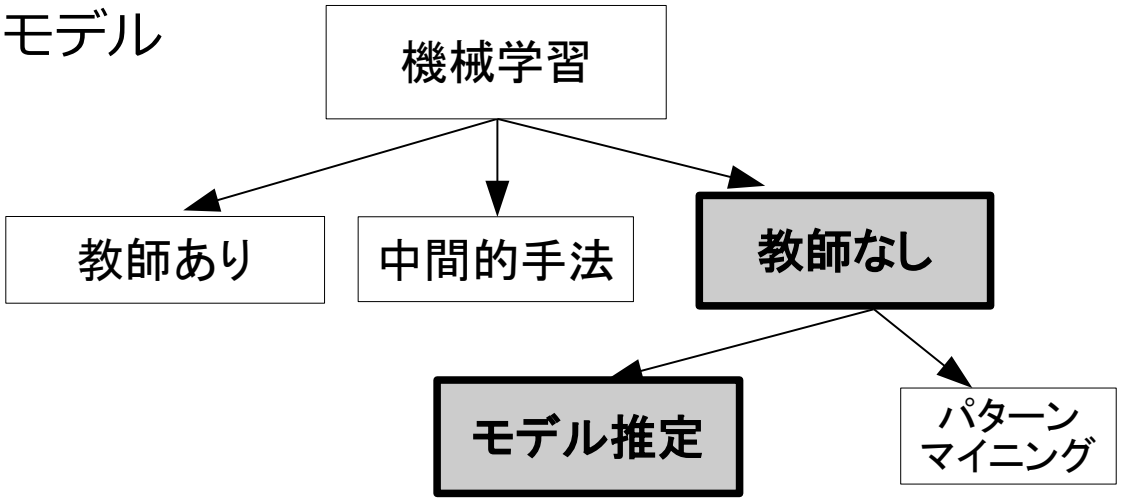
Section 3

- 教師なし学習(10,11章)

10. モデル推定

- 問題設定
  - 教師なし学習
  - 数値入力　→　クラスモデル
    - クラスモデルの例
      - クラスの分割結果
      - クラスの確率モデル

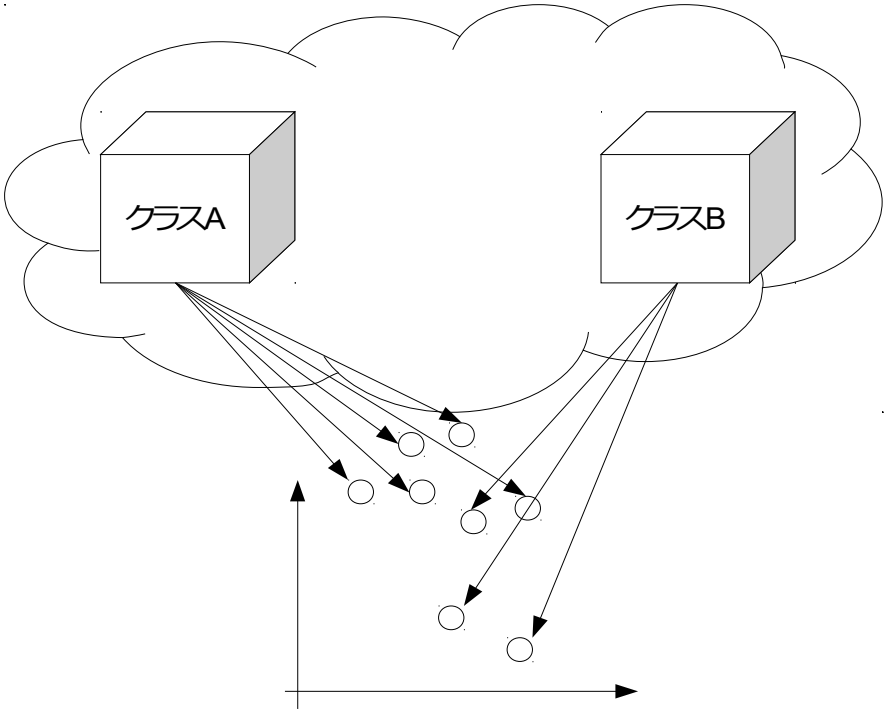


10.1 問題の定義

- 学習データ

$$\{\boldsymbol{x}^{(i)}\} \quad i = 1, .., N$$

- 問題設定
  - 特徴ベクトル $\boldsymbol{x}$ が生成された元のクラスの性質を推定する



10.2 クラスタリング

- クラスタリングとは
  - 対象のデータを、

    内的結合（同じ集合内のデータ間の距離は小さく）

    と

    外的分離（異なる集合間の距離は大きく）

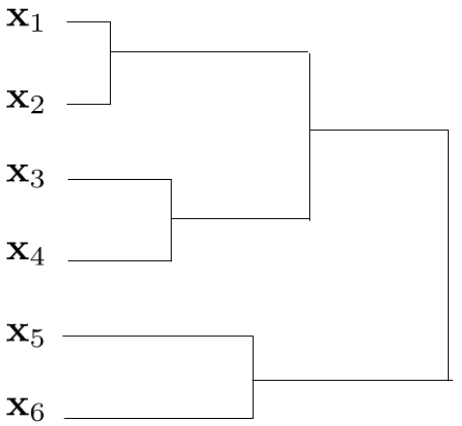
    が達成されるような部分集合に分割すること

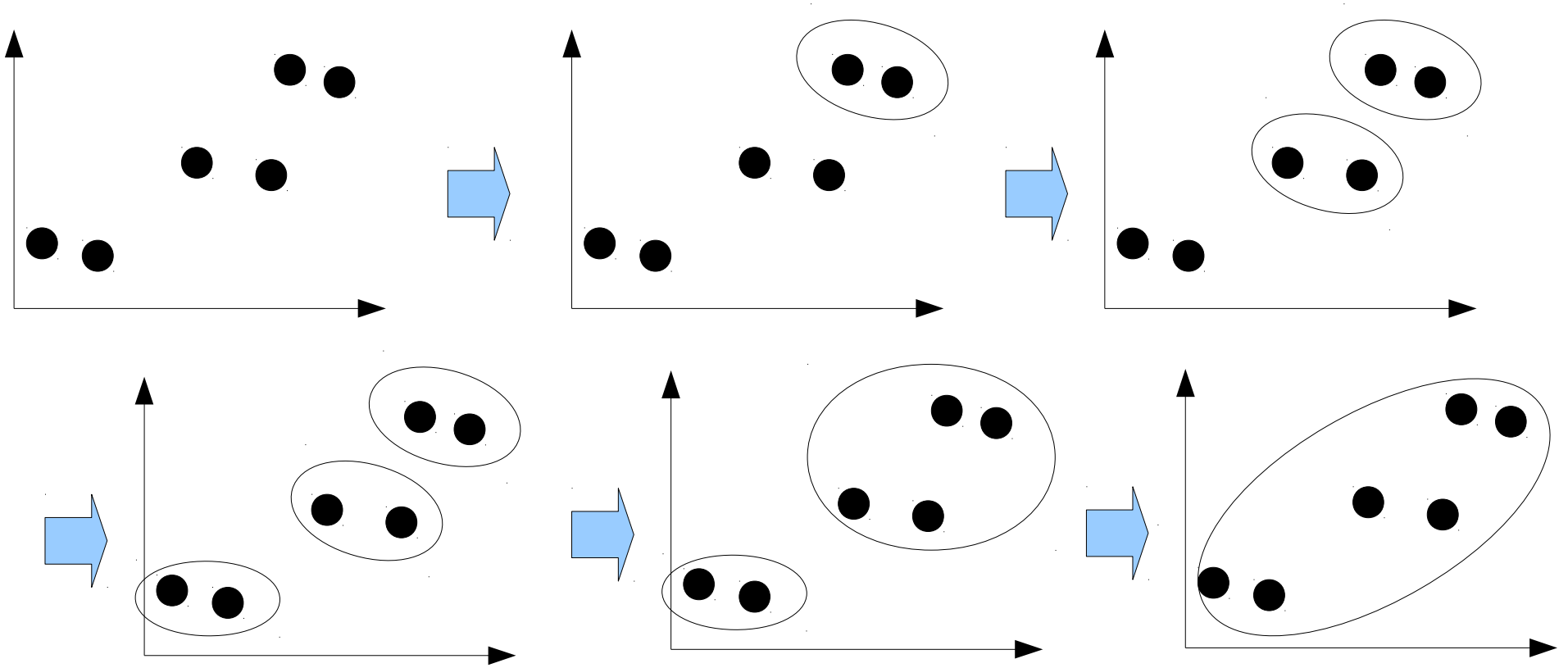
要するに塊を見つけること

- クラスタリング手法の分類
  - 階層的手法
    - ボトムアップ的にデータをまとめてゆく
  - 分割最適化手法
    - トップダウン的にデータ集合を分割してゆく

10.2.1 階層的クラスタリング

- 階層的クラスタリングとは
  1.1 データ 1 クラスタからスタート
  2.最も近接するクラスタをまとめる
  3.全データが 1 クラスタになれば終了





10.2.1 階層的クラスタリング

- 類似度 sim の定義
  - 単連結法
    - 最も近い事例対の距離を類似度とする。
    - クラスタが一方向に伸びやすくなる傾向がある。
  - 完全連結法
    - 最も遠い事例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がある。
  - 重心法
    - クラスタの重心間の距離を類似度とする。
    - クラスタの伸び方は、単連結と完全連結の間をとったようになる。
  - その他　群平均法、 Ward 法など

10.2.2 分割最適化クラスタリング ― k-means アルゴリズム―

- 分割最適化クラスタリングとは
  - 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
  - ただし、全ての可能な分割に対して評価値を求めることは、データ数 N が大きくなると、不可能
    - 2 分割で 2^N 通り
  - 探索によって、準最適解を求める

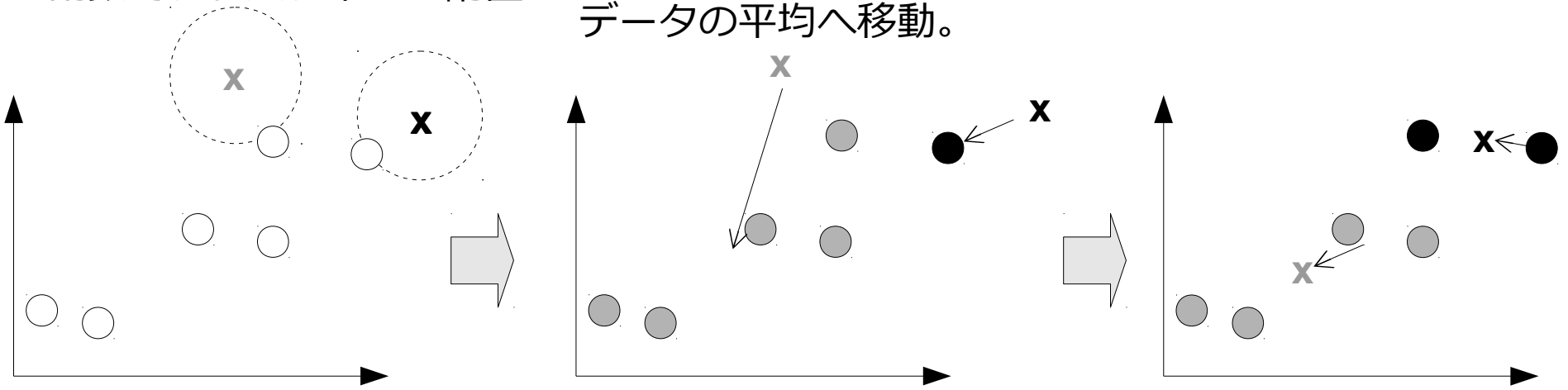
10.2.2 分割最適化クラスタリング ― k-means アルゴリズム―

- k-Means アルゴリズム
  1. 分割数 k を予め与える
  2. 乱数で k 個のクラスタ中心を設定し、逐次更新

k=2 とし、初期値として
乱数でクラスタ中心を配置

全データを近い方のクラスタ
中心に所属させる。そして、
クラスタ中心を所属している
データの平均へ移動。

左の処理を繰り返す。



10.2.3 自動分割最適化クラスタリング ― X-means アルゴリズム―

- k-means 法の問題点
  - 分割数 k を予め決めなければならない
- 解決法 ⇒ X-means アルゴリズム
  - 2 分割から始めて、分割数を適応的に決定する
  - 分割の妥当性の判断： BIC (Bayesian information criterion) が小さくなれば、分割を継続

$$BIC = -2\log L + q\log N$$

  - L: モデルの尤度
  - q: モデルのパラメータ数
  - N: データ数

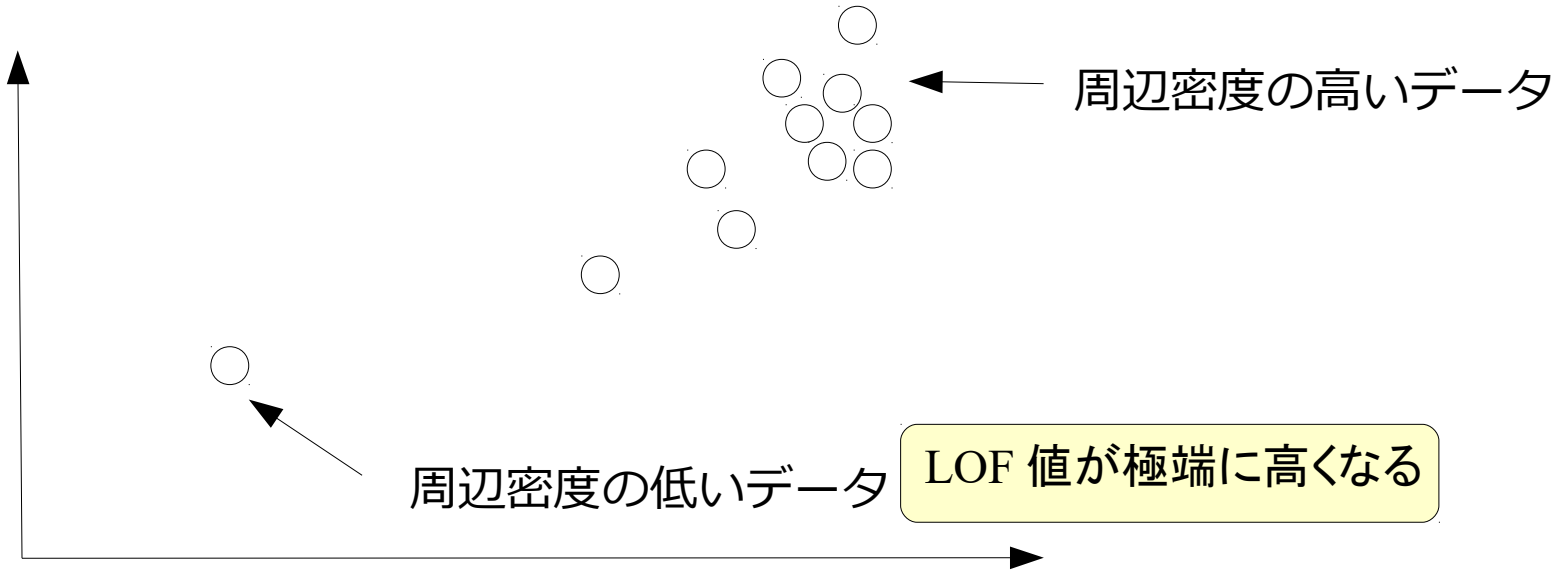
パラメータで表される
統計モデルの選択基準
(小さいほどよいモデル)

10.3 異常検出

- 異常検出とは
  - 正常クラスのデータと、それ以外のデータとのクラスタリング
  - 外れ値検知、変化点検出、異常状態検出など
  - 対象データが静的・動的で手法が異なる
- 外れ値検知（静的異常検出）
  - データの分布から大きく離れている値を見つける
  - 手法
    - 近くにデータがないか、あるいは極端に少ないものを外れ値とみなす
    - 「近く」の閾値を、予め決めておくことは難しい

10.3 異常検出

- 局所異常因子による外れ値検知
  - 周辺密度
    - あるデータの周辺の他のデータの集まり具合
  - 局所異常因子 (LOF: local outlier factor)
    - 近くの k 個のデータの周辺密度の平均と、あるデータの周辺密度との比



10.3 異常検出

- 局所異常因子の計算
  - 到達可能距離

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

$\boldsymbol{x}^{(k)}$ は、$\boldsymbol{x}$ に k 番目に近いデータ

近すぎる距離は、k 番目との距離に補正される

  - 局所到達可能密度

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

x の周りの密度が高い場合、大きな値になる

  - 局所異常因子

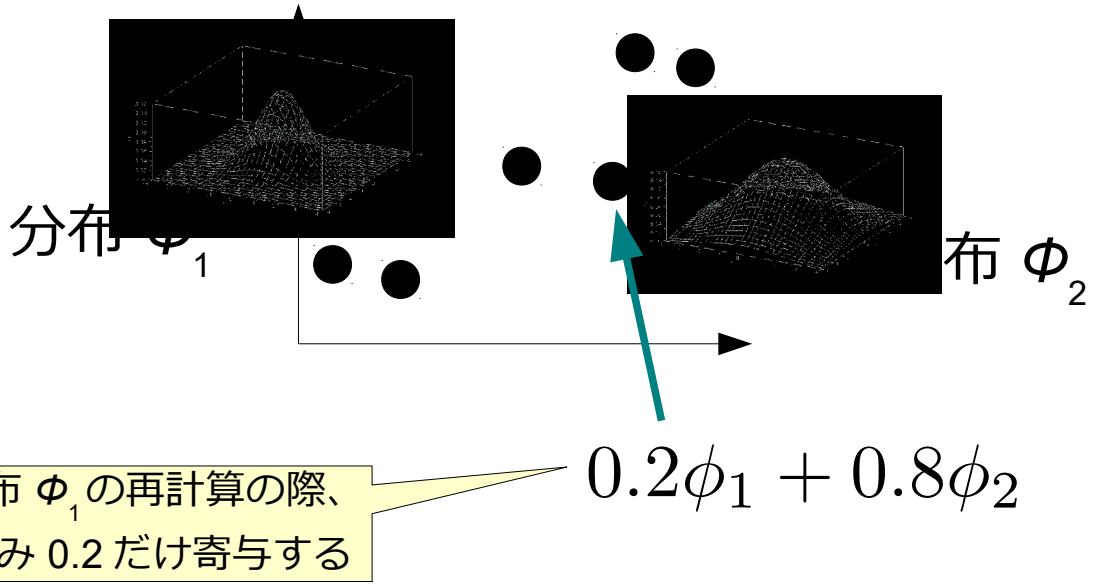
$$LOF_k(\boldsymbol{x}) = \frac{\frac{1}{k}\sum_{i=1}^{k} LRD_k(\boldsymbol{x}^{(i)})}{LRD_k(\boldsymbol{x})}$$

近くの k 個の密度の平均と自分の密度との比

10.4 確率密度推定

- 教師なし学習で識別器を作る問題
  - クラスタリング結果からは、 1 クラス 1 プロトタイプの単純な識別器しかできない
  - 各クラスの事前確率や確率密度関数も推定したい

    ⇨ EM アルゴリズム

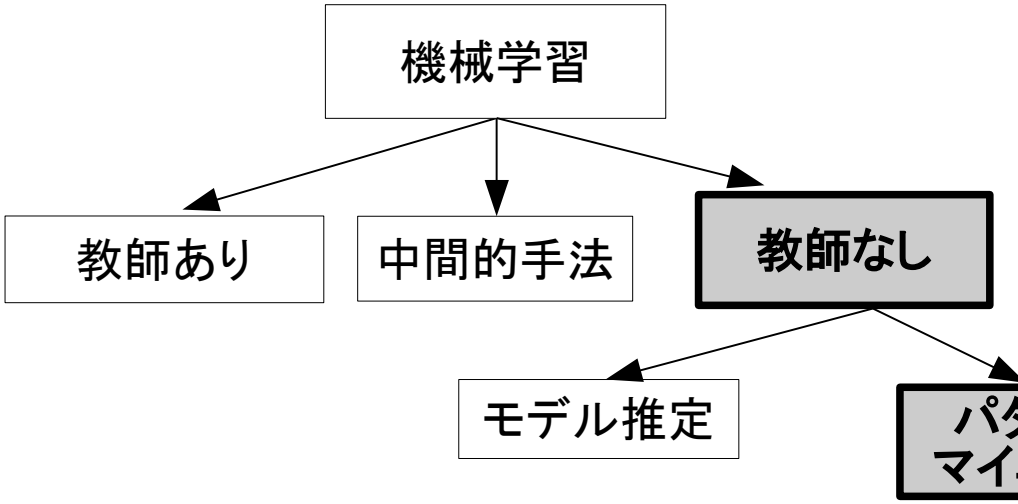


10.4 確率密度推定

- k-means 法の一般化
  - k 個の平均ベクトルを乱数で決める
    ⇒ k 個の正規分布を乱数で決める
  - 平均ベクトルとの距離を基準に、各データをいずれかのクラスタに所属させる
    ⇒各分布が各データを生成する確率を計算し、各クラスタにゆるやかに帰属させる
  - 所属させたデータをもとに平均ベクトルを再計算
    ⇒各データのクラスタへの帰属度に基づき各分布のパラメータ（平均値、共分散行列）を再計算

12 章 パターンマイニング

- パターンマイニングの問題設定
  - 入力：カテゴリ特徴の教師なしデータ
  - 出力：頻出項目、連想規則、未観測データ



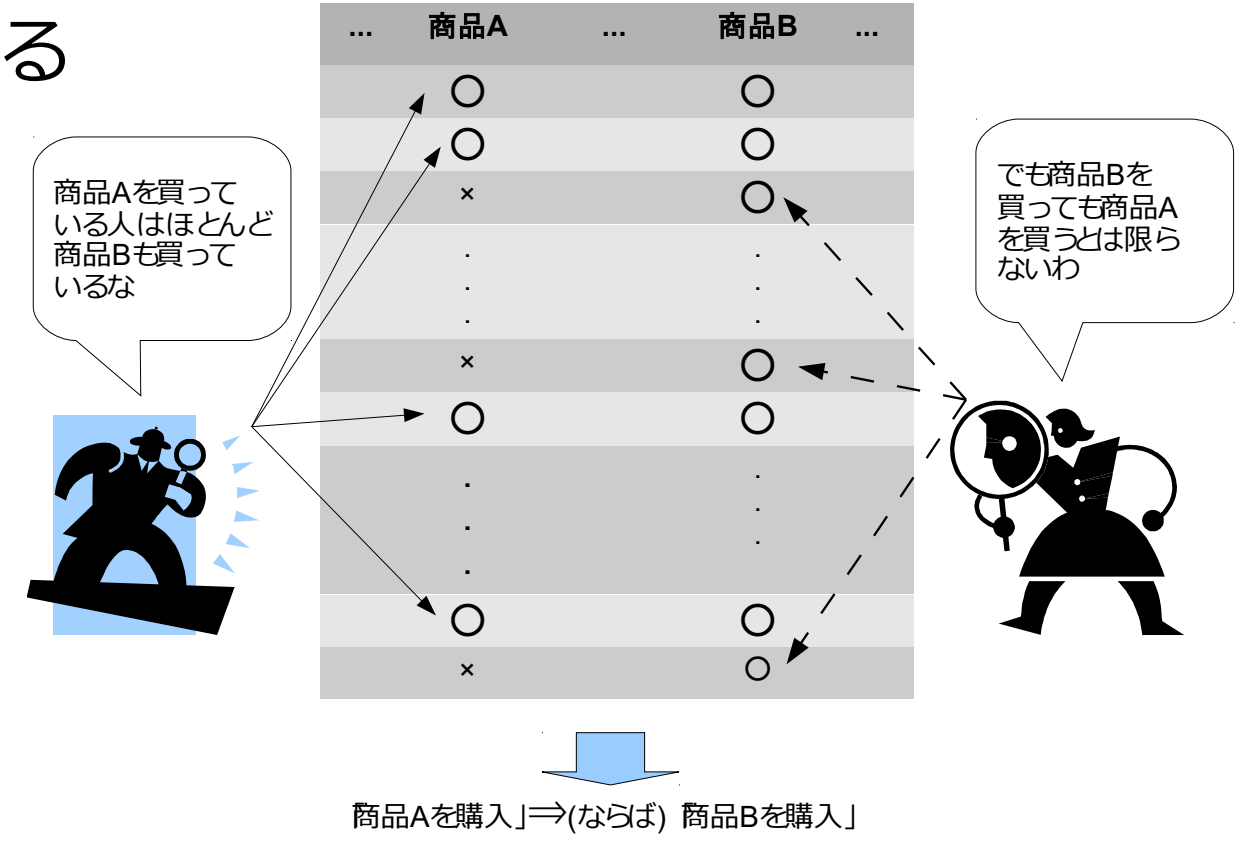
No.	ミルク	パン	バター	雑誌
1	t	t		
2		t		
3				t
4		t	t	
5	t	t	t	
6	t	t		

問題の定義

- 学習データ

$$\{\mathbf{x}^{(i)}\} \quad i = 1, .., N$$

- 問題設定
  - データ集合中で、一定頻度以上で現れるパターンを抽出する



11.2 Apriori アルゴリズムによる頻出項目抽出

- 例題：バスケット分析

No.	ミルク	パン	バター	雑誌
1	t	t		
2		t		
3				t
4		t	t	
5	t	t	t	
6	t	t		

バスケット分析では、 1 件分のデータをトランザクションとよぶ

- 支持度
  - 全トランザクション数 T に対して、ある項目集合 (items) が出現するトランザクションの割合

$$support(items) = \frac{T_{items}}{T}$$

11.2 Apriori アルゴリズムによる頻出項目抽出

- バスケット分析の目的
  - 支持度の値が閾値以上の項目集合を抽出したい
- バスケット分析の問題点
  - すべての可能な項目集合について、支持度を計算することは現実的には不可能

項目集合の種類数は 2 の商品数乗

商品数 1,000 の店なら 2^{1000}

高頻度の項目集合だけに絞って計算を行う必要がある

11.2 Apriori アルゴリズムによる頻出項目抽出

- a priori な原理

ある項目集合が頻出ならば、その部分集合も頻出である

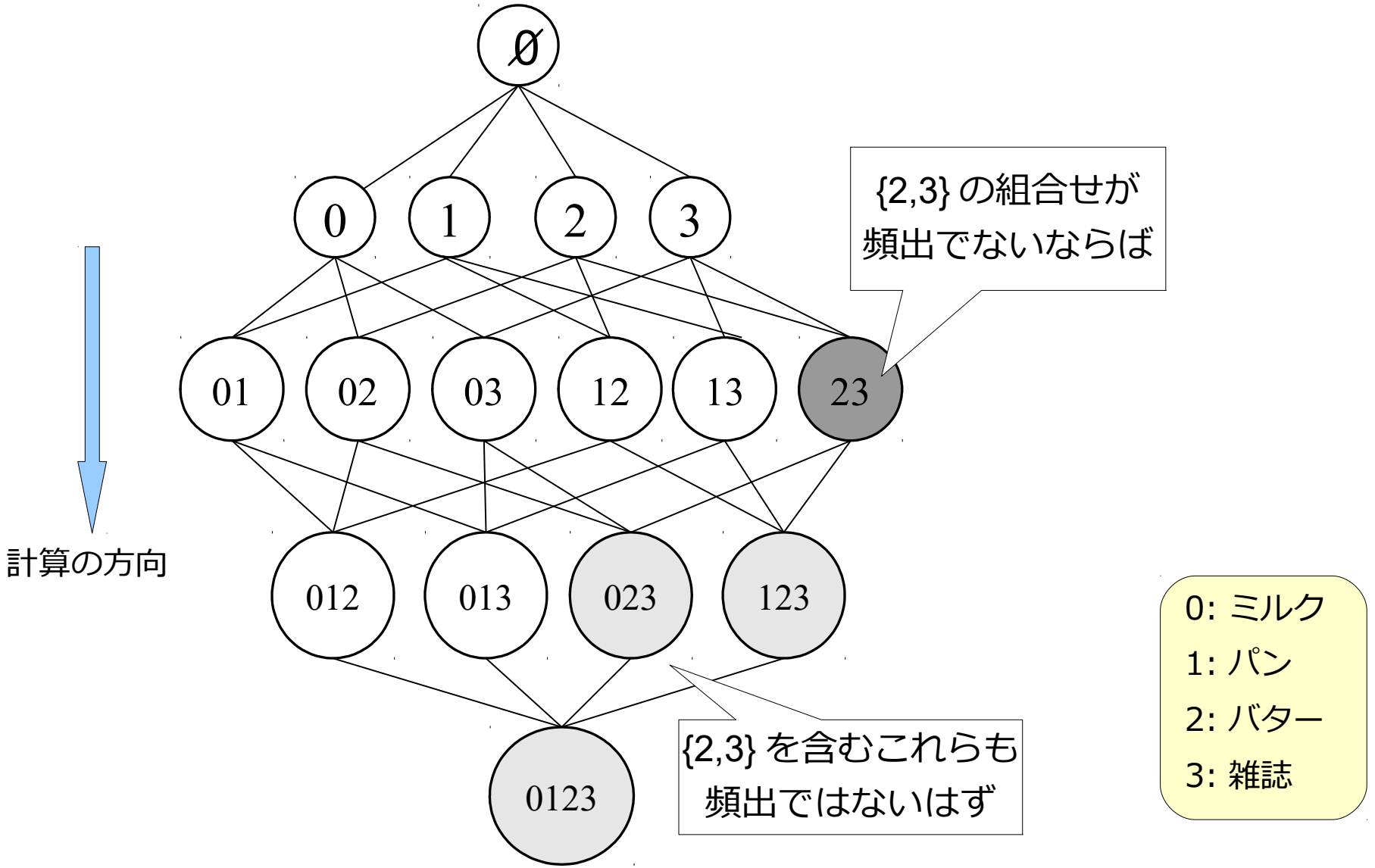
例）「パン・ミルク」が頻出ならば「パン」も頻出

対偶

ある項目集合が頻出でないならば、
その項目集合を含む集合も頻出でない

例）「バター・雑誌」が頻出でないならば「バター・雑誌・パン」も頻出でない

11.2 Apriori アルゴリズムによる頻出項目抽出



11.3 連想規則抽出

- 連想規則抽出の目的
  - 「商品 A を買った人は商品 B も買う傾向が強い」というような規則性を抽出したい
  - 確信度またはリフト値の高い規則を抽出

$$\mathrm{confidence}(\mathrm{A} \rightarrow B) = \frac{support(A \cup B)}{support(A)} = \frac{T_{A \cup B}}{T_A}$$

条件部 A が起こったときに結論部 B が起こる割合

$$\mathrm{lift}(\mathrm{A} \rightarrow B) = \frac{confidence(A \rightarrow B)}{support(B)}$$

B だけが単独で起こる割合と A が起こったときに B が起こる割合との比

11.3 連想規則抽出

- 連想規則抽出の手順
  - 頻出項目集合を求める
  - 項目集合を条件部、空集合を結論部とした規則を作成する
  - 条件部から結論部へ項目を 1 つずつ移動し、評価する

11.3 連想規則抽出

- a priori な原理

ある項目集合を結論部に持つ規則が頻出ならば、
その部分集合を結論部に持つ規則も頻出である

例）結論部が「パン・ミルク」の規則が頻出ならば、結論部が「パン」の規則も頻出である

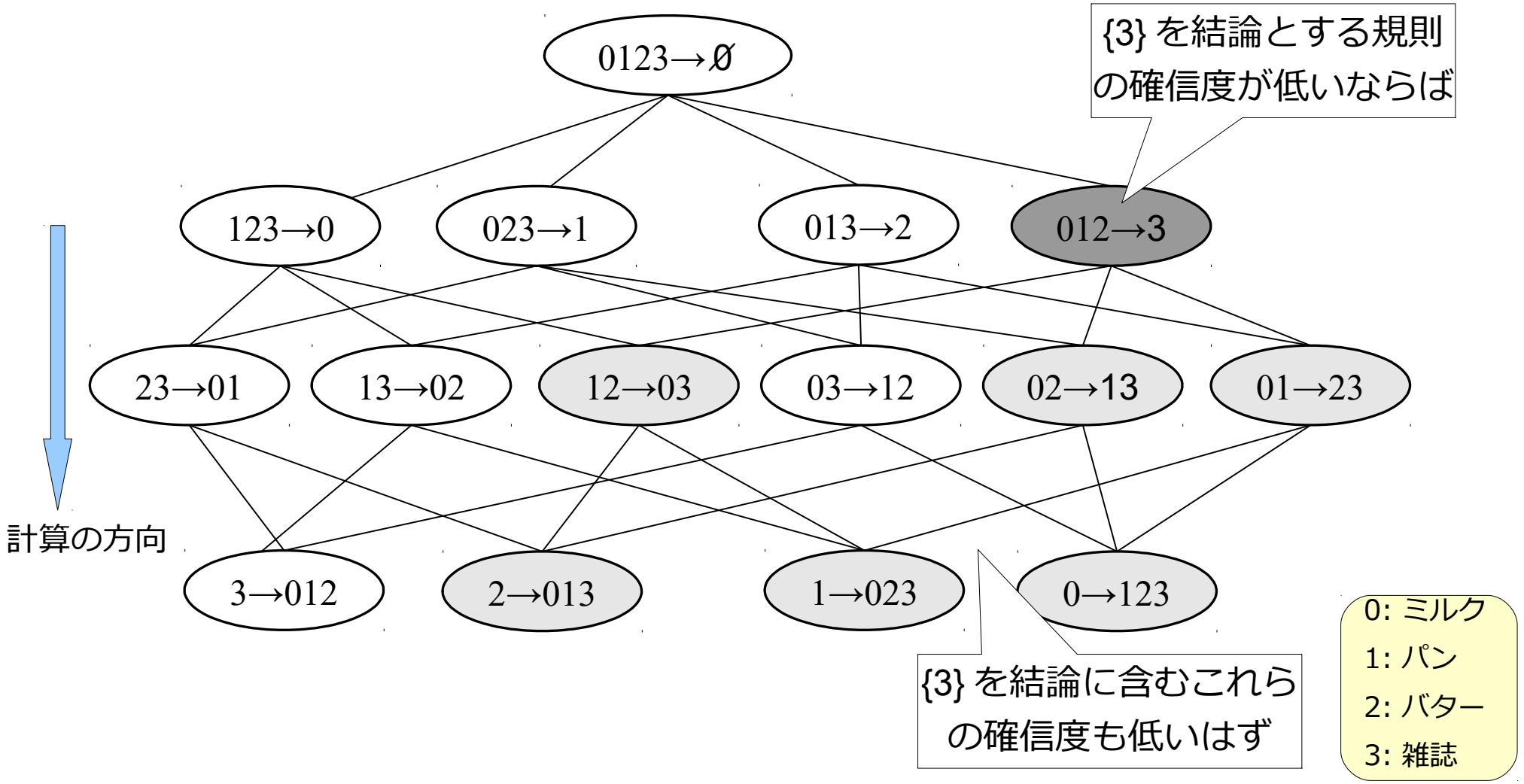
対偶

ある項目集合を結論部に持つ規則が頻出でないならば、
その項目集合を結論部に含む規則集合も頻出でない

例）結論部が「雑誌」の規則が頻出でないならば、結論部が「パン・雑誌」の規則も頻出でない

11.3 連想規則抽出

- a priori 原理に基づく探索



Section3 のまとめ

- モデル推定
  - クラスタリング
    - 階層的手法、 k-means 、 X-means
  - 異常検知
    - 周辺のデータとの違いを計算
  - 確率密度推定
    - 教師なしでクラスの確率分布を推定
- パターンマイニング
  - 頻出項目の効率的な数え上げ