はじめてのパターン認識

第５章　k最近傍法

11t4062a　広原裕亮

はじめに

- 最近傍法
  ー入力データとすべての学習データとの距離からクラスを識別
- K最近傍法
  ー最も近いk個の学習データを選ぶ方法
- 近似最近傍探索

５．１最近傍法とボロノイ境界（１）

- K個のクラスΩ＝$\{C_1, \cdots, C_k\}$
- i番目のクラスの学習データ数をN（i）
- 学習データの集合$S_i = \{X_1^{(i)}, \cdots, X^{(i)}_{N(i)}\}$

- 入力データと学習データの類似度
  $d(x, x_j^{(i)}) = ||x - x_j^{(i)}||$

５．１最近傍法とボロノイ境界（２）

- 識別クラス＝

$$\begin{cases} \arg\min d(x,x_j^{(i)}) & \min d(x,x_j^{(i)}) < t \text{の時} \\ \text{reject} & \min d(x,x_j^{(i)}) \geqq t \text{の時} \end{cases}$$

tは、どの学習データとも距離が大きい場合にリジェクトするためのしきい値

５．１最近傍法とボロノイ境界（３）

- ボロノイ領域

　－各学習データが隣接する学習データと等距離にある境界で囲まれた領域

- ボロノイ境界

　－各学習データから等距離の点の集合

　　$(\bar{x} - x)^T n=0,\ \bar{x}=(x_i + x_j)/2,\ n=(x_i - x_j)$

　　上式を満たすxが表す超平面が半空間に2分割する

5.1.2　学習データの数と識別性能

- 学習データを増やすと識別性能は上昇する
- しかし計算量が増加してしまう

５．２　K最近傍法（１）

- 識別クラス＝

{j　{k_j}=max{ k_1,…, k_K }

{リジェクト { k_i,…, k_j }=max{ k_1,…, k_K }

- K個のデータのうち最も多いクラスjと識別
- タイがある場合リジェクト

５．２　K最近傍法（２）

- 最近傍法の場合（k＝１）
  ー孤立した識別境界が見られる
  ー計算量が多い

- 11最近傍法（k＝１１）
  ー滑らかな境界となる
  ー識別精度が下がる

５．３　kNN法の計算量とその低減法

- データと全てのクラスの学習データとのユークリッド距離
- 距離を昇順にソート

↓

データが増えると多くの計算時間と記憶容量が必要。

これらを緩和する方法がある。

誤り削除型k最近傍法

- 正しくないクラスの領域に存在している学習データを削除する

↓　　不可能

ー誤ってるのに削除されない

ー正しいのに削除される

圧縮型k最近傍法

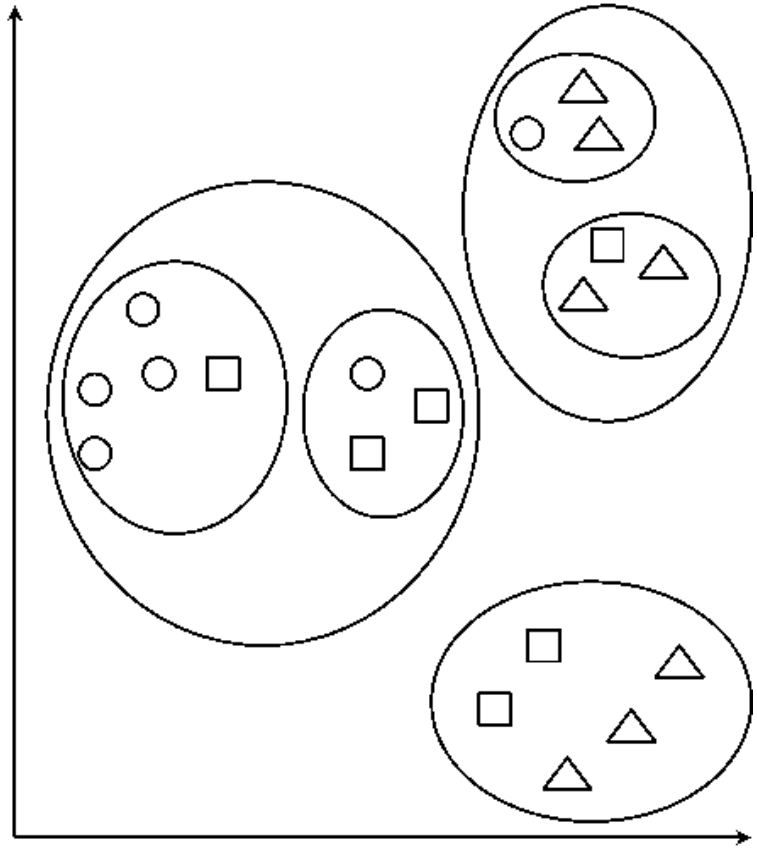
- 識別境界のデータが重要

  $\varepsilon(x) = \min[P(C_1|x), P(C_2|x)]$

  ーベイズ誤り率を小さくする
- 削除前と誤り率が同じになるようにする

分岐限定法（1）

・クラスタリングで
学習データを組織化
・平均ベクトルm_i
・最遠のデータ
までとの距離d_iを算出



分岐限定法（2）

・一番上の階層のクラスタから順に平均ベクトルの最も近いクラスタを探す

→ほかのクラスタからは？

$d(x,m_n) > d(x,x_i) + d_n$

が成り立てば、そのクラスタを探索する必要はない。

近似最近傍探索（1）

- 最近傍より少し遠くても許容する

ー$d(q,x) \leqq (1+\varepsilon)d(q,x^*)$

の距離にあるデータまで許容する。

x^*とは最適解である。

近似最近傍探索（2）

Step1:2分木を作成する。

Step2:入力データと同じ領域内のデータとの距離を算出

Step3:次に近い領域との距離を算出し比較

Step4:次にデータとの距離を算出し比較。以降それを繰り返す

近似最近傍探索（3）

- $d(q,x) / (1+\varepsilon) \leqq d(q,X)$を満たすと近似解

↓　なぜ言えるのか

上記の式を満たした領域までの距離をr'

$d(q,x) / (1+\varepsilon) \leqq r' \leqq d(q, x')$

↓

$d(q,x) \leqq (1+\varepsilon)r' \leqq (1+\varepsilon)d(q, x')$