パターン認識と機械学習
1.4 次元の呪い

吉田拓夢

次元の呪い(curse of dimensionality)

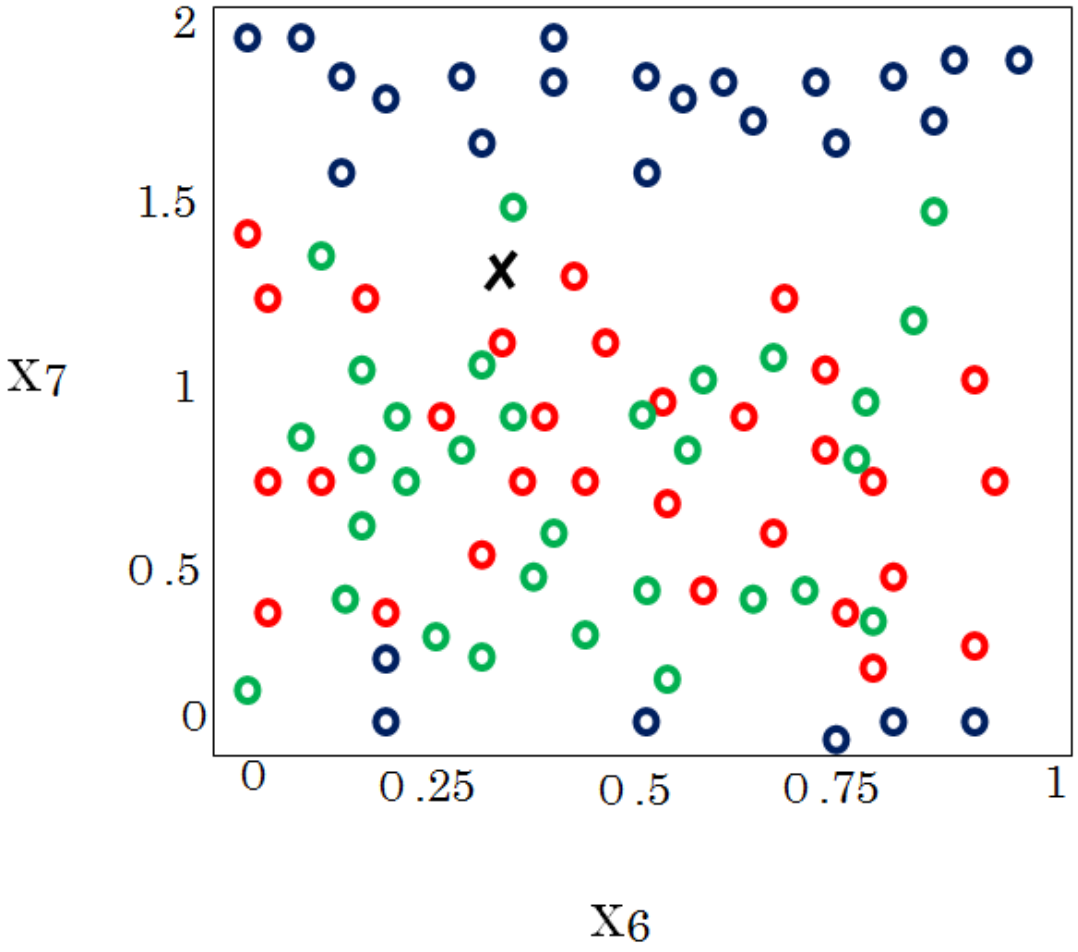
- 大きい次元の空間に伴う困難
- **低次元での直観が大きい次元で一般化できるとは限らない**

- 高次元に対する有効な手法が無い訳ではない

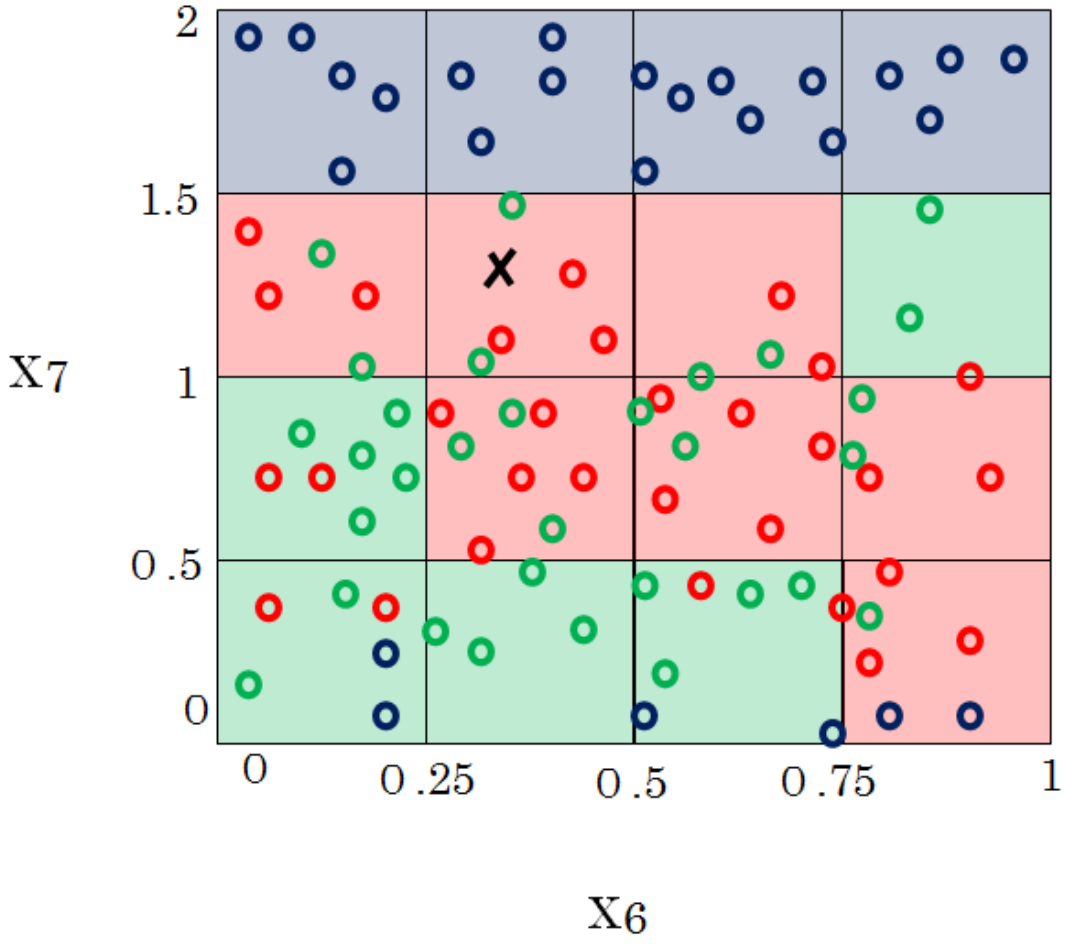
→実データは実質的には低次元の領域の場合が多い。
目標変数の重要な変化の方向が限定的

→実データが一般には滑らかで、大体は入力空間上の小さな変化は目標変数に大きな変化を与えない。故に、局所的な内挿等の手法で新たな値に対する予測が可能

送油データの散布図

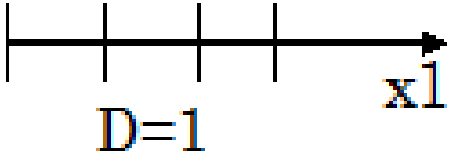


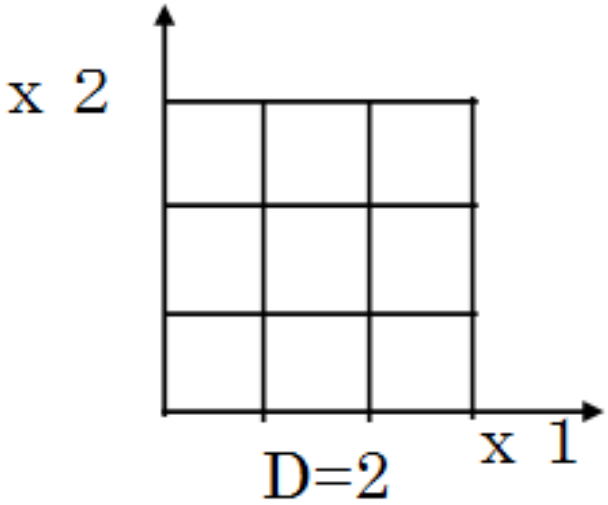
クラス分類

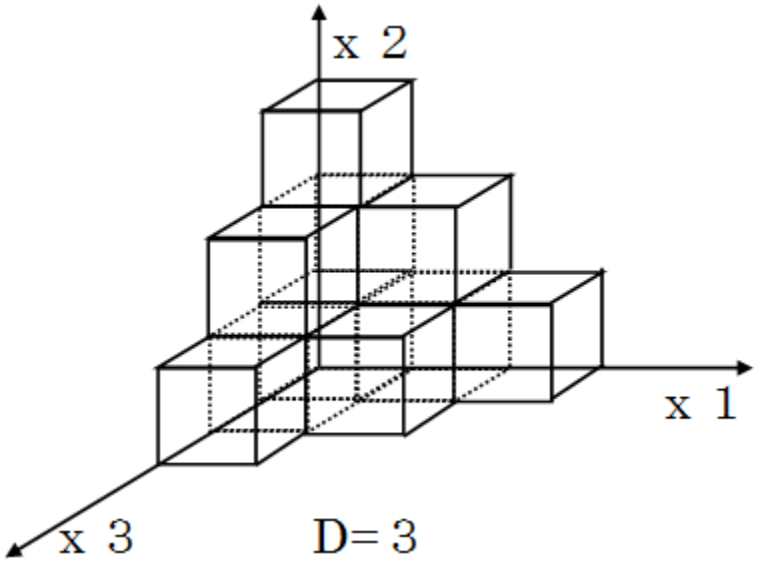


次元の呪いの図

- マス目は次元Dに対し指数的に増加







(立方体領域の一部)

高次元の問題　-多項式曲線フィッティングの例-

- 入力変数がD個ある時の3次までの多項式

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

→Dが増えると独立な係数はD^3に比例

- M次の多項式では係数の数はD^Mで増加

高次元に対する幾何的直観の相違

例：D次元空間の球(半径r=1)

- r=1-εとr=1の間にある体積の割合

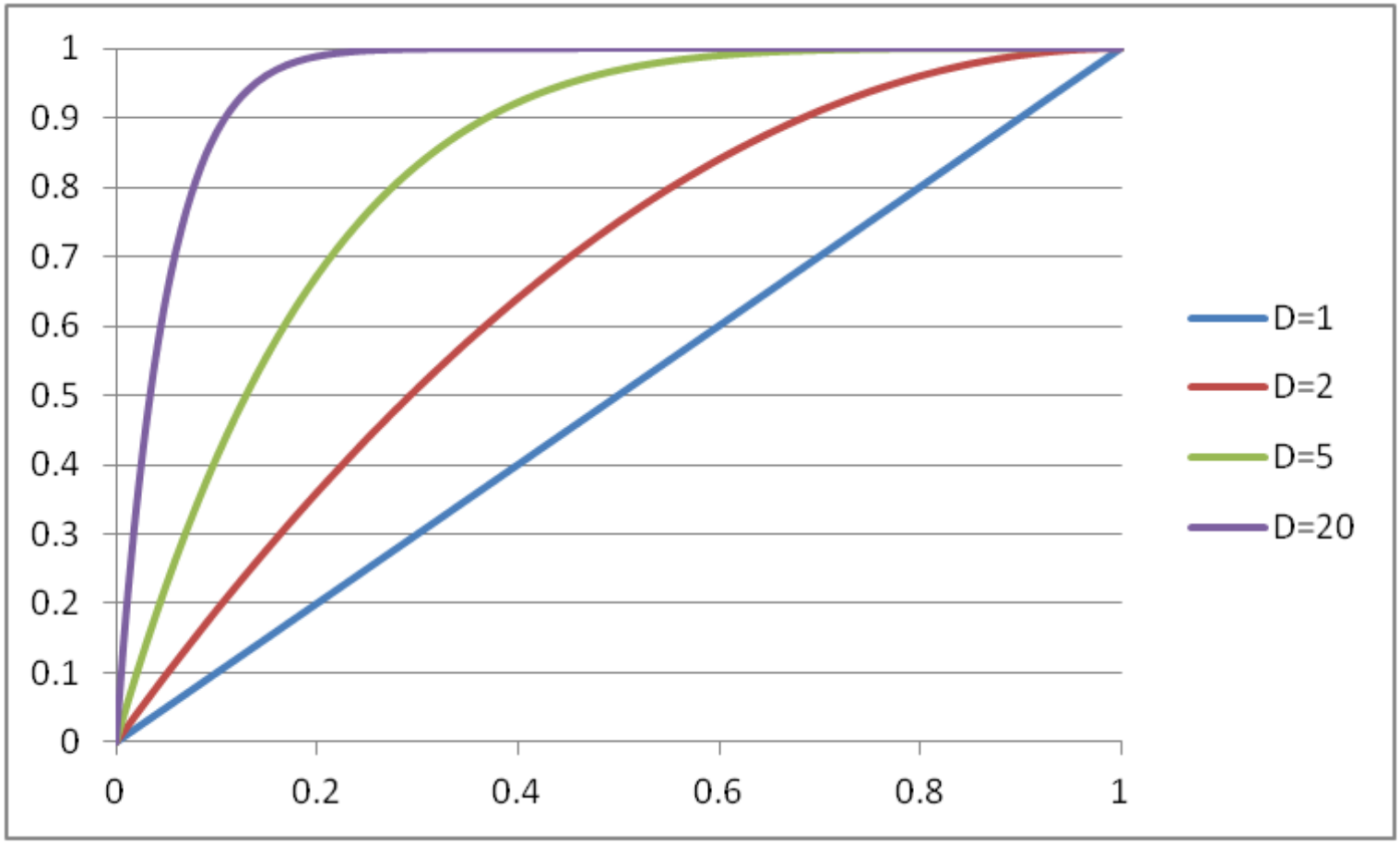
$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

*D次元の半径rの球の体積は、

$$V_D(r) = K_D r^D$$

高次元に対する幾何的直観の相違

- r=1-εとr=1の間にある体積の比率

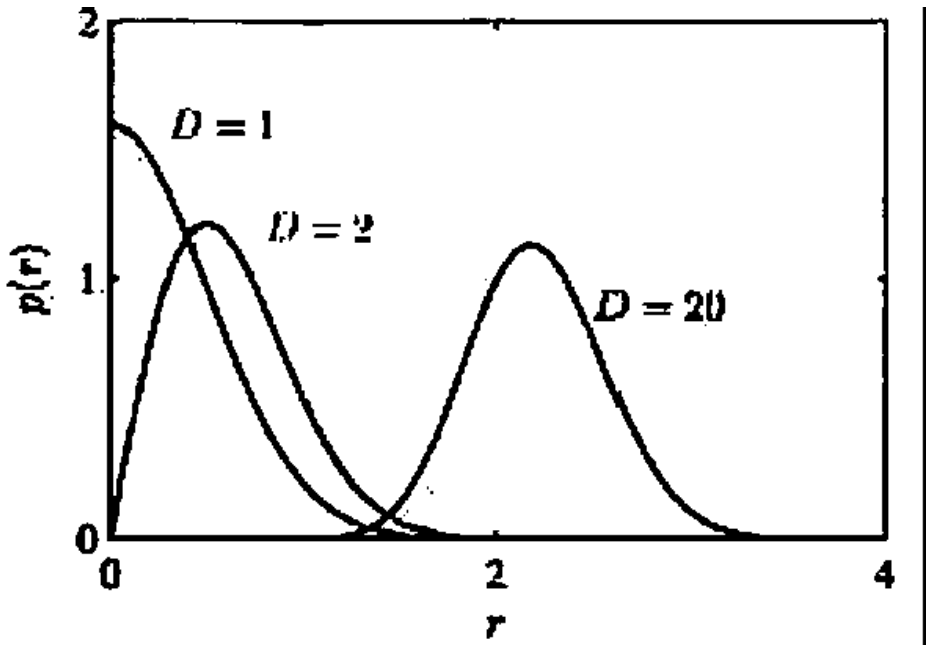


高次元では殆どの体積が表面近くの薄皮に集中

高次元に対する幾何的直観の相違

- 極座標系で原点からの半径rの関数とする密度p(r)

$$p(r) = \frac{S_D r^{D-1}}{(2\pi\sigma^2)^{D/2}} exp\left(-\frac{r^2}{2\sigma^2}\right)$$



p(r)δrは半径rの位置の
厚さδrの薄皮中の確率質量

←Dが大きくなるにつれて
ガウス分布の確率質量は
ある特定の半径における
薄皮に集中する