機械学習
練習問題①

管理工学科
篠沢佳久

練習問題①

- 多クラス分類問題
  - 学習用データ：train-1.csv
  - テストデータ：test-1.csv

- 特徴量の次元数：4
  - A列：特徴量1，B列：特徴量2，C列：特徴量3，D列：特徴量4
- クラス数：3
  - E列：クラス番号（0～2）
  - 1クラスにつき100個データ

練習問題①

- 学習データ（train-1.csv）を用いてロジスィック回帰モデルを学習しなさい.
- 学習後のモデルを用いて，テストデータ（test-1.csv）を予測しなさい.
- 評価指標（混同行列，accuracy，precision，recall，f値）も求めて下さい.

補足①（多クラス分類問題）

- 「1対その他」（one-versus-rest*）

- クラス数が3の場合 → 3個のモデルを学習
  - ① クラス1 ⇔ クラス2とクラス3 を分類するモデル
  - ② クラス2 ⇔ クラス1とクラス3 を分類するモデル
  - ③ クラス3 ⇔ クラス1とクラス2 を分類するモデル

- 予測する場合
  - 3個のモデルから予測値を求め，最大値を出力するモデルを分類結果とする

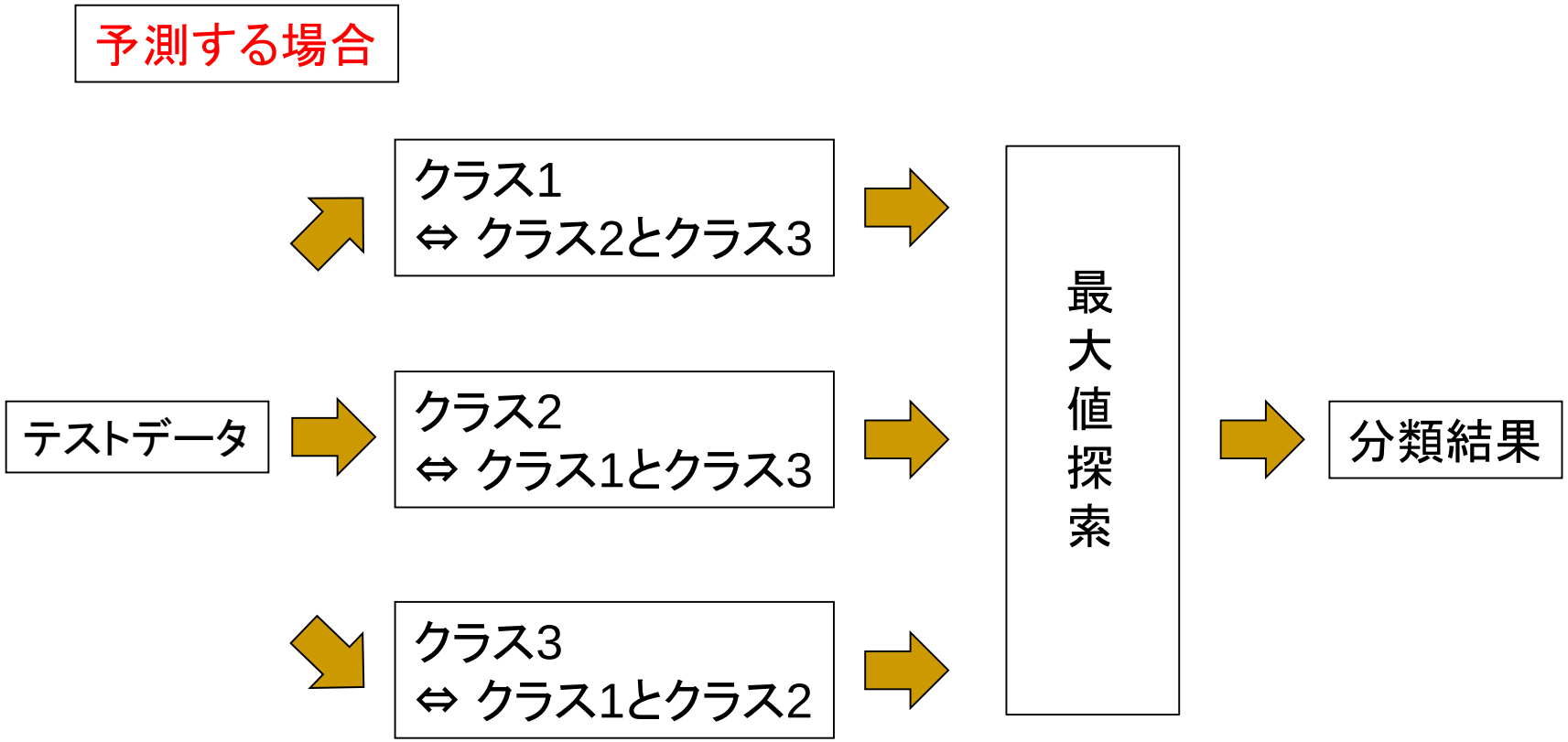
*one-versus-otherとも呼ばれます

補足①（多クラス分類問題）

3個のロジスティック回帰モデルの学習

クラス1 ⇔ クラス2とクラス3	⇒	正例：クラス1 負例：クラス1でない	クラス1である 確率を予測
クラス2 ⇔ クラス1とクラス3	⇒	正例：クラス2 負例：クラス2でない	クラス2である 確率を予測
クラス3 ⇔ クラス1とクラス2	⇒	正例：クラス3 負例：クラス3でない	クラス3である 確率を予測

補足①（多クラス分類問題）



*scikit-learnの場合，書き方（model，fit）はニクラス分類問題も多クラス分類問題も同じです

補足②

- csvファイルをnumpy配列に読み込むには，numpy.loadtxtを使うと便利です
  - https://docs.scipy.org/doc/numpy/reference/generated/numpy.loadtxt.html

- 練習問題の場合，
  - 学習用データの特徴量 → train_data
  - 学習用データのラベル → train_label
  - テスト用データの特徴量 → test_data
  - テスト用データのラベル → test_label

  に格納して下さい

補足③

- 多クラス分類問題の場合，評価指標（precision，recall，f値）を求める際に，averageオプションで，binary以外を選択して下さい．

- https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html

'micro'，'macro'，'weighted‘ のどれかを用いて下さい

- （例）
  - precision_score(test_label, predict, average='micro')
  - recall_score(test_label, predict, average='micro')
  - f1_score(test_label, predict, average='micro')

提出方法

- 提出場所：keio.jp
- 提出日：10/21（月）13時まで

- Pythonプログラム（.py形式）
- ワープロに，以下をまとめて下さい
  - 学籍番号，氏名
  - 実行結果
  - 学習時のパラメータ
  - 学習後のモデルの式（3個）

どういう学習をして，何がモデルとして学習されたかを理解しましょう