パターン認識

早稲田大学講義 – 平成19年度

(独) 産業技術総合研究所　脳神経情報研究部門
栗田多喜夫、赤穂昭太郎

I. パターン認識とベイズ識別

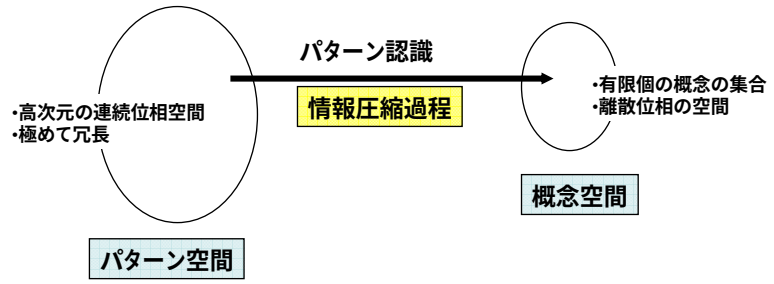
パターン認識、ベイズ決定理論、確率密度分布の推定

パターン認識問題の例

- スパムメイルを検出して、自動削除する
  - 特徴抽出
    - メイル本文やヘッダにどのような単語が現れているかの頻度を計測し、それらをまとめて特徴ベクトルとする
  - 訓練用のサンプルの作成
    - 過去のメイルのデータベースから特徴ベクトルを計測し、そのメイルがスパムかどうかを記録し、そのペアを訓練用サンプルデータとする
  - 識別器の学習
    - 訓練用のサンプルを用いて識別器のパラメータを学習する
  - 運用
    - 新たなメイルから特徴ベクトルを計測し、それを識別器に入力し、その結果がスパムであれば、そのメイルをスパムフォールだに移動する

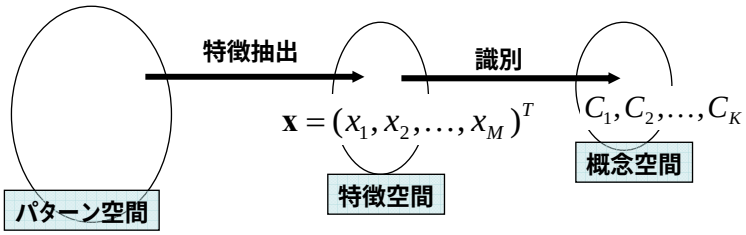
パターン認識とは

- パターン認識
  - 認識対象がいくつかの概念に分類出来るとき、観測されたパターンをそれらの概念（クラスあるいは類）のうちのひとつに対応させる処理
    - 数字の認識： 入力パターンを10種類の数字のいずれかに対応させる
    - 顔画像の識別： 顔画像から誰であるかを推定する



パターン認識過程

- 特徴抽出
  - 認識対象から何らかの特徴量を計測（抽出）する必要がある
  - 認識に有効な情報（特徴）を抽出し、次元を縮小した効率の良い空間を構成する過程
    - 文字認識： スキャナ等で取り込んだ画像から文字の識別に必要な本質的な特徴のみを抽出（例、文字線の傾き、曲率、面積など）
- 識別
  - 与えられた未知の対象を、特徴とクラスの関係に関する知識に基づいて、どのクラスに属するかを決定（判定）する過程



パターン認識の基本課題

- 識別方式の開発
  - 未知の認識対象を観測して得られる特徴ベクトルからその対象がどのクラスに属するかを判定する方法
- 一般的なアプローチ
  - 教師あり学習
    - クラスの帰属が既知の学習用のサンプル集合から特徴ベクトルとクラスとの確率的な対応関係を知識として学習
  - 識別
    - 学習された特徴ベクトルとクラスとの対応関係に関する確率的知識を利用して、与えられた未知の認識対象を特徴ベクトルからその認識対象がどのクラスに属していたかを推定（決定）

ベイズ決定理論

- ベイズ識別方式
  - 特徴ベクトルとクラスとの確率的な対応関係が完全にわかっている理想的な場合の理論
  - 未知の認識対象を誤って他のクラスに識別する確率（誤識別率）を出来るだけ小さくするような識別方式
  - 誤識別率の意味で理論的に最適な識別方式
- 例示：身長から男か女かを当てる

事前確率・条件付き確率

- 事前確率（先見確率）
  - クラス　の確率

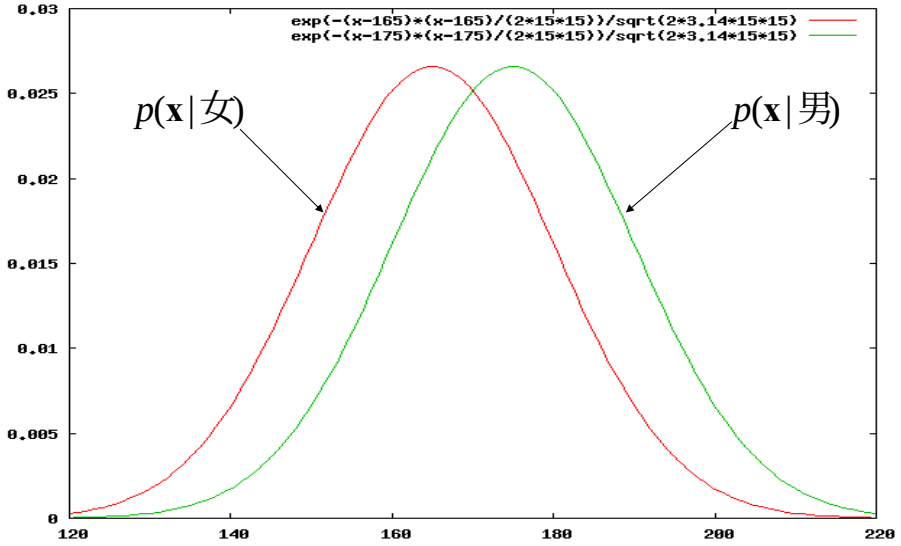
$$P(C_k) \qquad \sum_{k=1}^{K} P(C_k) = 1$$

- 特徴ベクトルの条件付き確率
  - あるクラスに属する対象を観測したとき、その特徴ベクトルが観測される確率密度分布

$$p(\mathbf{x} \mid C_k) \qquad \int p(\mathbf{x} \mid C_k) d\mathbf{x} = 1$$

  - これらの確率がわかれば、特徴ベクトルとクラスとの確率的な関係は全て計算できる。

身長に関する条件付密度分布



事後確率

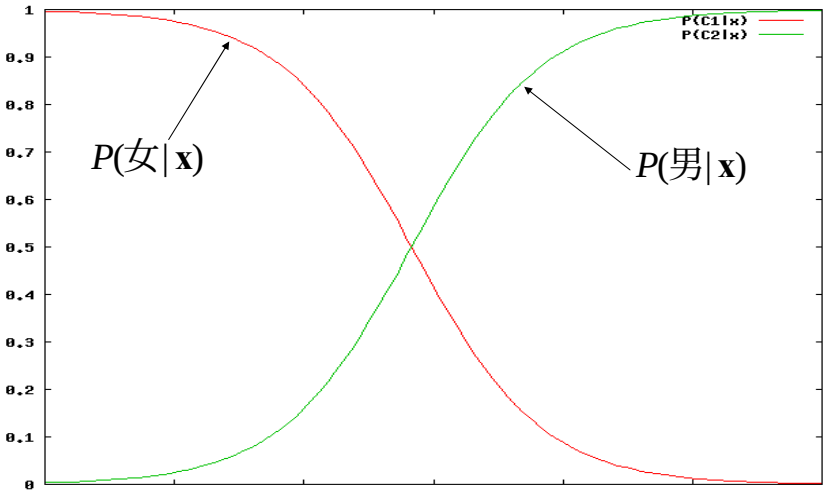
- 事後確率
  - ある対象から特徴ベクトルが観測されたとき、その対象がクラスCkに属している確率

$$P(C_k \mid \mathbf{x}) = \frac{P(C_k)p(\mathbf{x} \mid C_k)}{p(\mathbf{x})} \qquad \sum_{k=1}^{K} P(C_k \mid \mathbf{x}) = 1$$

ここで、特徴ベクトルの確率密度分布は、

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(C_k)p(\mathbf{x} \mid C_k) \qquad \int p(\mathbf{x}) = 1$$

身長に関する事後確率



期待損失

- **決定関数**
  - **特徴ベクトルに基づき対象がどのクラスに属するかを決定する関数**

$$d(\mathbf{x})$$

  - 損失関数
    - クラスC_kの対象をクラスC_jに決定したときの損失

$$r(C_j \mid C_k)$$

  - 期待損失（平均損失）

$$R[d] = \sum_{k=1}^{K} \int r(d(\mathbf{x}) \mid C_k) P(C_k \mid \mathbf{x}) \, p(\mathbf{x}) d\mathbf{x}$$

これを最小とする決定関数を求めるのがベイズ決定理論

0－1損失の場合

- 0-1損失
  - 誤った識別に対して均等な損失を与える

$$r(C_j \mid C_k) = 1 - \delta_{jk}$$

◆ 最適な識別関数（ベイズ識別方式）

■ 期待損失を最小とする最適な識別関数

$$d(\mathbf{x}) = C_k \quad \text{if} \quad P(C_k \mid \mathbf{x}) = \max_j P(C_j \mid \mathbf{x})$$

これは、事後確率が最大となるクラスに決定する識別方式

◆ 最小誤識別率

■ ベイズ識別方式により達成される最小誤識別率

$$P_e^* = 1 - \int \max_j P(C_j \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

2クラス（0－1損失）の場合

- 最適な識別方式
  - 事後確率の大小を比較すればよい

$$d(\mathbf{x}) = C_1 \quad \text{if} \quad P(C_1 \mid \mathbf{x}) \geq P(C_2 \mid \mathbf{x})$$

$$d(\mathbf{x}) = C_2 \quad \text{otherwise}$$

◆ 尤度比検定

$$d(\mathbf{x}) = C_1 \quad \text{if} \quad \frac{p(\mathbf{x} \mid C_1)}{p(\mathbf{x} \mid C_2)} \geq \theta$$

$$d(\mathbf{x}) = C_2 \quad \text{otherwise}$$

ここで、閾値は、$\theta = \dfrac{P(C_2)}{P(C_1)}$

正規分布の場合

- **確率密度分布**

$$p(\mathbf{x}|C_k)=\frac{1}{(\sqrt{2\pi})^M|\Sigma_k|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}$$

  - 2次の識別関数
    - 事後確率の対数

$$g_k(\mathbf{x})=\log P(C_k)-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)-\frac{1}{2}\log|\Sigma_k|$$

  - 線形識別関数
    - 各クラスの共分散行列が等しい場合

$$g_k(\mathbf{x})=\mu_k{}^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}\mu_k{}^T\Sigma^{-1}\mu_k+\log P(C_k)=\mathbf{w}_k^T\mathbf{x}-h_k$$

等方的な正規分布の場合

- クラスが2つで、各クラスの共分散行列が等しい場合

$$\phi(\mathbf{x})=g_1(\mathbf{x})-g_2(\mathbf{x})=(\mu_1-\mu_2)^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}(\mu_1{}^T\Sigma^{-1}\mu_1-\mu_2{}^T\Sigma^{-1}\mu_2)+\log\frac{P(C_1)}{P(C_2)}=\mathbf{w}^T\mathbf{x}-h$$

- クラスが2つで、各クラスの共分散行列が等しく、等方的な場合

$$g_k(\mathbf{x})=\log P(C_k)-\frac{\|\mathbf{x}-\mu_k\|^2}{2\sigma^2}$$

これは、先見確率が等しい場合には、特徴ベクトルと各クラスの平均ベクトルとの距離が最も近いクラスに決定する識別方式
つまり、各クラスの平均ベクトルをテンプレートと考えると、特徴ベクトルと各クラスのテンプレートとのマッチングによる識別

パターン認識問題の例

- ロボット
  - 顔、声から誰かを識別、音声から何を喋っているかを認識、手で触って、状態（柔らかい、硬い）を判定
- 車
  - 対向車や人の検出、運転者の状態（眠い、テンションがあがっている、、、）
- メイル
  - スパムメイルを検出して、自動削除する
- 医療
  - 検査結果から病気を推定（肺がん）
- 軍事
  - ソナーデータから潜水艦かどうかを識別
- ワイン
  - 成分からワインの種類を識別

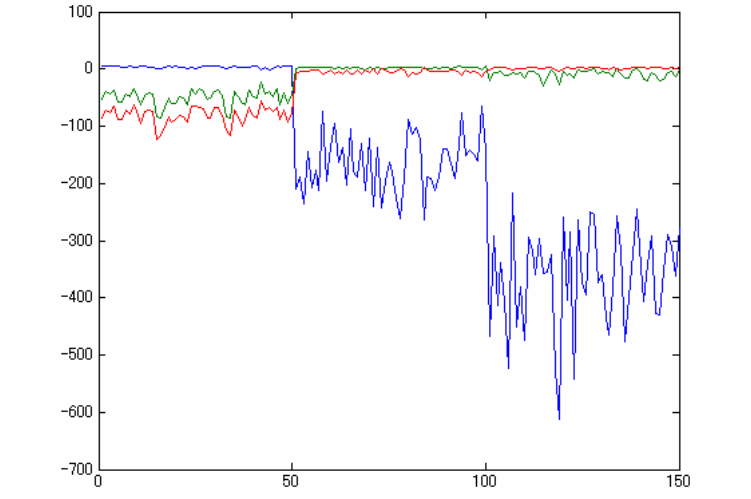
Fisherのアヤメのデータの識別課題

- 3種類のアヤメ
  - Setosa, Versicolor,Virginica
- 計測した特長
  - ガクの長さ、ガクの幅、花びらの長さ、花びらの幅
- 訓練用サンプル
  - 各アヤメそれぞれ50サンプルを収集
  - 合計150サンプル（50x3）
- 問題
  - ガクの長さ、ガクの幅、花びらの長さ、花びらの幅を計測して、どのアヤメかを推測する識別装置を設計すること

ベイズ決定則によるアヤメの識別

- データの表示
  - プログラム (testpca.m)
- ベイズ識別のための準備
  - 損失関数： 0－1識別の場合を考える
  - 確率分布の推定
  - 各クラスの事前確率は、等確率 (1／3) とする
  - 各アヤメから特徴ベクトルが得られる確率は正規分布と仮定
  - 正規分布のパラメータは、サンプル平均、サンプル分散共分散行列として推定
  - 識別関数の設計＜＝黒板で説明
  - ＝＞プログラム (bayes_iris.m)

アヤメのデータの識別結果



確率密度分布の推定

- ベイズ決定理論
  - 期待損失最小の意味で最適な識別方式
    しかし、、、
  - 各クラスと特徴ベクトルとの確率的な関係が完全にわかっていないと使えない！！！

  ＝＞ 訓練用のデータからデータの背後の確率的な関係を推定（確率密度分布の推定）

- 確率密度分布の推定法
  - パラメトリックモデルを用いる方法
    - 比較的少数のパラメータをもつモデル（パラメトリックモデル）を用いて確率分布を表現し、そのモデルをデータに当てはめ、データと尤も良く合うパラメータを推定
  - ノンパラメトリックモデルを用いる方法
    - 特定の関数型を仮定しないで、データに依存して分布の形を決める方法
  - セミパラメトリックな手法
    - 複雑な分布を表現するためにパラメータの数を系統的に増やせるようにすることで、パラメトリックモデルよりも一般的な関数型を表現できるようにする手法

パラメトリックモデル

- パラメトリックモデルによる確率密度分布の推定
  - モデル化
    - 確率密度分布をいくつかのパラメータを用いて表現
    - 正規分布：最も簡単で、最も広く用いられているパラメトリックモデル

$$p(\mathbf{x}|C_k)=\frac{1}{(\sqrt{2\pi})^M|\Sigma_k|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}$$

  - パラメータの推定法
    - 最尤推定法 (maximum likelihood method)
      - パラメータを未知の固定値だとみなし、実際に観測された訓練データが得られる確率を最大化するようにパラメータを推定
    - ベイズ推定 (Bayesian inference)
      - パラメータを既知の事前分布を持った確率変数だとみなし、パラメータの値の確信度をデータを観測した後の確率密度分布（事後確率密度分布）として表現

最尤推定

- パラメータを用いて表現された確率密度分布

$$p(\mathbf{x},\boldsymbol{\theta}) \qquad \boldsymbol{\theta}=(\theta_1,\ldots,\theta_P)^{\mathrm{T}}$$

- N個の独立なデータが与えられた時、そのデータがこの確率分布の独立なサンプルである尤もらしさ（尤度）

$$L(\boldsymbol{\theta})=\prod_{i=1}^{N} p(\mathbf{x}_i,\boldsymbol{\theta})$$

- 対数尤度（尤度の対数）

$$l(\boldsymbol{\theta})=\sum_{i=1}^{N}\log p(\mathbf{x}_i,\boldsymbol{\theta})$$

対数尤度を最大とするパラメータ（最尤解）に決定

最尤法（多変量正規分布の場合）

- 最尤解
  - 解析的に求めることが可能

$$\hat{\boldsymbol{\mu}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

$$\hat{\Sigma}=\frac{1}{\mathrm{N}}\sum_{i=1}^{N}(\mathbf{x}_i-\hat{\boldsymbol{\mu}})(\mathbf{x}_i-\hat{\boldsymbol{\mu}})^T$$

  - 平均ベクトルの最尤推定は、サンプル平均ベクトル
  - 分散共分散行列の最尤推定は、分散共分散行列のサンプル推定

ノンパラメトリックな方法

- 特徴
  - 任意の密度関数の推定に適用できる
  - 密度関数の形が未知でも良い
    - ＝＞確率密度関数の形が訓練データに依存して決まる。
- 最も簡単なノンパラメトリックな手法の例
  - ヒストグラム
    ただし、推定された密度関数が滑らかではない
    高次元への拡張が難しい
- 代表的な方法
  - 核関数に基づく方法 (kernel-based methods)
  - K-NN法 (K-nearest-neighbors methods)

ノンパラメトリックな確率密度関数の推定法

- ベクトルxがある領域Rの内側に入る確率

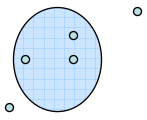
$$P = \int_R p(x')dx' \approx p(x)V$$

密度関数p(x)が連続で、領域R内でほとんど変化しない場合

- 独立なN個のサンプルが与えられた場合、N個のうちK個が領域Rに入る確率

$$\Pr(K) = \binom{N}{K} P^K (1-P)^{N-K}$$

  - Kの期待値は、E[K]=NP



- 確率密度関数は、

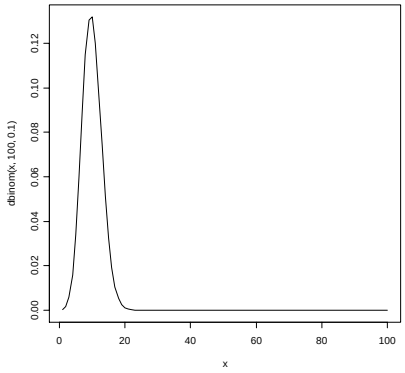
$$p(x) \approx \frac{K}{NV}$$

二項分布は平均付近で鋭いピークを持つので、比 K/N はPのよい近似

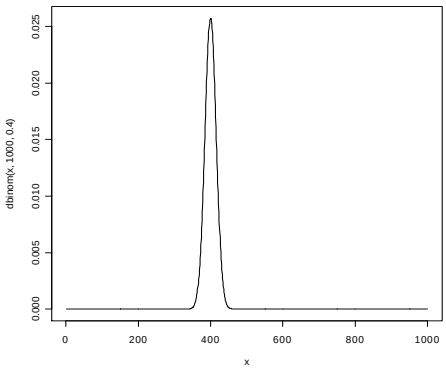
- 近似の成立の条件
  - 領域R内で確率密度関数があまり変化しないためには、領域は十分小さい
  - 二項分布がピークを持つためには、領域に入るサンプルはなるべく多くなければならず、領域はある程度大きい

二項分布とその期待値

$$\Pr(K) = \binom{N}{K} P^K (1-P)^{N-K}$$



N=100, P=0.1、E(K)=10



N=1000, P=0.4、E(K)=400

核関数に基づく方法

- 領域Rの体積Vを固定して、データからKを決定する
  - 点xを中心とする辺の長さがhの超立方体の体積: $V = h^M$
- 核関数
  - 原点を中心とする辺の長さが1の超立方体

$$\varphi(u) = \begin{cases} 1 & |u_j| < 1/2, \quad j = 1, \ldots, M \\ 0 & \text{otherwise} \end{cases}$$

  - 点uが点xを中心とする一辺hの超立方体の内部なら1: $\varphi\left(\frac{(x-u)}{h}\right)$
  - N個のデータのうち領域R内に入るデータの個数

$$K = \sum_{i=1}^{N} H(x_i) = \sum_{i=1}^{N} \varphi\left(\frac{(x-x_i)}{h}\right)$$

  - 確率密度分布

$$\hat{p}(x) \approx \frac{K}{NV} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h^M} H\left(\frac{(x-x_i)}{h}\right)$$

核関数に基づく方法（多変量正規分布）

- 超立方体以外の核関数は？
  - 核関数の条件1

$$\varphi(\mathbf{u}) \geq 0$$

  - 核関数の条件1

$$\int \varphi(\mathbf{u}) d\mathbf{u} = 1$$

- 滑らかな核関数（多変量正規分布）を用いた場合

$$\hat{p}(x) \approx \frac{K}{NV} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(2\pi h^2)^{M/2}} \exp\left(- \frac{\| x - x_i \|^2}{2h^2} \right)$$

滑らかさの制御

- 領域の大きさを変更することで、推定される密度関数の滑らかさが制御可能
  - 滑らかさを大きくしすぎる ＝＞ バイアスが大きくなる
  - 滑らかさが不十分 ＝＞ 個々の学習データに強く依存
  - 滑らかさのパラメータを適切に設定することが必要

- 滑らかさのパラメータの決定
  - 尤度：滑らかさの値が小さいほど尤度の値が大きくなる ＝＞ 使えない
  - Kullback-Leiblerの距離尺度

$$L = -\int p(x) \log \frac{\hat{p}(x)}{p(x)} dx$$

K-NN法

- Kを固定して、領域の大きさVを決定することで密度分布を推定
  - 点xを中心とする超球を考え、超球の半径をしだいに大きくして行き、その超球内に含まれるデータ点の数がちょうどK個になった時の超球の体積をV(x)とする

$$\hat{p}(x) \approx \frac{K}{NV(x)}$$

- 滑らかさの制御
  - データ点の個数Kを変更することで、推定される密度関数の滑らかさを制御可能
    - 滑らかさを大きくしすぎる ＝＞ バイアスが大きくなる
    - 滑らかさが不十分 ＝＞ ここの学習データに強く依存
    - 滑らかさのパラメータを適切に設定することが必要

K-NN (識別器の構成)

- K-NN法による条件付確率密度分布の推定
  - 学習データ
    - クラスCkからNk個の特徴ベクトルが得られているとする。全データ数は、N
    - 点xを中心とする超球を考え、その中にちょうどK個の学習データを含むまで超球の半径を大きくしていった時の超球の体積をV(x)とする。
    - 確率密度分布 $\hat{p}(x) \approx \frac{K}{NV(x)}$
    - その超球内、クラスCkのデータがKk個含まれているとすると、クラスCkの条件付確率密度分布 $\hat{p}(x \mid C_k) \approx \frac{K_k}{N_k V(x)}$
    - 事後確率 $\hat{P}(C_k \mid x) = \frac{\hat{P}(C_k)\hat{p}(x \mid C_k)}{\hat{p}(x)} = \frac{\frac{N_k}{N}\frac{K_k}{N_k V(x)}}{\frac{K}{NV(x)}} = \frac{K_k}{K}$

最近傍則 (NN-則、Nearest Neighbor Rule)

- NN-則
  - 訓練サンプル集合の中で、xに最も近いサンプルを見つけ、そのサンプルのラベルのクラス (属していたクラス) に識別
  - 最近傍則の誤り率
    - 訓練サンプルが無数にあれば、達成可能な最小の誤り率 (ベイズ誤り率) の2倍以下

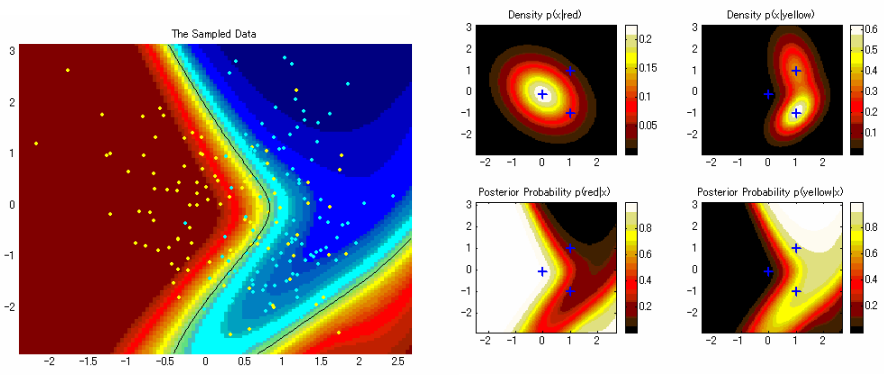
$$P^* \le P \le 2P^* - \frac{K}{K-1}(P^*)^2$$

- K－NN則
  - 入力ベクトルx に近いK個のサンプルの中で、最も頻度の高いラベルのクラスに識別

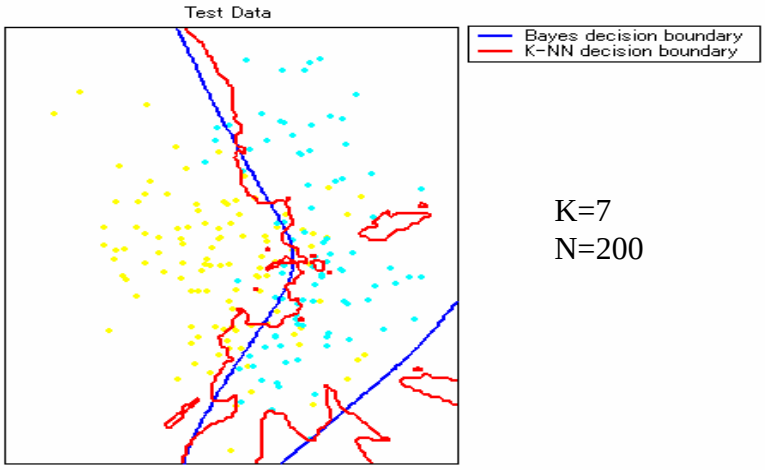
＝＞ xに近いK個のサンプルを用いた多数決

K-NN識別器によるパターン識別の例

- データ
  - Class 1: 2次元正規分布 N1=100
  - Class 2: 2つの正規分布の混合分布 N2=100



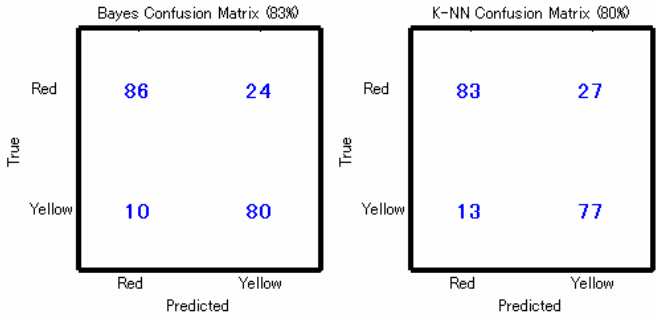
K-NN識別器による識別境界



K=7
N=200

K-NN識別器によるテストサンプルの識別結果

新たに生成したテストサンプル (N=200)**の識別**



セミパラメトリックな手法

- パラメトリックモデルに基づく方法とノンパラメトリックな方法の中間的手法
  - パラメトリックモデルに基づく方法
    - 利点： 新しいデータに対する確率密度の計算が比較的簡単
    - 欠点： 真の分布と仮定したモデルが異なる場合には必ずしも良い推定結果が得られない
  - ノンパラメトリックな手法
    - 利点： 真の分布がどんな関数系であっても推定できる
    - 欠点： 新しいデータに対して確率密度を評価するための計算量が学習用のデータが増えるとどんどん増加してしまう
  - 両方の良い点を取り入れ、欠点を改善するような手法

- 代表例
  - 混合分布モデル(Mixture models)に基づく方法
  - ニューラルネットワーク

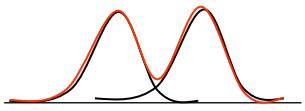
混合分布モデル

- 混合分布

$$p(x) = \sum_{j=1}^{O} \omega_j p(x \mid j)$$

- 混合パラメータの条件

$$\sum_{j=1}^{O} \omega_j = 1, \qquad 0 \le \omega_j \le 1$$



- 各確率密度分布の条件

$$\int p(x \mid j) dx = 1$$

- 各確率密度分布が正規分布の場合 (混合正規分布モデル)

$$p(x \mid j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left\{ -\frac{\| x - \mu_j \|^2}{2\sigma_j^2} \right\}$$

混合正規分布の最尤推定

- N個の学習データに対する対数尤度

$$l = \log L = \log \prod_{n=1}^{N} p(x_n) = \sum_{n=1}^{N} \log p(x_n) = \sum_{n=1}^{N} \log \left\{ \sum_{j=1}^{O} \omega_j p(x_n \mid j) \right\}$$

- 各確率密度分布のパラメータ推定（正規分布の場合）
  - 非線形最適化手法を利用

$$\frac{\partial l}{\partial \mu_j} = \sum_{n=1}^{N} \frac{\omega_j p(x_n \mid j)}{p(x_n)} \frac{(x_n - \mu_j)}{\sigma_j^2} = \sum_{n=1}^{N} P(j \mid x_n) \frac{(x_n - \mu_j)}{\sigma_j^2}$$

$$\frac{\partial l}{\partial \sigma_j} = \sum_{n=1}^{N} \frac{\omega_j p(x_n \mid j)}{p(x_n)} \left\{ -\frac{d}{\sigma_j} + \frac{\| x_n - \mu_j \|^2}{\sigma_j^3} \right\} = \sum_{n=1}^{N} P(j \mid x_n) \left\{ -\frac{d}{\sigma_j} + \frac{\| x_n - \mu_j \|^2}{\sigma_j^3} \right\}$$

ただし、 $$P(j \mid x) = \frac{\omega_j p(x \mid j)}{\sum_{k=1}^{O} \omega_k p(x \mid k)}$$

混合正規分布の最尤推定（つづき）

- 混合パラメータの推定
  - 補助パラメータを利用（softmax関数）

$$\omega_j = \frac{\exp(\gamma_j)}{\sum_{k=1}^{O} \exp(\gamma_k)}$$

  - 対数尤度の補助パラメータに関する微分

$$\frac{\partial l}{\partial \gamma_j} = \sum_{k=1}^{O} \frac{\partial l}{\partial \omega_j} \frac{\partial \omega_j}{\partial \gamma_j} = \sum_{n=1}^{N} \{ P(j \mid x_n) - \omega_j \}$$

混合正規分布の最尤推定（つづき）

- 最尤解の性質
  - 対数尤度の微分＝0とおくと

$$\hat{\omega}_j = \frac{1}{N}\sum_{n=1}^{N} P(j \mid x_n)$$

$$\hat{\mu}_j = \frac{\sum_{n=1}^{N} P(j \mid x_n) x_n}{\sum_{n=1}^{N} P(j \mid x_n)}$$

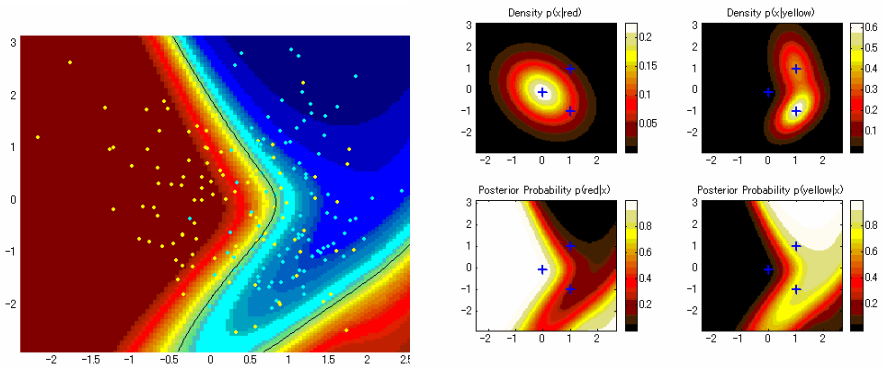
$$\hat{\sigma}_j^2 = \frac{1}{d}\frac{\sum_{n=1}^{N} P(j \mid x_n) \| x_n - \hat{\mu}_j \|^2}{\sum_{n=1}^{N} P(j \mid x_n)}$$

$$P(j \mid x) = \frac{\omega_j p(x \mid j)}{\sum_{k=1}^{O} \omega_k p(x \mid k)}$$

  - 各要素への帰属度を表す事後確率P(j|x)を重みとして計算される

混合正規分布モデルを用いた識別の例

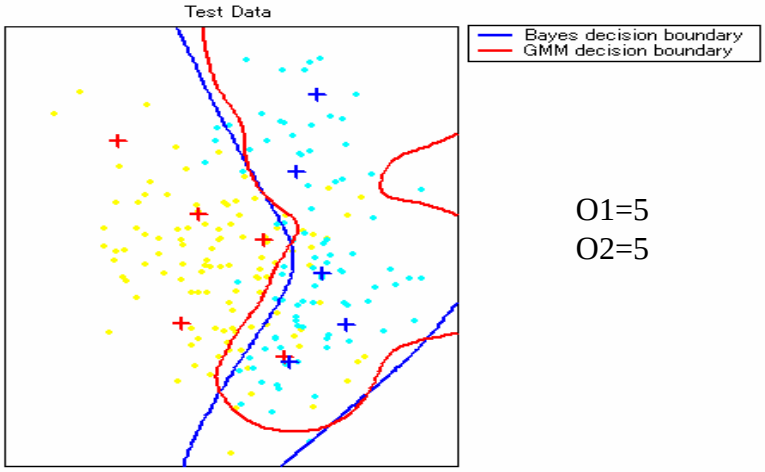
- データ
  - Class 1: 2次元正規分布 N1=100
  - Class 2: 2つの正規分布の混合分布 N2=100



識別器の構成と学習

- 各クラスの分布を正規混合分布により推定
  - Class 1： O=5個の正規混合分布
  - Class 2： O=5個の正規混合分布
- 訓練サンプル
  - N=200サンプル（各クラス100サンプル）
- パラメータの学習法
  - EMアルゴリズムを利用

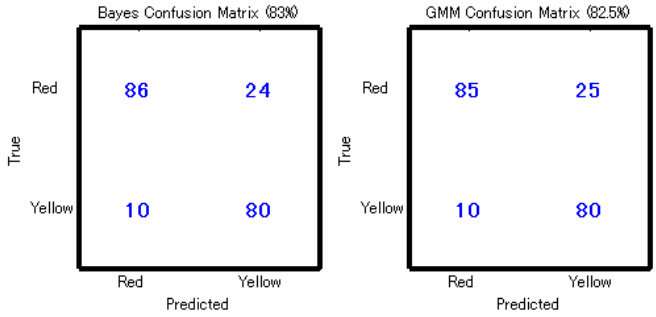
混合正規分布推定による識別境界



O1=5
O2=5

混合正規分布推定によるテストサンプルの識別結果

新たに生成したテストサンプル (N=200)の識別



2限目終了