The Elements of Statistical Learning
統計的学習の基礎

nukui

2019 年 10 月 5 日

1 序章

2 教師あり学習の概要

2.1

K クラス分類問題の目標変数 t_k が、第 k 要素のみが 1 で他の要素は 0 である K 次元ベクトルによって表されているとする。予測結果 $\hat{y}$ を全ての要素の和が 1 となるように正規化したとき、$\hat{y}$ の最大要素を持つクラスへの分類と $||t_k - \hat{y}||$ を最小化するクラスへの分類が等価であることを示せ。

$\hat{y}$ の最大要素が第 j 要素 $\hat{y_j}$ であるとする。

$$
\begin{aligned}
&||t_k - \hat{y}||^2 - ||t_j - \hat{y}||^2 \\
&= \{\sum_{i \neq k} \hat{y_i}^2 + (1 - \hat{y_k})^2\} - \{\sum_{i \neq j} \hat{y_i}^2 + (1 - \hat{y_j})^2\} \\
&= \hat{y_j}^2 + (1 - \hat{y_k})^2 - \hat{y_k}^2 - (1 - \hat{y_j})^2 \\
&= 2(\hat{y_j} - \hat{y_k}) \geq 0
\end{aligned}
$$

よって、$k = j$ のとき、かつその時に限り、$||t_k - \hat{y}||$ が最小値 $||t_j - \hat{y}||$ になることがわかる。以上より、$\hat{y}$ の最大要素を持つクラスへの分類 j と、$||t_k - \hat{y}||$ を最小化するクラスへの分類が等価であることがわかる。

2.2

図 2.5 の試行の例においてベイズ決定境界を求めよ。

（青色クラス）2 次元ガウス分布 $N((1,0)^T, \mathbf{I})$ から生成された 10 個の平均ベクトル（青色クラス）を $\{m_1, m_2, ..., m_{10}\}$ とし、2 次元ガウス分布 $N((0,1)^T, \mathbf{I})$ から生成された 10 個の平均ベクトル（オレンジ色クラス）を $\{n_1, n_2, ..., n_{10}\}$ とする。

2.2.1

$\{m_1, m_2, ..., m_{10}\}$ と $\{n_1, n_2, ..., n_{10}\}$ の値がすでにわかっていると仮定する。このとき、ベイズ決定境界上の点 x の条件は

$$\Pr(\text{青色} \, |x) = \Pr(\text{オレンジ色} \, |x)$$

となる。

$$\frac{\Pr(\text{青色} \, |x)}{\Pr(\text{オレンジ色} \, |x)} = \frac{\Pr(\text{青色} \, |x) \Pr(x)}{\Pr(\text{オレンジ色} \, |x) \Pr(x)} = \frac{\Pr(x| \, \text{青色}) \Pr(\text{青色})}{\Pr(x| \, \text{オレンジ色}) \Pr(\text{オレンジ色})}$$

$\Pr$(青色) $= \Pr$(オレンジ色) なので、結局、ベイズ決定境界の式は

$$\Pr(x|\text{ 青色}) = \Pr(x|\text{ オレンジ色})$$

と表せる。10 個の平均ベクトルのどれが選ばれるかは等確率であり、i 番目のベクトルが選ばれた時には、平均 m_i(または n_i) で、分散 $\mathbf{I/5}$ の 2 変数ガウス分布に従うので、ベイズ決定境界上の点 x の条件式は

$$\sum_{i=1}^{10} \frac{1}{10} \frac{\sqrt{5}}{2\pi} \exp\{-\frac{5(x-m_i)^T(x-m_i)}{2}\} = \sum_{i=1}^{10} \frac{1}{10} \frac{\sqrt{5}}{2\pi} \exp\{-\frac{5(x-n_i)^T(x-n_i)}{2}\}$$

と表される。

2.2.2

$\{m_1, m_2, ..., m_{10}\}$ と $\{n_1, n_2, ..., n_{10}\}$ の値が未知だと仮定した場合、観測された値を頼りに確率を計算することができる。N 個の p 次元ベクトル x_i $(i = 1, ..., N)$ が青色の点として観測されていて、y_i $(i = 1, ..., N)$ がオレンジ色の点として観測されているとする。平均 μ で、分散 σ の 2 変数ガウス分布の x における確率密度関数を $f(x; \mu, \sigma)$ と表すとすると、

$$\begin{aligned}
&\Pr(x \mid \{x_1, x_2, ..., x_N\}, \text{青色}) \\
&= \sum_{m_1,...,m_{10}} \Pr(\{m_1, ..., m_{10}\} \mid \{x_1, x_2, ..., x_N\}) \Pr(x|\{m_1, ..., m_{10}\}) \\
&= \sum_{m_1,...,m_{10}} \frac{\Pr(\{m_1, ..., m_{10}\}) \Pr(\{x_1, ..., x_N\}|\{m_1, ..., m_{10}\})}{\Pr(\{x_1, x_2, ..., x_N\})} \Pr(x|\{m_1, ..., m_{10}\}) \\
&= \sum_{m_1,...,m_{10}} \frac{\{\prod_{i=1}^{10} f(m_i; (1,0)^T, \mathbf{I})\}\{\prod_{k=1}^{N} \sum_{j=1}^{10} \frac{1}{10} f(x_k; m_j, \mathbf{I}/5)\}}{\int \{\prod_{i=1}^{10} f(m_i'; (1,0)^T, \mathbf{I})\}\{\prod_{k=1}^{N} \sum_{j=1}^{10} \frac{1}{10} f(x_k; m_j', \mathbf{I}/5)\} dm_1' ... dm_{10}'} \Pr(x|\{m_1, ..., m_{10}\}) \\
&= \int \{ \frac{\{\prod_{i=1}^{10} f(m_i; (1,0)^T, \mathbf{I})\}\{\prod_{k=1}^{N} \sum_{j=1}^{10} \frac{1}{10} f(x_k; m_j, \mathbf{I}/5)\}\{\sum_{j=1}^{10} \frac{1}{10} f(x; m_j, \mathbf{I}/5)\}}{\int \{\prod_{i=1}^{10} f(m_i'; (1,0)^T, \mathbf{I})\}\{\prod_{k=1}^{N} \sum_{j=1}^{10} \frac{1}{10} f(x_k; m_j', \mathbf{I}/5)\} dm_1' ... dm_{10}'} \} dm_1 ... dm_{10}
\end{aligned}$$

となる。また、ベイズ決定境界上の点 x の条件式は

$$\Pr(x|\{x_1, x_2, ..., x_N\}, \text{青色}) = \Pr(x|\{y_1, y_2, ..., y_N\}, \text{オレンジ色})$$

と表される。これを計算機でいい感じに求めることができるのかどうか、知りませんが。。。

2.3

式 (2.24) を導け。

$$d(p, N) = (1 - (\frac{1}{2})^{\frac{1}{N}})^{\frac{1}{p}}$$

p 次元空間の半径 r の球の体積を Cr^p とおく (C は定数)。半径 1 の球に N 個の点が均一に散らばっているとすると、原点に最も近い点 x が半径 r の中に入っている確率は

$$\begin{aligned}
\Pr(X < r) &= 1 - \Pr(X \geq r) \\
&= 1 - (\frac{C - Cr^p}{C})^N \\
&= 1 - (1 - r^p)^N
\end{aligned}$$

よって、X の中央値 d は $\Pr(X < d) = \frac{1}{2}$ となる境界なので、

$$\begin{aligned} 1 - (1 - d^p)^N &= \frac{1}{2} \\ \frac{1}{2} &= (1 - d^p)^N \\ 1 - d^p &= (\frac{1}{2})^{\frac{1}{N}} \\ d &= (1 - (\frac{1}{2})^{\frac{1}{N}})^{\frac{1}{p}} \end{aligned}$$

以上より、$d(p, N) = (1 - (\frac{1}{2})^{\frac{1}{N}})^{\frac{1}{p}}$ と求められた。

2.4

Later

2.5

(a) 式 (2.27) を導出せよ。
(b) 式 (2.28) を導出せよ。

(a)

$$\begin{aligned} \mathrm{EPE}(x_0) &= \mathbb{E}_{y_0|x_0}[\mathbb{E}_{\mathcal{T}}[(y_0 - \hat{y}_0)^2]] \\ &= \mathbb{E}_{y_0|x_0}[\mathbb{E}_{\mathcal{T}}[y_0^2 - 2y_0\hat{y}_0 + \hat{y}_0^2]] \end{aligned}$$

ここで、y_0 はトレーニングデータ $\mathcal{T}$ には依存せず、予測値 $\hat{y}_0$ は真の値 y_0 には依存しないので、

$$\mathrm{EPE}(x_0) = \mathbb{E}_{y_0|x_0}[y_0^2] - 2\mathbb{E}_{y_0|x_0}[y_0]\mathbb{E}_{\mathcal{T}}[\hat{y}_0] + \mathbb{E}_{\mathcal{T}}[\hat{y}_0^2]$$

分散と平均の関係 $\mathbb{E}(x^2) = \mathrm{Var(x)} + \mathbb{E}(x)^2$ を用いることで、

$$\begin{aligned} \mathrm{EPE}(x_0) &= \mathrm{Var}(y_0|x_0) + \mathbb{E}_{y_0|x_0}[y_0]^2 + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathbb{E}_{\mathcal{T}}[\hat{y}_0]^2 - 2\mathbb{E}_{y_0|x_0}[y_0]\mathbb{E}_{\mathcal{T}}[\hat{y}_0] \\ &= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + (\mathbb{E}_{y_0|x_0}[y_0] - \mathbb{E}_{\mathcal{T}}[\hat{y}_0])^2 \end{aligned}$$

ここで $\mathrm{Var}(y_0|x_0)$ は真の値の分散 σ^2 である。
$|\mathbb{E}_{y_0|x_0}[y_0] - \mathbb{E}_{\mathcal{T}}[\hat{y}_0]|$ は Bias と呼ばれる値で、真の値の期待値 $\mathbb{E}_{y_0|x_0}[y_0]$ とモデルの予測した値の期待値 $\mathbb{E}_{\mathcal{T}}[\hat{y}_0]$ の間の乖離を示す。最小二乗推定は不偏であることが知られており、Bias は 0 になる。

以下では、$\mathrm{Var}_{\mathcal{T}}(\hat{y}_0)$ を計算する。

トレーニングデータ $\mathcal{T}$ は入力 $\mathbf{X}$ と出力 y の組み合わせとして表現できるので、$\mathcal{T} = (\mathbf{X}, y)$ と書ける。

$$
\begin{aligned}
\mathrm{Var}_{\mathcal{T}}(\hat{y}_0) &= \mathbb{E}_{(\mathbf{X},y)}[\hat{y}_0^2] - \mathbb{E}_{(\mathbf{X},y)}[\hat{y}_0]^2 \\
&= \int \{\Pr(\mathbf{X}, y)\hat{y}_0^2\} d\mathbf{X}dy - (\int \{\Pr(\mathbf{X}, y)\hat{y}_0\} d\mathbf{X}dy)^2 \\
&= \int \{\Pr(\mathbf{X})(\int \{\Pr(y|\mathbf{X})\hat{y}_0^2\} dy)\} d\mathbf{X} - (\int \{\Pr(\mathbf{X})(\int \{\Pr(y|\mathbf{X})\hat{y}_0\} dy)\} d\mathbf{X})^2 \\
&= \int \{\Pr(\mathbf{X})\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0^2]\} d\mathbf{X} - (\int \{\Pr(\mathbf{X})\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]\} d\mathbf{X})^2 \\
&= \int \{\Pr(\mathbf{X})(\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0) + \mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]^2)\} d\mathbf{X} - (\int \{\Pr(\mathbf{X})\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]\} d\mathbf{X})^2 \\
&= \int \{\Pr(\mathbf{X})\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)\} d\mathbf{X} + \int \{\Pr(\mathbf{X})\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]^2\} d\mathbf{X} - (\int \{\Pr(\mathbf{X})\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]\} d\mathbf{X})^2 \\
&= \mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)] + \mathbb{E}_{\mathbf{X}}[(\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0])^2] - (\mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]])^2 \\
&= \mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)] + \mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0])
\end{aligned}
$$

以下では、$\mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)]$ と $\mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0])$ を計算していく。まず、$\mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)]$ を展開すると

$$
\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)] &= \mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(x_0^T\hat{\beta})] \\
&= \mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)^2] - (\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)])^2] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)^T] \\
&\qquad - (\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)])(\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)^T])] \\
&= \mathbb{E}_{\mathbf{X}}[\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)(y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0)] \\
&\qquad - (\mathbb{E}_{\Pr(y|\mathbf{X})}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty)])(\mathbb{E}_{\Pr(y|\mathbf{X})}[y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0])] \\
&= \mathbb{E}_{\mathbf{X}}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_{\Pr(y|\mathbf{X})}[yy^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0) \\
&\qquad - (x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbb{E}_{\Pr(y|\mathbf{X})}[y]\mathbb{E}_{\Pr(y|\mathbf{X})}[y^T]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0)] \\
&= \mathbb{E}_{\mathbf{X}}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbb{E}_{\Pr(y|\mathbf{X})}[yy^T] - \mathbb{E}_{\Pr(y|\mathbf{X})}[y]\mathbb{E}_{\Pr(y|\mathbf{X})}[y^T])\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0)]
\end{aligned}
$$

観測値 y_i が互いに独立で、分散 σ^2 を持つとすると $\mathbb{E}_{\Pr(y|\mathbf{X})}[yy^T] - \mathbb{E}_{\Pr(y|\mathbf{X})}[y]\mathbb{E}_{\Pr(y|\mathbf{X})}[y^T] = \sigma^2 I$ なので、

$$
\begin{aligned}
\mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)] &= \mathbb{E}_{\mathbf{X}}[(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\sigma^2 I)\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}x_0)] \\
&= \mathbb{E}_{\mathbf{X}}[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2]
\end{aligned}
$$

ここで、$\mathcal{T} = (\mathbf{X}, y)$ であり、またカッコの中に y に依存する項目が出てこないので、

$$
\mathbb{E}_{\mathbf{X}}[\mathrm{Var}_{\Pr(y|\mathbf{X})}(\hat{y}_0)] = \mathbb{E}_{\mathbf{X}}[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2] = \mathbb{E}_{\mathcal{T}}[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2]
$$

次に、$\mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0])$ を計算していく。

$$
\begin{aligned}
\mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]) &= \mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[x_0^T\hat{\beta}]) \\
&= \mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^Ty]) \\
&= \mathrm{Var}_{\mathbf{X}}(x_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta) \\
&= \mathrm{Var}_{\mathbf{X}}(x_0^T\beta)
\end{aligned}
$$

これは $\mathbf{X}$ に依存しない定数なので、$\mathrm{Var}_{\mathbf{X}}(\mathbb{E}_{\Pr(y|\mathbf{X})}[\hat{y}_0]) = 0$
結局、$\mathrm{EPE}(x_0) = \sigma^2 + \mathbb{E}_{\mathcal{T}}[x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0\sigma^2]$ という (2.27) 式が導かれた。

(b) N が大きく $\mathcal{T}$ がランダムに選択されており、$\mathbb{E}_{\mathcal{T}}(X)=0$ であるとする[*1]。$\mathbf{X}^T\mathbf{X}\to N\mathrm{Cov}(X)$ となるので、

$$\begin{aligned}\mathbb{E}_{x_0}[\mathrm{EPE}(x_0)] &= \mathbb{E}_{x_0}[x_0^T\mathbb{E}_{\mathcal{T}}[(\mathbf{X}^T\mathbf{X})^{-1}]x_0]\sigma^2+\sigma^2\\ &\sim \mathbb{E}_{x_0}[x_0^T(N\mathrm{Cov}(X))^{-1}x_0]\sigma^2+\sigma^2\\ &= \mathbb{E}_{x_0}[x_0^T(\mathrm{Cov}(X))^{-1}x_0]\sigma^2/N+\sigma^2\end{aligned}$$

トレース演算の巡回性 $[\mathrm{trace}(AB)=\mathrm{trace}(BA)]$ および線形性により

$$\begin{aligned}\mathbb{E}_{x_0}[\mathrm{EPE}(x_0)] &\sim \mathrm{trace}[\mathbb{E}_{x_0}[x_0^T(\mathrm{Cov}(X))^{-1}x_0]]\sigma^2/N+\sigma^2\\ &= \mathbb{E}_{x_0}[\mathrm{trace}[x_0^T(\mathrm{Cov}(X))^{-1}x_0]]\sigma^2/N+\sigma^2\\ &= \mathbb{E}_{x_0}[\mathrm{trace}[(\mathrm{Cov}(X))^{-1}x_0x_0^T]]\sigma^2/N+\sigma^2\\ &= \mathrm{trace}[(\mathrm{Cov}(X))^{-1}\mathbb{E}_{x_0}[x_0x_0^T]]\sigma^2/N+\sigma^2\end{aligned}$$

$\mathbb{E}_{\mathcal{T}}[X]=0$ より $\mathbb{E}_{x_0}[x_0]=0$ である。$\mathbb{E}_{x_0}[x_0x_0^T]=\mathbb{E}_{x_0}[(x_0-\mathbb{E}_{x_0}[x_0])(x_0-\mathbb{E}_{x_0}[x_0])^T]=\mathrm{Cov}(x_0)$ となるので、

$$\begin{aligned}\mathbb{E}_{x_0}[\mathrm{EPE}(x_0)] &\sim \mathrm{trace}[(\mathrm{Cov}(X))^{-1}\mathrm{Cov}(x_0)]\sigma^2/N+\sigma^2\\ &= \mathrm{trace}[(\mathrm{Cov}(X))^{-1}\mathrm{Cov}(x_0)]\sigma^2/N+\sigma^2\\ &= \mathrm{trace}[\mathrm{I}]\sigma^2/N+\sigma^2\\ &= p\sigma^2/N+\sigma^2\end{aligned}$$

以上により、式 (2.28) が導かれた。

2.6

Later

2.7

Later

2.8

Later

2.9

$$\begin{aligned}\mathbb{E}_{\{X,y\}}[R_{\mathrm{tr}}(\hat{\beta})] &= \mathbb{E}_{\{X,y\}}[\min_{\beta} R_{\mathrm{tr}}(\beta)]\\ &\le \min_{\beta}\mathbb{E}_{\{X,y\}}[R_{\mathrm{tr}}(\beta)]\\ &= \min_{\beta}\mathbb{E}_{\{\tilde{X},\tilde{y}\}}[R_{\mathrm{te}}(\beta)]\\ &\le \mathbb{E}_{\{\tilde{X},\tilde{y}\}}[R_{\mathrm{te}}(\hat{\beta})]\end{aligned}$$

[*1] $\mathbf{X}$ は観測された値を行とする行列を表しており、X は一つの観測値の確率変数を表す。X は p 次元のベクトル。

上記のそれぞれの等号や不等号の変形を証明すれば良い。

1. $\mathbb{E}_{\{X,y\}}[R_{\mathrm{tr}}(\hat{\beta})] = \mathbb{E}_{\{X,y\}}[\min_\beta R_{\mathrm{tr}}(\beta)]$ は最小 2 乗法の定義より成り立つ。
2. それぞれの値を β で最小化してから期待値（平均値）を取った方が、平均値を取った後に β で最小化するよりも小さいか等しくなるので、$\mathbb{E}_{\{X,y\}}[\min_\beta R_{\mathrm{tr}}(\beta)] \le \min_\beta \mathbb{E}_{\{X,y\}}[R_{\mathrm{tr}}(\beta)]$ が成り立つ。
3.

$$\begin{aligned}\min_\beta \mathbb{E}_{\{X,y\}}[R_{\mathrm{tr}}(\beta)] &= \min_\beta \mathbb{E}_{\{X,y\}}[\frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^{\mathrm{T}} x_i)^2] \\ &= \min_\beta \mathbb{E}_{\{x,y\}}[(y - \beta^{\mathrm{T}} x)^2] \\ &= \min_\beta \mathbb{E}_{\{\tilde{x},\tilde{y}\}}[(\tilde{y} - \beta^{\mathrm{T}} \tilde{x})^2] \\ &= \min_\beta \mathbb{E}_{\{\tilde{X},\tilde{y}\}}[\frac{1}{M}\sum_{i=1}^{M}(\tilde{y_i} - \beta^{\mathrm{T}} \tilde{x_i})^2] \\ &= \min_\beta \mathbb{E}_{\{\tilde{X},\tilde{y}\}}[R_{\mathrm{te}}(\beta)]\end{aligned}$$

   より $\min_\beta \mathbb{E}_{\{X,y\}}[R_{\mathrm{tr}}(\beta)] = \min_\beta \mathbb{E}_{\{\tilde{X},\tilde{y}\}}[R_{\mathrm{te}}(\beta)]$ が成り立つ。
4. $\min_\beta$ という関数の定義から $\min_\beta \mathbb{E}_{\{\tilde{X},\tilde{y}\}}[R_{\mathrm{te}}(\beta)] \le \mathbb{E}_{\{\tilde{X},\tilde{y}\}}[R_{\mathrm{te}}(\hat{\beta})]$ が成り立つ。

3　回帰のための線形手法

Later

4　分類のための線形手法

Later

5　基底展開と正則化

5.1

式 (5.3) 中の切断べき基底関数が、文中で述べられていたような二つの節点を持つ 3 次スプラインの基底を表すことを示せ。

まず、$h_1(X) = 1,\ h_2(X) = X,\ h_3(X) = X^2,\ h_4(X) = X^3,\ h_5(X) = (X-\xi_1)_+^3,\ h_6(X) = (X-\xi_2)_+^3$ と定義すると、任意の $\beta_m\ (m = 1, 2, .., 6)$ に対して、$f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$ とその 1 階導関数、および 2 階導関数がそれぞれ連続になることを示す。そのためには、各 i に対して関数 h_i とその 1 階導関数,2 階導関数がそれぞれ連続になることを示せば十分。h_1, h_2, h_3, h_4 は明らかに無限回微分可能である。h_5 が $X = \xi_1$ 以外の点で連続であることは明らかである。よって、$X = \xi_1$ の点で、$h_5,\ h_5', h_5''$ がそれぞれ連続であることを確かめる。

$$\lim_{X \to \xi_1 + 0} h_5(X) = \lim_{X \to \xi_1 + 0} (X - \xi_1)^3 = 0$$

$$\lim_{X \to \xi_1 - 0} h_5(X) = \lim_{X \to \xi_1 - 0} 0 = 0$$

よって、$h_5(X)$ は $X = \xi_1$ で連続である。また、

$$\lim_{X \to \xi_1 + 0} h_5'(X) = \lim_{X \to \xi_1 + 0} 3(X - \xi_1)^2 = 0$$

$$\lim_{X \to \xi_1 - 0} h_5'(X) = \lim_{X \to \xi_1 - 0} 0 = 0$$

よって、$h_5'(X)$ は $X = \xi_1$ で連続である。さらに、

$$\lim_{X \to \xi_1 + 0} h_5''(X) = \lim_{X \to \xi_1 + 0} 6(X - \xi_1) = 0$$

$$\lim_{X \to \xi_1 - 0} h_5''(X) = \lim_{X \to \xi_1 - 0} 0 = 0$$

よって、$h_5''(X)$ も $X = \xi_1$ の点で連続である。同様に $h_6,\ h_6', h_6''$ の連続性も証明できる。以上より、$f(X)$ とその 1 階導関数、および 2 階導関数がそれぞれ連続になり、3 次スプラインの条件を満たす。

次に、このように定義された f が 3 次スプラインの基底であることを確かめる。$h_1, h_2, h_3, h_4, h_5, h_6$ が 1 次独立であることは明らかである[*2]。よって、3 次スプラインの条件を満たす任意の関数 $g(X)$ が $f(X) = \sum_{m=1}^{M} \beta_m h_m(X)$ の形で表現できることを示せば十分である。3 次関数 $g : (\xi_0, \xi_3) \to \mathbb{R}$ が 3 次関数 g_1, g_2, g_3 を用いて各領域ごとに以下のように定義されるとしても g は一般性を失わない。

$$g(X) = \begin{cases} g_1(X) \quad (\xi_0 \le X \le \xi_1) \\ g_1(X) + g_2(X) \quad (\xi_1 \le X \le \xi_2) \\ g_1(X) + g_2(X) + g_3(X) \quad (\xi_2 \le X \le \xi_3) \end{cases}$$

このとき、$g(X)$ の連続性により、$g_2(\xi_1) = 0,\ g_3(\xi_2) = 0$ が成り立つ。よって、

$$\begin{aligned} g_1(X) &= aX^3 + bX^2 + cX + d \\ g_2(X) &= e(X - \xi_1)(X^2 + hX + i) \\ g_3(X) &= j(X - \xi_2)(X^2 + kX + l) \end{aligned}$$

と表現できるはずである。さらに、$g'(X)$ の連続性により、$g_2'(\xi_1) = 0,\ g_3'(\xi_2) = 0$ が成り立つはずなので、$e = 0$ または $\xi_1^2 + h\xi_1 + i = 0$ が成り立ち、さらに $j = 0$ または $\xi_2^2 + k\xi_2 + l = 0$ が成り立つ。よって、

$$\begin{aligned} g_1(X) &= aX^3 + bX^2 + cX + d \\ g_2(X) &= e(X - \xi_1)^2(X + m) \\ g_3(X) &= j(X - \xi_2)^2(X + n) \end{aligned}$$

と表現できることがわかった。さらに、$g''(X)$ の連続性により、$g_2''(\xi_1) = 0,\ g_3''(\xi_2) = 0$ が成り立つはずなので、$e = 0$ または $\xi_1 + m = 0$ が成り立ち、さらに $j = 0$ または $\xi_2 + n = 0$ が成り立つ。結局、

$$\begin{aligned} g_1(X) &= aX^3 + bX^2 + cX + d \\ g_2(X) &= e(X - \xi_1)^3 \\ g_3(X) &= j(X - \xi_2)^3 \end{aligned}$$

と表現できることになり、この $g(X)$ は関数 $f(X)$ と同じ形になった。以上より、関数 f の形で任意の 3 次スプライン関数が表現できることになり、関数 $h_1, ..., h_6$ は 3 次スプラインの基底であることがわかった。

[*2] $\{h_n\}_{n=1}^{6}$ の中のどの元も残りの元の線型結合で表現できない。