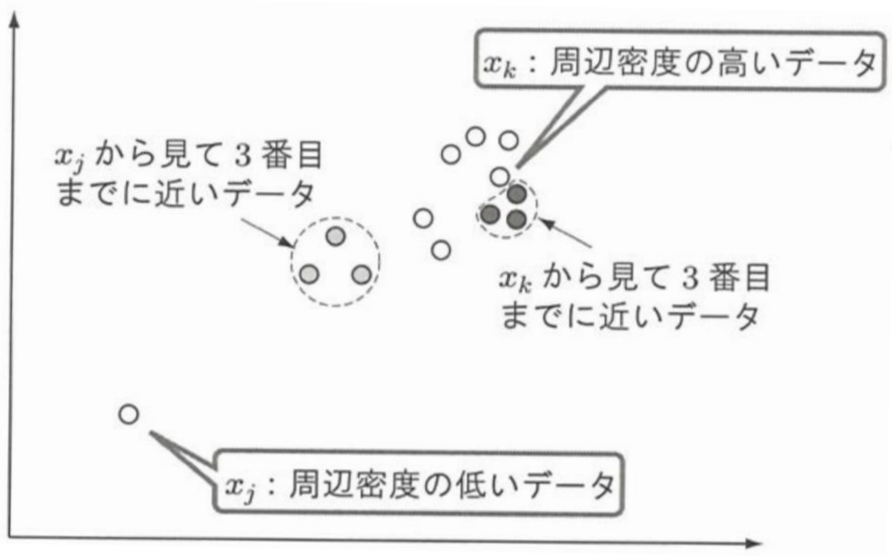
□ 外れ値（はずれち）の検出

- ■ 最も基礎（きそ）的な異常検出
- ■ 外れ値
  - 学習データに含まれるデータの中で、他と大きく異なるデータ
    - – 全体的なデータのまとまりから極端（きょくたん）に離れたデータ
    - – 正解付きデータの中で一つだけ他のクラスのデータに紛（まぎ）れ込（こ）んだデータ
  - 計測誤りや、教師信号付与作業上でのミスが原因（げんいん）で生じ（しょう）たと考えられるデータは、学習前に除去（じょきょ）しておくのが望ましい

異常検出：局所異常因子

□ 局所異常因子（いんし） (Local Outlier Factor; LOF)

- 近くにデータが無い（あるいは極端に少ない）ものを外れ値とみなす方法
  - 「近く」という概念を表現するために**周辺密度**（しゅうへん）を定義する
    - 周辺（k番目までに近いデータがある範囲）にあるデータまでの距離の平均
  - あるデータの周辺密度が、近くのk個のデータの周辺密度の平均と比べて極端に低いときに、そのデータを外れ値とみなす



周辺密度の考え方

異常検出：局所異常因子

- 局所異常因子LOFの算出手順
  1. データの個数に応じて、kを適当な値に定める
  2. あるデータ$\boldsymbol{x}$から、別のデータ$\boldsymbol{x}'$への
     到達可能距離 (Reachability Distance) を定義

$$RD_k(\boldsymbol{x}, \boldsymbol{x}') = \max\left(\left\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\right\|, \|\boldsymbol{x} - \boldsymbol{x}'\|\right)$$

- $\boldsymbol{x}^{(k)}$： $\boldsymbol{x}$にk番目に近いデータ
- $\boldsymbol{x}$と$\boldsymbol{x}'$が十分に遠い：通常の距離
- $\boldsymbol{x}'$が$\boldsymbol{x}^{(k)}$よりも$\boldsymbol{x}$に近ければ、$\left\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\right\|$に補正（ほせい）

異常検出：局所異常因子

局所異常因子LOFの算出手順（つづき）

3. **局所到達可能密度** (Local Reachability Density) を計算

到達可能距離を用いて定義された $\boldsymbol{x}$の周辺密度

$$LRD_k(\boldsymbol{x}) = \left\{\frac{1}{k}\sum_{i=1}^{k} RD_k\left(\boldsymbol{x}^{(i)}, \boldsymbol{x}\right)\right\}^{-1}$$

- $LRD_k(\boldsymbol{x})$は、$\boldsymbol{x}$からk番目までに近いデータとの到達可能距離の平均を求め、その逆数をとったもの
- k番目までのデータが近くにあるとき、
  - 到達可能距離の平均は小さい値
  - 局所到達可能密度（到達可能距離の平均の逆数）は大きい値

異常検出：局所異常因子

局所異常因子の算出手順（つづき）

4. 局所到達可能密度LRDを用いて、
以下のように$\boldsymbol{x}$の**局所異常因子**LOFを定義

$$LOF_k(\boldsymbol{x}) = \frac{\frac{1}{k}\sum_{i=1}^{k} LRD_k\left(\boldsymbol{x}^{(i)}\right)}{LRD_k(\boldsymbol{x})}$$

- $LOF_k(\boldsymbol{x})$は、$\boldsymbol{x}$に対してk番目までに近いデータの
局所到達可能密度の平均と、$\boldsymbol{x}$の局所到達可能密度の比
- $LOF_k(\boldsymbol{x})$が1に近い値 → $\boldsymbol{x}$が正常なデータ
  - k番目までに近いデータのLRDの平均と$\boldsymbol{x}$のLRDに大きな違いがない場合
- $LOF_k(\boldsymbol{x})$が大きな値 → $\boldsymbol{x}$が外れ値
  - k番目までに近いデータのLRDの平均よりも、$\boldsymbol{x}$のLRDが極端に低い場合

確率密度推定

- 教師なし学習で識別器を作ることを考える
  - ここまで説明してきたクラスタリングの結果を用いて新たなデータが観測されたときに、そのデータが属するクラスタを決める
- 事後確率が最大となる識別器を作る
  - 事後確率最大となる識別結果
    - 事前確率と尤度（各クラスごとの確率密度関数が観測されたデータを生成する確率）の積を最大とするクラス
  - 事前確率：クラスタリング結果のデータ数の分布から算出
  - 尤度：与えられた教師なし学習データから計算モデルを構築する方法を考える

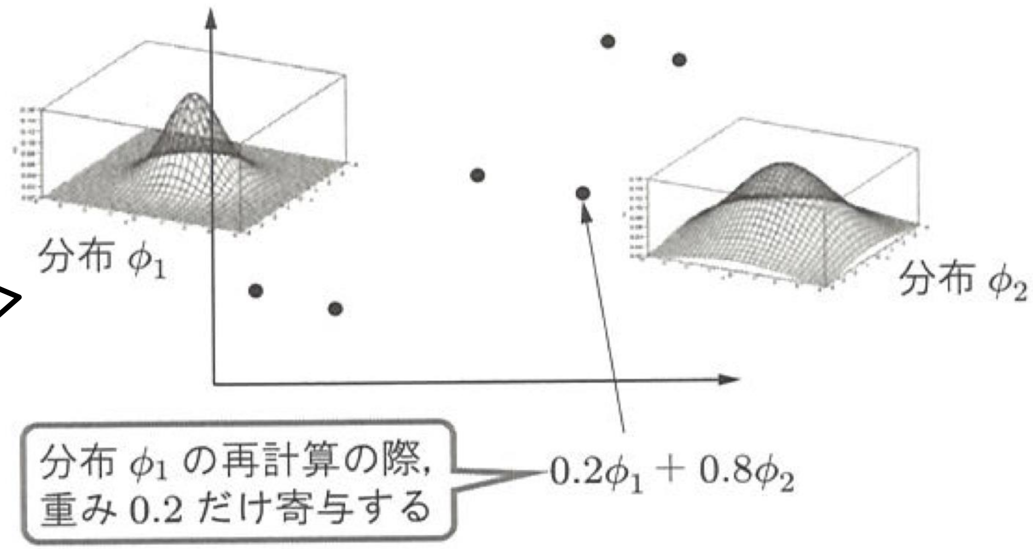
確率密度推定

- 各クラスタの確率分布の形を仮定して、そのパラメータを学習データから推定する問題を設定
  - 確率分布を正規分布と仮定とし、教師なし学習データから、クラスタc_mの平均$\boldsymbol{\mu}_m$と分散Σ_mを推定する問題にする

$$p(\boldsymbol{x}|c_m) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_m|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_m)^T\Sigma_m^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_m)\right\}$$

$$= \phi(\boldsymbol{x};\boldsymbol{\mu}_m,\Sigma_m)$$

データの所属クラスタを一意(いちい)に決めるのではなく、「クラスタ1に属する確率が0.2、クラスタ2に属する確率が0.8」といった表現（混合(こんごう)分布による表現）となる



確率密度推定

□ EMアルゴリズム（Expectation-Maximization）

- 個々の学習データに対して、EステップとMステップを順に繰り返して
パラメータの最尤推定量を得るアルゴリズム
- Eステップ
  - ある時点の分布から各データがそのクラスタに属する確率を計算

$$p\left(c_m|\boldsymbol{x}^{(i)}\right)=\frac{p(c_m)p\left(\boldsymbol{x}^{(i)}|c_m\right)}{p\left(\boldsymbol{x}^{(i)}\right)}=\frac{p(c_m)p\left(\boldsymbol{x}^{(i)}|c_m\right)}{\sum_{j=1}^{k}p\left(c_j\right)p\left(\boldsymbol{x}^{(i)}|c_j\right)}=\frac{p(c_m)\phi\left(\boldsymbol{x}^{(i)};\boldsymbol{\mu}_m,\Sigma_m\right)}{\sum_{j=1}^{k}p\left(c_j\right)\phi\left(\boldsymbol{x}^{(i)};\boldsymbol{\mu}_j,\Sigma_j\right)}$$

- Mステップ
  - Eステップでの確率をデータの重みとして分布のパラメータを再計算

$$\boldsymbol{\mu}_m=\frac{1}{|D|}\sum_{\boldsymbol{x}^{(i)}\in D}\boldsymbol{x}^{(i)}=\frac{\sum_{\boldsymbol{x}^{(i)}\in D}p\left(c_m|\boldsymbol{x}^{(i)}\right)\cdot\boldsymbol{x}^{(i)}}{\sum_{\boldsymbol{x}^{(i)}\in D}p\left(c_m|\boldsymbol{x}^{(i)}\right)}\qquad\Sigma_m=\frac{1}{|D|}\sum_{\boldsymbol{x}^{(i)}\in D}\left\{\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_m\right\}\left\{\boldsymbol{x}^{(i)}-\boldsymbol{\mu}_m\right\}^T$$

確率密度推定

- EMアルゴリズム

入力：正解なしデータ D
出力：各クラスを表す確率密度関数のパラメータ

入力空間上にk個の分布ϕ_jをランダムに設定

repeat

/* Eステップ */
for all 学習データ $\boldsymbol{x}^{(i)}$ **do**
$p\left(\boldsymbol{x}^{(i)}|c_j\right) = \phi_j\left(\boldsymbol{x}^{(i)}\right)$ $(j = 1, \ldots, k)$ を計算
end for

/* Mステップ */
Eステップの確率$p\left(\boldsymbol{x}^{(i)}|c_j\right)$を使って分布$\phi_j$のパラメータを再計算

until 分布のパラメータの変化量が閾値以下

演習問題11-2（10分間）

- k-meansアルゴリズムのクラスタリング結果を用いて新たなデータが観測されたときに、そのデータが属するクラスタを決める方法を考えなさい

- k-meansアルゴリズムのクラスタリング結果に基づく教師なし学習では、良い識別器を作ることができない理由を考えなさい