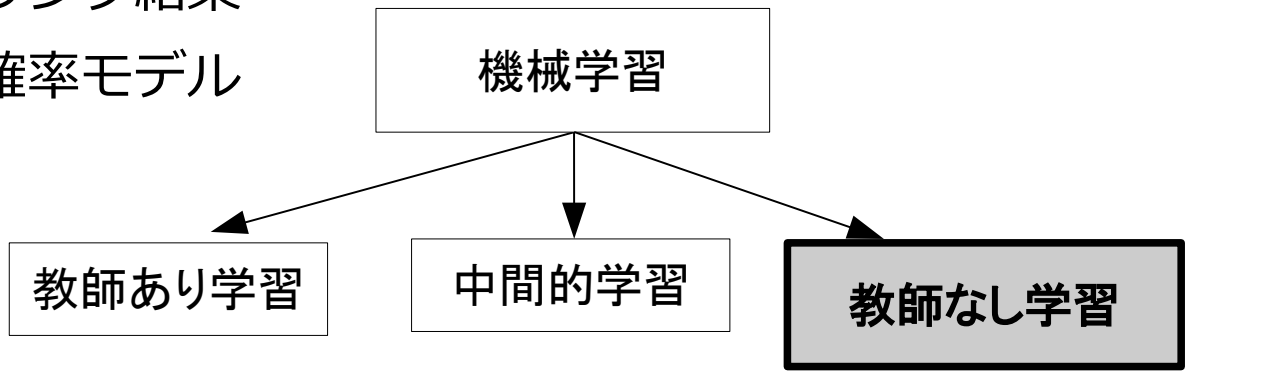
本日の予定

- 9:30-10:30　教師なし学習 ―モデル推定―（ 11 章）
- 10:45-11:45 教師なし学習 ―パターンマイニング―（ 12 章）

  （昼休憩）
- 13:00-14:00 系列データの学習、半教師あり学習（ 13, 14 章）
- 14:15-15:15 強化学習（ 15 章）
- 15:30-16:30　機械学習エンジニアへの道

11. 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
      - クラスタリング結果
      - クラスの確率モデル



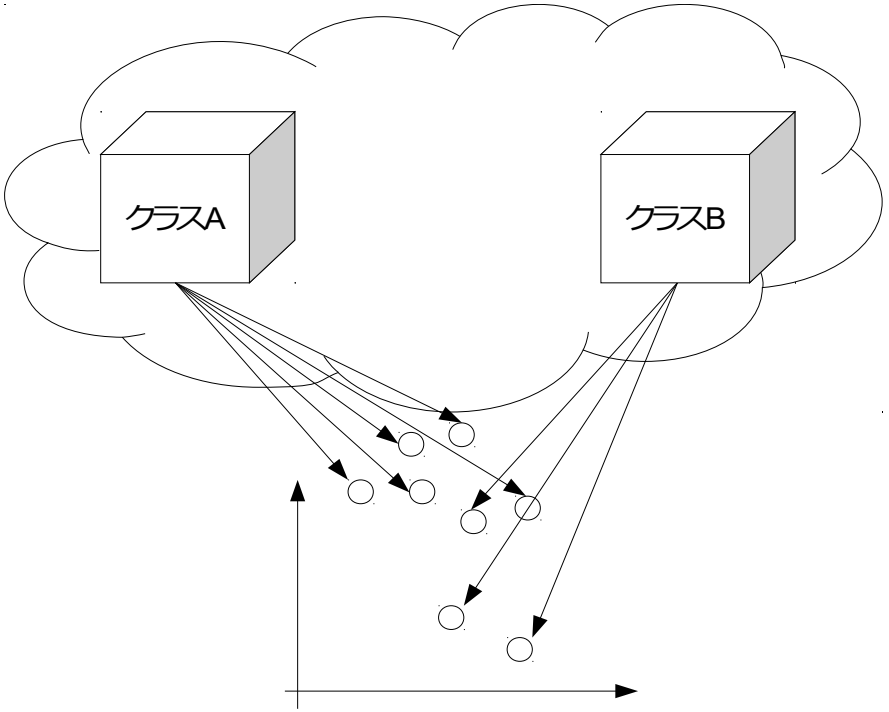
データ全体を対象に
その構造を発見する

11.1 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



11.2 クラスタリング

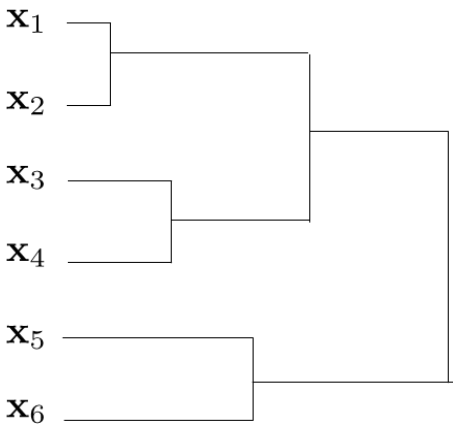
- クラスタリングとは
  - 対象のデータを、
    内的結合（同じ集合内のデータ間の距離は小さく）
    と
    外的分離（異なる集合間の距離は大きく）
    が達成されるような部分集合に分割すること

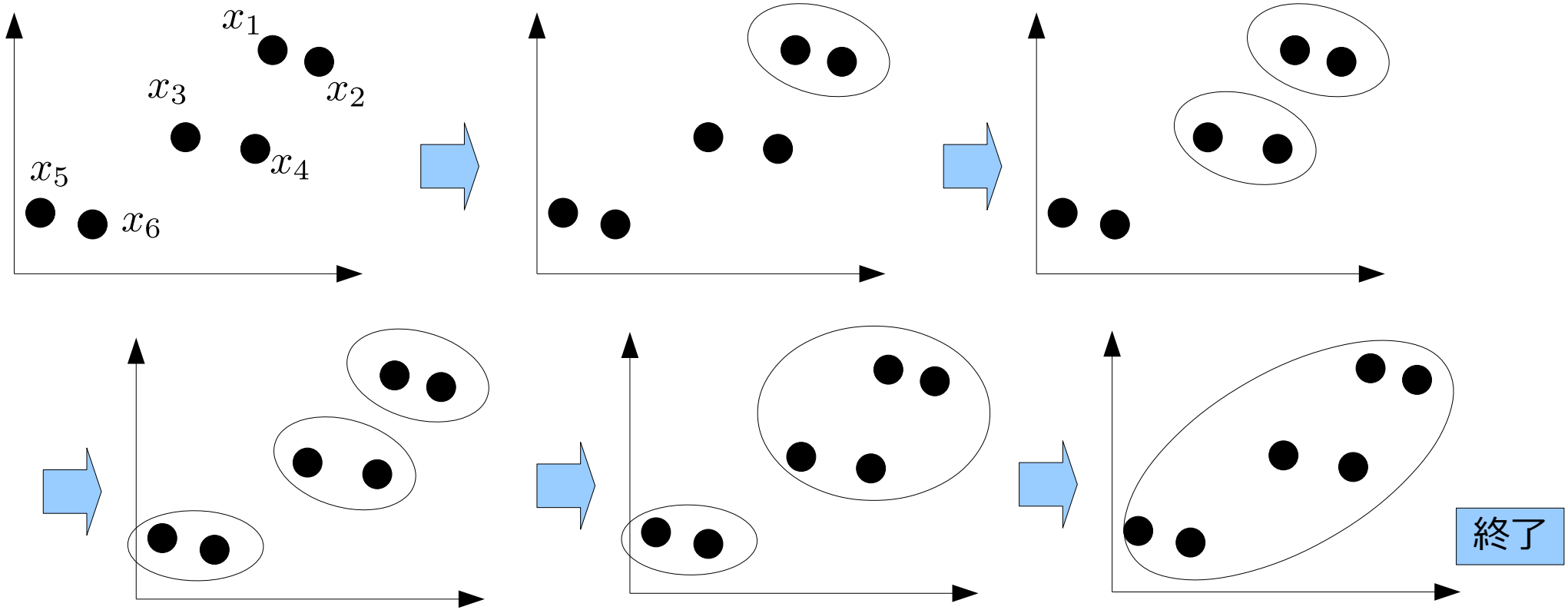
要するに塊を見つけること

- 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

11.2.1 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





11.2.1 階層的クラスタリング

- 類似度 sim の定義
  - 単連結法
    - 類似度：最も近い事例対の距離
    - 傾向：クラスタが一方向に伸びやすくなる
  - 完全連結法
    - 類似度：最も遠い事例対の距離
    - 傾向：クラスタが一方向に伸びるのを避ける
  - 重心法
    - 類似度：クラスタの重心間の距離
    - 傾向：単連結と完全連結の中間的な形
  - Ward 法
    - 類似度：融合前後の「平均ベクトルとの距離の二乗和」の差
    - 傾向：極端な形になりにくく、よく用いられる手法

11.2.2 分割最適化クラスタリング

- 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

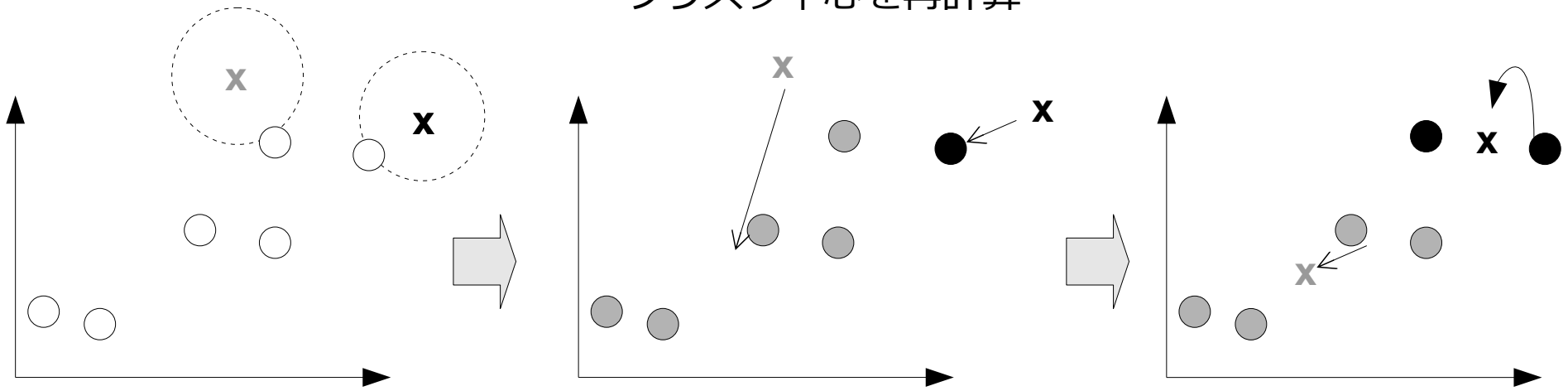
11.2.2 分割最適化クラスタリング

- 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

① 初期値として乱数でクラスタ中心を配置

② 各データを、最も近いクラスタ中心に配属
③ 所属しているデータからクラスタ中心を再計算

④ ②, ③ の処理をクラスタ中心が動かなくなるまで繰り返す



11.2.2 分割最適化クラスタリング

- 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

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

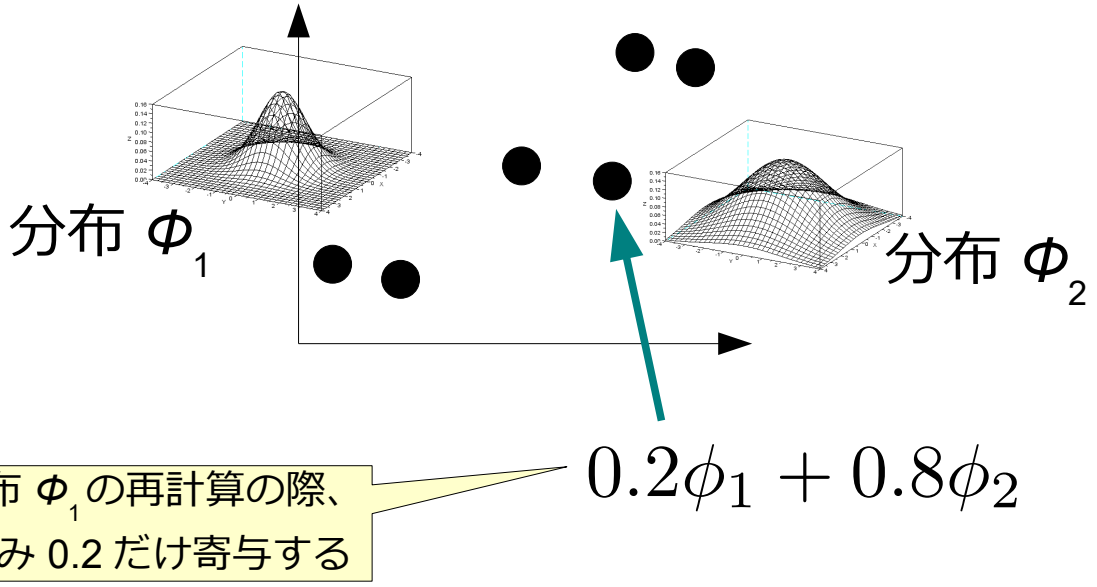
- L: モデルの尤度
- q: モデルのパラメータ数
- N: データ数

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）

11.4 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



11.4 確率密度推定

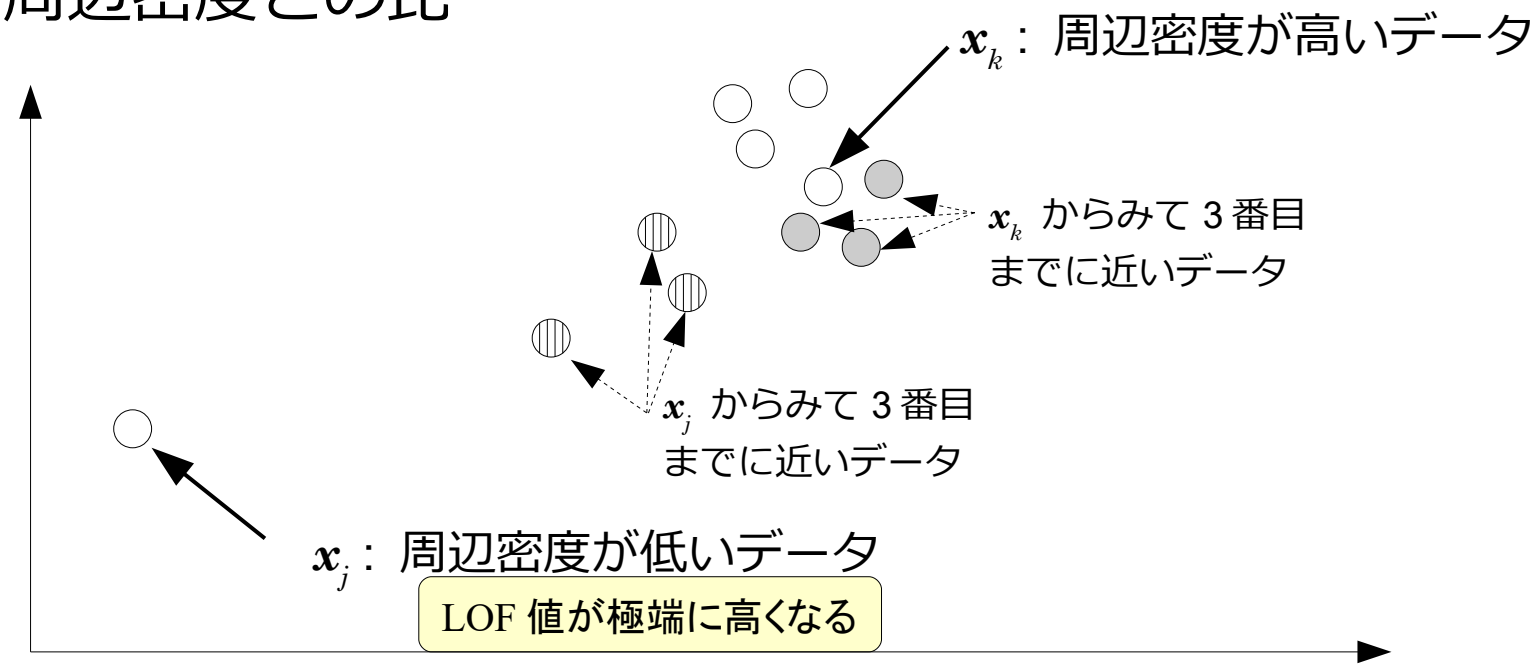
- k-means 法の一般化
  - k 個の平均ベクトルを乱数で決める
    ⇒ k 個の正規分布を乱数で決める
  - 平均ベクトルとの距離を基準に、各データをいずれかのクラスタに所属させる
    ⇒各分布が各データを生成する確率を計算し、各クラスタにゆるやかに帰属させる
  - 所属させたデータをもとに平均ベクトルを再計算
    ⇒各データのクラスタへの帰属度に基づき各分布のパラメータ（平均値、共分散行列）を再計算

11.3 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

11.3 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



11.3 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

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

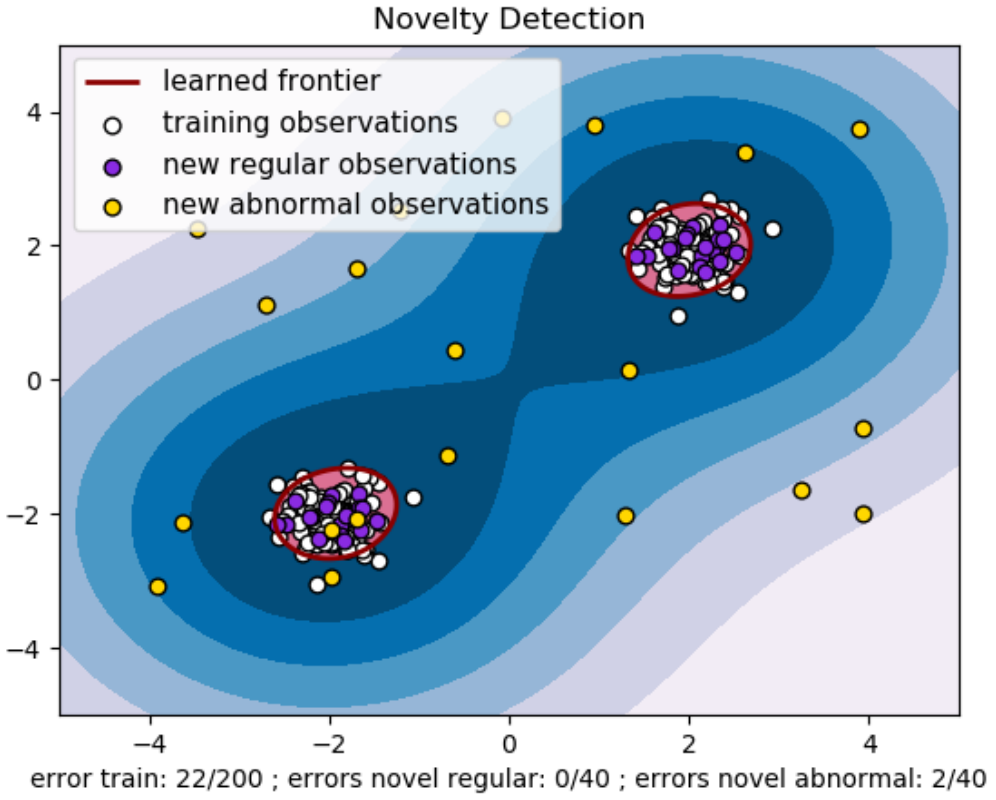
x の周りの密度が高い場合、大きな値になる

  - 局所異常因子

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

One-class SVM

- One-class SVM による新規検知
  - RBF カーネルによる写像後の空間における学習データを正例、原点を負例とみなして境界を得る
  - 新規データに対して、境界の外の場合は異常とみなす



まとめ

- Weka デモ
  - iris.2D データのクラスタリング
  - iris データの LOF フィルタによる異常検出
- クラスタリング
  - 同じ性質を持つデータのまとまりを見つける方法
  - クラスタ数が特定できれば良い結果が得やすい
- 異常検出
  - 正常なデータのまとまりを推定し、そこからはずれるデータを検出