

初めてのパターン認識

第一章:はじめに

1. 1 パターン認識とは
1. 2 特徴の方
1. 3 特徴ベクトル空間と次元の呪い

12月13日(金)

広原裕亮

はじめに

人間は五感を用いてパターン認識を行っている。

物が何か判断するために有効な特徴を抽出する段階が存在する

- 識別

「何か」を判断すること

- 特徴抽出

「有効な特徴」を抽出すること

- 識別規則

「有効な特徴」と「何か」を結びつける規則

人間はこの規則を学習している

1.1 パターン認識とは

パターン認識の流れ

- 切符の券売機の場合

硬貨や紙幣を観測し、その材質やサイズなどの特徴に注目する(特徴抽出)



特徴を並べてベクトルの形にまとめる(特徴ベクトル)



特徴ベクトルにより10円か100円かなどのクラスに分類する

この分類するための規則を識別規則という

切符券売機の例

- クラス = {10円、50円、100円、500円}
- 硬貨や紙幣を観察



識別に有効な特徴を抽出

特徴・・・サイズ、重さ、穴の有無...

特徴ベクトル(1cm, 1g, 無)

↓ 識別規則によりクラスが決まる

「1円玉」

パターン認識の成否

- 識別対象は多岐に渡るため識別に有効な特徴をいかに早く抽出できるかが鍵となる。特徴ベクトルの形になってしまえば、同じ識別規則が使える
- 識別規則の多さ。つまり特徴ベクトルとクラスの対応関係を学習することが重要

1. 2 特徴の型

特徴

- 定性的特徴(非数値データ)
 - 名義尺度(分類のための単なる名前)
 - 順序尺度(順序関係を表す。比較はできるが加減算などの演算はできない)
- 定量的特徴(数値データ)
 - 間隔尺度(一定の単位で量られた量で、等間隔性がある。加減算が意味を持つ)
 - 比例尺度(原点が定まっている量。比が意味を持つ)

例題 1. 1

次の特徴の型は何か。

(1) 試験の点数

→ 間隔尺度

(2) 成績表のA,B,C,D

→ 順序尺度

(3) 偏差値

→ 間隔尺度

(4) 単語の出現頻度

→ 比例尺度

符号化

定性的な特徴を計算機上で扱うために符号を用いる(符号化)。クラス数が $K (> 2)$ 個ある場合、 K 個の2値変数(ダミー変数)を用意することで符号化が可能。

例)「大中小」の場合 $A = (A_1, A_2, A_3)^T$

| | A1 | A2 | A3 |
|---|----|----|----|
| 大 | 1 | 0 | 0 |
| 中 | 0 | 1 | 0 |
| 小 | 0 | 0 | 1 |

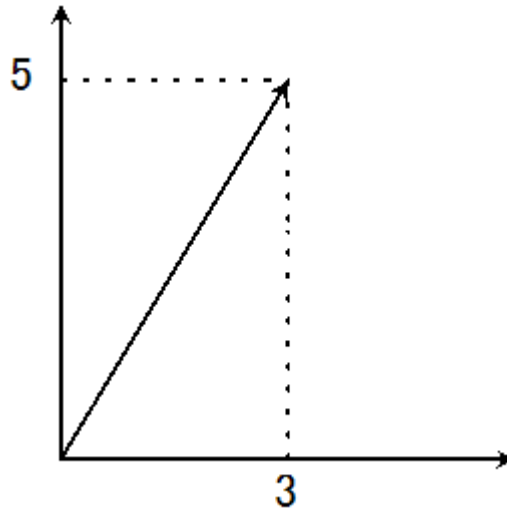
1. 3 特徴ベクトル空間と次元の呪い

特徴ベクトル空間

- 特徴数をdとすれば、特徴ベクトルはd次元線形空間を張る。

(例) 特徴数2の特徴ベクトルの場合

$$A=(3,5)^T$$



次元の呪い

(数学的)空間の次元が増えるのに対して指数関数的に問題に必要なデータが増加すること。

例)

$16 \times 16 = 256$ の画素値からなる画像データの場合256次元ベクトルだが、各画素が16レベルの階調をもっていると区画の数は16の256乗となり、必要なデータが格段に増加する。

超立方体

- ・d次元単位超立方体の中心から頂点までの距離は
 $D(d)=1/2\sqrt{d}$
- ・中心から単位超立方体の面までの最短距離は次元によらず1/2。頂点までの距離との比は \sqrt{d} となる。
- ・頂点の数は 2^d 個ある。
- ・d次元超立方体の面(ファセット)の数は $2d$ 個である。

まとめ

- 識別するものを素早く特徴ベクトルの形にするが重要
- 特徴ベクトルとクラスを対にしたたくさんの事例を使い、対応関係を学習する
- 特徴には定性的な物と定量的な物があり、定性的な特徴は符号化を用いる必要がある
- 特徴数を d とすれば、特徴ベクトルは d 次元空間を張る