

パターン認識と学習

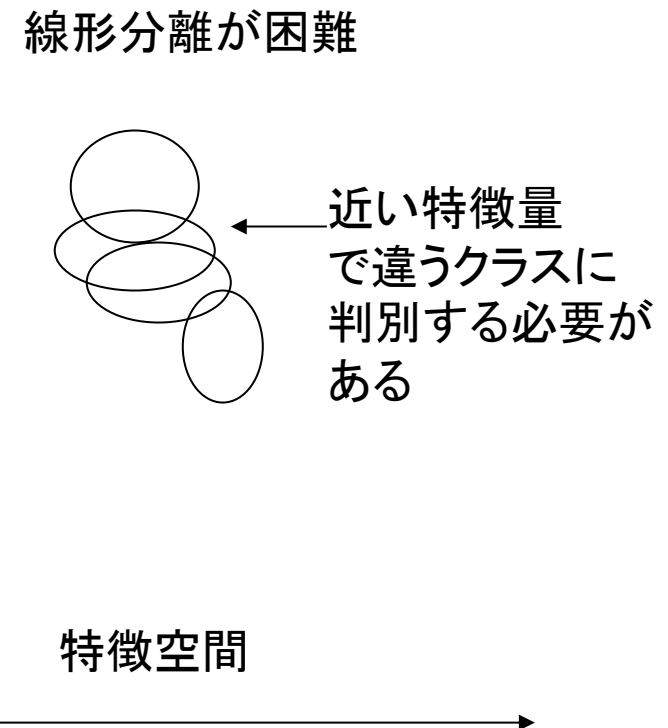
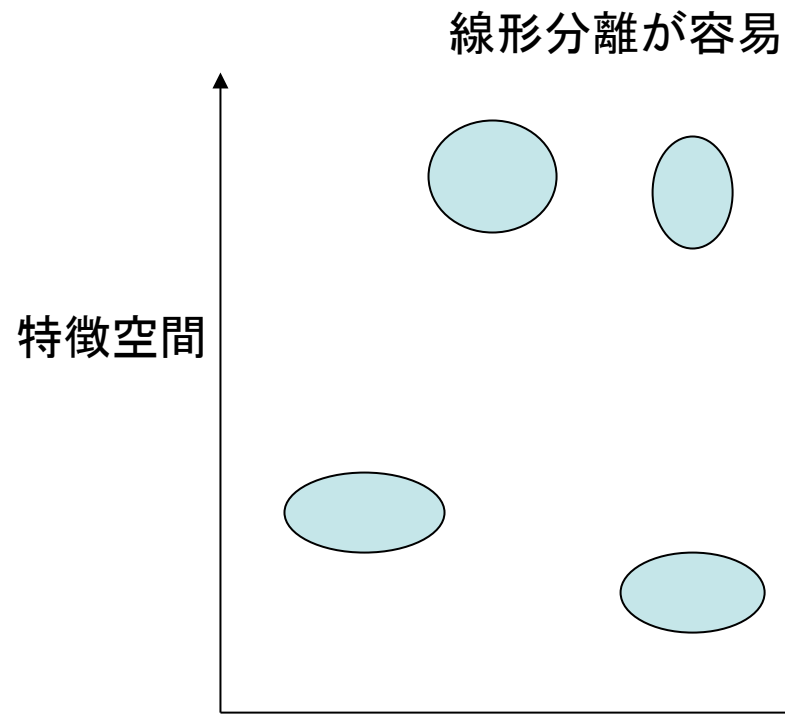
竹内孔一

本日の内容

- ベイズ誤り

特徴量の良し悪し

- いい特徴量とは?
 - 平均が離れている
 - 各クラスの分散が小さい



ベイズ誤り確率

- あるデータ $\mathbf{x}$ に対して誤ったクラスを付与する可能性を積算した物
 - 特徴空間で重なってる部分では必ず起こる

2クラス分類を事例に

$$P_{error}(\mathbf{x}) = \begin{cases} P(A_1 | \mathbf{x}) & \mathbf{x} \text{を} A_2 \text{と判定した場合} \\ P(A_2 | \mathbf{x}) & \mathbf{x} \text{を} A_1 \text{と判定した場合} \end{cases}$$

判別結果と異なるクラスになる確率

Bayes Error

$$e_{Bayes} = \int \min \{P(A_1 | \mathbf{x}), P(A_2 | \mathbf{x})\} p(\mathbf{x}) d\mathbf{x}$$

多クラス分類

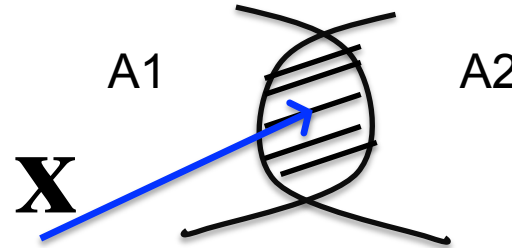
$$e_{Bayes} = \int \min_k \{1 - P(A_k | \mathbf{x})\} p(\mathbf{x}) d\mathbf{x}$$

ベイズ誤り確率の式の解釈

- 例 ある物体の色判定(赤(A1)か白(A2))
ある入力 $\mathbf{x}$ に対して下記のような確率であったとしよう

$$P(A_1 | \mathbf{x}) = 0.8$$

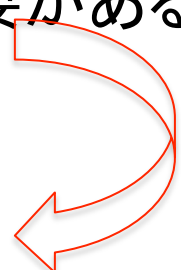
$$P(A_2 | \mathbf{x}) = 0.2$$



これは同じ入力 $\mathbf{x}$ のところは曖昧で8割のデータは赤だけど2割のデータは白のデータが混じってしまうことを示している

いま大きい確率値のクラスを選択するので, $\mathbf{x}$ では判別器は常に赤(A1)としてしまう. そのとき, 2割は白(A2)のデータもいるので, 誤り確率として $P(A_2 | \mathbf{x}) = 0.2$ は覚悟しておく必要がある

$$P_{error}(\mathbf{x}) = \begin{cases} P(A_1 | \mathbf{x}) & \mathbf{x} \text{を} A_2 \text{と判定した場合} \\ P(A_2 | \mathbf{x}) & \mathbf{x} \text{を} A_1 \text{と判定した場合} \end{cases}$$



ベイズ誤り確率

- 最近傍法との比較
 - ベイズ誤り確率の2倍を超えない
 - ただし高次元空間では成立しない

近年の学習モデル

	Support vector machine (1995-)	Conditional random field (2001-)	Ensemble learning (集団学習)	
			Bagging	AdaBoost
特徴	• kernel法による高次元化 • 線形識別による高速な学習 • 初期値による依存性が無い	• 特徴量をexp関数で評価 • 連続列に対する予測に組み合わせ最適の結果を出力できる	• 学習データを復元抽出 • 複数のモデルによる投票で全体の精度を向上	• 学習データを復元抽出 • 3つのモデルの投票 • 先に作成したモデルの不得意な事例を学習する

予備

- SVM

$$\mathbf{w} = \sum_k^n \beta_k y_k \Phi(\mathbf{x}_k)$$

$$f(\Phi(\mathbf{x}_i)) = \sum_k^n \beta_k y_k K(\mathbf{x}, \mathbf{x}_k) + b$$

$K(\mathbf{x}, \mathbf{x}_i)$
はカーネル

- CRF

$$P(y \mid x) = \frac{\exp\langle \Theta, \Phi(x, y) \rangle}{\sum_y \exp\langle \Theta, \Phi(x, y) \rangle}$$

Down load

- SVM
 - 多項式カーネルの計算が速い
 - tinysvm
 - <http://chasen.org/~taku/software/TinySVM/>
- CRF
 - CRF++
 - <http://crfpp.sourceforge.net/>