「言語処理のための機械学習入門」
第6章　実験の仕方など
6-1　プログラムとデータの入手
6-2　分類問題の実験の仕方

茨城大学工学部
佐々木稔

プログラムとデータの入手

- 各種機械学習手法をデータに適用
  - 手法が動作するプログラムを自作する
  - 各手法のプログラムを入手する
    - WEKA
    - R, Matlab, Octave
    - Cluto, Cluster 3.0
    - C4.5, libSVM
- 入力データのフォーマットを調査する
  - 1行1事例、クラス名と組成ベクトルの組合せ

WEKA

- フリーのデータマイニングツール
  - ニュージーランドのWaikato大学で開発
  - プログラムは Java で記述
    - Java の SDK が必要
  - インストール（Windows版）
    - ウィザードに従ってインストール
  - 手軽にデータの分析が可能
    - 分類、クラスタリング、相関ルール
    - 属性選択、データの可視化

WEKAのデータフォーマット

- @RELATION
  - データの名前
- @ATTRIBUTE
  - 素性の名前と値の型を記述
  - 型： NUMERIC, STRING, DATE, 集合({YES, NO})
- @DATA
  - 各インスタンスの値

入力データの例

% 1. Title: Iris Plants

@RELATION iris

@ATTRIBUTE sepallength NUMERIC

@ATTRIBUTE sepalwidth NUMERIC

@ATTRIBUTE petallength NUMERIC

@ATTRIBUTE petalwidth NUMERIC

@ATTRIBUTE class {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA

5.1,3.5,1.4,0.2,Iris-setosa

4.9,3.0,1.4,0.2,Iris-setosa

4.7,3.2,1.3,0.2,Iris-setosa

4.6,3.1,1.5,0.2,Iris-setosa

WEKAの使用例

1. 環境変数CLASSPATHの確認
2. WEKAを起動
3. Explorerを選択
4. “Open file …” で入力データを選択
5. 実行したい処理のタブを選択
6. 分類、フィルタリングなどの分析手法を選択
7. Start ボタンで実行

実験データ

- 必要とする実験データ
  - 自作
  - 他の研究で使われたデータセットの利用
    - 20Newsgroup（文書分類）
    - Reuters-21578（文書分類）
    - Enron E-mail Analysis Project（電子メール分類）
    - MovieReview Data（レビュー分類）
    - MovieLens Datasets（レビュー分類）
  - データセットの加工はperlやrubyなどを使う

分類問題の実験の仕方

- データセット入手後の確認事項
  - 事例数（文書数）
  - １事例あたりの平均パターン数（単語数）

データの分け方

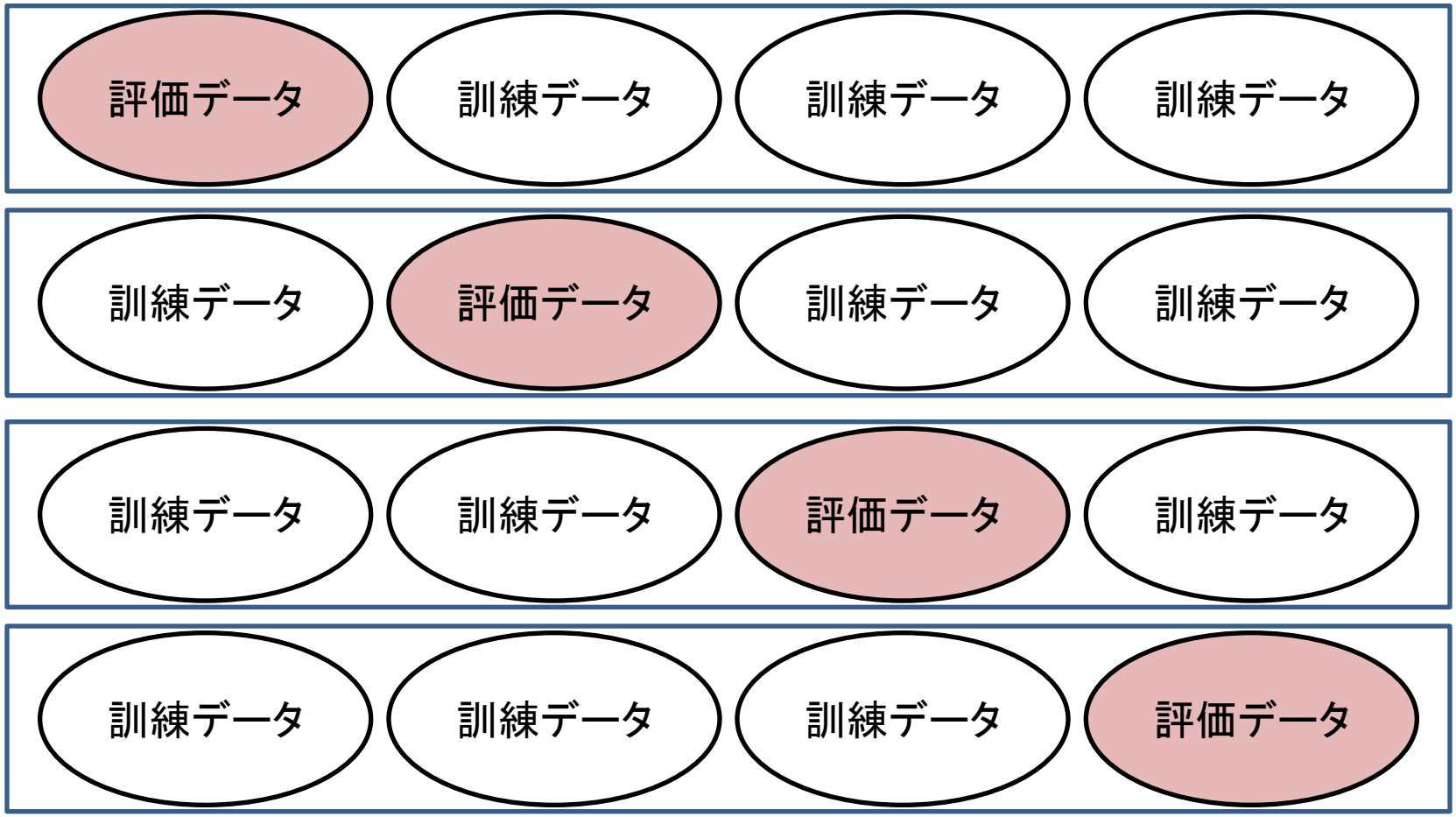
- データセット
  - 訓練データ：モデル学習に使うデータ
  - 評価データ：モデルの評価に使うデータ
- 開発データ
  - モデルのパラメータを決めるためのデータ
    - ナイーブベイズ分類器の事前確率 α
    - 緩和制約下SVMモデルの定数 C
  - いくつかのパラメータ値を設定し実験
  - 最も評価指標の高いときのパラメータを採用

交差検定

- データセットが分かれてない場合の評価方法
  - ひとつのデータセットを分割
- 交差検定（Cross-Validation）
  - データセットを *n* 個に分割（*n*-fold cross-validation）
    - そのうち 1 個が評価データ
    - 残りの *n*-1 個が訓練データ
  - 各分割データを評価データとし *n* 回繰り返す
  - 得られた *n* 個の評価指標値の平均値を求める

N-fold Cross-Validation

- *N*=4 の場合の評価データの種類



多クラスと複数ラベル

- データセットのクラスの種類
  - 2種類の場合は「2値クラスデータ」
  - 2種類より大きい場合は「多クラスデータ」
- 各データに割り当てたクラスの数
  - ひとつの場合は「単一ラベルデータ」
  - ふたつ以上の場合は「複数ラベルデータ」
- 分類問題
  - 「二値分類」: 2つのクラスのどちらか分類
  - 「多値分類」: 3つ以上のクラスでどれかに分類