9. 本当にすごいシステムができたの？

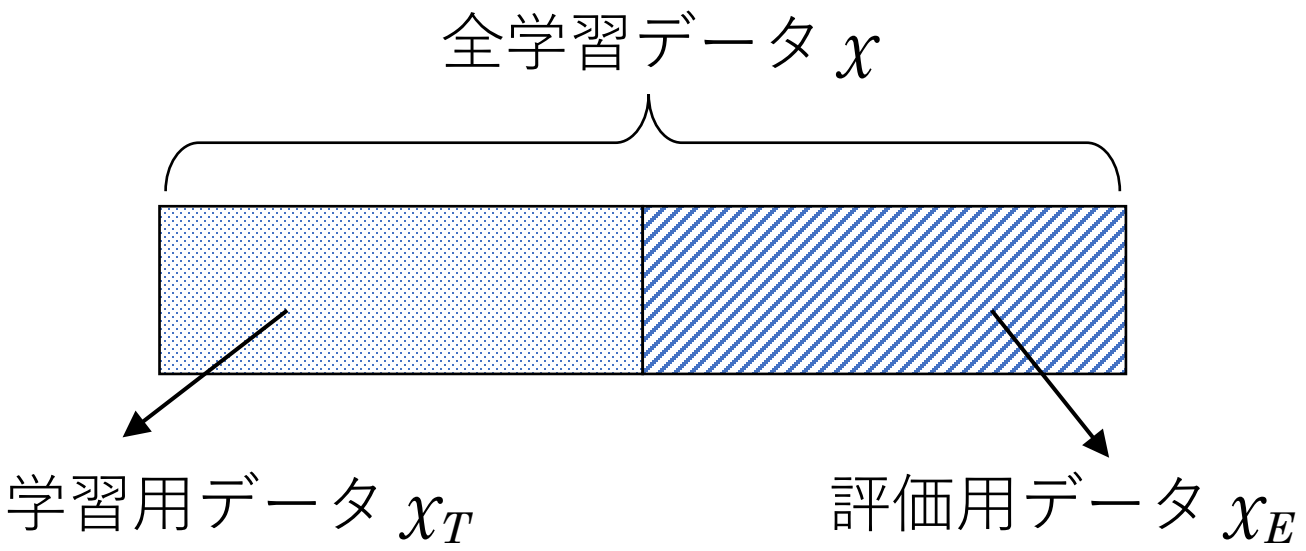
9.1　未知データに対する認識率の評価

- 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 を用いて識別率を推定する



9.1.1 分割学習法

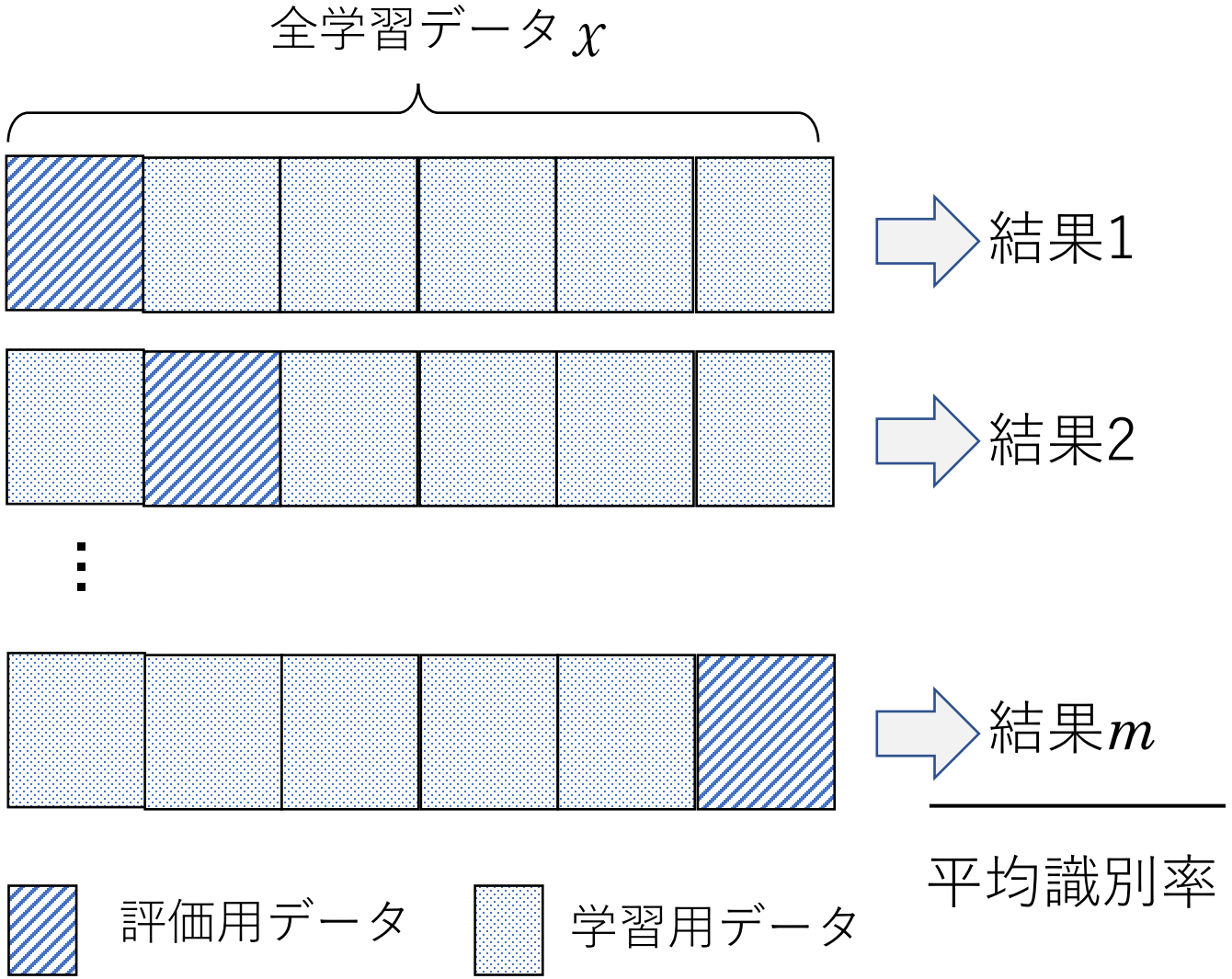
- 利点
  - 評価が容易
- 欠点
  - 学習に用いるデータ数が減るので、識別性能が低く見積もられる
  - 評価に用いるデータ数が少ない場合、識別率の推定精度は低い

9.1.2 交差確認法

- 手順

1. x をm個のグループ $x_1,\ldots,x_m$ に分割する
2. x_i を除いた$(m\text{-}1)$個のグループで学習し、x_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識別率の推定値とする

9.1.2 交差確認法

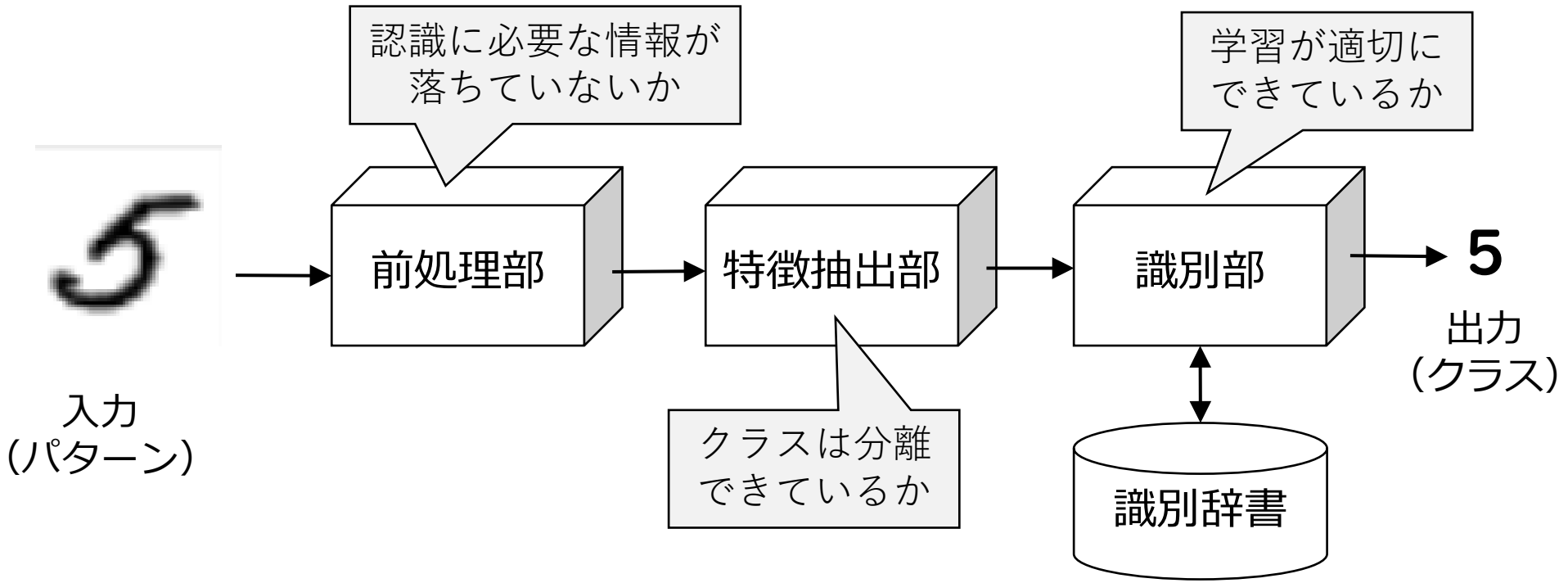


9.1.2 交差確認法

- 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、分割が異なると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
- 例題9.1

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



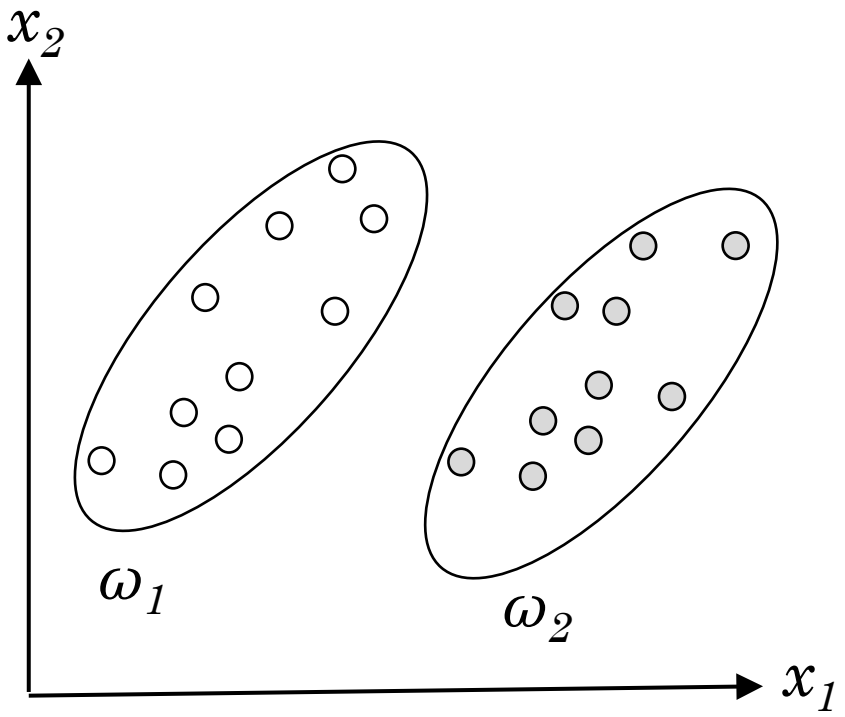
9.2.1 前処理部の確認

- 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認識率が低い場合

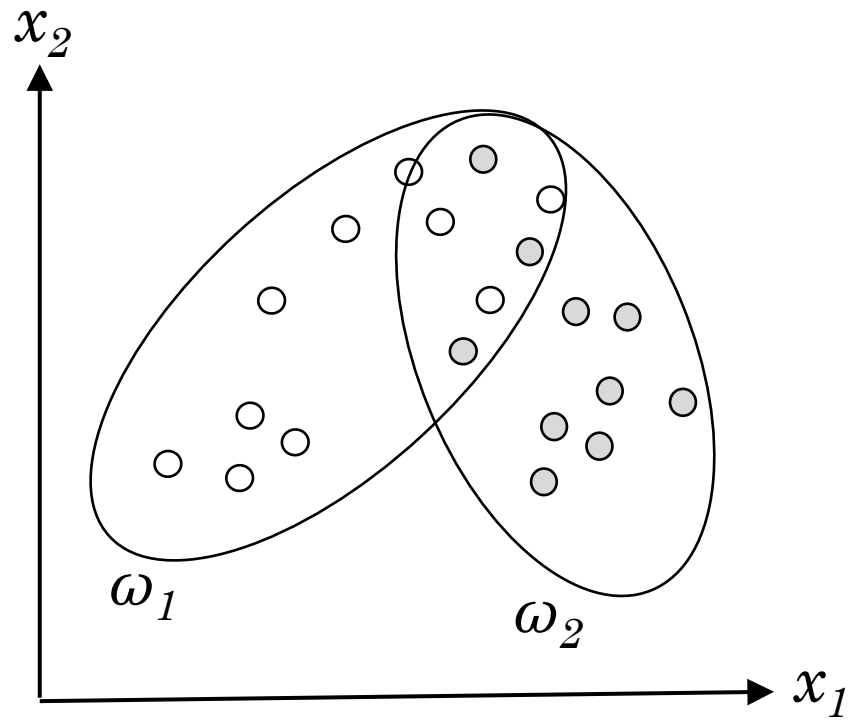
→ 識別部を再設計（識別関数の学習）



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）



9.2.2 特徴空間の評価

- クラス内分散・クラス間分散比
  - 選択した特徴の評価法
  - 特徴空間の評価法でクラス毎のデータの広がり方を評価する尺度
  - 同じクラスのデータはなるべく接近し、異なるクラスのデータ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

- クラス内分散

$$\sigma_W^2 = \frac{1}{n}\sum_{i=1}^{c}\sum_{\boldsymbol{x}\in\chi_i}(\boldsymbol{x}-\boldsymbol{m}_i)^T(\boldsymbol{x}-\boldsymbol{m}_i)$$

$\boldsymbol{m}_i$:クラス i の平均

- クラス間分散

$$\sigma_B^2 = \frac{1}{n}\sum_{i=1}^{c} n_i(\boldsymbol{m}_i-\boldsymbol{m})^T(\boldsymbol{m}_i-\boldsymbol{m})$$

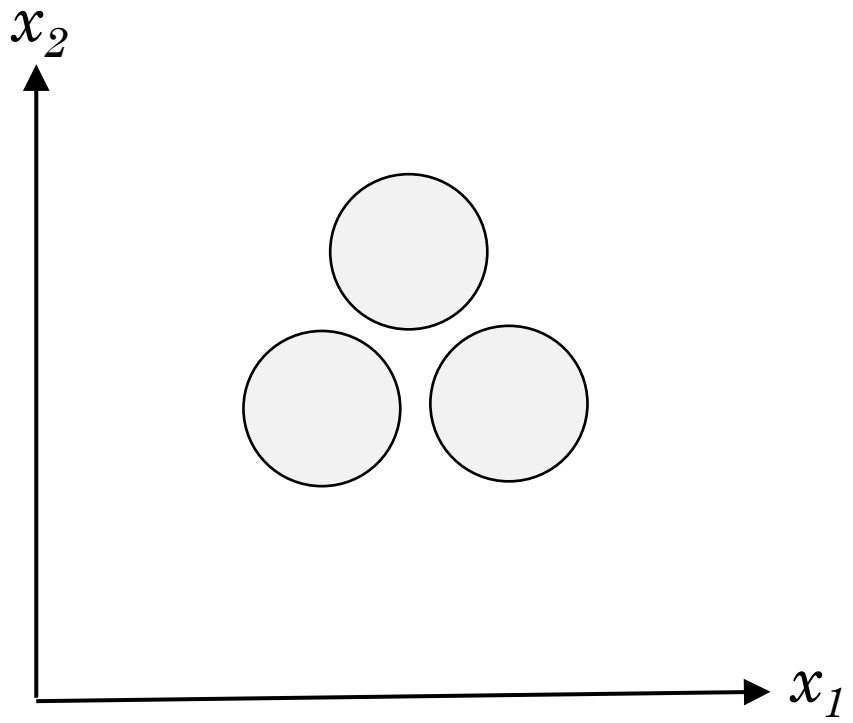
$\boldsymbol{m}$:全データの平均
n_i :クラス i のデータ数

- クラス内分散･クラス間分散比（大きいほど良い）

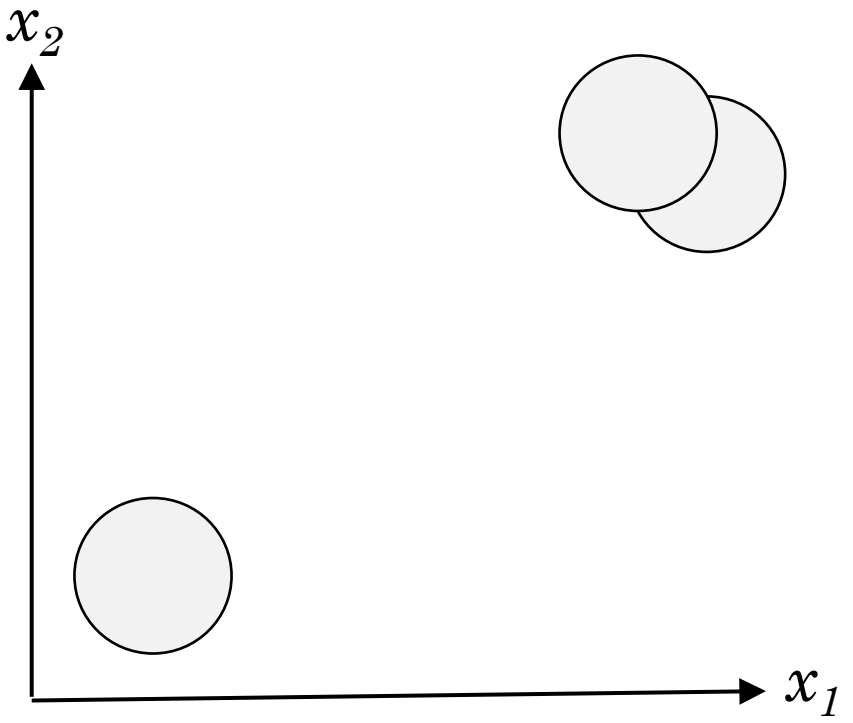
$$J_\sigma = \frac{\sigma_B^2}{\sigma_W^2}$$

9.2.2 特徴空間の評価

- 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



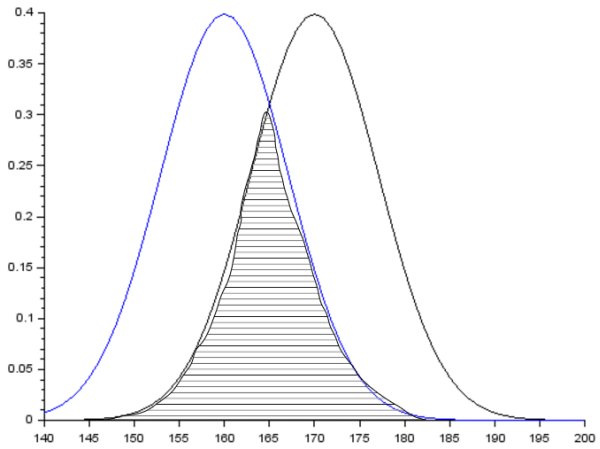
(a) クラス間分散：小



(a) クラス間分散：大

9.2.2 特徴空間の評価

- 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

- 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

- ベイズ誤り確率 e_B

$$e_B = \int e_B(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

$$= \int \min\{P(\omega_1|\boldsymbol{x}), P(\omega_2|\boldsymbol{x})\} p(\boldsymbol{x}) d\boldsymbol{x}$$

  - e_B は誤り確率をこれより小さくはできないという限界、すなわち分布の重なりを表す

9.2.2 特徴空間の評価

- ベイズ誤り確率：特徴の評価基準
  - 分布は一般に未知であるため、ベイズ誤り確率を直接推定することは困難

  →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

$$e_B \leq e_N \leq 2e_B$$

e_N：1-NN法の誤り確率

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α
- 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

- 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、学習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

9.2.3 識別部の調整

- 学習結果に影響を与えるパラメータ
  - モデルの複雑さに連続的に影響を与える

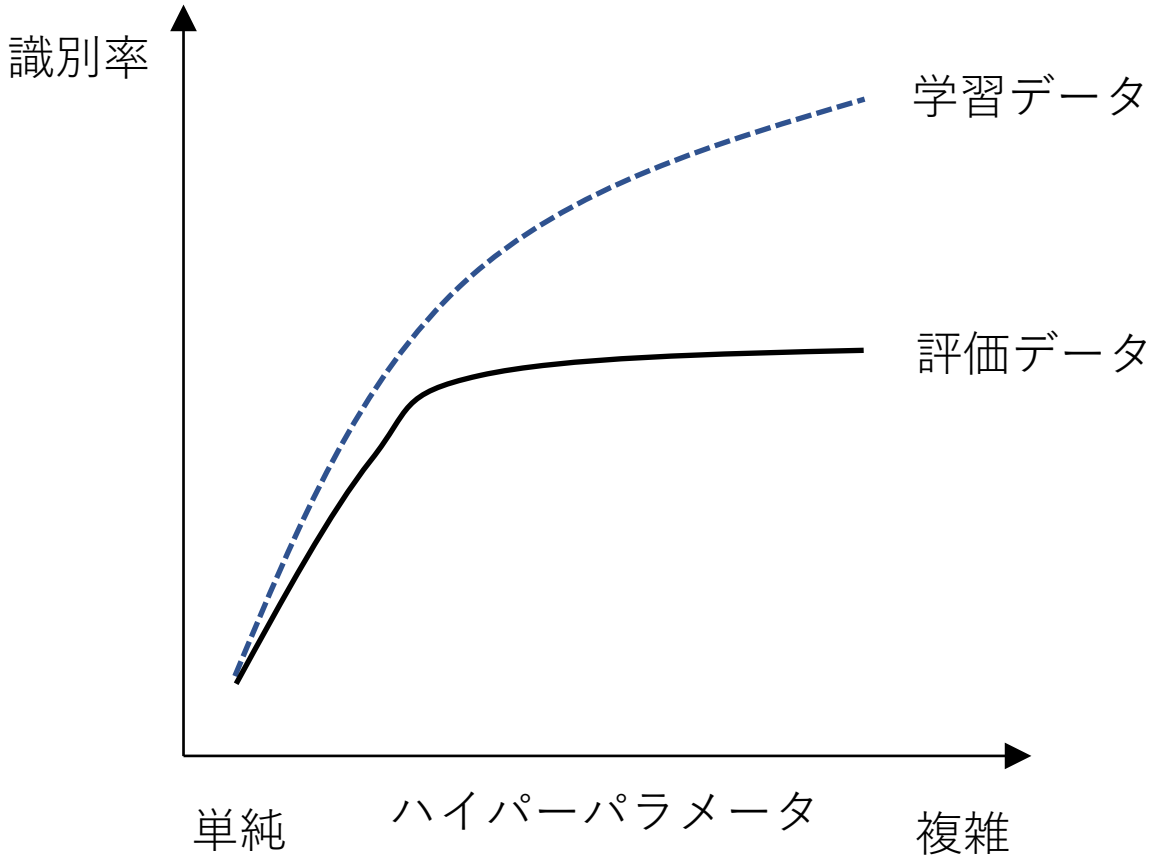
    → 性能に直結する
  - 例）SVMの多項式カーネルの次数、ガウシアンカーネルの半径γ
  - いくつかの異なる値で性能を評価する必要がある

9.2.3 識別部の調整

- 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。
  - 分割学習法や交差確認法で e_λ を求める。

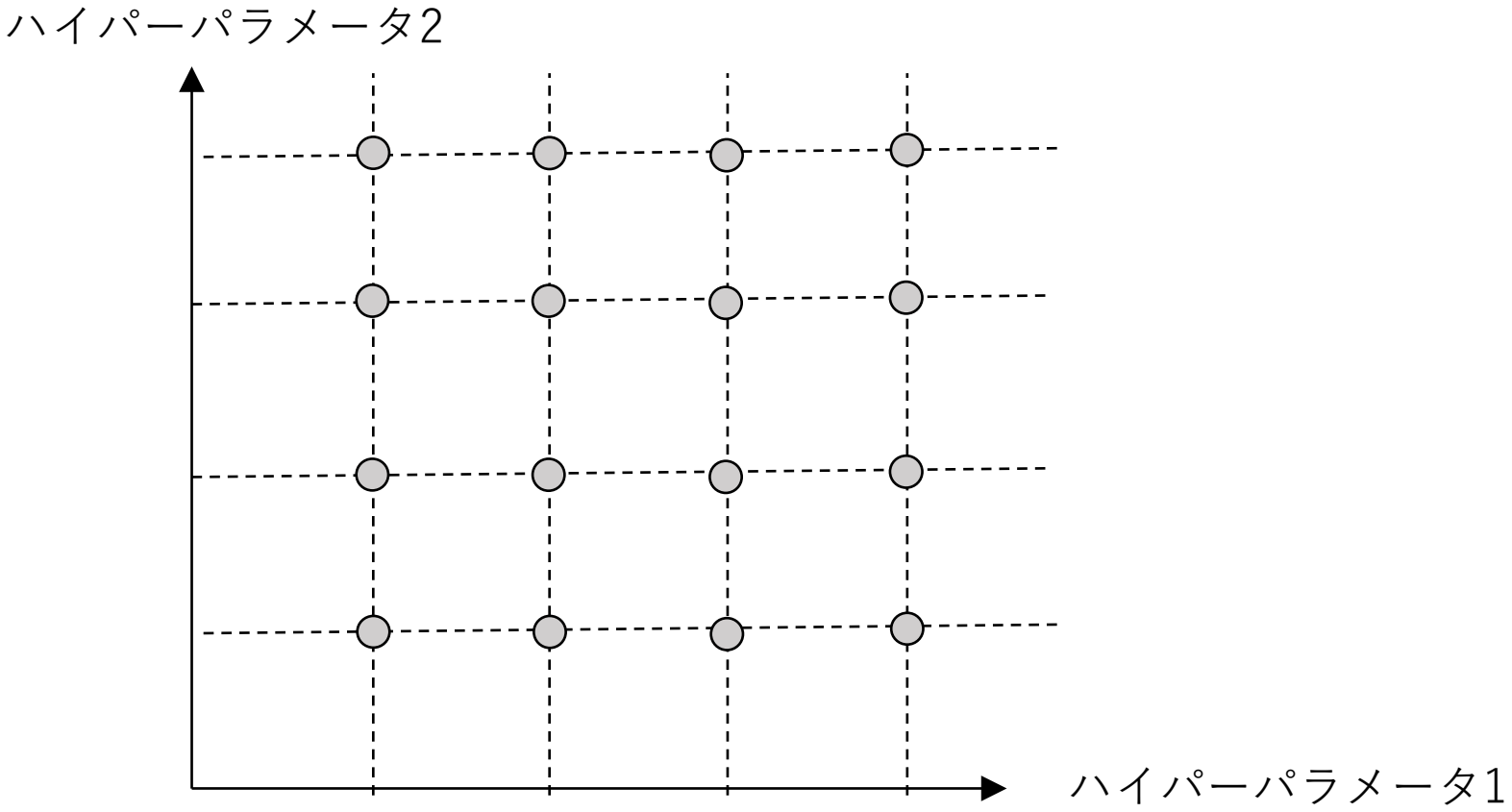
9.2.3 識別部の調整

- ハイパーパラメータの性質
  - 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

- 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



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