

参考書:

- [1] C. M. Bishop: “*Neural Networks for Pattern Recognition*,” Oxford University Press, 1995.
- [2] R. O. Duda, P. E. Hart, and D. G. Stork: “*Pattern Classification*,” Second Edition, John Wiley & Sons, 2001.
- [3] K. Fukunaga: “*Introduction to Statistical Pattern Recognition*,” Second Edition, Academic Press, 1990.
- [4] 石井, 上田, 前田, 村瀬: 「わかりやすいパターン認識」, オーム社, 1998年.

講義ノート: <http://cis.k.hosei.ac.jp/~wakahara>

講義內容：

1. Bayes' Decision Theory
2. Parametric Probability Density Estimation
3. Mixture Models and The EM Algorithm
4. The Multi-layer Perceptron
5. Radial Basis Functions
6. Support Vector Machines
7. Parameter Optimization in Error Functions
8. Learning and Generalization
9. Presentation of Your Project

講義の進め方:

1. パターン認識の理論的枠組みを把握する
→ 計算問題を解く
2. パターン認識の動作を確認する
→ コーディングする
3. 手書き数字認識の識別系を構築する
→ プロジェクト

評価方法:

出席点 20点 宿題 40点

プロジェクト 40点

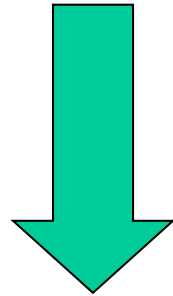
Bayes' Decision Theory

Toru Wakahara

- [1] パターン認識とは何か
- [2] 一つの例: 文字認識
- [3] 変数削減としての特徴選択
- [4] 統計的アプローチ
- [5] ベイズの定理
- [6] 決定境界とベイズの定理
- [7] 識別関数
- [8] 生成モデル vs. 識別モデル

パターン認識とは何か

人間が主観的／概念的に把握している
「パターン」を計算機で「認識」すること
“人工知能研究の夢”



工学的アプローチ
cf. 心理／神経生理学的

- ① 「パターン」を数量的に定義すること
- ② 「認識」する数学的手段を提供すること

パターン認識理論の枠組み

統計論的アプローチ

vs.

決定論的アプローチ

確率論的性質:

有限のパターン集合

確率的な認識判断

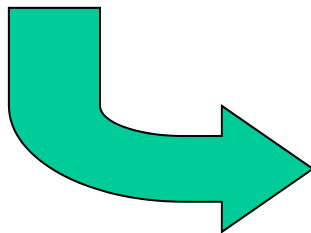
cf. 量子力学

決定論的振舞い:

パターン生成規則

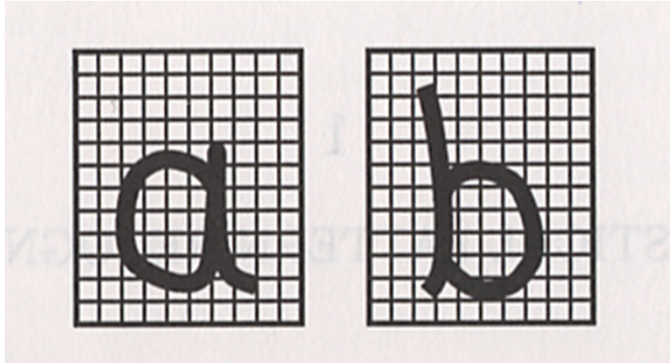
パターン変形規則

cf. 古典力学



最も汎用的かつ自然な枠組み
“統計的パターン認識理論”

一つの例：文字認識 ‘a’ と ‘b’ の識別

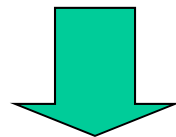


C_1 : ‘a’ C_2 : ‘b’

$$\mathbf{x} = (x_1 \ x_2 \ \dots \ x_d)^t$$
$$x_i \in [0, 255]$$

256×256 サイズ画像で 8-bit 濃淡レベル

→ 可能な画像の数 $2^{8 \times 256 \times 256} \cong 10$ の？乗



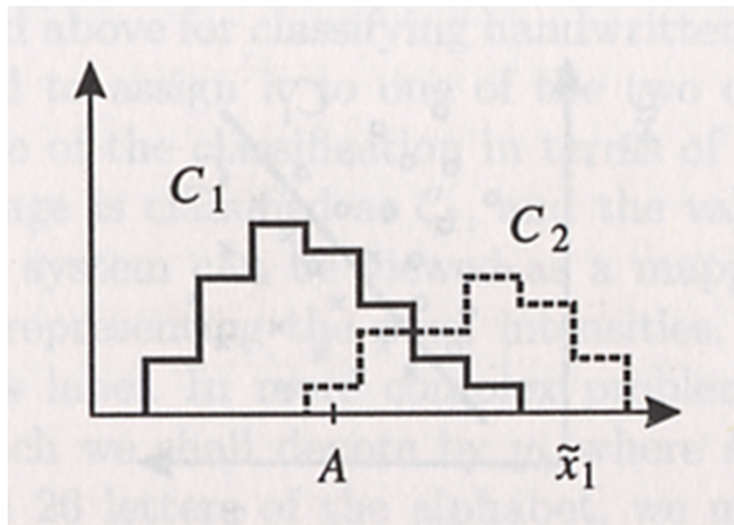
可能な全ての画像を記憶できない

限られた標本集合から未知サンプルを
識別する分類器をつくる → 汎化

変数削減としての特徴選択

features の選択:

- ① 対象の知識に基づく発見法的な手法
- ② 数学的な変数変換による次元圧縮



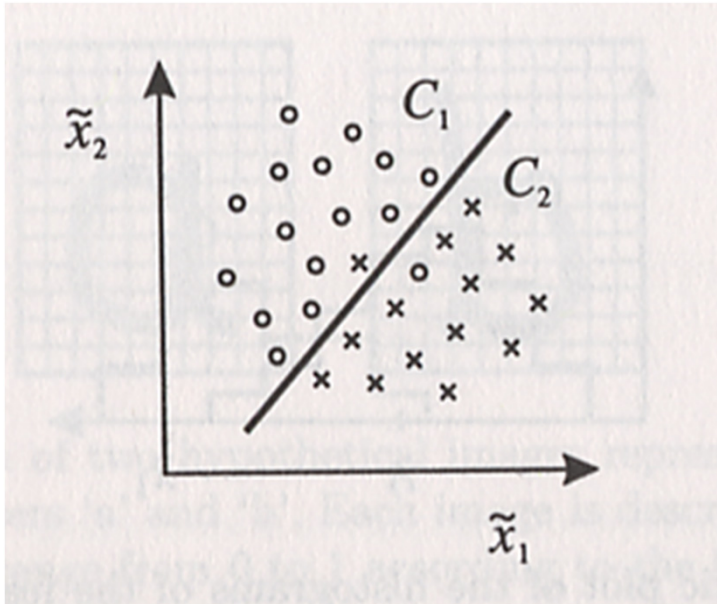
\tilde{x}_1 : 文字の縦/横比

← 実現値のヒストグラム

$$\tilde{x}_1 = A \quad \longrightarrow$$

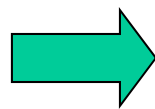
C_1, C_2 のどちらに属すると判断するか？ その根拠は？

変数増加の効果－Feature selection



- ① \tilde{x}_2 の追加
→ 安定な決定境界
- ② \tilde{x}_1, \tilde{x}_2 単独では
識別困難

変数増加で識別能力は単調に向上するか？

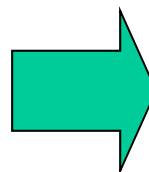


飽和点あり → 次元の呪い
最小誤り確率による判断が本質

パターン認識への統計的アプローチ

文字認識 → 誤り確率が最小となるように
入力文字のカテゴリ 'a', 'b' を決定

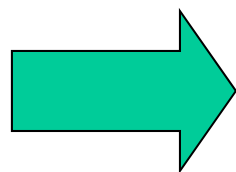
膨大な数のサンプルを収集
→ 'a' と 'b' の出現頻度
ex. 'a' は 'b' の1.5倍出現



事前確率

$$P(C_1) = 0.60$$

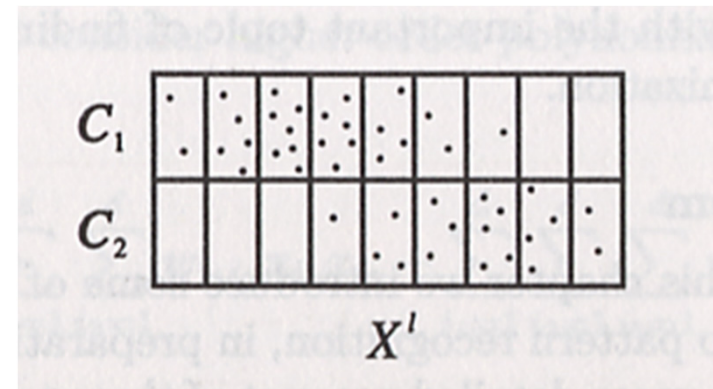
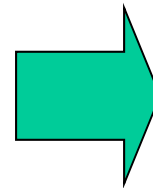
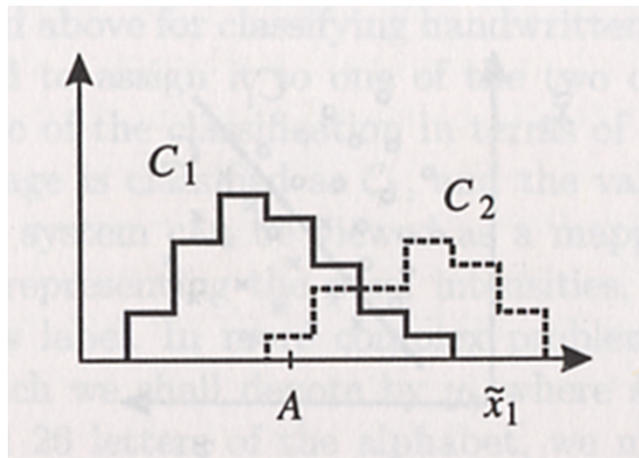
$$P(C_2) = 0.40$$



もし入力文字を観測せずに
認識しろと強制されたら最善策は？
その根拠は？

事後確率に基づく認識

もし入力文字を観測して $\tilde{x}_1 = A$ であることが分かったら、 C_1, C_2 のどちらに属すると判断するのが最善策か？



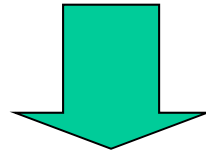
この図に含まれる情報から事後確率

$P(C_1 | \tilde{x}_1), P(C_2 | \tilde{x}_1)$ が推論できるだろうか？

ベイズの定理 (Bayes' theorem)

$$P(C_k, X^l) = P(X^l | C_k) P(C_k)$$

$$P(C_k, X^l) = P(C_k | X^l) P(X^l)$$



$$P(C_k | X^l) = \frac{P(X^l | C_k) P(C_k)}{P(X^l)}$$

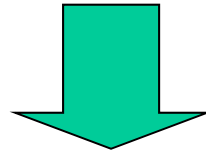
事後確率

Bayes' theorem

ベイズの定理 (Bayes' theorem)

$$P(C_1 | X^l) + P(C_2 | X^l) = 1$$

$$\therefore P(X^l) = P(X^l | C_1)P(C_1) + P(X^l | C_2)P(C_2)$$

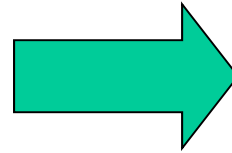
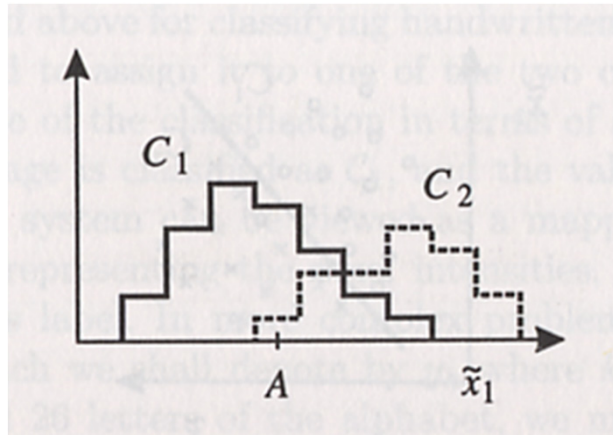


$$P(C_k | X^l) = \frac{P(X^l | C_k)P(C_k)}{P(X^l | C_1)P(C_1) + P(X^l | C_2)P(C_2)}$$

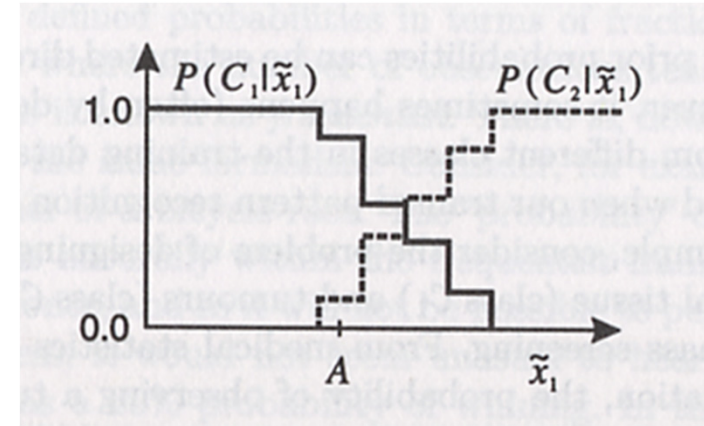
事後確率

Bayes' theorem

推論と決定 — Inference and decision



$$P(C_1) = 0.6$$
$$P(C_2) = 0.4$$



$$P(\tilde{x}_1 | C_1), P(\tilde{x}_1 | C_2)$$

$$P(C_1 | \tilde{x}_1), P(C_2 | \tilde{x}_1)$$

- ① 推論段階: 事前確率とクラス条件付確率を用いて
→ 事後確率を推論する
- ② 決定段階: 事後確率最大のクラスに決定すると
→ 誤り確率が最小となる

確率から確率密度へ

離散変数 \tilde{x}_1 から連続変数 x_1 への移行:

$$P(x \in [a, b]) = \int_a^b p(x) dx$$

$\mathbf{x} = (x_1, \dots, x_d)^t$ の場合:

$$P(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}) d\mathbf{x}$$

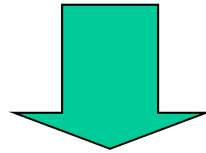
期待値:

$$E[Q] \equiv \int Q(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \cong \frac{1}{N} \sum_{n=1}^N Q(\mathbf{x}_n)$$

ベイズの定理の一般形

$$P(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) P(C_k)}{\sum_{j=1}^c p(\mathbf{x} | C_j) P(C_j)}$$

事後確率



$$posterior = \frac{likelihood \times prior}{normalization factor}$$

Bayes' theorem in general

最適な決定境界 (1)

入力文字の \mathbf{x} を観測して 'a', 'b' に分類する際
誤り確率を最小とする最適な決定境界は？

$$P(\text{error}) = P(\mathbf{x} \in \mathfrak{R}_2, C_1) + P(\mathbf{x} \in \mathfrak{R}_1, C_2)$$

$$= P(\mathbf{x} \in \mathfrak{R}_2 | C_1) P(C_1) + P(\mathbf{x} \in \mathfrak{R}_1 | C_2) P(C_2)$$

$$= \int_{\mathfrak{R}_2} p(\mathbf{x} | C_1) P(C_1) d\mathbf{x} + \int_{\mathfrak{R}_1} p(\mathbf{x} | C_2) P(C_2) d\mathbf{x}$$

→ min for $\mathfrak{R}_1, \mathfrak{R}_2$

最適な決定境界 (2)

$$P(\text{error}) = P(\mathbf{x} \in \mathfrak{R}_2, C_1) + P(\mathbf{x} \in \mathfrak{R}_1, C_2)$$

$$= \int_{\mathfrak{R}_2} p(\mathbf{x} | C_1) P(C_1) d\mathbf{x} + \int_{\mathfrak{R}_1} p(\mathbf{x} | C_2) P(C_2) d\mathbf{x}$$

$$= \int_{\mathfrak{R}_1 + \mathfrak{R}_2} p(\mathbf{x} | C_1) P(C_1) d\mathbf{x}$$

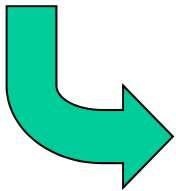
$$+ \int_{\mathfrak{R}_1} [p(\mathbf{x} | C_2) P(C_2) - p(\mathbf{x} | C_1) P(C_1)] d\mathbf{x}$$

$$= P(C_1) + \int_{\mathfrak{R}_1} [p(\mathbf{x} | C_2) P(C_2) - p(\mathbf{x} | C_1) P(C_1)] d\mathbf{x}$$

→ min for $\mathfrak{R}_1, \mathfrak{R}_2$

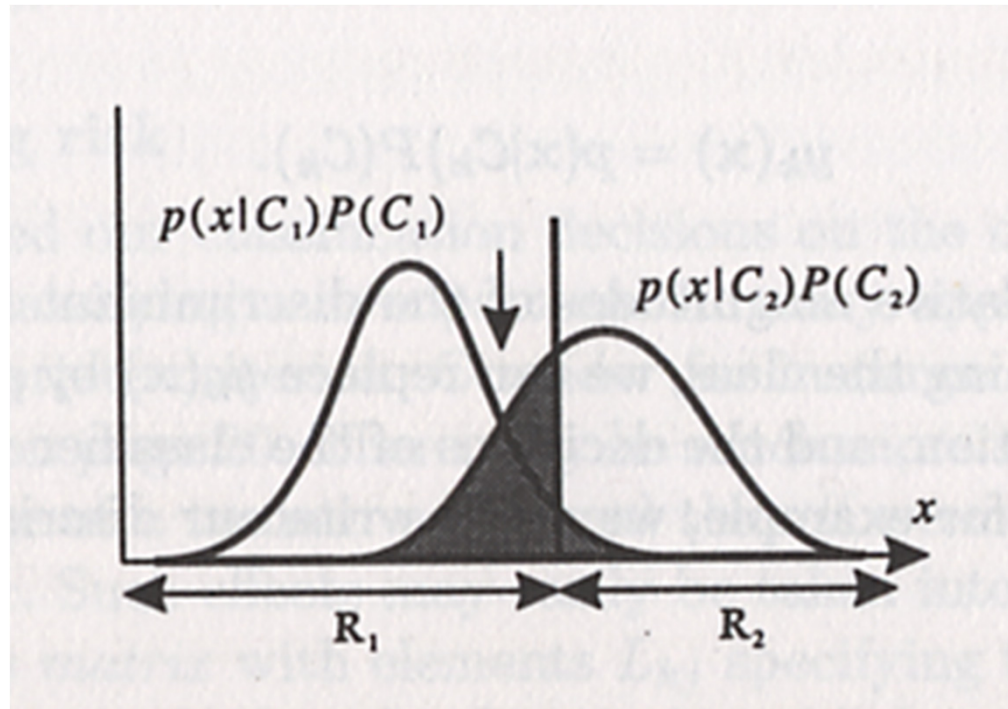
最適な決定境界 (3)

$$\begin{aligned} P(\text{error}) &= P(\mathbf{x} \in \mathfrak{R}_2, C_1) + P(\mathbf{x} \in \mathfrak{R}_1, C_2) \\ &= P(C_1) + \int_{\mathfrak{R}_1} [p(\mathbf{x} | C_2) P(C_2) - p(\mathbf{x} | C_1) P(C_1)] d\mathbf{x} \\ &= P(C_1) + \int_{\mathfrak{R}_1} [P(C_2 | \mathbf{x}) - P(C_1 | \mathbf{x})] p(\mathbf{x}) d\mathbf{x} \\ &\rightarrow \min \text{ for } \mathfrak{R}_1, \mathfrak{R}_2 \end{aligned}$$

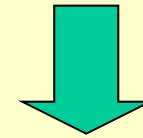


$$\begin{aligned} P(C_1 | \mathbf{x}) > P(C_2 | \mathbf{x}) &\rightarrow \mathbf{x} \in \mathfrak{R}_1 \\ P(C_2 | \mathbf{x}) > P(C_1 | \mathbf{x}) &\rightarrow \mathbf{x} \in \mathfrak{R}_2 \end{aligned}$$

決定境界とベイズ決定理論



$P(\text{error})$ は
影部分の面積



最小化する

$\mathcal{R}_1, \mathcal{R}_2$

$$p(\mathbf{x} | C_1)P(C_1) > p(\mathbf{x} | C_2)P(C_2) \rightarrow \mathbf{x} \in \mathcal{R}_1$$

$$p(\mathbf{x} | C_2)P(C_2) > p(\mathbf{x} | C_1)P(C_1) \rightarrow \mathbf{x} \in \mathcal{R}_2$$

多クラスでの決定境界

$$\begin{aligned} P(\text{correct}) &= \sum_{k=1}^c P(\mathbf{x} \in \mathfrak{R}_k, C_k) \\ &= \sum_{k=1}^c P(\mathbf{x} \in \mathfrak{R}_k | C_k) P(C_k) \\ &= \sum_{k=1}^c \int_{\mathfrak{R}_k} p(\mathbf{x} | C_k) P(C_k) d\mathbf{x} \rightarrow \text{max for } \{\mathfrak{R}_k\} \end{aligned}$$

$$p(\mathbf{x} | C_k) P(C_k) > p(\mathbf{x} | C_j) P(C_j)$$

$$\text{for } \forall j \neq k \rightarrow \mathbf{x} \in \mathfrak{R}_k$$

識別関数

識別関数: $y_1(\mathbf{x}), \dots, y_c(\mathbf{x})$

$$y_k(\mathbf{x}) > y_j(\mathbf{x}) \text{ for } \forall j \neq k$$

→ \mathbf{x} is assigned to class C_k

誤り確率最小化:

ベイズ決定理論より識別関数は次式でOK

$$y_k(\mathbf{x}) = p(\mathbf{x} | C_k) P(C_k)$$

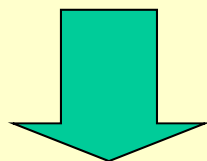
$$y_k(\mathbf{x}) = \ln p(\mathbf{x} | C_k) + \ln P(C_k)$$

2クラス問題での識別関数

$$y_1(\mathbf{x}) = \ln p(\mathbf{x} | C_1) + \ln P(C_1)$$

$$y_2(\mathbf{x}) = \ln p(\mathbf{x} | C_2) + \ln P(C_2)$$

$$y(\mathbf{x}) = y_1(\mathbf{x}) - y_2(\mathbf{x}) = \ln \frac{p(\mathbf{x} | C_1)}{p(\mathbf{x} | C_2)} + \ln \frac{P(C_1)}{P(C_2)}$$



$y(\mathbf{x}) > 0 \rightarrow \mathbf{x}$ is assigned to class C_1

$y(\mathbf{x}) < 0 \rightarrow \mathbf{x}$ is assigned to class C_2

リスクの導入

L_{kj} : 本当は C_k に属するパターンを
誤って C_j に分類した際の損失
 C_k に属するパターンに対する期待損失:

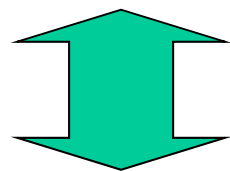
$$R_k = \sum_{j=1}^c L_{kj} \int_{\mathcal{R}_j} p(\mathbf{x} | C_k) d\mathbf{x}$$

$$R = \sum_{k=1}^c R_k P(C_k) = \sum_{j=1}^c \int_{\mathcal{R}_j} \left\{ \sum_{k=1}^c L_{kj} p(\mathbf{x} | C_k) P(C_k) \right\} d\mathbf{x}$$

リスク最小化での決定境界

$$R = \sum_{j=1}^c \int_{\mathfrak{R}_j} \left\{ \sum_{k=1}^c L_{kj} p(\mathbf{x} | C_k) P(C_k) \right\} d\mathbf{x}$$

$\rightarrow \min \text{ for } \{ \mathfrak{R}_j \}$



$$\sum_{k=1}^c L_{kj} p(\mathbf{x} | C_k) P(C_k) < \sum_{k=1}^c L_{ki} p(\mathbf{x} | C_k) P(C_k)$$

for $\forall i \neq j \rightarrow \mathbf{x} \in \mathfrak{R}_j$

統計的パターン認識における 生成モデル vs. 識別モデル(1)

■ クラス分類問題における推論段階と決定段階

① 推論段階

学習データを用いて**事後確率** $P(C_k | \mathbf{x})$ を学習する。
そのため、**クラス条件付き確率密度** $p(\mathbf{x} | C_k)$ と
事前確率 $P(C_k)$ を推論する問題を扱う。

② 決定段階

観測ベクトル \mathbf{x} に対して、**事後確率** $P(C_k | \mathbf{x})$ が
最大となる最適なクラス C_j に割り当てを行う。

統計的パターン認識における 生成モデル vs. 識別モデル(2)

■ 統計的パターン認識における生成モデル

クラス条件付き確率密度 $p(x|C_k)$ と
事前確率 $P(C_k)$ のそれぞれを, 学習データを用いて, 直接, 推論する問題を解く。

■ 生成モデルの得失

1. x は高次元なので $p(x|C_k)$ を高精度で求めるには多くの学習データ(次元数の2桁以上)が必要となる。
2. 周辺分布 $p(x)$ から外れ値が検出できる。

統計的パターン認識における 生成モデル vs. 識別モデル(3)

■ 統計的パターン認識における識別モデル

推論と決定の段階を同時に扱い、観測ベクトル x を割り当てるクラスを直接決定する**識別関数**を学習する。理想的には**事後確率** $P(C_k | x)$ の直接推論になる。

■ 識別モデルの得失

1. 学習データは比較的少なくても済む。ただし、クラス決定境界の近傍データが必要となる。
2. リスク最小化や棄却オプションが扱えない。

課題1

x を M 個のクラス C_i ($i = 1, \dots, M$) のいずれかに分類する問題は、ベイズの定理によれば

$$P(C_{max} | \mathbf{x}) \geq P(C_i | \mathbf{x}) \quad \text{for } i = 1, 2, \dots, M$$

を満たすクラス C_{max} に分類すれば、誤り確率が最小となる。

このとき、平均誤り確率 P_e の上限を求める。

[1] まず次式(1)が成り立つことを示し、等号が成り立つのはどんな場合か述べなさい。

$$P(C_{max} | \mathbf{x}) \geq \frac{1}{M} \quad (1)$$

課題1 (続き)

[2] 平均誤り確率 P_e が, \mathbf{x} の確率密度関数 $p(\mathbf{x})$ と最大事後確率 $P(C_{max} | \mathbf{x})$ を用いて, 次式で表せることを示しなさい。

$$P_e = 1 - \int P(C_{max} | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad (2)$$

[3] 式(1), (2) を用いて, 平均誤り確率 P_e の上限が次式で与えられることを示し, 等号が成り立つのはどんな場合か述べなさい。

$$P_e \leq \frac{M-1}{M} \quad (3)$$