Section 4

- 統計的識別手法 (8章)
- 学習の評価 (9章)

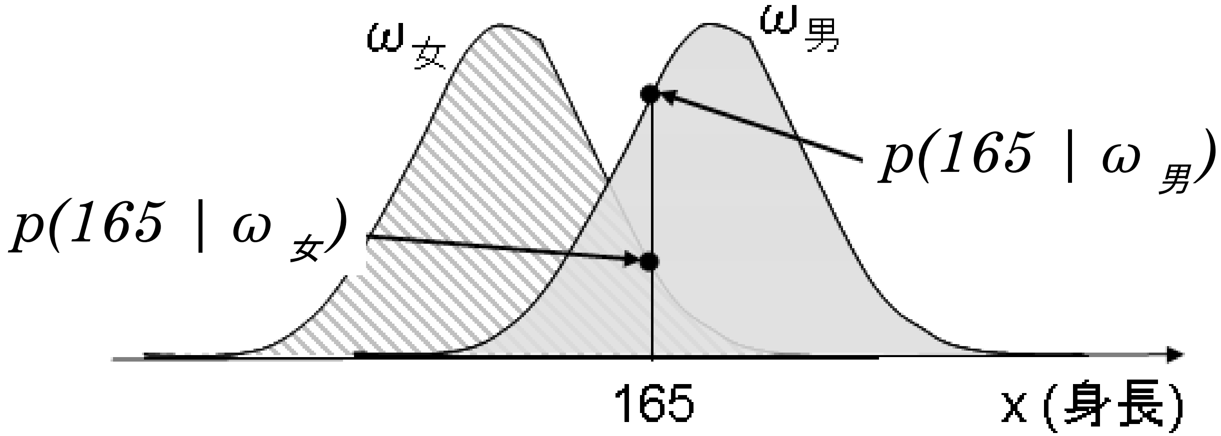
8. 未知データを推定しよう －統計的方法－

8.1　間違う確率を最小にしたい

- 本当に作りたいシステムは？
  - 誤り 0 のシステム
    - 現実的には不可能
  - 出力誤差最小のシステム ← Widrow-Hoff の学習規則 ニューラルネットワーク
    - 学習データに対して最適化してしまうかもしれない
  - 誤り確率最小のシステム ← **この時間のテーマ**
    - 未知データに対する誤り確率最小
  - (期待損失最小のシステム)

8.1　間違う確率を最小にしたい

- 例）身長を特徴量として成人男女を識別する
  - 身長が与えられたときの、確率の高い方（＝誤り確率の低い方）を識別結果とすればよい



8.1.1 誤り確率最小の判定法

- 確率を用いたパターン認識
  - 事後確率最大法 (ベイズ決定則)
    - $P(\omega_i|\boldsymbol{x})$ を最大にするクラスω_i を識別結果とする

$$\mathop{\arg\max}_{i=1,\ldots,c} P(\omega_i|\boldsymbol{x}) = k \quad \Rightarrow \quad \boldsymbol{x} \in \omega_k$$

- 身長による成人男女の判別システムの場合
  - 入力が 185.0cm の時、 P(男 |185.0) と P(女 |185.0) の大きい方に判定する

8.1.2 事後確率の求め方

- 一般に $P(\omega_i|\boldsymbol{x})$ は直接求めることができない
  - 例）185.0cm の人を何人集めれば P(男 |185.0) の値が推定できる？
  - 標本誤差（*p*: 調査対象の比率、 *n*: 標本数）

$$2 \cdot \sqrt{\frac{p(1-p)}{n}}$$

    - p=0.5 のとき、真の値が 95% の確率で存在する範囲
      - n=100: 50.0 % ±10.0
      - n=2,000: 50.0 % ±2.2
  - それを 130.0 ～ 200.0 まで行うと？

8.1.3 事後確率の間接的な求め方

- ベイズの定理

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})}$$

P: 離散変数に対する確率関数

p: 連続変数に対する確率密度関数

証明 $$\begin{aligned} P(A,B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$

8.1.3 事後確率の間接的な求め方

- 事後確率　$P(\omega_i|\boldsymbol{x})$
  - $\boldsymbol{x}$ が生起したとき、そのクラスが ω_i である確率
- 事前確率　$P(\omega_i)$
  - クラス ω_i の生起確率
- クラスによらない $\boldsymbol{x}$ の生起確率　$p(\boldsymbol{x})$
- 尤度
  - 認識対象としているパターンの生起確率を示したもの

    $$p(\boldsymbol{x}|\omega_i) \quad (i = 1, \ldots, c)$$
  - クラス ω_i に属する $\boldsymbol{x}$ が出現する確率

8.1.4 厄介者 $p(\boldsymbol{x})$ を消そう

- $p(\boldsymbol{x})$ は全クラスに共通であり、 最大となる $P(\omega_i|\boldsymbol{x})$ を決めるのに関与しない

$$
\begin{aligned}
&\mathop{\arg\max}_{i=1,\ldots,c} P(\omega_i|\boldsymbol{x}) \\
&= \mathop{\arg\max}_{i=1,\ldots,c} \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} \\
&= \mathop{\arg\max}_{i=1,\ldots,c} p(\boldsymbol{x}|\omega_i)P(\omega_i)
\end{aligned}
$$

8.1.5 事前確率 $P(\omega_i)$ を求める

- 事前確率 $P(\omega_i)$ の求め方
  - 本当は全ての可能なデータを集めて、それぞれのクラスのデータ数を集計しなければ求まらないが ...
- 最尤推定
  - 学習データ数：N
  - クラス ω_i のデータ数：n_i
  - 事前確率の最尤推定値

$$P(\omega_i) = \frac{n_i}{N}$$

8.1.6 最後の難敵「クラス分布 $P(x\,|\,\omega_i)$ 」

- クラス分布 $p(\boldsymbol{x}|\omega_i)$ の求め方
  - クラス分布とは
    - あるクラスのデータ集合からある特徴ベクトルが観測される確率をあらわす確率密度関数
    - ある値は観測されやすく、それから遠くなるに従って観測されにくくなるような性質を持つ
  - 確率分布の形を仮定して、そのパラメータを学習データから推定
  - 例）正規分布：平均と共分散行列を推定

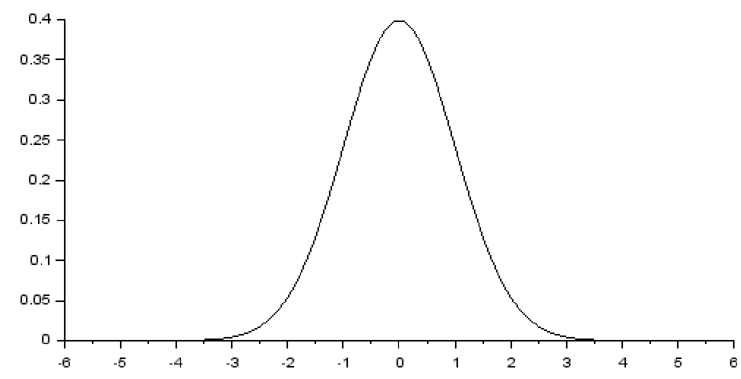
8.2　データの広がりを推定する

8.2.1 未知データの統計的性質を予測する

- 確率密度関数 $p(\boldsymbol{x}|\omega_i)$ の例
  - 正規分布

$$p(\boldsymbol{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_i)^t\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{m}_i)\}$$

- $\boldsymbol{m}_i$ ：平均ベクトル
- $\boldsymbol{\Sigma}_i$ ：共分散行列



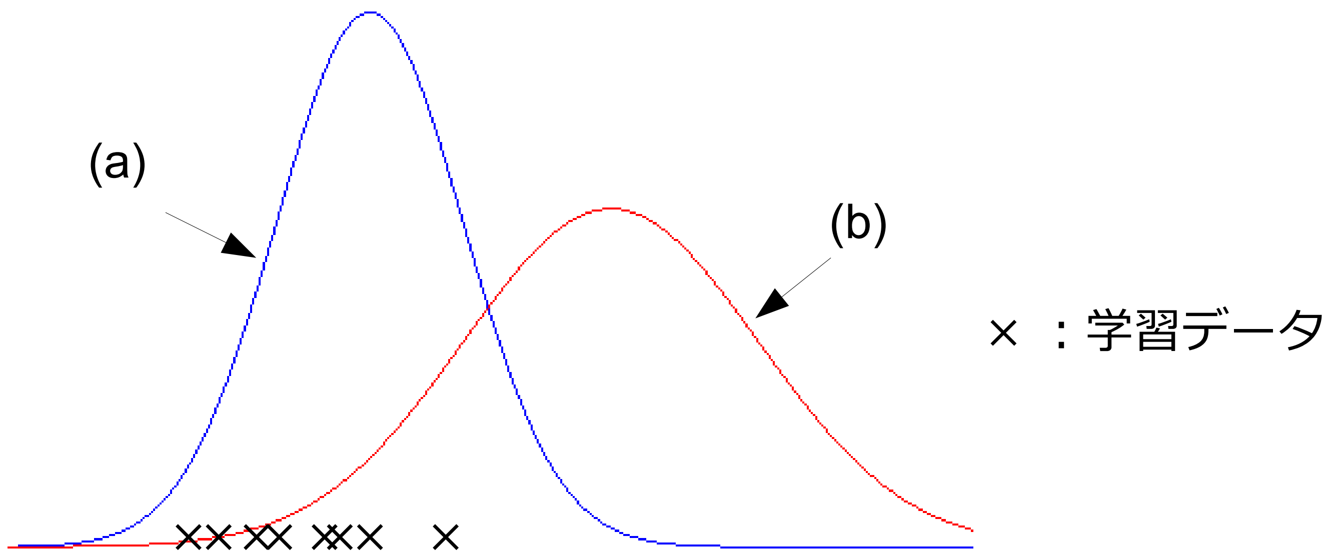
正規分布とは

- 離散型二項分布の例
  - n 枚のコインを投げた時の、表の枚数の度数
    - n=1　　1 1
    - n=2　　1 2 1
    - n=3　　1 3 3 1
    - n=4　　1 4 6 4 1
    - n=5　　1 5 10 10 5 1

    …
  - n→∞　の時の分布が正規分布

8.2.2 最尤推定

- データを最もうまく説明できる分布を探す



＝対数尤度が最大となる分布を探す

$$p(\chi|\boldsymbol{\theta}) = \prod_{x_p \in \chi} p(x_p|\boldsymbol{\theta}) \qquad \boldsymbol{\theta} \ : \ 分布のパラメータ$$

8.2.2 最尤推定

- 最尤推定の結果
  - 平均ベクトル

$$\boldsymbol{m}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \chi_i} \boldsymbol{x}$$

  - 共分散行列

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \chi_i} (\boldsymbol{x} - \boldsymbol{m}_i)(\boldsymbol{x} - \boldsymbol{m}_i)^t$$

対角要素はその次元の分散を、それ以外の要素は交差する次元間の相関を表す

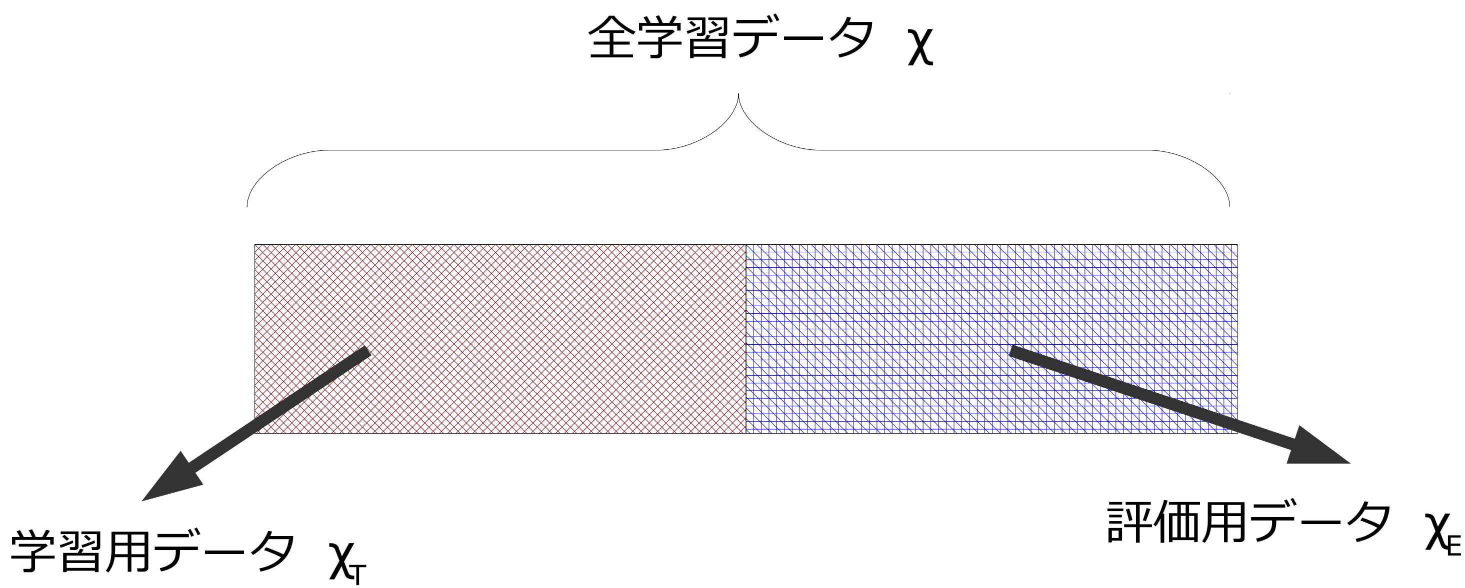
9. 本当にすごいシステムができたの？
9.1　未知データに対する認識率の評価

- 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
  - 全学習データ χ を学習用データ集合 χ_T と評価用データ集合 χ_E に分割する
  - χ_T を用いて識別機を設計し、 χ_E を用いて誤識別率を推定する



9.1.1 分割学習法

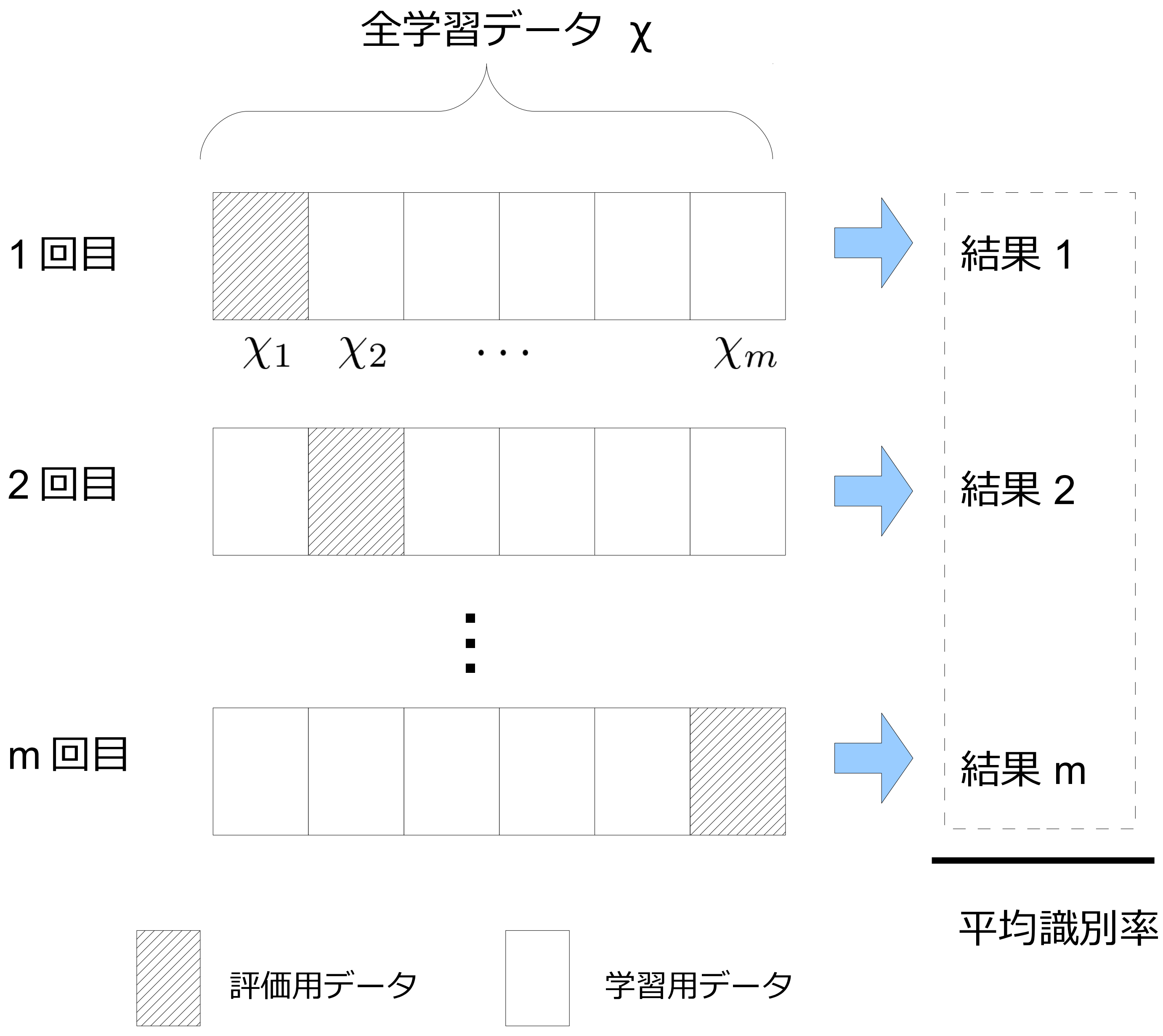
- 問題点
  - 学習に用いるデータ数が減るので、識別性能が劣化する
  - 誤識別率の推定精度は良くない

9.1.2 交差確認法

- 手順

1. χ を m 個のグループ $\chi_1,\ldots,\chi_m$ に分割する
2. χ_i を除いた $(m\text{-}1)$ 個のグループで学習し、χ_i を用いて識別率を算出する
3. この手順をすべての i について行い、m 個の識別率の平均を誤識別率の推定値とする

9.1.2 交差確認法



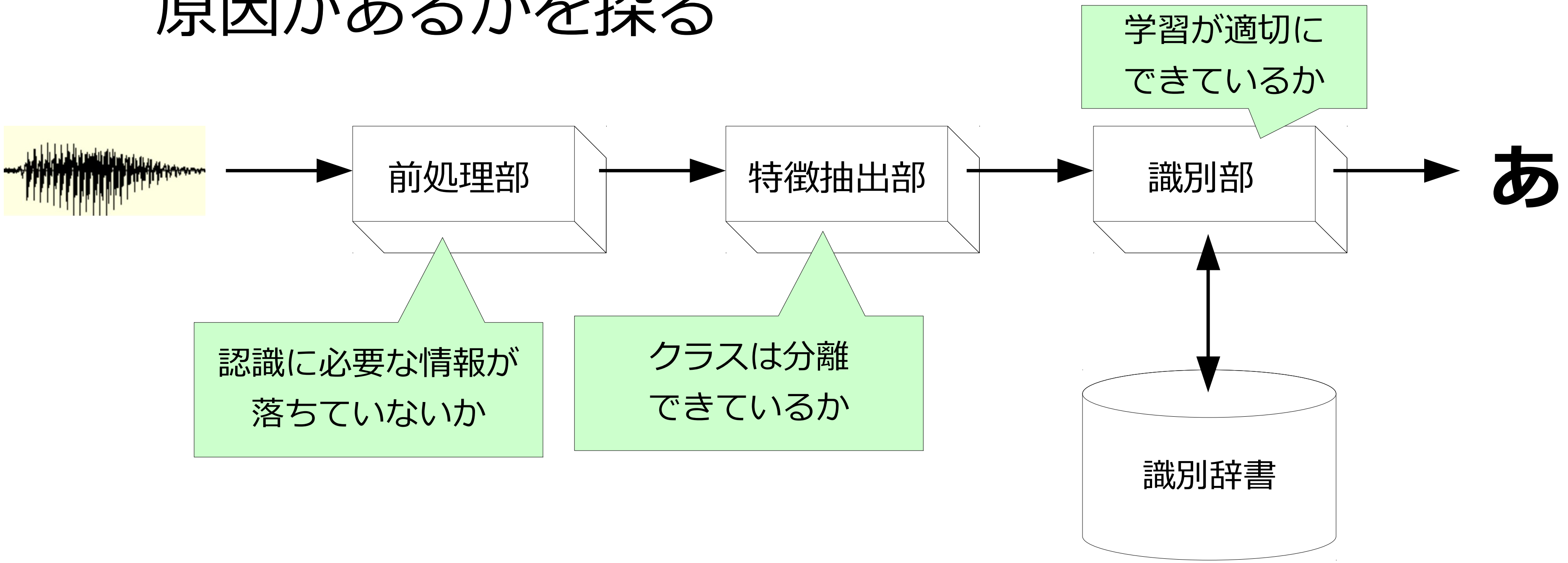
全学習データ χ
1 回目
2 回目
m 回目
χ_1
χ_2
$\cdots$
χ_m
結果 1
結果 2
結果 m
平均識別率
評価用データ
学習用データ

9.1.2 交差確認法

- 特徴
  - 分割学習法に比べ、誤識別率の推定精度は高い
  - 評価に時間がかかるのが欠点
- 要素数が１となるように分割する方法を一つ抜き法 と呼ぶ

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

- 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
- 信号取り込み部のチェック
  - マイクの入力レベルやカメラのキャリブレーション

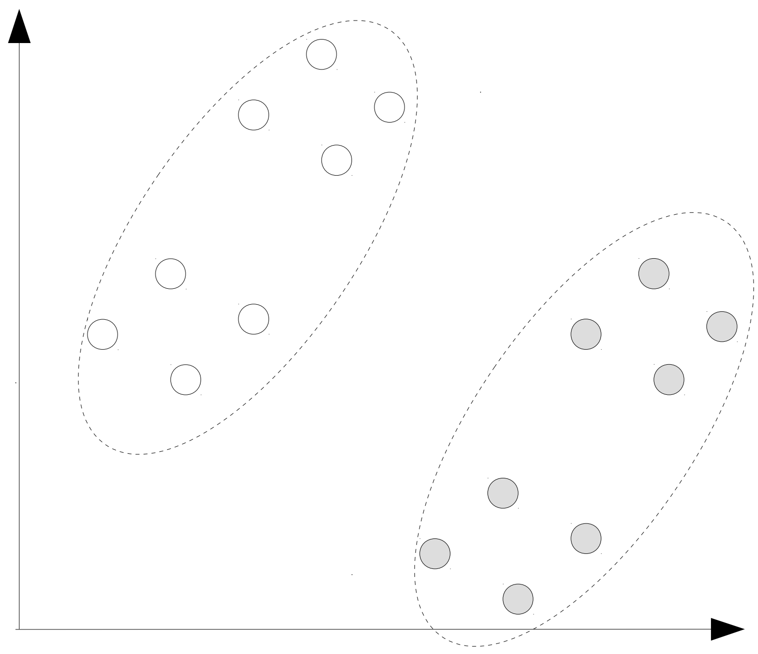
突発的な異常入力に対しては誤動作の防止が必要

- 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

- クラスが特徴空間上で完全に分離されているのに誤認識率が高い場合

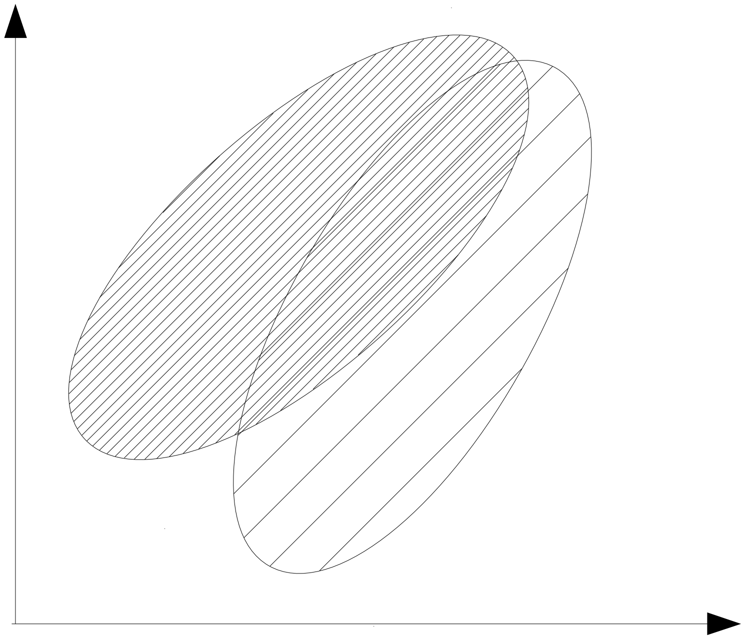
→　識別部を再設計（識別関数の学習）



9.2.2 特徴空間の評価

- クラスの分布間に重なりがある場合

→　特徴抽出部を再設計（特徴の評価）



9.2.3　識別部の調整

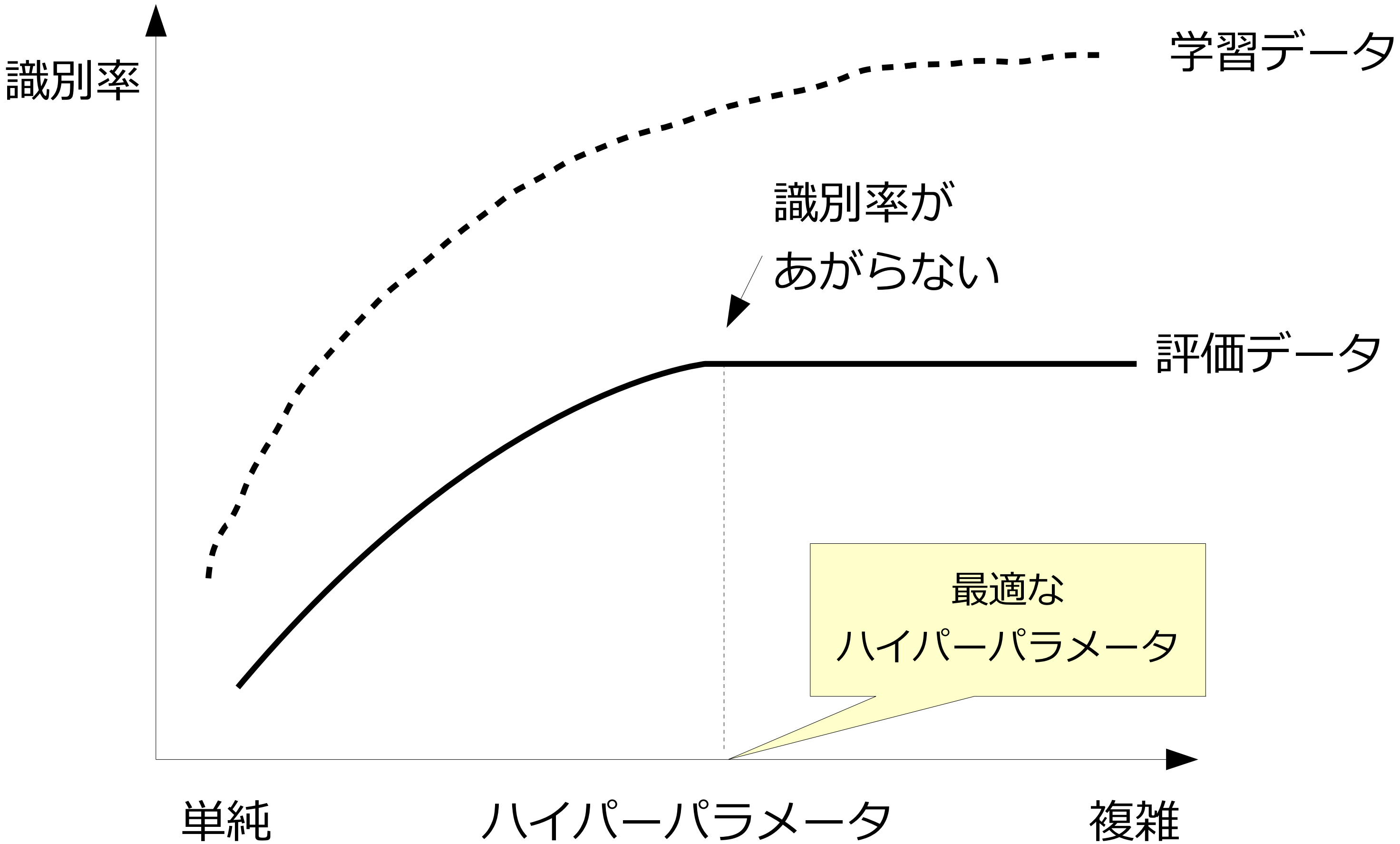
- パラメータ　➡ 学習可能
  - 識別関数の重み
  - ニューラルネットワークの結合の重み
  - SVM の $\boldsymbol{\alpha}$
- ハイパーパラメータ　➡　学習結果によって調整
  - 識別関数の次数
  - ニューラルネットワークの中間ユニット数
  - SVM　多項式カーネルの次数

9.2.3 識別部の調整

- ハイパーパラメータ λ の決定手順
  - 未知パターンに対する誤識別率 e_λ が低い λ が望ましい。
  - 実際は分布が未知なので、単純に e_λ を計算することはできない。
  - 多くの場合、交差確認法で e_λ を求める。

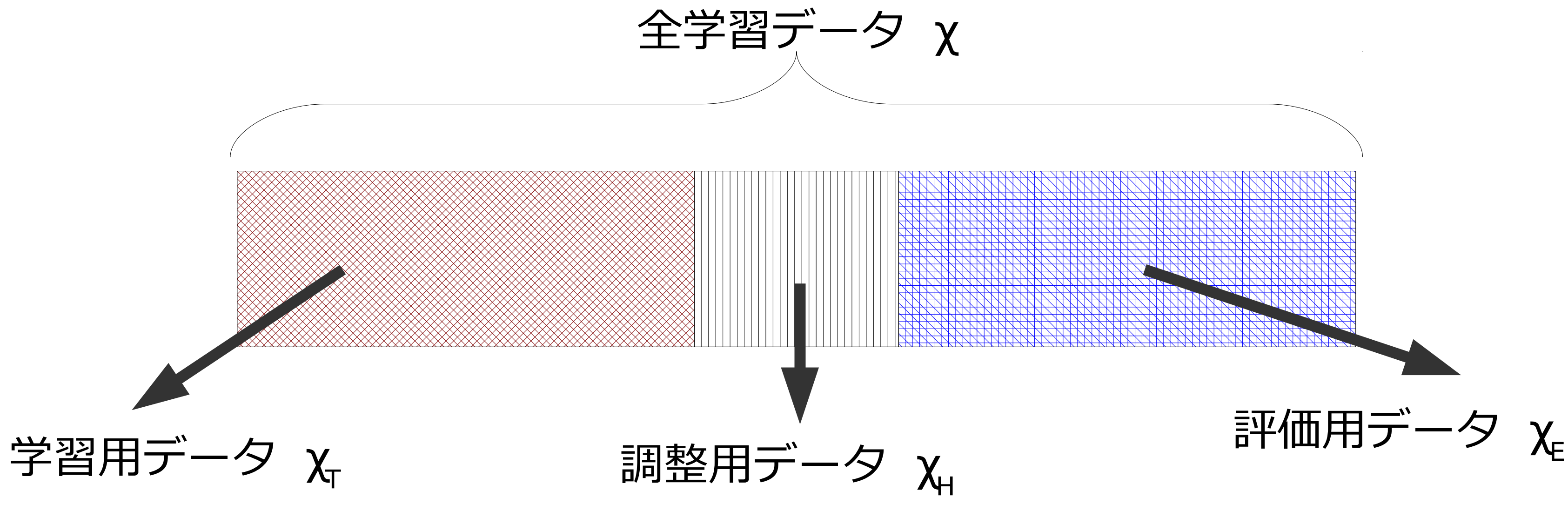
9.2.3 識別部の調整

- ハイパーパラメータの性質



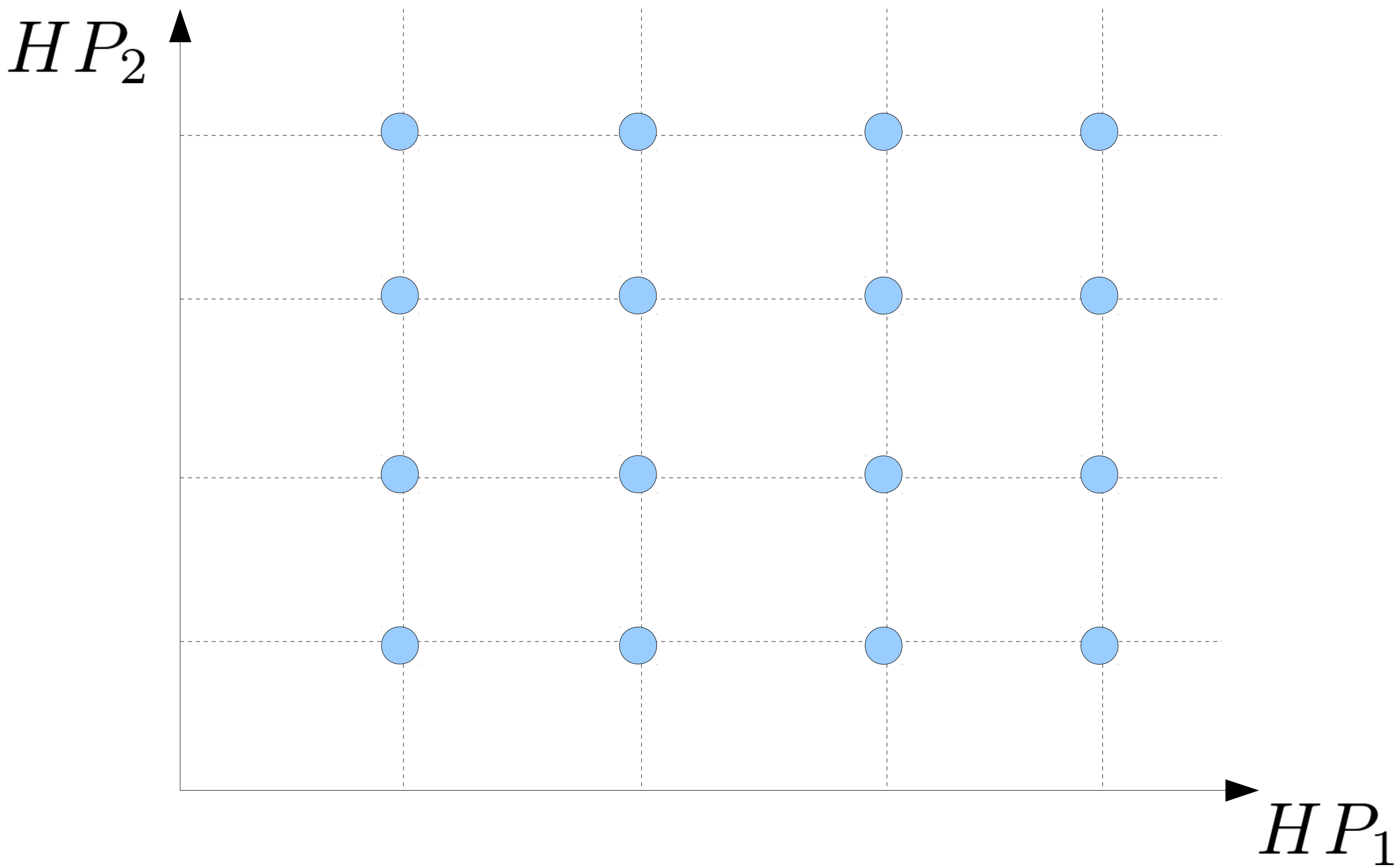
9.2.3 識別部の調整

- ハイパーパラメータ調整を含む分割学習法
  - 全学習データ χ を学習用データ集合 χ_T、調整用データ集合 χ_H、評価用データ集合 χ_E に分割する
  - χ_T を用いて識別機を設計、 χ_H を用いてハイパーパラメータを調整、 χ_E を用いて誤識別率を推定する



9.2.3 識別部の調整

- ハイパーパラメータが複数ある場合
  - グリッドサーチ：各格子点で e_λ を求める



Section4 のまとめ

- 統計的識別手法
  - 事後確率が最大となるクラスを計算する
  - ベイズの定理を用いて、事前確率とクラス分布の積に変換する
  - クラス分布の関数形を仮定し、そのパラメータを学習データから推定する
- 学習の評価
  - 分割法や交差確認法で未知データでの評価に近づける
  - ハイパーパラメータの調整はグリッドサーチ