

# The Elements of Statistical Learning

## 統計的学習の基礎

nukui

2019 年 2 月 10 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。

$$\begin{aligned} & \|t_k - \hat{y}\|^2 - \|t_j - \hat{y}\|^2 \\ &= \left\{ \sum_{i \neq k} \hat{y}_i^2 + (1 - \hat{y}_k)^2 \right\} - \left\{ \sum_{i \neq j} \hat{y}_i^2 + (1 - \hat{y}_j)^2 \right\} \\ &= \hat{y}_j^2 + (1 - \hat{y}_k)^2 - \hat{y}_k^2 - (1 - \hat{y}_j)^2 \\ &= 2(\hat{y}_j - \hat{y}_k) \geq 0 \end{aligned}$$

よって、 $k = j$  のとき、 $\|t_k - \hat{y}\|$  が最小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  $\{m_1, m_2, \dots, m_{10}\}$  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  $\{n_1, n_2, \dots, n_{10}\}$  とする。

#### 2.2.1

$\{m_1, m_2, \dots, m_{10}\}$  と  $\{n_1, n_2, \dots, n_{10}\}$  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。

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

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

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

と表される。

### 2.2.2

$\{m_1, m_2, \dots, m_{10}\}$  と  $\{n_1, n_2, \dots, n_{10}\}$  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  $x_i$  ( $i = 1, \dots, N$ ) が青色の点として観測されていて、 $y_i$  ( $i = 1, \dots, N$ ) がオレンジ色の点として観測されているとする。平均  $\mu$  で、分散  $\sigma$  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、

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

となる。また、ベイズ決定境界上の点  $x$  の条件式は

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

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。

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

$p$  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

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

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

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

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

2.4

2.5

(a) 式 (2.27) を導出せよ。

(b) 式 (2.28) を導出せよ。

(a)

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

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、

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

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

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

ここで  $\text{Var}(y_0|x_0)$  は真の値の分散  $\sigma^2$  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、 $\text{Var}_{\mathcal{T}}(\hat{y}_0)$  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。

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

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

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

観測値  $y_i$  が互いに無相関で、分散  $\sigma^2$  を持つとすると  $\mathbb{E}_{\text{Pr}(y|\mathbf{X})}[y y^T] - \mathbb{E}_{\text{Pr}(y|\mathbf{X})}[y] \mathbb{E}_{\text{Pr}(y|\mathbf{X})}[y^T] = \sigma^2 I$  なので、

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

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で、

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

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

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

これは  $\mathbf{X}$  に依存しない定数なので、 $\text{Var}_{\mathbf{X}}(\mathbb{E}_{\text{Pr}(y|\mathbf{X})}[\hat{y}_0]) = 0$

結局、 $\text{EPE}(x_0) = \sigma^2 + \mathbb{E}_{\mathcal{T}}[x_0^T (\mathbf{X}^T \mathbf{X})^{-1} x_0 \sigma^2]$  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<sup>\*1</sup>。  $\mathbf{X}^T \mathbf{X} \rightarrow N \text{Cov}(X)$  となるので、

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

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

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

<sup>\*1</sup>  $\mathbf{X}$  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。

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

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

以上により、式 (2.28) が導かれた。