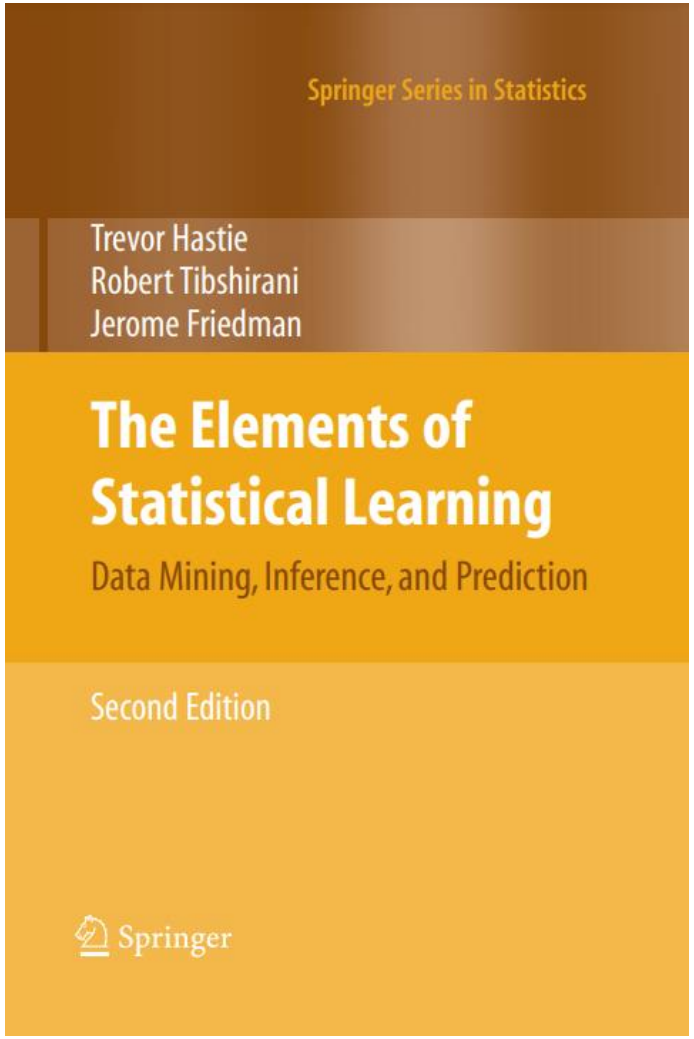
THE ELEMENTS OF STATICAL LEARNING

2023 英語ゼミ　橋本 響

第二章　教師あり学習の概要

- 最小２乗法
- 最近傍法
- 高次元での局所的手法



▪ 線形モデル

線形モデルでは，入力ベクトル $X^T = (X_1, X_2, \ldots, X_p)$
が与えられたとき出力Yをこのように予測する

$$\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{p} X_j \hat{\beta}_j.$$

切片

$$\hat{Y} = X^T \hat{\beta},$$

$\hat{\beta}_0$ を $\hat{\beta}$ にかえたら
表記がカンタンになる.

線形モデルにどのように訓練データを当てはめればよいだろうか？
⇒頻繁に使われるのは**最小２乗法**

- **最小２乗法**

残差 = 実際の観察値 − 予測値

残差２乗和（RSS）：予測モデルの精度を評価し、モデルの良さを評価する指標

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

最小２乗法：残差２乗和(RSS)を最小化するパラメータを見つけるための手法

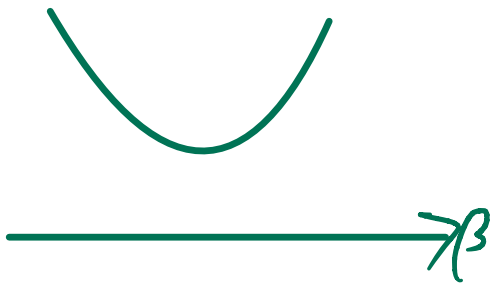
- **最小２乗法**

RSSはβの2次関数

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{N} (y_i - x_i^T \beta)^2.$$

行列表記に

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta)$$



βで偏微分

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = 2(X^T X)\beta - 2X^T y$$

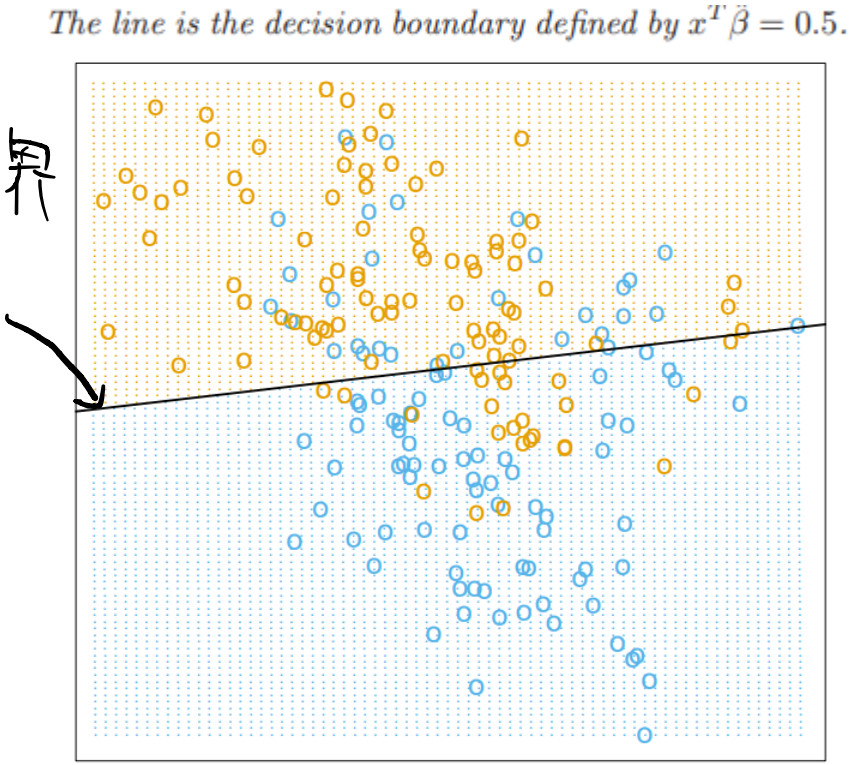
最小なとき傾きが0

$$2(X^T X)\beta - 2X^T y = 0$$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- **分類問題の線形モデル**

応答変数Yを青色が０，オレンジ色が１として
線形モデルを当てはめると…



決定境界の両側に誤分類された点が存在している

⇒線形モデルの柔軟性の欠如？ or 避けられないノイズ？

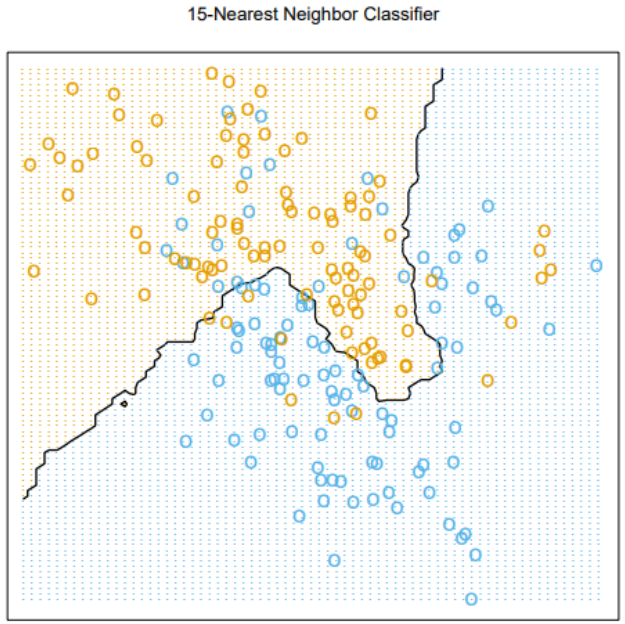
$$\hat{G} = \begin{cases} \text{ORANGE} & \text{if } \hat{Y} > 0.5, \\ \text{BLUE} & \text{if } \hat{Y} \leq 0.5. \end{cases}$$

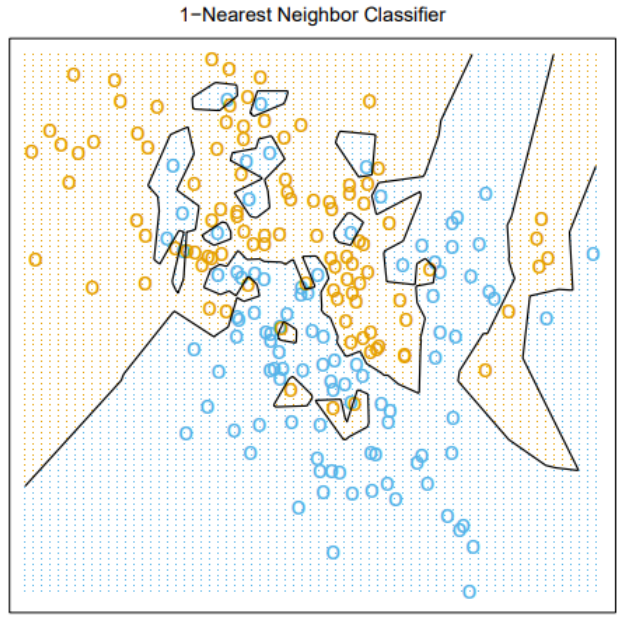
- **k最近傍法**

$\hat{Y}$を予測する際に，入力xに最も近い訓練データ集合の観測値を利用する

⇒k個の近くの点が 青色 or オレンジ色 で多数決を取る

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$





$N_k(x)$

⇒訓練データのうち x に近い k 個の点

- **損失関数**

入力Xが与えられたとき，Yを予測する関数$f(X)$を見つけるための予測に罰則を設ける

⇒予測に対する罰則を定義するための**損失関数**を導入

損失関数のうち最も頻繁に利用される，2乗誤差損失(squared error loss)は

$$L\big(Y, f(X)\big) = (Y - f(X))^2$$ で定義され，

fを選ぶ基準として，期待値予測誤差(expected prediction error)は

$$\mathrm{EP}E(f) = E(Y - f(X))^2$$ で定義される

⇒EPEを最小化することで良い予測を構築することが目標

- **期待予測誤差**

$$E[x] = \int_{-\infty}^{\infty} x\, p(x)\, dx$$

$$= \int x\, p(dx)$$

$$\begin{aligned} \mathrm{EPE}(f) &= \mathrm{E}(Y - f(X))^2 \\ &= \int [y - f(x)]^2 \Pr(dx, dy), \end{aligned}$$

同時分布 $\Pr(dx, dy)$ を X で条件づけ

$$\mathrm{EPE}(f) = \mathrm{E}_X \mathrm{E}_{Y|X} \left([Y - f(X)]^2 | X \right)$$

EPEを最小化するにはそれぞれの x について最小化を行う

$$f(x) = \mathrm{argmin}_c \mathrm{E}_{Y|X} \left([Y - c]^2 | X = x \right)$$

最小化するのは条件つき期待値

$$f(x) = \mathrm{E}(Y | X = x),$$

$$E[(Y-c)^2]$$

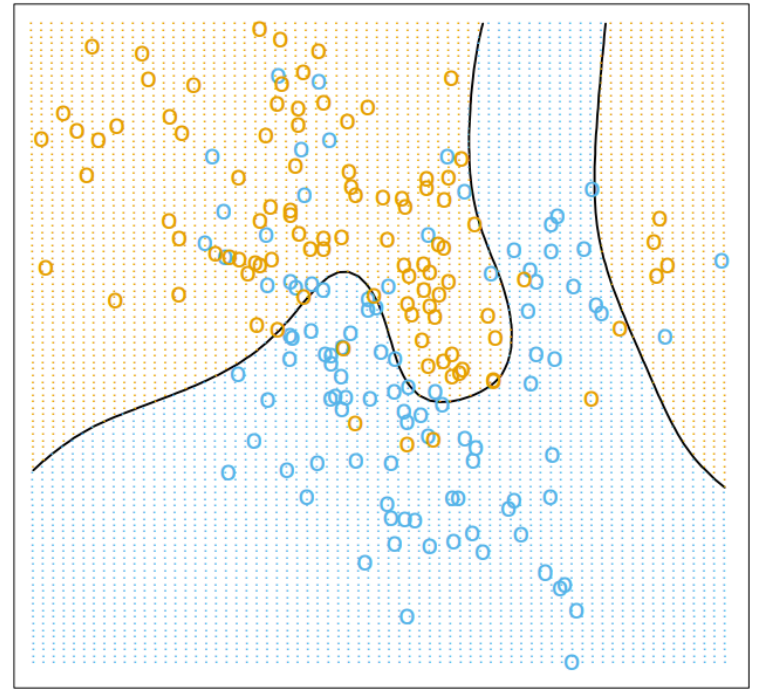
$$= E[Y^2 - 2cY + c^2]$$

$$= E[Y^2] - 2cE[Y] + c^2$$

$$= E[Y^2] - 2cE[Y] + c^2$$

ベイズ分類器

条件付確率を用いて最も確からしいクラスに分類する
⇒理論的に最適な決定境界を示す



出力がカテゴリ型変数Gである場合に
損失関数と期待予測誤差を考えると…

$$\mathrm{EPE} = \mathrm{E}[L(G, \hat{G}(X))],$$

同時分布 $\Pr(X, Y)$ を X で条件づけ

$$\mathrm{EPE} = \mathrm{E}_X \sum_{k=1}^{K} L[\mathcal{G}_k, \hat{G}(X)]\Pr(\mathcal{G}_k|X)$$

最小化したい

$$\hat{G}(x) = \mathrm{argmin}_{g\in\mathcal{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, g)\Pr(\mathcal{G}_k|X = x).$$

または

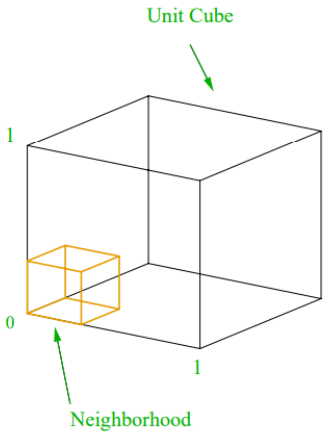
$$\hat{G}(x) = \mathrm{argmin}_{g\in\mathcal{G}}[1 - \Pr(g|X = x)]$$

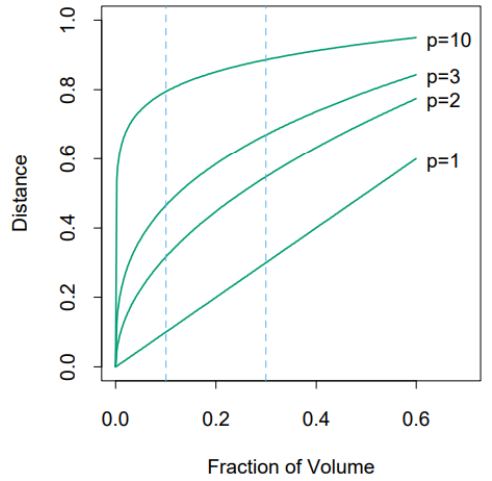
と表わすことができる

- # 高次元での局所的手法

訓練データが多ければ多いほど,k最近傍法を使えばより
多くの近傍データを利用し最適な予測ができる…？

⇒高次元では上手く近似できなくなってしまう!
⇒この現象を次元の呪いとよぶ





例)右のp次元単位立方体

近傍のデータ数が全データの割合rであるとすると
立方体の１辺の長さの期待値は$e_p(r) = r^{1/p}$で表される

$p = 10$の場合，　$e_{10}(0.01) = 0.63$となる

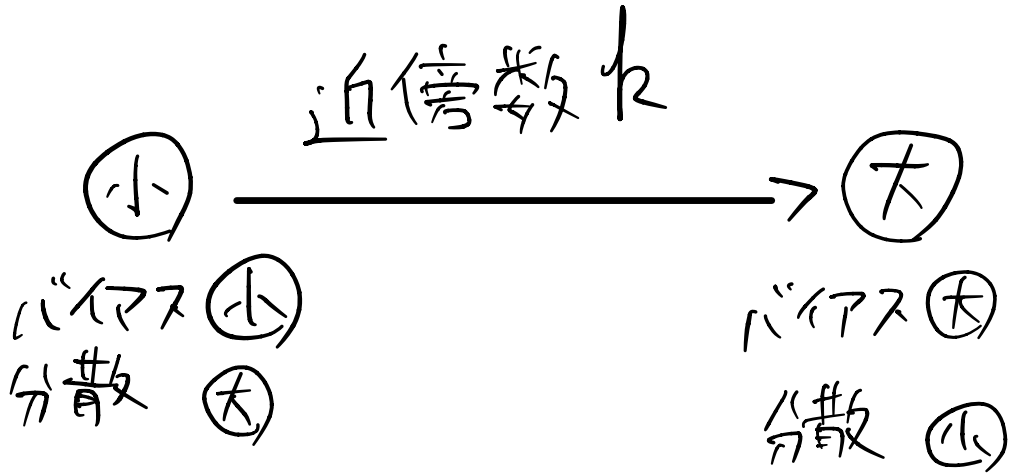
⇒ 全データの0.01%を近傍にするためには入力変数の63%

- **バイアスと分散のトレードオフ**

バイアス：極めて複雑な実際の事象を単純なモデルで近似したために生じる誤差
分散　　：異なる訓練データを使ったときにどの程度$\hat{f}$が変化するかを表す量

$$Y = f(X) + \varepsilon,$$ からデータが生成されたとして，

$$\begin{cases} \mathrm{E}(\varepsilon) = 0 \\ \mathrm{Var}(\varepsilon) = \sigma^2 \end{cases}$$



$$\begin{aligned} \mathrm{EPE}_k(x_0) &= \mathrm{E}[(Y - \hat{f}_k(x_0))^2 | X = x_0] \\ &= \sigma^2 + [\mathrm{Bias}^2(\hat{f}_k(x_0)) + \mathrm{Var}_{\mathcal{T}}(\hat{f}_k(x_0))] \\ &= \sigma^2 + \left[f(x_0) - \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})\right]^2 + \frac{\sigma^2}{k}. \end{aligned}$$

σ^2：削減できない → 新しいテスト点による誤差

$\left[f(x_0) - \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})\right]^2 + \frac{\sigma^2}{k}$：削減可能

▪ バイアスと分散のトレードオフ

訓練誤差はモデルの複雑度を増やすほど（データに強く適合させるほど）減少する
⇒過度に適合させると訓練データに特化したモデルになり
　大きな分散をもち，汎用性が悪化

逆にモデルが十分の複雑度をもっていない
⇒過小適合となって大きなバイアスをもち，汎用性が悪化