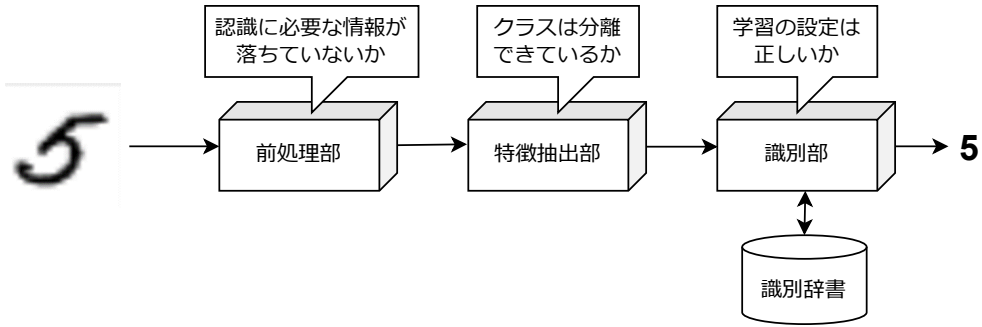
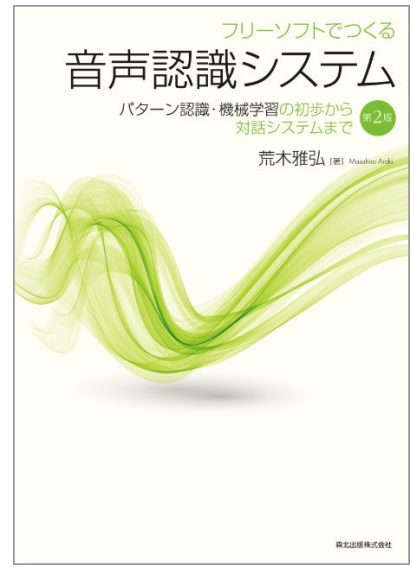
9. 本当にすごいシステムができたの？



- 9.1 未知データに対する認識率の評価
- 9.2 システムを調整する方法



- 荒木雅弘:『フリーソフトでつくる音声認識システム(第2版)』(森北出版, 2017年)
- スライドとJupyter notebook
- サポートページ

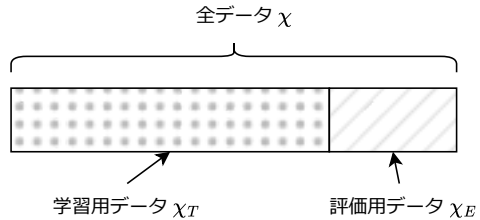
9.1 未知データに対する認識率の評価

- パターン認識システムの評価
  - 学習データに対して識別率100%でも意味がない
  - 未知データに対してどれだけの識別率が期待できるかが評価のポイント

    → どうやって未知データで評価する？

9.1.1 分割学習法

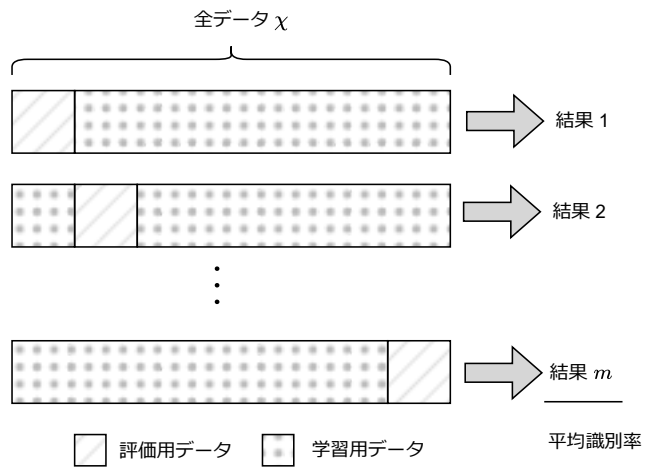
- 手順
  - 全データ χ を学習用データ χ_T と評価用データ χ_E に分割
  - χ_T を用いて識別器を設計し、χ_E を未知データとみなして識別率を推定



- 利点
  - 評価が容易
- 欠点
  - 学習に用いるデータ数が減るので、識別性能が低く見積もられる
  - 学習データの割合を高くすると評価データ数が少なくなり、識別率の推定精度が低くなる

9.1.2 交差確認法 (1/2)

- 手順
  1. 全データ χ を m 個のグループ $\chi_1, \ldots, \chi_m$ に分割する
  2. 以下の手順を $i = 1, \ldots, m$ について行い、m 個の識別率の平均を推定値とする
     - χ_i を除いた $m - 1$ 個のグループで学習し、χ_i を用いて識別率を算出する

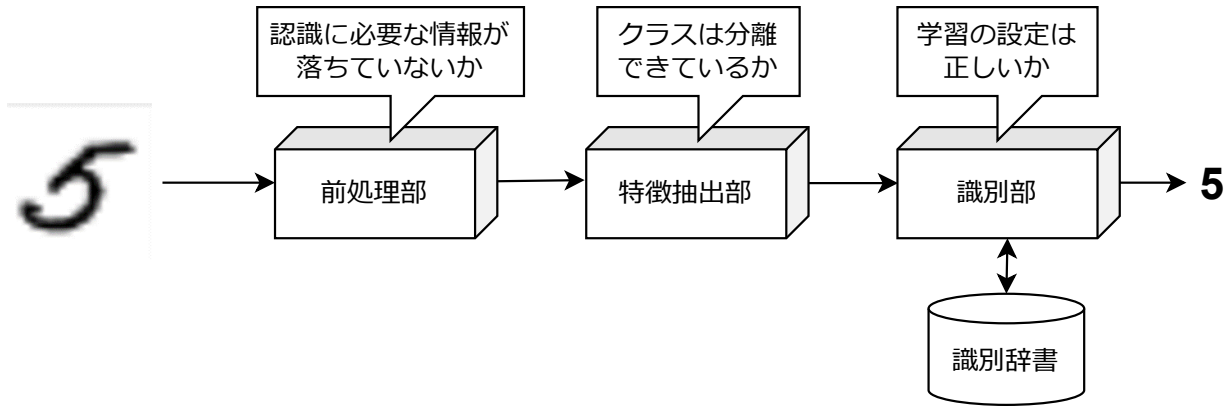


9.1.2 交差確認法（2/2）

- 利点
  - 分割学習法に比べ、識別率の推定精度は高い
- 欠点
  - 評価に時間がかかる
  - 分割数が少ない場合、分割方法の違いによって評価値が大きくぶれる
- 一つ抜き法
  - 要素数が 1 となるように分割する方法
  - 時間はかかるが最も信頼できる交差確認法

9.2 システムを調整する方法

- システムの性能向上のために
  - 前処理部、特徴抽出部、識別部のどこに性能低下の原因があるかを探る

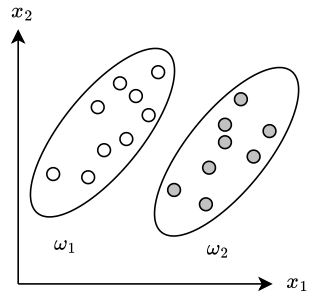


9.2.1 前処理部の確認

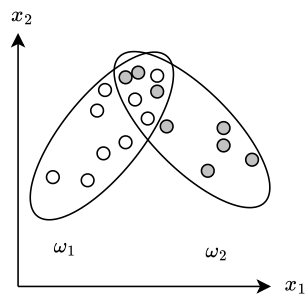
- 信号取り込み部のチェック
  - マイクの入力レベル調整やカメラのキャリブレーションが必要
  - 自動運転のように識別と動作が連動している場合、突発的な異常入力を検知して、誤動作を防止する機構が必要
- デジタル化に伴う情報劣化のチェック
  - サンプリング周波数や量子化ビット数が適切か
- ノイズ除去のチェック
  - 原信号への影響を確認

9.2.2 特徴空間の評価 (1/7)

- 次元削減による可視化を通じて評価
  - クラスが適切に分離されているのに認識率が低い場合 → 識別部の設定ミスが疑われる



  - クラス分布が大きく重なっている場合 → 特徴抽出部を再設計 → 評価基準が必要



9.2.2 特徴空間の評価（2/7）

- クラス内分散・クラス間分散比
  - 特徴空間の評価法
    - クラス毎のデータのまとまり具合と、クラス間の離れ具合を評価する尺度
  - 同じクラスのデータ同士はなるべく接近し、異なるクラスのデータの塊はなるべく離れているものが高い値を取るようにする

9.2.2 特徴空間の評価（3/7）

- クラス内分散 σ_W^2

$$\sigma_W^2 = \frac{1}{n}\sum_{i=1}^{c}\sum_{\boldsymbol{x}\in\chi_i}(\boldsymbol{x}-\boldsymbol{m}_i)^T(\boldsymbol{x}-\boldsymbol{m}_i) \quad \boldsymbol{m}_i：クラス i(\chi_i)の平均，n：全データ数$$

- クラス間分散 σ_B^2

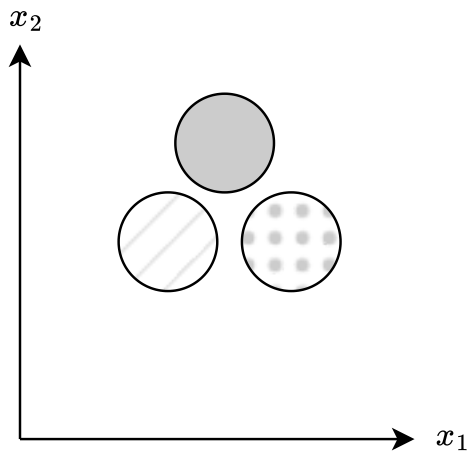
$$\sigma_B^2 = \frac{1}{n}\sum_{i=1}^{c} n_i(\boldsymbol{m}_i-\boldsymbol{m})^T(\boldsymbol{m}_i-\boldsymbol{m}) \quad \boldsymbol{m}：全データの平均，n_i：クラス i のデータ数$$

- クラス内分散・クラス間分散比 J_σ（大きいほど良い）

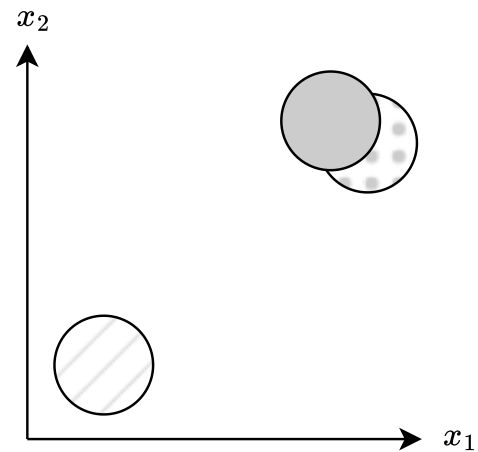
$$J_\sigma = \frac{\sigma_B^2}{\sigma_W^2}$$

9.2.2 特徴空間の評価（4/7）

- 多クラスのクラス内分散・クラス間分散比
  - 分布の重なりを考慮できないので、あまりよい評価尺度とはいえない
    - 大きいクラス間分散がよい特徴空間と対応しない例（クラス内分散は同一と仮定）



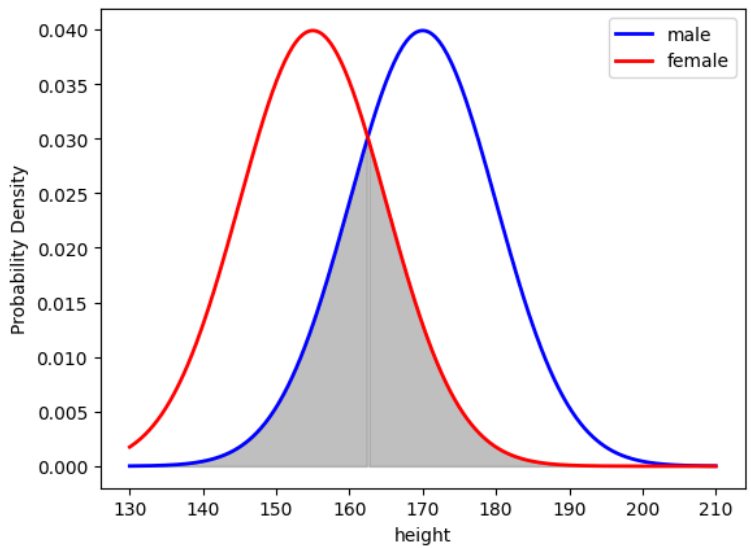
(a) クラス間分散：小



(b) クラス間分散：大

9.2.2 特徴空間の評価（5/7）

- ベイズ誤り確率
  - 特徴空間上での分布の重なりの度合いを評価
  - 例）身長による成人男女の判別
    - 一般に同一の特徴が男女両方にあてはまるので、性別を確実に決定することはできない。



9.2.2 特徴空間の評価（6/7）

- ベイズ決定則
  - 誤識別率を最小にするために事後確率 $P(\omega_i|\boldsymbol{x})$ が最大となるような ω_i を出力する判定方法
- 条件付きベイズ誤り確率: $e_B(\boldsymbol{x})$
  - $\boldsymbol{x}$ が与えられたときの誤り確率の最小値
  - 2クラス識別問題の場合

$$e_B(\boldsymbol{x}) = \min\{P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\}$$

9.2.2 特徴空間の評価（7/7）

- ベイズ誤り確率 e_B

$$
\begin{aligned}
e_B &= \int e_B(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int \min\{P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\} p(\boldsymbol{x}) d\boldsymbol{x}
\end{aligned}
$$

  - e_B は誤り確率をこれよりは小さくできないという限界、すなわち分布の重なりを表す
  - 分布は一般に未知であるため、ベイズ誤り確率を直接計算することは困難

    $\rightarrow$ 学習パターンに基づいてベイズ誤り確率を間接的に推定
    - 近似的な計算：1-NN法の誤り確率 e_N との関係

      $e_B \leq e_N \leq 2e_B$ （e_N はベイズ誤り確率の2倍を超えない）

9.2.3 識別部の調整 (1/5)

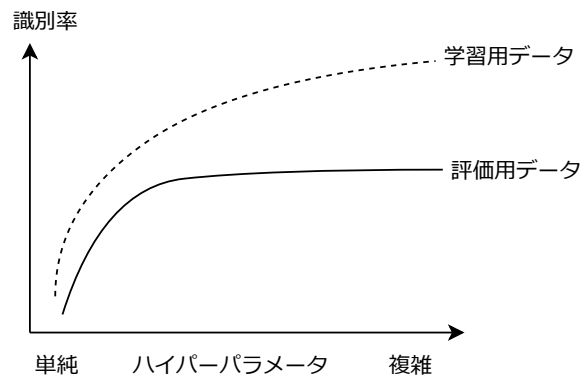
- パラメータ：学習可能
  - 識別関数の重み
  - ニューラルネットワークの結合の重み
  - SVMの $\boldsymbol{\alpha}$
- ハイパーパラメータ：学習結果によって調整
  - 識別関数の次数
  - ニューラルネットワークの層数や隠れ層のユニット数
  - SVM 多項式カーネルの次数

9.2.3 識別部の調整 (2/5)

- 学習過程に影響を与えるパラメータ
  - 例)ニューラルネットワークの学習係数、EMアルゴリズムの収束判定に用いる値
  - 設定値が不適切な場合、不必要に多くの時間がかかったり、学習が途中で終わったりする
  - 適切な値の設定は機械学習の know-how
    - 特徴を標準化することによって、ある程度は経験的に設定可能
- 学習結果に影響を与えるパラメータ(= ハイパーパラメータ)
  - モデルの複雑さに連続的に影響を与える $\rightarrow$ 性能に直結する
  - 例)SVMの多項式カーネルの次数、ガウシアンカーネルの半径 γ
  - いくつかの異なる値で性能を評価する必要がある

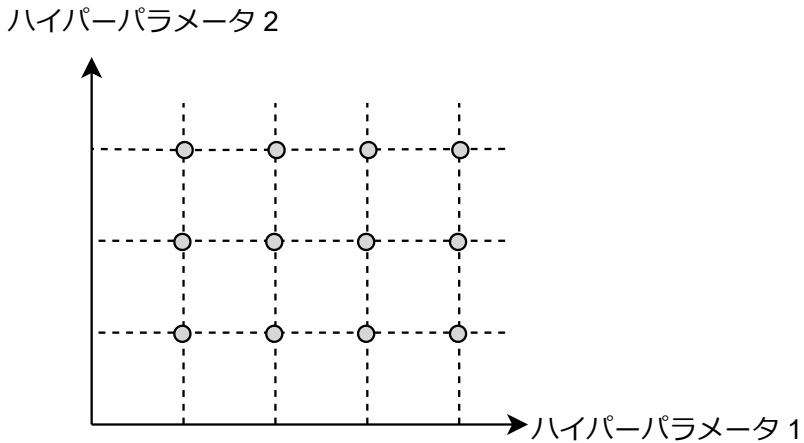
9.2.3 識別部の調整（3/5）

- ハイパーパラメータ λ の決定手順
  - 未知データに対する誤識別率 e_λ が低い λ が望ましい
  - 分割学習法や交差確認法を用いて未知データに対する e_λ を推定する
- ハイパーパラメータの性質
  - 複雑な分布を示す学習データに対しては、複雑なモデルにする必要がある
  - モデルを複雑にしても、あるところで識別率が上がらなくなる（下がることもある）



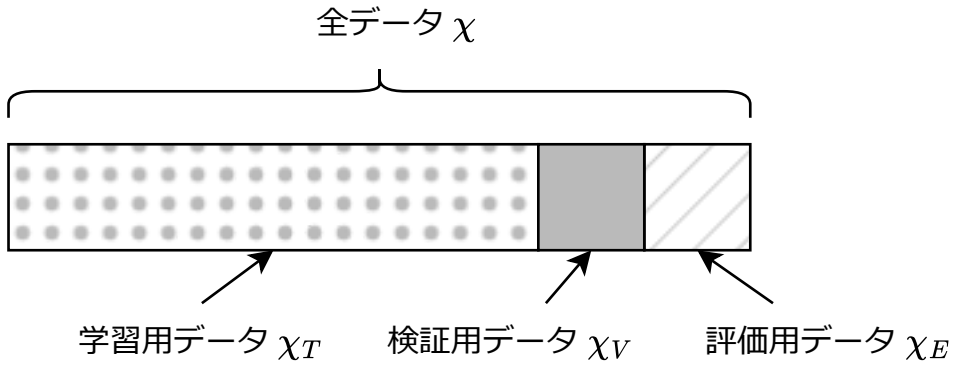
9.2.3 識別部の調整 (4/5)

- ハイパーパラメータが複数ある場合
  - 例)SVMの多項式カーネルの次数 d と誤りの重み C
  - グリッドサーチ:各格子点で e_λ を求める



9.2.3 識別部の調整（5/5）

- ハイパーパラメータ調整のためのデータ分割
  - ハイパーパラメータ選択に用いたデータに対する識別率は、そのハイパーパラメータの性能を過大評価するので信用できない
  - ハイパーパラメータを選択するための検証用データ χ_V を分割に加える
  - χ_V での性能が最も高くなる識別器の識別率を χ_E を用いて推定



まとめ

- 未知データに対する認識率の評価
  - 分割学習法
  - 交差確認法
- パターン認識システム全体の調整
  - 前処理の結果の確認
  - 特徴空間の評価
  - ハイパーパラメータの調整
- Jupyter notebook