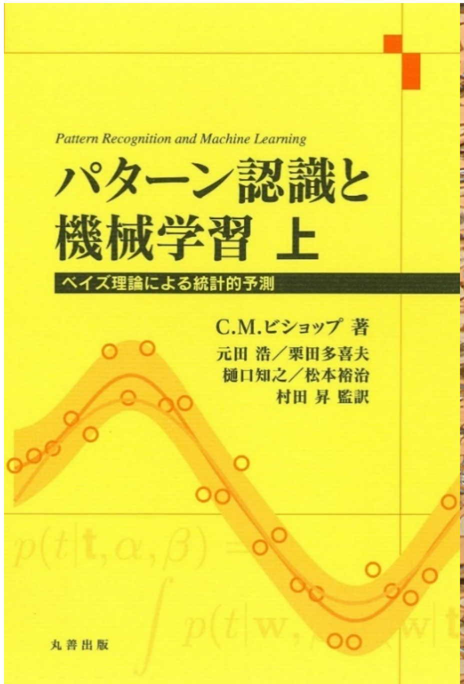
PRMLゼミ

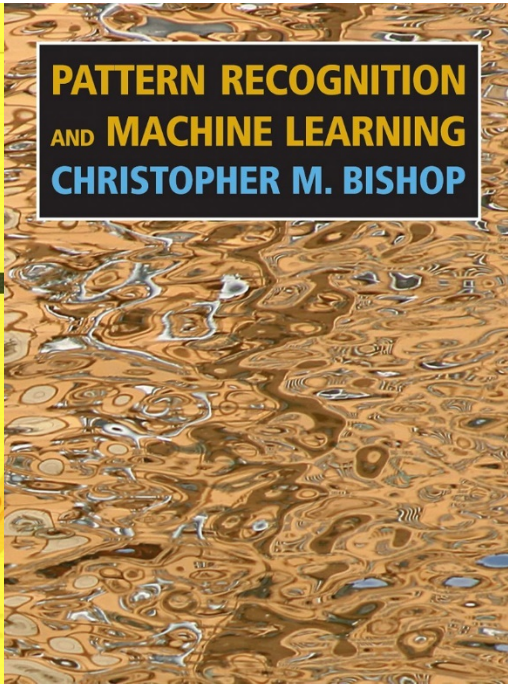
1.4節：次元の呪い

anmitsu48

本資料について

- 本資料は、『パターン認識と機械学習　上　－ベイズ理論による統計的予測－』（丸善出版）を用いてゼミを行った際に、私が使用した発表資料を再編集したものである。
- 再編集の際は、私が持っている他の資料も利用した。
参考にした資料は最後にまとめて紹介する。



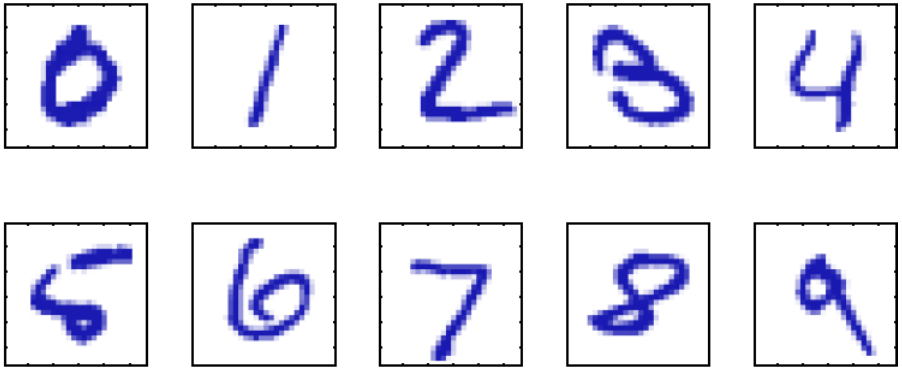


高次元データ

- 1.1節や1.3節の多項式近似の議論では、入力が1次元の場合を考えた。
- しかし、実際の機械学習の問題は多次元のデータを取り扱う。1.4節では、高次元になると生じる問題を考える。

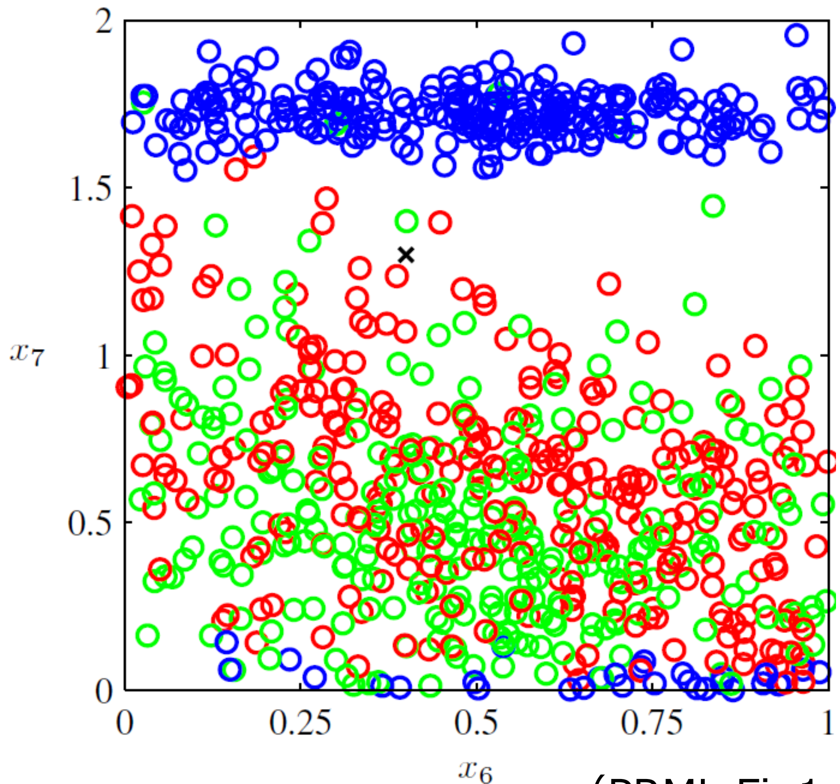
（例） MNISTデータセット（手書き文字のデータセット）

- ➢ 各画像は 28×28＝784ピクセルからなる。
  ⇒ 各画像は784次元の実数値ベクトルで表せる。
- ➢ 784次元ベクトルから、0～9のどの数字かを推定する。



例：３値分類問題 [Bishop and James(1993)]

- オイル流れに関する12次元の入力データを考えるから、３値分類が可能な学習器を作成することを考える。
  - 下図の×印が赤、青、緑のどのグループに属するかを判定できる学習器を作成したい。



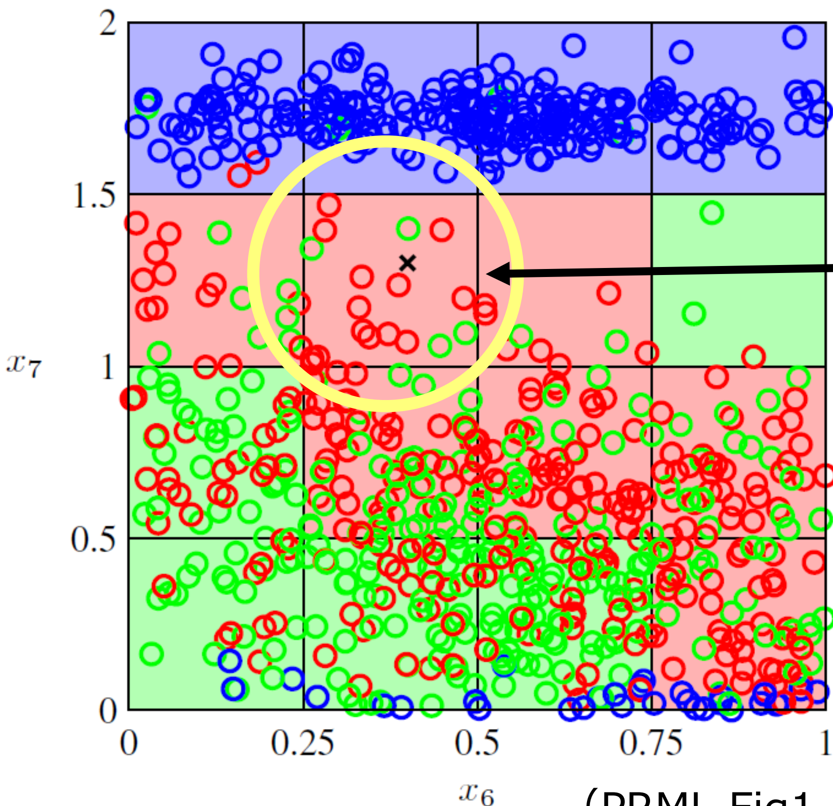
（PRML Fig1.19 より）

- 左図から、×印は赤あるいは緑のグループに属すると判定するのが妥当である。

- その根拠としては、「赤と緑が×印の近くにある」ことが挙げられる。

メッシュ状に空間を分割する

- 「未知のデータ点に近い既知データと同じグループに属する」と考えることは自然。
- 例えば、空間を分割化して、未知のデータを含む領域中で一番多いものと同じグループに属すると考える。

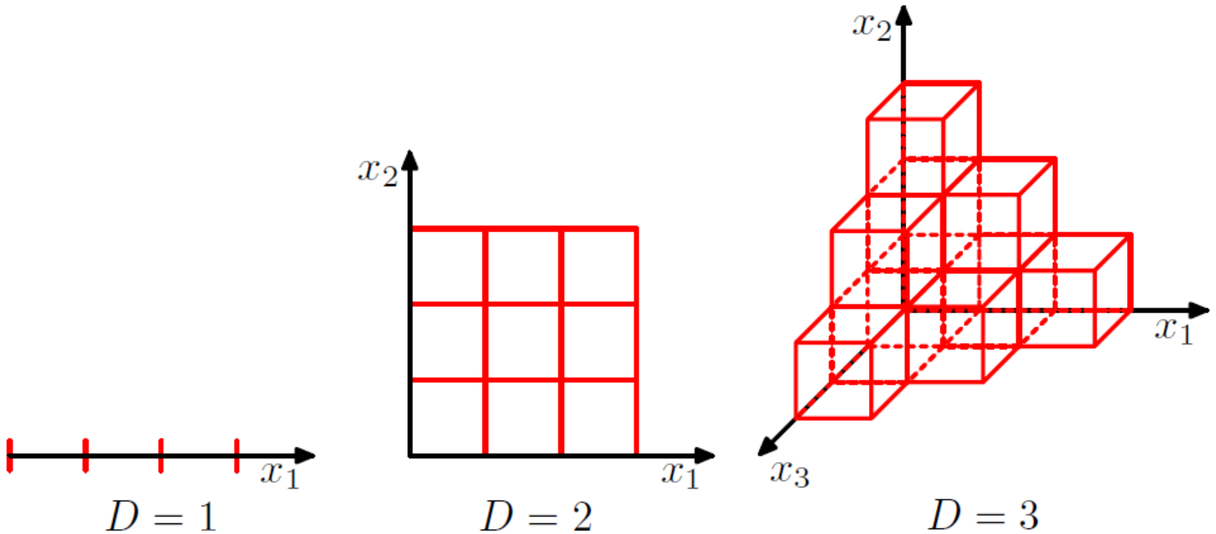


（PRML Fig1.20 より）

- ×印がある領域には、赤丸が一番多い。
⇒「×印は赤グループに属する」と考えるのが妥当。

空間分割の限界

- 空間を分割して、各セル内のデータの中で一番多いものと同じグループに属するという考え方は、各セルに1つ以上のデータが存在することが要求される。
- 次元の増加に対応して、セルの数は指数関数的に増加。
  ⇒ 必要なデータ数も指数関数的に増加。
  ⇒「空間分割」の方法は高次元データには向かない。



（PRML Fig1.21 より）

多項式近似を再度考える

- D個のデータによる多項式近似について、
３次式でデータ間の相互作用も考慮する。

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

- 上記の式のパラメータの数: $O(D^3)$
⇒ M次多項式のパラメータの数は $O(D^M)$。

- より正確にデータ間の相互作用を把握するには、
高次の多項式を使用する必要があるが、
パラメータ数が$O(D^M)$で増加する。
  - Dについてべき乗関数で、Mについて指数関数。
  - D、Mどちらかが大きくなると、実用的なレベルをすぐに逸脱する。

高次元超球の体積

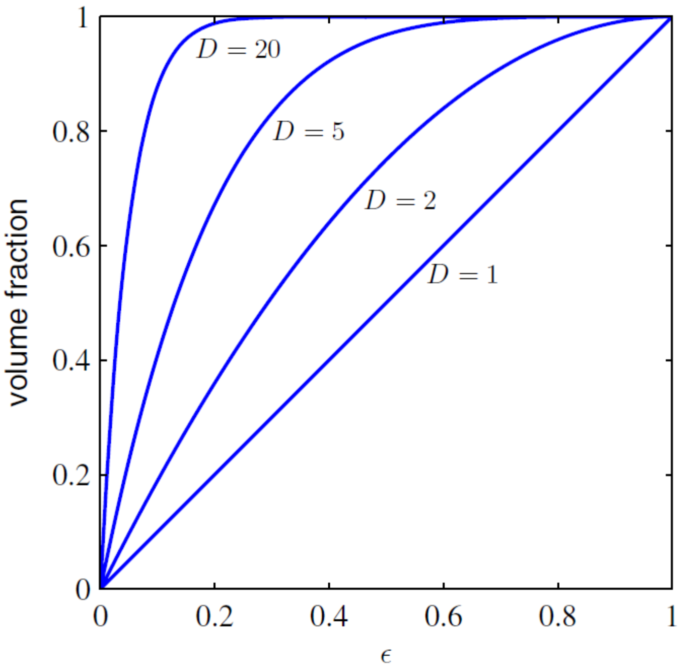
- 半径 R の n 次元超球の体積の公式

$$V_n(R) = \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)} R^n$$

- 同一中心の半径1と半径1 − εのD次元超球の体積の差と、半径1の超球の体積の比。

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

- 次元 D が大きいとき、ε が小さくても、上の比は1に近くなる。
- 高次元空間では、球の体積のほとんどが、球の表面付近に集中していることを意味する。



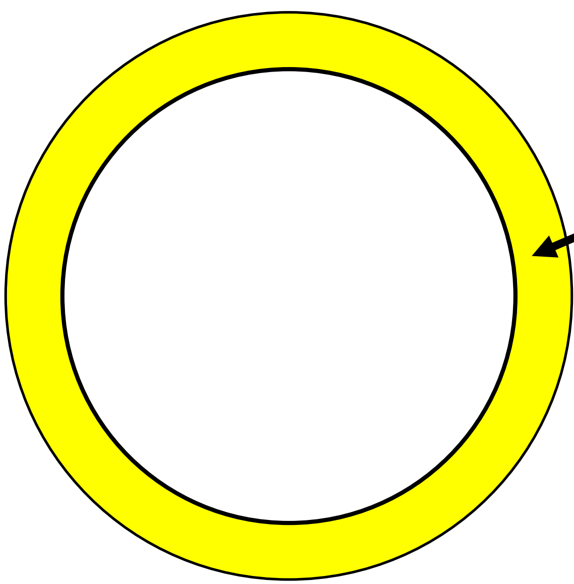
（PRML Fig1.22 より）

高次元超球の体積

- 同一中心の半径1と半径 $1-\varepsilon$ のD次元超球の体積の差と、半径1の超球の体積の比。

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

- 上の式は、「高次元空間では、球の体積のほとんどが、球の表面付近に集中している」ことを表す。



この部分の体積が
$V_D(1)-V_D(1-\varepsilon)$

「ε が0に近いときの黄色部分の領域」
≒「半径1の超球の表面付近の領域」
であり、その体積が $V_D(1)$ と近い。
⇒ 表面付近に体積が集中している