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

nukui

2019 年 1 月 3 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。

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

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 $\{m_1, m_2, ..., m_{10}\}$ 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。