第1章　統計的学習の考え方

1.1　導入

データサイエンスにおいて, データの視覚化や整理[1]は重要であるが, 主な課題はデータの数理的な解析である. データの不確実性 (ランダム性) のモデル化や定量化を目的とするとき, これらのことを扱う分野を**統計的推測**/**統計的学習**といい, 大規模データを用いた予測に重点をおく分野を**機械学習**または**データマイニング**という. 何に重きをおくかの違いで, ふたつの文化[2]が存在する.

データのモデリングにはふたつの目的がある.

(1) ある変量の将来の値を正確に予測すること.

(2) データの中にある未知の面白いパタンを発見すること.

この目的を達成するために, 数理科学の 3 つの重要な支柱からの知識を利用する.

(1) **関数の近似**: データに対して数理モデルを作ることは, ある変数が他の変数にどのように依存/関係しているかを理解することである. 変数間の関連を表現する最も自然なやり方は, 関数もしくは写像を用いることである. この関数ないしは写像が解析者[3]には完全には特定できないとき, データに基づきこの関数の近似を試みる.

(2) **最適化**: 統計的モデル[4]が与えられたとき, この族の中で最もよい確率分布をみつけたい. このために, 効率的な探索と最適化の手続きが必要となる. 最適化のステップは観測データへの関数のあてはめや校正と考えることができる. このステップでは, 最適化アルゴリズムや効率的なプログラミングが必要となる.

---

[1] いわゆる記述統計学の重要な内容.

[2] Breiman (2001): Statist Sci **16**(3), pp. 199-231 を参照のこと.

[3] 統計家, データサイエンティストなどデータを分析する人々の総称として解析者とよぶことにする.

[4] 統計的推測では, 確率分布の族を統計的モデルという. 「モデル」という言葉が何をさしているかを文脈から正確に理解することが大切である.

(3) **確率論と統計的推測**: モデルをあてはめるデータはランダムな過程の実現値とみなす．ここでは, 分布や確率が重要になる．将来の予測を行うときに内包する不確実性を定量化するために, 確率論や統計的推測論の知識が肝要である．

多次元のデータを扱う場合には, これらの知識に加えて, 線型代数の知識が重要となる．

1.2 教師あり学習と教師なし学習

入力ベクトル $\boldsymbol{x}$ が与えられたとき, 出力 y の予測をすることが機械学習の主な目的である．たとえば, $\boldsymbol{x}$ をデジタル化された署名データとし, y を 2 値の出力で, 署名が「本人のもの」か「本人のものではない」のいずれかを表すとする．別の例では, 入力 $\boldsymbol{x}$ は妊婦の体重と喫煙嗜好とし, 出力 y を生まれてくる新生子の体重とする．入力 $\boldsymbol{x}$ による出力 y の予測値を関数 g を用いて, $g(\boldsymbol{x})$ と表す．g の定義域は, 入力 $\boldsymbol{x}$ の取りうる値の集合を含むものであり, 終域は y の取りうる値の集合である．関数 g のことを**予測関数**とよぶことにする．関数 g は, データがもつランダム性を除いた入力 $\boldsymbol{x}$ と出力 y の関連性についての情報を内包するものである．出力 y が実数値をとるとき, 出力から入力を予測することを**回帰問題**といい, 出力 y が 2 値もしくは有限集合に値をとるとき, **判別問題**[5]という．

入力 $\boldsymbol{x}$ による出力 y の予測値を $\widehat{y} := g(\boldsymbol{x})$ と書いたとき, この予測値の精度を**損失関数** $\mathsf{Loss}(y, \widehat{y})$ で測る．損失関数は非負値関数で, 精度がわるいほどおおきな値をとるものである．2 値の判別問題では, 損失関数を **0-1 損失**とすることが基本的である．すなわち

$$\mathsf{Loss}(y, \widehat{y}) = \mathbb{1}\{y \neq \widehat{y}\} = \begin{cases} 1 & (y \neq \widehat{y}) \\ 0 & (y = \widehat{y}) \end{cases}$$

である．後で 2 値の判別問題でも別の損失関数を考えることになる．回帰問題では, 2 乗損失

$$\mathsf{Loss}(y, \widehat{y}) = \big(y - \widehat{y}\big)^2$$

が最もよく使われる．

すべての入力と出力の組 $(\boldsymbol{x}, y)$ に対して正確な予測をすることができる予測関数をみつけることは一般に不可能である．これはデータのもつランダム性に由来する．さきほどの例において同じ値の入力に対して異

[5]分類問題ともいう．

なる出力をもつ組が存在することを考えればよい. すなわち各 $(\boldsymbol{x}, y)$ はある同時分布に従う確率ベクトル $(\boldsymbol{X}, Y)$ の実現値と考えることにする. このことで予測関数 g の精度を**期待損失**(**誤差**/**リスク**)で定量化する. すなわち

$$\mathsf{Err}(g) := \mathsf{E}\big[\mathsf{Loss}\big(Y, g(X)\big)\big]$$

である. ただし期待値 $\mathsf{E}[\cdot]$ は $(\boldsymbol{X}, Y)$ の同時分布に関するものである. $0-1$ 損失を用いた判別問題では, 誤差は誤差判別の確率となる. すなわち

$$\mathsf{Err}(g) = \mathsf{Pr}\big[Y \neq g(\boldsymbol{X})\big]$$

である. この文脈では予測関数のことを**判別器**とよぶ.

確率ベクトル $(\boldsymbol{X}, Y)$ の同時分布と損失関数が与えられると, 最良の判別器を求めることができる. $c \in \mathbb{N}$, $c \geq 2$ とし, Y は $\{0, 1, \ldots, c-1\}$ に値を取るとする. このとき

$$g^* \in \arg\min_{g} \mathsf{E}\big[\mathsf{Loss}(Y, g(\boldsymbol{X}))\big]$$

を誤差を最小とする判別器とする. すなわち, 任意の判別器 g に対して

$$\mathsf{E}\big[\mathsf{Loss}(Y, g(\boldsymbol{X}))\big] \geq \mathsf{E}\big[\mathsf{Loss}(Y, g^*(\boldsymbol{X}))\big]$$

が成立している.

$X = \boldsymbol{x}$ が与えられたとき, $g^*(\boldsymbol{x})$ を具体的に求めてみる. そのために, $X = \boldsymbol{x}$ が与えられたときの Y の条件付き p.d.f. (または p.m.f.) を $\mathsf{p}^*(y|\,\boldsymbol{x})$ と書く. すなわち

$$\mathsf{p}^*(y|\,\boldsymbol{x}) = \mathsf{Pr}\big(Y = y \big|\, \boldsymbol{X} = \boldsymbol{x}\big)$$

である. このとき

$$\mathsf{E}[\mathsf{Loss}(Y, g(\boldsymbol{X}))] = \mathsf{Pr}[Y \neq g(\boldsymbol{X}))] = 1 - \mathsf{Pr}[Y = g(\boldsymbol{X}))]$$

なので

$$\begin{aligned} g^*(\boldsymbol{x}) &= \mathop{\arg\min}_{y \in \{0,1,\ldots,c-1\}} \mathsf{E}[\mathsf{Loss}(Y, y)|\, \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathop{\arg\min}_{y \in \{0,1,\ldots,c-1\}} \big\{1 - \mathsf{Pr}(Y = y|\, \boldsymbol{X} = \boldsymbol{x})\big\} \\ &= \mathop{\arg\max}_{y \in \{0,1,\ldots,c-1\}} \mathsf{Pr}(Y = y|\, \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathop{\arg\max}_{y \in \{0,1,\ldots,c-1\}} \mathsf{p}^*(y|\, \boldsymbol{x}) \end{aligned}$$

となる. 上式の 3 番目の等号は $1 - \mathsf{Pr}(Y = y|\, \boldsymbol{X} = \boldsymbol{x})$ を最小にする y の値は $\mathsf{Pr}(Y = y|\, \boldsymbol{X} = \boldsymbol{x})$ を最大にする y の値に一致することからわかる. 回帰問題で 2 乗損失を用いたとき最適の予測関数 g^* を**回帰関数**とよぶ.

定理 1.1. 回帰問題において 2 乗損失 $\mathsf{Loss}(y, \widehat{y}) = \big(y - \widehat{y}\big)^2$ を用いる. $\mathsf{E}[Y^2] < \infty$ を仮定する. このとき最適予測関数 g^* は次で与えられる.

$$g^*(\boldsymbol{x}) = \mathsf{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}].$$

すなわち $\boldsymbol{X} = \boldsymbol{x}$ を与えたときの Y の条件付き期待値である.

Proof. $g^*(\boldsymbol{x}) = \mathsf{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}]$ とする. 任意の (2 乗可積分な) 関数 g を考える. すると

$$\begin{aligned}
\mathsf{E}\big[\big\{Y - g(\boldsymbol{X})\big\}^2\big] &= \mathsf{E}\big[\{Y - g^*(\boldsymbol{X}) + g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}^2\big] \\
&= \mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}^2\big] + \underbrace{\mathsf{E}\big[\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}^2\big]}_{\geq 0} \\
&\quad + 2\,\mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}\big] \\
&\geq \mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}^2\big] + 2\,\mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}\big]
\end{aligned}$$

となる. 上の不等式の等号は $g(\boldsymbol{x}) = g^*(\boldsymbol{x})$ のとき成立する. しかし

$$\begin{aligned}
\mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}\big] &= \mathsf{E}\big[\mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}\big|\boldsymbol{X}\big]\big] \\
&= \mathsf{E}\big[\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}\mathsf{E}\big[Y - g^*(\boldsymbol{X})\big|\boldsymbol{X}\big]\big] \\
&= \mathsf{E}\big[\{g^*(\boldsymbol{X}) - g(\boldsymbol{X})\}\underbrace{\big\{\mathsf{E}\big[Y\big|\boldsymbol{X}\big] - g^*(\boldsymbol{X})\big\}}_{=0}\big] \\
&= 0
\end{aligned}$$

となる. よって

$$\mathsf{E}\big[\big\{(Y - g(\boldsymbol{X})\big\}^2\big] \geq \mathsf{E}\big[\{Y - g^*(\boldsymbol{X})\}^2\big]$$

がわかる. □

定理 1.1 により $\boldsymbol{X} = \boldsymbol{x}$ を与えたときの出力 Y を次のように表現できる.

$$Y = g^*(\boldsymbol{x}) + \epsilon(\boldsymbol{x}).$$

すると

$$\mathsf{E}[\epsilon(\boldsymbol{x})|\, \boldsymbol{X} = \boldsymbol{x}] = \mathsf{E}\big[Y - g^*(\boldsymbol{x})|\, \boldsymbol{X} = \boldsymbol{x}\big] = \mathsf{E}[Y|\, \boldsymbol{X} = \boldsymbol{x}] - g^*(\boldsymbol{x}) = 0.$$

さらに

$$\mathsf{Var}[\epsilon(\boldsymbol{X})|\, \boldsymbol{X} = \boldsymbol{x}] =: \nu^2(\boldsymbol{x})$$

と書ける. $\nu(\boldsymbol{x})$ は未知であり $\epsilon(\boldsymbol{x})$ の分布は平均が 0 であること以外は未知である.

一般に, 最適予測関数 g^* は $(\boldsymbol{X}, Y)$ の同時分布に依存し, その分布は未知である. したがって g^* を実際の問題では用いることができない. $(\boldsymbol{X}, Y)$ の同時分布 $\mathsf{p}_{(X,Y)}(\boldsymbol{x}, y)$ に従う n 個の確率変数列を

$$\mathfrak{T}_n = \{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), \ldots, (\boldsymbol{X}_n, Y_n)\}$$

と書き $\mathfrak{T}_n$ を**訓練集合**という. この訓練集合の実現値を

$$\mathfrak{t}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$$

と書く.

目標は訓練集合 $\mathfrak{T}_n$ が与えられたとき, 未知の最適予測関数 g^* を推定/学習することである. 訓練集合 $\mathfrak{T}$ に基づく g^* の近似/推定量を $g_{\mathfrak{T}}$ と書くこと[6]にする. ここで $g_{\mathfrak{T}}$ はランダムな関数であることに注意せよ. 実現値 $\mathfrak{t}_n$ により計算されたものが実現値になり, $g_{\mathfrak{t}}$ と書く. 訓練集合 $\mathfrak{T}_n$ により未知の関数

$$g^* : \boldsymbol{x} \mapsto \widehat{y}$$

を学習する学習器が $g_{\mathfrak{T}}$ である. 教師が真の関係に基づき n 個の出力と入力の対 $\mathfrak{T}_n$ を見本として与え, その見本で学習器 $g_{\mathfrak{T}}$ を訓練する. 新しい入力 $\boldsymbol{X}$ に対して教師によって与えられていない出力 Y を $g_{\mathfrak{T}}(\boldsymbol{X})$ で予測することになる. この設定を**教師あり学習**といい, 入力のことを**説明変数**, 出力のことを**応答変数**ということもある.

一方, **教師なし学習**では変数に説明変数と応答変数の区別はない. データの同時分布 $\mathsf{p}^*(\boldsymbol{x}, y)$ を学習することになる. この変数を $(\boldsymbol{X}, Y)$ と書くことにする. $(\boldsymbol{X}, Y)$ の分布 p^* の推定量 p に対して誤差は

$$\mathsf{Err}(\mathsf{p}) = \mathsf{E}\big[\mathsf{Loss}\big(\mathsf{p}^*(\boldsymbol{X}, Y), \mathsf{p}(\boldsymbol{X}, Y)\big)\big]$$

となる. 教師なし学習の例としては, コンビニエンスストアの客の購買行動の解析を考える. 100 個の商品があるとする. ある客が, $i\,(i = 1, 2, \ldots, 100)$ 番目の商品を買うか買わないの結果を $0, 1$ に対応させる. 客の購買パタンは $\boldsymbol{x} \in \{0, 1\}^{100}$ となる. 訓練集合の実現値 $\mathfrak{t}_{100} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, , \ldots, \boldsymbol{x}_{100}\}$ に基づいて, 客の購買パタンを見つけたい. 教師なし学習の手法としては, **クラスタリング**, **主成分分析**, **カーネル密度推定**, **決定木**等がある.

以下の節では, 教師あり学習についてみていくことにする.

[6]本来ならば, $g_{\mathfrak{T}_n}$ と書くべきだろうが, 煩雑になるので, 簡単に $g_{\mathfrak{T}}$ と書くことにした.

1.3 教師あり学習における訓練誤差と汎化誤差

任意の予測関数 (以後は学習器とよぶことにする.) g が与えられたとき**予測誤差**を

$$\mathsf{Err}(g) := \mathsf{E}\big[\mathsf{Loss}\big(Y,\, g(\boldsymbol{X})\big)\big]$$

で定義する. これを求めることは解析者は一般にはできない. なぜならば統計家にとって $(\boldsymbol{X},\, Y)$ の同時分布は未知だからである. 学習器 g の予測誤差 $\mathsf{Err}(g)$ を近似/推定するために訓練集合 $\mathfrak{T}_n$ に基づく**訓練誤差**

$$\widehat{\mathsf{Err}} := \frac{1}{n}\sum_{j=1}^{n} \mathsf{Loss}\big(Y_j,\, g(\boldsymbol{X}_j)\big)$$

を用いる. 訓練集合 $\mathfrak{T}_n$ は $(\boldsymbol{X},\, Y)$ の同時分布に独立同一に従う確率変数列なので

$$\mathsf{Err}(g) = \mathsf{E}\big[\widehat{\mathsf{Err}}(g)\big]$$

が成立していることが簡単にわかる. すなわち $\widehat{\mathsf{Err}}(g)$ は $\mathsf{Err}(g)$ の不偏推定量である.

最適予測関数を g^* を近似/学習/推定するために, 学習器の候補の集まりである関数族を $\mathcal{G}$ を選択する. すなわち, 訓練誤差を最小にする学習器をみつけることを目指す. 新しいデータ $\boldsymbol{X}$ の学習器の予測精度を測る尺度を導入する. 訓練集合の実現値 $\mathfrak{t}_n$ を固定する. すなわち, $(\boldsymbol{X}_j,\, Y_j) = (\boldsymbol{x}_j,\, y_j)\,(j = 1,\, 2,\, \ldots,\, n)$ とする.

訓練集合の実現値 $\mathfrak{t}_n$ の基づく学習器 $g_{\mathfrak{t}}^{\mathcal{G}}$ を

$$g_{\mathfrak{t}}^{\mathcal{G}} \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n} \mathsf{Loss}\big(y_j,\, g(\boldsymbol{x}_j)\big)$$

で定義する. $(\boldsymbol{X},\, Y)$ を訓練集合 $\mathfrak{T}_n$ とは独立に同時分布 $\mathsf{p}_{(X,Y)}(\boldsymbol{x},\, y)$ に従う確率ベクトルとしたとき学習器 $g_{\mathfrak{t}}^{\mathcal{G}}$ の**汎化誤差** $\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big)$ を

$$\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) := \mathsf{E}\big[\mathsf{Loss}\big(Y,\, g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X})\big)\big]$$

で定義する. ただし期待値 $\mathsf{E}[\,\cdot\,]$ は $(\boldsymbol{X},\, Y)$ の同時分布 $\mathsf{p}_{(X,Y)}(\boldsymbol{x},\, y)$ に関して取ったものである.

ランダムな訓練集合 $\mathfrak{T}_n$ に基づく学習器 $g_{\mathfrak{T}}^{\mathcal{G}}$ の汎化誤差 $\mathsf{Err}\big(g_{\mathfrak{T}}^{\mathcal{G}}\big)$ はランダムになるので, $\mathfrak{T}_n$ の同時分布に関して期待値をとったものを**期待汎化誤差**という.

$$\mathsf{E}\big[\mathsf{Err}\big(g_{\mathfrak{T}}^{\mathcal{G}}\big)\big] = \mathsf{E}\big[\mathsf{Loss}\big(Y,\, g_{\mathfrak{T}}^{\mathcal{G}}(\boldsymbol{X})\big)\big].$$

右辺の $\mathsf{E}[\,\cdot\,]$ は $\mathfrak{T}_n$ の同時分布に関する期待値であり, 左辺の $\mathsf{E}[\,\cdot\,]$ は $(\boldsymbol{X},\, Y)$ と $\mathfrak{T}_n$ の同時分布に関する期待値である.

注意 1.2. (多項式回帰) (U, Y) の同時分布を以下のように定める. $U \sim \mathsf{Unif}(0, 1)$ とし $U = u\,(0 < u < 1)$ が与えられたときの Y の条件付き分布は $\mathsf{N}(\theta, \sigma^2)$ とする. ただし,

$$\theta = 10 - 140u + 400u^2 - 250u^3, \quad \sigma^2 = 25$$

である. 2 乗損失を用いたとき最良予測関数は

$$g^*(u) = \mathsf{E}[Y \,|\, U = u] = 10 - 140u + 400u^2 - 250u^3$$

となる. これは以下の議論からわかる.

$$\begin{aligned}
\mathsf{Err}(g) &= \mathsf{E}[(Y - g(U))^2] = \mathsf{E}[\{Y - \mathsf{E}[Y \,|\, U] + \mathsf{E}[Y \,|\, U] + g(U)\}^2] \\
&= \mathsf{E}[\{\mathsf{E}[Y \,|\, U] + g(U)\}^2] + \mathsf{E}[\{\mathsf{E}[Y \,|\, U] + g(U)\}^2] \\
&\quad + 2\mathsf{E}[\{Y - \mathsf{E}[Y \,|\, U]\}\{\mathsf{E}[Y \,|\, U] + g(U)\}] \\
&= \mathsf{E}[\{\mathsf{E}[Y \,|\, U] + g(U)\}^2] + \mathsf{E}[\{\mathsf{E}[Y \,|\, U] + g(U)\}^2] \\
&\quad + 2\mathsf{E}\left[\mathsf{E}\left[\{Y - \mathsf{E}[Y \,|\, U]\}\{\mathsf{E}[Y \,|\, U] + g(U)\}\middle| U\right]\right] \\
&= \mathsf{E}[\{\mathsf{E}[Y \,|\, U] + g(U)\}^2] + \mathsf{E}[\{\mathsf{E}[Y \,|\, U] + g(U)\}^2] \\
&\quad + 2\mathsf{E}\Bigg[\{\mathsf{E}[Y \,|\, U] + g(U)\} \underbrace{\mathsf{E}\left[Y - \mathsf{E}[Y \,|\, U]\middle| U\right]}_{=0}\Bigg] \\
&= \mathsf{E}\big[\{Y - \mathsf{E}[Y \,|\, U]\}^2\big] + \mathsf{E}\big[\{\mathsf{E}[Y \,|\, U] - g(U)\}^2\big] \\
&\geq \mathsf{E}\big[\{\mathsf{E}[Y \,|\, U] - g(U)\}^2\big].
\end{aligned}$$

$d \in \mathbb{N}$, $d \geq 2$ とする. $d - 1$ 次元の u の多項式がなす関数族を考える.

$$\mathcal{G}_d := \{g(u) = \beta_0 + \beta_1 u + \beta_2 u^2 + \cdots + \beta_{d-1} u^{d-1};\, \beta_0, \beta_1, \ldots, \beta_{d-1} \in \mathbb{R}\}.$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{d-1})^\top, \quad \boldsymbol{x} = (1, u, u^2, \ldots, u^{d-1})^\top$$

とおくと

$$\mathcal{G}_d = \{g(u) = \boldsymbol{x}^\top \boldsymbol{\beta};\, \boldsymbol{\beta} \in \mathbb{R}^d\}$$

と書ける. $d \geq 4$ とする.

$$\boldsymbol{\beta}^* = (10, -140, 400, -250, 0, \ldots, 0)^\top$$

としたとき $g^*(u) = \boldsymbol{x}^\top \boldsymbol{\beta}^*$ が最良予測関数となる. 訓練データを

$$(u_1, y_1), (u_2, y_2), \ldots, (u_n, y_n)$$

とし

$$\boldsymbol{x}_j = (1, u_j, u_j^2, \ldots, u_j^{d-1})^\top$$

と書く．これらを行列とベクトルで表記する．

$$\underbrace{\mathbf{X}}_{n\times d\text{ 行列}} = \begin{pmatrix} \boldsymbol{x}_1^\top \\ \boldsymbol{x}_2^\top \\ \vdots \\ \boldsymbol{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^{d-1} \\ 1 & u_2 & u_2^2 & \cdots & u_2^{d-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & u_n & u_n^2 & \cdots & u_n^{d-1} \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

とおく．すると学習器 $g(u) = \boldsymbol{x}^\top\boldsymbol{\beta}$ の訓練誤差は

$$\widehat{\mathsf{Err}}(g) = \frac{1}{n}\big|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\big|_{2,n}^2$$

となる．ただし $|\cdot|_{2,n}$ は $\mathbb{R}^n$ の Euclid ノルムである．すると

$$\widehat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathcal{G}_d}\big|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\big|_{2,n}^2$$

とすれば $\widehat{\boldsymbol{\beta}}$ は最小 2 乗推定量となる．$\mathbf{X}$ の列で張られた線型部分空間を $\mathrm{span}(\mathbf{X})$ と書く．$\mathrm{span}(\mathbf{X})$ への射影行列を $\mathbf{P}$ とすれば $\widehat{\boldsymbol{\beta}}$ は

$$\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \mathbf{P}\boldsymbol{y}$$

をみたすことがわかる．すると $d\times n$ の行列 $\mathbf{X}^\dagger$ が存在[7]して

$$\mathbf{P} = \mathbf{X}\mathbf{X}^\dagger$$

と書けることが知られている．したがって

$$\widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\boldsymbol{y}$$

と書けることがわかる．よって

$$g_{\mathfrak{t}}^{\mathcal{G}_d}(u) = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}, \quad \widehat{\boldsymbol{\beta}} = \mathbf{X}^\dagger\boldsymbol{y}$$

となる．特に $\mathbf{X}^\top\mathbf{X}$ が正則のとき

$$\widehat{\boldsymbol{\beta}} = \big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top\boldsymbol{y}$$

となる． □

[7] $\boldsymbol{X}^\dagger$ は $\mathbf{X}$ の Moore-Penrose の一般逆行列と呼ばれるものである．補遺の B 章を参照のこと．

1.4 教師あり学習における兼ね合い (trade-off)

教師ありの機械学習において汎化誤差

$$\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) = \mathsf{E}\big[\mathsf{Loss}\big(Y,\, g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X})\big)\big]$$

もしくは

$$\mathsf{E}\big[\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big)\big] = \mathsf{E}\big[\mathsf{E}\big[\mathsf{Loss}\big(Y,\, g_{\mathfrak{T}}^{\mathcal{G}}(\boldsymbol{X})\big)\big]\big]$$

の最小化問題を考える. この問題を解くために関数族 $\mathcal{G}$ を適切に選択することが肝要である. 選択の最には以下の点を考慮する必要がある.

- 関数族の複雑さ (最適予測関数を近似するために十分な豊かさを持つことが望ましい. 最適予測関数を含む族であればなおさらよい).
- 最適問題
$$g_{\mathfrak{T}}^{\mathcal{G}} \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \frac{1}{n}\sum_{j=1}^{n} \mathsf{Loss}\big(Y_j,\, g(\boldsymbol{X}_j)\big)$$
を解く学習器の訓練が容易である.
- 訓練誤差
$$\widehat{\mathsf{Err}}(g) = \frac{1}{n}\sum_{j=1}^{n} \mathsf{Loss}\big(Y_j,\, g(\boldsymbol{X}_j)\big)$$
が予測誤差
$$\mathsf{Err}(g) = \mathsf{E}\big[\mathsf{Loss}\big(Y,\, g(\boldsymbol{X})\big)\big]$$
の推定量としてよい精度をもっているか?
- 特徴量のタイプは何か?

適切な関数族 $\mathcal{G}$ の選択には衝突する因子間の兼ね合いが伴う. 簡単な関数族からの学習器の訓練は容易であるが, 最良予測関数 g^* をうまく近似/推定していないかもしれない. 逆に g^* を含むかもしれない豊かな関数族 $\mathcal{G}$ からの学習器の訓練には計算コストがかかる. 関数族の複雑さ, 計算コスト, 推定精度の関係を理解するために汎化誤差を複数の因子に分解する. 分解には二通りある. ひとつは近似と推定精度の兼ね合いを説明するもので, 他方はバイアスと分散の兼ね合いを説明する分解である.

まず近似と推定誤差の兼ね合いを理解するための分解を考える.

$$\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) = \mathsf{Err}^* + \underbrace{\mathsf{Err}\big(g^{\mathcal{G}}\big) - \mathsf{Err}^*}_{\text{近似の精度}} + \underbrace{\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) - \mathsf{Err}\big(g^{\mathcal{G}}\big)}_{\text{推定の精度}}. \tag{1.1}$$

ただし

$$\mathsf{Err}^* = \mathsf{Err}\big(g^*\big), \quad g^* \in \operatorname*{arg\,min}_{g} \mathsf{Err}(g), \quad g^{\mathcal{G}} \in \operatorname*{arg\,min}_{g\in\mathcal{G}} \mathsf{Err}(g).$$

(1.1) の右辺の第 2 項目を**近似の精度**といい, 誤差の下限と関数族 $\mathcal{G}$ の中の最適な予測関数の誤差との差である. 関数族 $\mathcal{G}$ の選択と $\mathcal{G}$ 上の $\mathsf{Err}(g)$ の最小化問題は関数解析と数値解析の問題である. 訓練データ $\mathfrak{t}$ は関わらない. 近似精度をよくするためには, 関数族 $\mathcal{G}$ を大きくとればよい. 第 3 項は統計的誤差になる. これは訓練データにかかわるものである. $g_{\mathfrak{t}}^{\mathcal{G}}$ がよりうまく $g^{\mathcal{G}}$ を推定するかどうかである. 標本数 n が無限大に近づくとこの精度は 0 に収束する. 近似と推定の精度をよくするためには互いに相反することが要求される. 推定精度をよくするためには $\mathcal{G}$ を小さくすればよい. 一方, 近似精度をよくするためには関数族 $\mathcal{G}$ を大きく取ればよい. したがって兼ね合いが肝要となる.

2 乗損失を考える. この場合には汎化誤差は

$$\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) = \mathsf{E}\big[\big\{Y - g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X})\big\}^2\big].$$

最良予測関数は

$$g^*(\boldsymbol{X}) = \mathsf{E}[Y \,|\, \boldsymbol{X} = \boldsymbol{x}]$$

であることを思い出そう. すると

(1) 誤差の下限: $\mathsf{Err}^* = \mathsf{E}\big[\big(Y - g^*(\boldsymbol{X})\big)^2\big]$.

(2) 第 2 因子の近似精度: $\mathsf{Err}\big(g^{\mathcal{G}}\big) - \mathsf{Err}\big(g^*\big) = \mathsf{E}\big[\big(g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})\big)^2\big]$.

(3) 第 3 因子の推定精度: $\mathcal{G} = \{g(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta};\, \boldsymbol{\beta} \in \mathbb{R}^d\}$ としたとき,

$$\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) - \mathsf{Err}\big(g^{\mathcal{G}}\big) = \mathsf{E}\big[\big(g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\big)^2\big].$$

よって線型関数族 $\mathcal{G}$ に対して

$$\begin{aligned}\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) &= \mathsf{E}\big[\big(g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - Y\big)^2\big] \\ &= \mathsf{Err}^* + \mathsf{E}\big[\big(g^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})\big)^2\big] + \mathsf{E}\big[\big(g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - g^{\mathcal{G}}(\boldsymbol{X})\big)^2\big]\end{aligned}$$

となる.

次に 2 乗損失の場合についてバイアスと分散の兼ね合いについて説明する. まず

$$\begin{aligned}\mathsf{Err}\big(g_{\mathfrak{t}}^{\mathcal{G}}\big) &= \mathsf{E}\big[\big(g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - Y\big)^2\big] \\ &= \mathsf{E}\big[\big(g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X}) + g^*(\boldsymbol{X}) - Y\big)^2\big] \\ &= \mathsf{E}\big[\big(Y - g^*(\boldsymbol{X})\big)^2\big] + \mathsf{E}\big[\big(g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})\big)^2\big] \\ &= \mathsf{Err}^* + \mathsf{E}\big[D^2(\boldsymbol{X}, \mathfrak{t})\big].\end{aligned}$$

ただし

$$D(\boldsymbol{X}, \mathfrak{t}) = g_{\mathfrak{t}}^{\mathcal{G}}(\boldsymbol{X}) - g^*(\boldsymbol{X})$$

の $\mathfrak{t}$ に $\mathfrak{T}$ を代入すると

$$\begin{aligned}\mathsf{E}\big[\big(g_{\mathfrak{T}}^{\mathcal{G}}(\boldsymbol{x})-g^*(\boldsymbol{x})\big)^2\big|\boldsymbol{X}=\boldsymbol{x}\big] &= \mathsf{E}\big[D^2(\boldsymbol{x},\mathfrak{T})\big|\boldsymbol{X}=\boldsymbol{x}\big]\\ &= \big\{\mathsf{E}[D(\boldsymbol{x},\mathfrak{T})|\,\boldsymbol{X}=\boldsymbol{x}]\big\}^2+\mathsf{Var}\big[D(\boldsymbol{x},\mathfrak{T})|\,\boldsymbol{X}=\boldsymbol{x}\big]\\ &\quad\big(\because \text{注意 } A.47\big)\\ &= \big\{\mathsf{E}\big[g_{\mathfrak{T}}^{\mathcal{G}}(\boldsymbol{x})-g^*(\boldsymbol{x})\big|\,\boldsymbol{X}=\boldsymbol{x}\big]\big\}^2+\mathsf{Var}\big[g_{\mathfrak{T}}^{\mathcal{G}}|\,\boldsymbol{X}=\boldsymbol{x}\big].\end{aligned}$$

よって

$$\mathsf{E}\big[\mathsf{Err}\big(g_{\mathfrak{T}}^{\mathcal{G}}\big)\big]=\mathsf{Err}^*+\underbrace{\big\{\mathsf{E}\big[g_{\mathfrak{T}}^{\mathcal{G}}(\boldsymbol{X})\big|\boldsymbol{X}\big]-g^*(\boldsymbol{X})\big\}^2}_{\text{バイアス}}+\underbrace{\mathsf{E}\big[\mathsf{Var}\big[g_{\mathfrak{T}}^{\mathcal{G}}(\boldsymbol{X})\big|\boldsymbol{X}\big]\big]}_{\text{分散}}$$

と書ける.

1.5 教師なし学習における兼ね合い (trade-off)

応答変数 Y と説明変数ベクトル $\boldsymbol{X}$ の区別がある教師あり学習とは対照的に教師なし学習ではこのような変数の区別がない.

$n\in\mathbb{N}$ を標本 (観測データ) 数とする. 各標本 (観測データ) の次元を $k\in\mathbb{N}$ とし, 縦ベクトルで表すことにする. したがって, n 個の標本を $\boldsymbol{x}_j\,(j=1,2,\ldots,n)$ と書いたとき, $k\times n$ の観測データ行列 $[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]$ から有用情報やパタンを抽出することが教師なし学習の主目的になる.

標本 (観測データ) $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n$ はある未知の確率分布からの n の独立同一標本の実現値と考え, データ行列からこの未知の確率分布を学習/推測することが教師なし学習は本質的に目標とする.

観測データはある未知の確率分布からの n 個のランダム標本の実現値であるときデータの経験分布は未知の確率分布に関する重要な情報を含んでいる. 経験分布の概念とカーネル型密度推定については後ほど説明する.

教師なし学習における損失と誤差を導入する. 訓練データを

$$\mathfrak{T}_n:=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\ldots,\boldsymbol{X}_n\}$$

とする. ここで $\boldsymbol{X}_j\,(j=1,2,\ldots,n)$ は確率ベクトル $\boldsymbol{X}\in\mathbb{R}^k$ の独立複製とする. また $\boldsymbol{X}$ の未知の同時 p.d.f. を p^* と書くことにする. この p を**真の分布**ということにする. さらに, 訓練データ $\mathfrak{T}_n$ の実現値 $\mathfrak{t}_n=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$ と書くことにする.

教師なし学習では未知の同時 p.d.f.(真の分布) p^* をうまく近似する p.d.f. p をみつけることが目標である. すなわち, $\mathsf{Loss}\big(\mathsf{p}^*,\mathsf{p}\big)$ を損失関

数としたときその汎化誤差

$$\begin{aligned}\mathsf{Err}(\mathsf{p}) :&= \mathsf{E}\big[\mathsf{Loss}\big(\mathsf{p}^*(\boldsymbol{X}),\, \mathsf{p}(\boldsymbol{X})\big)\big] \\ &= \int \mathsf{Loss}\big(\mathsf{p}^*(\boldsymbol{x}),\, \mathsf{p}(\boldsymbol{x})\big)\mathsf{p}^*(\boldsymbol{x})\, d\boldsymbol{x}\end{aligned}$$

を小さくする p.d.f. p をみつけることである. ただし $\mathsf{E}[\cdot]$ は未知の真の確率分布 p^* に関する期待値である. 損失関数を

$$\mathsf{Loss}\big(\mathsf{p}^*,\, \mathsf{p}\big) = \log \frac{\mathsf{p}^*}{\mathsf{p}} = \begin{cases} \log \dfrac{\mathsf{p}^*(\boldsymbol{x})}{\mathsf{p}(\boldsymbol{x})} & \big(\mathsf{p}^*(\boldsymbol{x})\mathsf{p}(\boldsymbol{x}) > 0\big), \\ 0 & \big(\mathsf{p}^*(\boldsymbol{x})\mathsf{p}(\boldsymbol{x}) = 0\big) \end{cases}$$

と取れば, 汎化誤差は **Kullback-Leibler 情報量** (K-L 情報量とかくことにする) と呼ばれるものになる.

以下では K-L 情報量のもとで考える. K-L 情報量を KL と書くことにする. すなわち, K-L 情報量 KL を

$$\mathsf{KL}(\mathsf{p}) = \mathsf{E}\left[\log \frac{\mathsf{p}^*(X)}{\mathsf{p}(X)}\right] = \mathsf{E}\big[\log \mathsf{p}^*(X)\big] - \mathsf{E}\big[\log \mathsf{p}(X)\big]$$

で定める. 上の式の最右辺の第 1 項目は p に関係ないので $\mathsf{KL}(\mathsf{p})$ の最小化は

$$-\mathsf{E}\big[\log \mathsf{p}(X)\big] = -\int \mathsf{p}^*(\boldsymbol{x}) \log \mathsf{p}(\boldsymbol{x})\, d\boldsymbol{x}$$

の最小化となる. ここで

$$\mathsf{CE}(\mathsf{p}) := -\mathsf{E}\big[\log \mathsf{p}(\boldsymbol{X})\big] \tag{1.2}$$

と書くことにする. これを**クロス・エントロピー誤差**と呼ぶことにする.

次に探索の対象となる確率密度関数の集まりを $\mathcal{P}$ となくことにする. この確率分布の集まりを**統計的モデル**という. 統計的モデル $\mathcal{P}$ が真の分布 p を含んでいる ($\mathsf{p}^* \in \mathcal{P}$ が成立する) とき, クロス・エントロピー誤差 $\mathsf{CE}(\mathsf{p})$ を最小とするものは p^* となる. すなわち, 任意の $\mathsf{p} \in \mathcal{P}$ に対して

$$\mathsf{CE}(\mathsf{p}) \geq \mathsf{CE}(\mathsf{p}^*)$$

が成立する. しかし (1.2) の最小化問題は一般に実行不可能である. なぜならばこの誤差は未知の分布 p^* に依存するからである. $\mathsf{CE}(\mathsf{p})$ の代わりに訓練クロス・エントロピー誤差

$$\widehat{\mathsf{CE}}_{\mathsf{t}} := -\frac{1}{n}\sum_{j=1}^{n} \log \mathsf{p}(\boldsymbol{x}_j)$$

の $\mathsf{p} \in \mathcal{P}$ に関する最小化問題を考える.

ここで重要なステップは探索の対象となる統計的モデル $\mathcal{P}$ の選択である. まず $\mathcal{P}$ の元 p を添え字 $\boldsymbol{\theta}$ で添え字付ることにする. 添え字 $\boldsymbol{\theta}$ の全体の成す集合 (この集合を添え字集合) を Θ と書くことにする. すなわち集合 $\mathcal{P}$ を集合 Θ により母数化する. 添え字の $\boldsymbol{\theta}$ を**母数**といい, Θ を**母数空間** と呼ぶことにする. すると $\mathcal{P}$ の選択は Θ の選択となる.

いま, $d \in \mathbb{N}$ とし, $\Theta \subset \mathbb{R}^d$ とする. すなわち

$$\mathcal{P} := \{\mathsf{p}(\,\cdot\,|\,\boldsymbol{\theta});\, \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\}. \tag{1.3}$$

(1.3) のような形で与えられる統計的モデルを**母数モデル**と統計学では伝統的に呼んでいる. すなわち, 母数モデルとは有限次元の母数空間によって添え字づけられている分布の集まり (統計的モデル) である.

以降では, (1.3) の中から未知の真の分布 p^* をよく近似する確率分布を学習することにする.

まず, データから近似モデルを探すときの情報な概念を導入する. データの実現値 $\boldsymbol{X} = \boldsymbol{x}$ が与えられたときに関数

$$\boldsymbol{\theta} \mapsto \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta}) \tag{1.4}$$

を**尤度関数**という. この関数の $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^\top$ に関する勾配を**スコア関数**といい, $\boldsymbol{S}(\boldsymbol{x}|\,\boldsymbol{\theta})$ と書くことにする.

$$\boldsymbol{S}(\boldsymbol{x}|\,\boldsymbol{\theta}) := \frac{1}{\mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})} \frac{\partial \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} := \frac{1}{\mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})} \left(\frac{\partial \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})}{\partial \theta_2}, \ldots, \frac{\partial \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})}{\partial \theta_d} \right)^\top.$$

$\boldsymbol{S}(\boldsymbol{x}|\,\boldsymbol{\theta})$ の $\boldsymbol{x}$ に $\boldsymbol{X}$ を代入しランダムにしたものを $\boldsymbol{S}(\boldsymbol{X}|\,\boldsymbol{\theta})$ と書くことにする. すると

$$\begin{aligned} \mathsf{E}_{\boldsymbol{\theta}}\big[\boldsymbol{S}(\boldsymbol{X}|\,\boldsymbol{\theta})\big] &= \int \frac{1}{\mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})} \left(\frac{\partial \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})\, d\boldsymbol{x} = \int \frac{\partial \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\boldsymbol{x} \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \int \mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})\, d\boldsymbol{x} = \mathbf{0} \end{aligned} \tag{1.5}$$

となる. ただし最後の等号は, 微分記号と積分記号の交換が可能であることを仮定[8]している. すなわち $\mathsf{p}(\,\cdot\,|\,\boldsymbol{\theta})$ の関する $\boldsymbol{S}(\boldsymbol{X}|\,\boldsymbol{\theta})$ の期待値は零ベクトルとなる. さらに $\mathbf{p}(\,\cdot\,|\,\boldsymbol{\theta})$ の関する $\boldsymbol{S}(\boldsymbol{X}|\,\boldsymbol{\theta})$ の共分散行列を考える.

[8]無条件でこの交換が可能なわけではない. 通常は, 交換が可能になるような正則条件を母数モデル $\mathcal{P}$ や真の分布 p^* に課す. ここの議論では, 厳密性に欠くが, 正則条件を気にせずに議論をする.

これを Fisher 情報量行列といい, $\mathbf{I}(\boldsymbol{\theta})$ と記すことにする.

$$
\begin{aligned}
\mathbf{I}(\boldsymbol{\theta}) &:= \mathsf{E}_{\boldsymbol{\theta}}\bigl[\boldsymbol{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta})\boldsymbol{S}^{\top}(\boldsymbol{X}\,|\,\boldsymbol{\theta})\bigr] \\
&= \mathsf{E}_{\boldsymbol{\theta}}\begin{bmatrix}
\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1} & \dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d} \\
\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1} & \dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d} \\
\vdots & \vdots & \ddots & \vdots \\
\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1} & \dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2} & \cdots & \dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d}\dfrac{\partial \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d}
\end{bmatrix}
\end{aligned}
$$

で定義する. さらに $-\log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})$ の Hesse 行列の分布 p に関する期待値を $\mathbf{J}(\boldsymbol{\theta})$ を

$$
\begin{aligned}
\mathbf{J}(\boldsymbol{\theta}) &:= -\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial \mathsf{S}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \\
&:= -\mathsf{E}_{\boldsymbol{\theta}}\begin{bmatrix}
\dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_d} \\
\dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} & \cdots & \dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_d} \\
\vdots & \vdots & \ddots & \vdots \\
\dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d \partial \theta_1} & \dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d \partial \theta_2} & \cdots & \dfrac{\partial^2 \log \mathsf{p}(\boldsymbol{X}\,|\,\boldsymbol{\theta})}{\partial \theta_d \partial \theta_d}
\end{bmatrix}
\end{aligned}
$$

で定義する. ある $\boldsymbol{\theta}^* \in \Theta$ が存在して $\mathsf{p}^* = \mathsf{p}(\,\cdot\,|\,\boldsymbol{\theta}^*)$ となるとき

$$
\mathbf{J}(\boldsymbol{\theta}^*) = \mathbf{I}(\boldsymbol{\theta}^*)
$$

となること[9]がわかる.

[9] $d = 1$ として確認する. $d \geq 2$ の場合も本質的には同じである.

$$
\begin{aligned}
J(\theta) &= -\mathsf{E}\left[\frac{\partial^2 \log \mathbf{p}(X\,|\,\theta)}{\partial \theta^2}\right] = -\mathsf{E}\left[\frac{\partial}{\partial \theta}\left(\frac{1}{\mathbf{p}(X\,|\,\theta)}\frac{\partial \mathbf{p}(X\,|\,\theta)}{\partial \theta}\right)\right] \\
&= \mathsf{E}\left[\frac{1}{\mathbf{p}^2(X\,|\,\theta)}\left(\frac{\partial \mathbf{p}(X\,|\,\theta)}{\partial \theta}\right)^2 - \frac{1}{\mathbf{p}(X\,|\,\theta)}\frac{\partial^2 \mathbf{p}(X\,|\,\theta)}{\partial^2 \theta}\right] \\
&= \mathsf{E}\left[\left(\frac{\partial \log \mathbf{p}(X\,|\,\theta)}{\partial \theta}\right)^2\right] - \mathsf{E}\left[\frac{1}{\mathbf{p}(X\,|\,\theta)}\frac{\partial^2 \mathbf{p}(X\,|\,\theta)}{\partial \theta^2}\right] \\
&= I(\theta) - \mathsf{E}\left[\frac{1}{\mathbf{p}(X\,|\,\theta)}\frac{\partial^2 \mathbf{p}(X\,|\,\theta)}{\partial \theta^2}\right].
\end{aligned}
$$

しかし

$$
\begin{aligned}
\mathsf{E}\left[\frac{1}{\mathbf{p}(X\,|\,\theta)}\frac{\partial^2 \mathbf{p}(X\,|\,\theta)}{\partial \theta^2}\right] &= \int \frac{1}{\mathbf{p}(x\,|\,\theta)}\frac{\partial^2 \mathbf{p}(x\,|\,\theta)}{\partial \theta}\mathbf{p}(x\,|\,\theta)\,dx \\
&= \int \frac{\partial^2 g(x\,|\,\theta)}{\partial \theta}\,dx = \frac{1}{\partial \theta^2}\int g(x\,|\,\theta)\,dx = 0.
\end{aligned}
$$

n が十分大きいとき行列 $\mathbf{I}(\boldsymbol{\theta})$, $\mathbf{J}(\boldsymbol{\theta})$ はクロス・エントロピー誤差の近似において重要な役割を担う. 設定を記述するために, $\mathsf{p}^{\mathcal{P}} = \mathbf{p}(\cdot\,|\,\boldsymbol{\theta}^*)$ をクロス・エントロピー誤差

$$\mathsf{CE}(\boldsymbol{\theta}) := -\mathsf{E}\big[\mathbf{p}(\boldsymbol{X}|\,\boldsymbol{\theta})\big]$$

を最小にする点とする. すなわち任意の $\boldsymbol{\theta} \in \Theta$ に対して

$$\mathsf{CE}\big(\boldsymbol{\theta}\big) \geq \mathsf{CE}\big(\boldsymbol{\theta}^*\big)$$

である. CE は $\boldsymbol{\theta}$ の関数として「なめらか」な関数と仮定する. 特に, これらの行列が $\boldsymbol{\theta}^*$ の近傍で狭義凸で 2 回連続微分可能[10]なときこの仮定をみたしていることがわかる. $\boldsymbol{\theta}^*$ の定義より

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\theta}}\mathsf{CE}\big(\boldsymbol{\theta}\big)\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = -\frac{\partial}{\partial \boldsymbol{\theta}}\mathsf{E}\big[\log \mathbf{p}(\boldsymbol{X}|\,\boldsymbol{\theta})\big]\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} = -\mathsf{E}\bigg[\frac{\partial}{\partial \boldsymbol{\theta}}\log \mathbf{p}(\boldsymbol{X}|\,\boldsymbol{\theta})\bigg]\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$$

となる. ただし微分記号と積分記号の交換が可能であることを仮定している. 同じように $\boldsymbol{J}\big(\boldsymbol{\theta}\big)$ は Err の Hesse 行列であることがわかる. $\mathbf{p}\big(\cdot\,\big|\widehat{\boldsymbol{\theta}}_n\big)$ を訓練誤差

$$\mathsf{CE}_{\mathfrak{T}_n}\big(\boldsymbol{\theta}\big) := -\frac{1}{n}\sum_{j=1}^{n}\log \mathsf{p}\big(\boldsymbol{X}_j\big|\,\boldsymbol{\theta}\big)$$

を最小にする点とする. ただし $\mathfrak{T}_n = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n\}$ はランダム標本の集合である. CE^* をすべての確率分布に関するクロス・エントロピーの最小値とする. 明らかに $\boldsymbol{X} \sim \mathsf{p}^*$ のとき最小となる. すなわち $\mathsf{CE}^* = -\mathsf{E}\big[\log \mathsf{p}^*(\boldsymbol{X})\big]$ である. 教師あり学習の場合と同様に汎化誤差 $\mathsf{CE}\big(\widehat{\boldsymbol{\theta}}_n\big)$ を分解する.

$$\begin{aligned}\mathsf{CE}\big(\widehat{\boldsymbol{\theta}}_n\big) &= \mathsf{CE}^* + \underbrace{\mathsf{CE}\big(\boldsymbol{\theta}^*\big) - \mathsf{CE}^*}_{\text{近似精度}} + \underbrace{\mathsf{CE}\big(\widehat{\boldsymbol{\theta}}_n\big) - \mathsf{CE}\big(\boldsymbol{\theta}^*\big)}_{\text{推定精度}} \\ &= \mathsf{CE}\big(\boldsymbol{\theta}^*\big) + \mathsf{E}\bigg[\frac{\mathbf{p}\big(\boldsymbol{X}\big|\boldsymbol{\theta}^*\big)}{\mathbf{p}\big(\boldsymbol{X}\big|\widehat{\boldsymbol{\theta}}_n\big)}\bigg].\end{aligned}$$

上の式の最左辺と最右辺の第 2 項目は $\mathfrak{T}_n$ に依存しているのでランダムな量であることに注意せよ.

1.6 最尤法

$d \in \mathbb{N}$ とする. 母数モデル $\mathcal{P} = \big\{\mathsf{p}(\boldsymbol{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\big\}$ を考える. ただし $\mathsf{p}(\boldsymbol{x}|\,\boldsymbol{\theta})$ は p.d.f. もしくは p.m.f. とする. 真の母数 $\boldsymbol{\theta}^*$ は母数空間 Θ に含まれるとする. すなわち, $\boldsymbol{\theta}^* \in \Theta$ である.

[10] たとえば g が正規分布のときにはこの条件をみたしている.

いま, $n \in \mathbb{N}$ とし,

$$\boldsymbol{X}_1,\, \boldsymbol{X}_2,\, \ldots,\, \boldsymbol{X}_n \sim \text{i.i.d. } \mathsf{p}(\boldsymbol{x}|\, \boldsymbol{\theta}^*)$$

とする.

定義 1.3. $\boldsymbol{X}_1 = \boldsymbol{x}_1,\, \boldsymbol{X}_2 = \boldsymbol{x}_2,\, \ldots,\, \boldsymbol{X}_n = \boldsymbol{x}_n$ を観測したとき $\boldsymbol{\theta}$ の **尤**(ゆう) **度関数** $\mathsf{lik}_n(\boldsymbol{\theta})$ を

$$\mathsf{lik}_n\big(\boldsymbol{\theta}\big) = \prod_{j=1}^{n} \mathsf{p}(\boldsymbol{x}_j|\, \boldsymbol{\theta})$$

で定義し, **対数尤度** $\ell_n(\boldsymbol{\theta})$ を

$$\ell_n(\boldsymbol{\theta}) = \log \mathsf{lik}_n(\boldsymbol{\theta})$$

で定義する.

注意 1.4. 尤度関数は

$$\mathsf{lik}_n :\, \Theta \ni \boldsymbol{\theta} \mapsto \mathsf{lik}_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathsf{p}(\boldsymbol{x}_i|\, \boldsymbol{\theta}) \in [0,\, \infty)$$

であることに注意をせよ. 一方, $\mathbb{X} = \big\{\boldsymbol{x} \in \mathbb{R}^k;\, \mathsf{p}(\boldsymbol{x}|\, \boldsymbol{\theta}^*) > 0\big\}$ とおいたとき, $(\boldsymbol{X}_1,\, \boldsymbol{X}_2,\, \ldots,\, \boldsymbol{X}_n)$ の同時 p.d.f. または p.m.f. $\prod_{i=1}^{n} \mathsf{p}(\boldsymbol{x}_i|\, \boldsymbol{\theta}^*)$ は

$$\mathbb{X} \times \mathbb{X} \times \cdots \times \mathbb{X} \ni (\boldsymbol{x}_1,\, \boldsymbol{x}_2,\, \ldots,\, \boldsymbol{x}_n) \mapsto \prod_{i=1}^{n} \mathsf{p}(\boldsymbol{x}_i|\, \boldsymbol{\theta}^*) \in (0,\, \infty)$$

である. □

ここで, $\arg\max$ の記号を導入する. 関数 $g(x)$ の最大値を取る点を表す集合を

$$\operatorname*{arg\,max}_{x \in \mathbb{R}} g(x)$$

と書く. たとえば $g(x) = -(x-1)^2$ のとき

$$\operatorname*{arg\,max}_{0 \le x \le 4\pi} g(x) = \{1\}$$

となる. $g(x) = \sin x$ のとき

$$\operatorname*{arg\,max}_{0 \le x \le 4\pi} g(x) = \{\pi/2,\, 5\pi/2\}$$

となる.

定義 1.5. $\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n$ を観測したとき $\boldsymbol{\theta}^*$ の**最尤推定値** (maximum likelihood estimate) を $\mathsf{lik}_n(\boldsymbol{\theta})$ を最大にする値 $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ で定義する. すなわち

$$\widehat{\boldsymbol{\theta}}_n(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} \mathsf{lik}_n(\boldsymbol{\theta}).$$

$(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ に $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ を代入したもの $\widehat{\boldsymbol{\theta}}_n(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n)$ を $\boldsymbol{\theta}$ の**最尤推定量** (maximum likelihood estimator=m.l.e.) という.

注意 1.6. $X_1, X_2, \ldots, X_n \sim$ i.i.d. $\mathsf{Ber}(\theta^*)$ とする. ただし $(0, 1) =: \Theta \ni \theta^*$ は未知の真の母数とする. すなわち

$$\mathsf{p}(x|\,\theta^*) = (\theta^*)^x\big(1-\theta^*\big)^{1-x}, \qquad (x = 0, 1)$$

である. このとき

$$\mathsf{lik}_n\big(\theta\big) = \prod_{j=1}^n \mathsf{p}(x_j|\,\theta) = \prod_{j=1}^n \theta^{x_j}(1-\theta)^{1-x_j} = \theta^s(1-\theta)^{n-s}$$

となる. ただし $t_n = \sum_{j=1}^n x_j$ である. よって対数尤度は

$$\ell_n(\theta) = t_n \log\theta + (n - t_n)\log(1-\theta) \qquad (0 < \theta < 1)$$

となる. このことから, $0 < t_n < n$ のとき,

$$\frac{t_n}{n} \in \operatorname*{arg\,max}_{\theta\in(0,1)} \ell_n(\theta)$$

がわかる. したがって, $0 < t_n < n$ のとき, θ の最尤推定量は $\widehat{\theta}_n = \frac{\sum_{j=1}^n X_j}{n}$ となる. 一方, $t_n = 0, n$ のとき, θ^* の最尤推定値は存在しない. □

注意 1.7. $\boldsymbol{\theta}^* = (\mu^*, \sigma^*) \in \mathbb{R}\times(0,\infty)$ を真の母数とし,

$$X_1, X_2, \ldots, X_n \sim \text{ i.i.d. } \mathsf{N}\big(\mu^*, (\sigma^*)^2\big)$$

とする. $X_j = x_j\,(j = 1, 2, \ldots, n)$ を観測したとき, 尤度関数は

$$\begin{aligned}
\mathsf{lik}_n(\mu, \sigma) &= \prod_{j=1}^n \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_j-\mu)^2}{\sigma^2}\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n}\exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^n (x_j-\mu)^2\right\} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sigma^n}\exp\left\{-\frac{ns_n^2}{2\sigma^2}\right\}\exp\left\{-\frac{n(\overline{x}_n-\mu)^2}{2\sigma^2}\right\}.
\end{aligned}$$

ただし

$$\overline{x}_n = \frac{1}{n}\sum_{j=1}^{n} x_j, \qquad s_n^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x}_n)^2.$$

最後の等号は

$$\sum_{j=1}^{n}(x_j - \mu)^2 = ns_n^2 + n(\overline{x}_n - \mu)^2 \tag{1.6}$$

からわかる. 対数尤度は

$$\ell_n(\mu, \sigma) = -n\log\sigma - \frac{ns_n^2}{2\sigma^2} - \frac{n(\overline{x}_n - \mu)^2}{2\sigma^2}.$$

よって

$$\begin{cases} \dfrac{\partial \ell_n(\mu, \sigma)}{\partial \mu}(\mu, \sigma) = -\dfrac{n(\overline{x}_n - \mu)}{2\sigma^2} = 0 \\ \dfrac{\partial \ell_n}{\partial \sigma}(\mu, \sigma) = -\dfrac{n}{\sigma} + \dfrac{ns_n^2}{\sigma^3} + \dfrac{n(\overline{x}_n - \overline{x}_n)^2}{\sigma^3} = 0 \end{cases}$$

を解くと

$$\mu = \overline{x}_n, \qquad \sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\big(x_j - \overline{x}_n\big)^2}$$

となる. $\ell_n(\mu, \sigma)$ の Hessian を求める.

$$\mathbf{H} := \begin{pmatrix} \dfrac{\partial \ell_n^2}{\partial \mu^2}\big(\overline{x}_n, s_n\big) & \dfrac{\partial \ell_n^2}{\partial \mu \partial \sigma}\big(\overline{x}_n, s_n\big) \\ \dfrac{\partial \ell_n^2}{\partial \partial \sigma \mu}\big(\overline{x}_n, s_n\big) & \dfrac{\partial \ell_n^2}{\partial \sigma^2}\big(\overline{x}_n, s_n\big) \end{pmatrix} = \begin{pmatrix} -\dfrac{1}{2\sigma^2} & 0 \\ 0 & -\dfrac{2n}{s_n^2} \end{pmatrix} \tag{1.7}$$

より, $-\mathbf{H}$ は正定値行列となるので, 関数

$$\Theta \ni (\mu, \sigma) \mapsto \ell_n(\mu, \sigma)$$

は $(\mu, \sigma) = (\overline{x}_n, s_n)$ で最大となる. よって (μ, σ) の最尤推定量は

$$\widehat{\mu}_n = \frac{1}{n}\sum_{j=1}^{n} X_j =: \overline{X}_n, \quad \widehat{s}_n = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\big(X_j - \overline{X}_n\big)^2}$$

となる. □

問 1.1. (1.6) と (1.7) を確認せよ.