パターン認識と機械学習

１．４　次元の呪い

5月9日（金）

豊川　幸秀

次元の呪いとは

- 高次元空間を扱う際に伴う困難のこと

  ーこの本では主に低次元の例を扱う

- 具体的にどのような問題が発生するか例によって示していく

具体例（１）

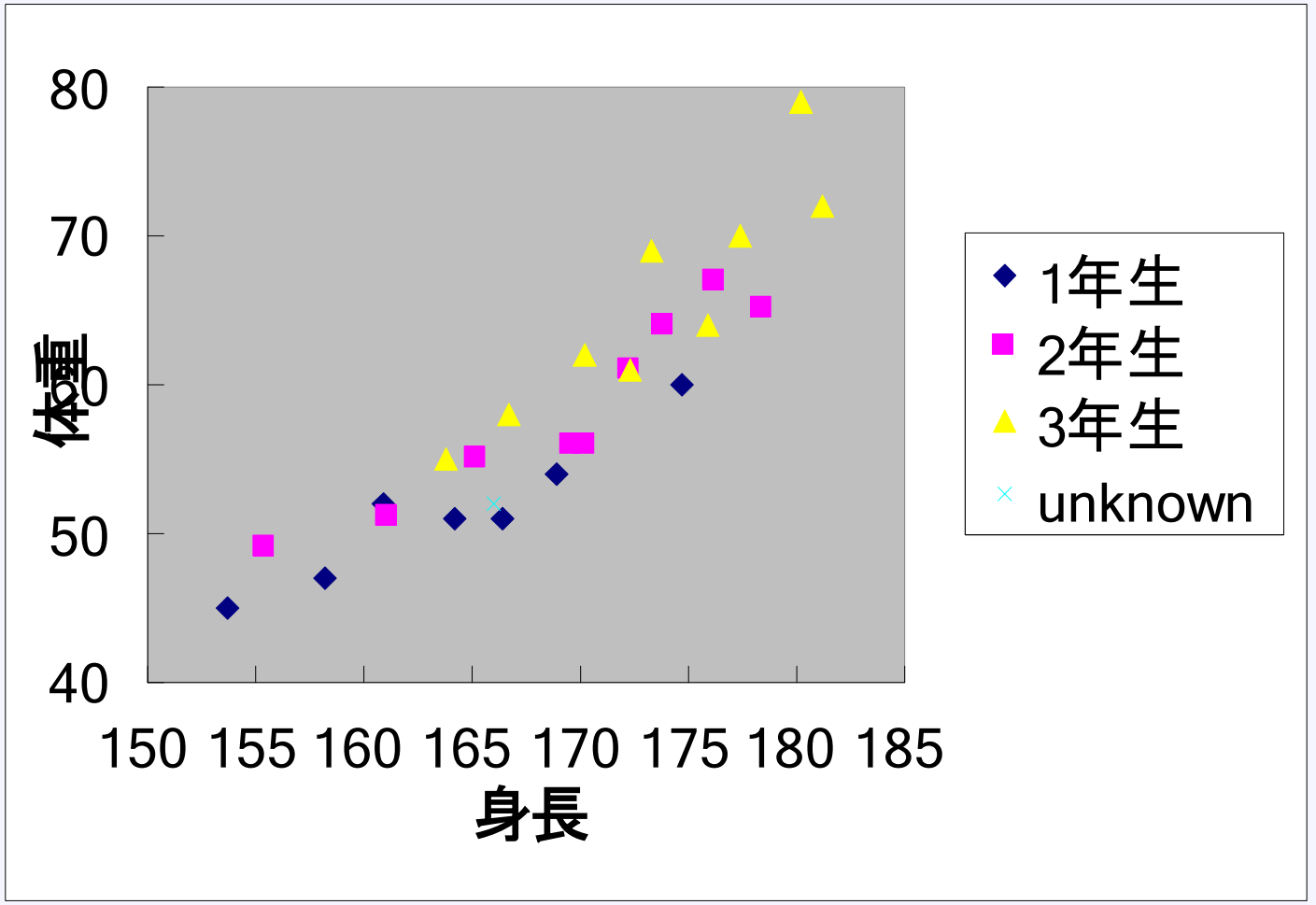
- ある高校の男子生徒に対する様々な計測値のデータ集合を考える。
- その中から、身長と体重の２つの計測値をプロットしたものを示す。

↓

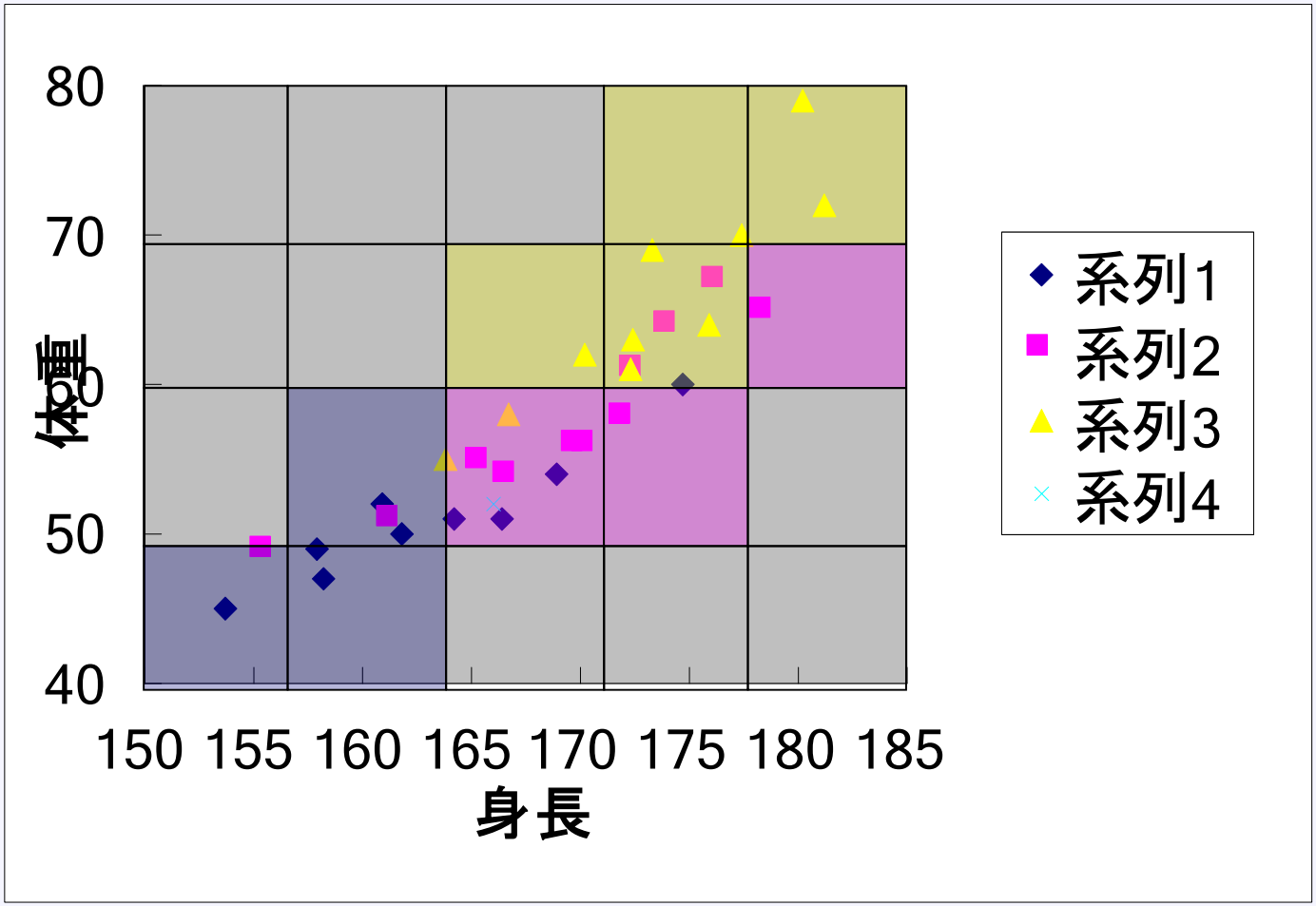
「1年生」・「2年生」・「3年生」の３つに

大別できるとする

散布図


80
70
60
50
40
体重
150 155 160 165 170 175 180 185
身長
1年生
2年生
3年生
unknown

散布図(マス目分け)


80
70
60
50
40
体重
150 155 160 165 170 175 180 185
身長
系列1
系列2
系列3
系列4

問題点

- 入力空間が1次元増えていく毎に、マス目の数が指数的に増加する。

↓

必要な訓練データも増加する。

1次元：3個

2次元：3×3＝9個

3次元：3×3×3＝27個

：

具体例（２）

- 多項式曲線フィッティングの例について、入力変数を複数個に拡張してみる。

入力変数をＤ個としたとき、

３次までの多項式は

$$y(x,w) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

で表すことが出来る。

具体例（２）の続き

- 入力変数Dが増えると、独立な係数の数は
  D^3に比例して増加する。

  ↓M次元の多項式では

  D^Mに比例して増加する。（べき乗の増加）

先程の指数的な増加に比べて増加速度は
遅いが、これでもまだ実用的ではない。

具体例(3)

- D次元空間上に存在する半径 r=1 の球について考え、r=1- ε と r=1 の球の体積の割合を求める。

D次元空間上の半径rの球の体積は、

$$V_D(r) = K_D r^D \quad (K_D : \text{定数})$$

で表すことができる。

具体例（3）の続き

- よって、求める割合は、

$$\frac{V_D(1)-V_D(1-\varepsilon)}{V_D(1)}=1-(1-\varepsilon)^D$$

  となる。
- 高次元になるほど、球の表面に近い部分に体積が集中している。

我々が普段過ごしている3次元空間
での常識は、高次元では通じない

まとめ

- 以上の例などから、我々の低次元における常識は、高次元では必ずしも通じないことが分かるので、注意が必要。
- パターン認識においてこの点は問題ではあるが、以下の2点により問題解決が容易となる。

－１：実際に扱う実データは、低次元の領域のものがほとんどである。

－２：一般的に実データは滑らかで、多少の変化は目標変数に大きな影響を及ぼさない。