ベイズ推定に基づく線形回帰

佐伯　雄飛

2020 年 5 月 6 日

1　背景

回帰問題の目的は，N 個の観測値と対応する目標値からなる訓練データ集合が与えられたとき，新しい観測値に対する目標値の値を予測することである．今回扱う線形回帰モデルは，多項式は調節可能なパラメータの線形結合という特徴を利用した最も単純なモデルである．固定された基底関数の入力変数に関して非線型な関数の固定された集合結合をとることにより，有用な関数のクラスが得られる．

観測されたデータ $D = \{(x_i, y_i); i = 1, 2, ..., n\}$ に対して，基底関数の線形結合に基づく回帰関数モデルを以下のように定義する．ここで Φ を x の基底関数，ϵ を誤差項とする．

$$y = \Phi w + \epsilon \tag{1}$$

2　ベイズ線形回帰について

2.1　最小二乗推定

最小二乗推定は，回帰モデルによる予測誤差の二乗和 $S(w)$ を最小化する $\hat{w}$ を求める手法である．$S(w)$ を w で偏微分し，$\hat{w}$ を求める．

$$S(w) = \epsilon^T \epsilon = (y - \Phi w)^T (y - \Phi w) \tag{2}$$

$$\frac{dS(w)}{dw} = -\Phi^T y + \Phi^T \Phi w \tag{3}$$

$\frac{dS(w)}{dw} = 0$ のときを考えると，

$$\hat{w} = (\Phi^T \Phi)^{-1} \Phi^T y \tag{4}$$

従って，式 (4) より最小二乗推定による予測モデルは以下のようになる．

$$\hat{y} = \Phi \hat{w} = \Phi (\Phi^T \Phi)^{-1} y \tag{5}$$

2.2　最尤推定

最尤推定は，尤度 $P(y, w)$ を最大化する $\hat{w}$ を求める手法である．誤差項に正規分布を仮定したモデルを考える．このとき観測値 y は平均 Φw，分散行列 $\sigma^2 I_n$ の n 次元正規分布に従う．よって尤度は，以下のように与えられる．

$$y = \Phi w + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I_n) \tag{6}$$

$$P(y \mid w, \sigma^2) = \mathcal{N}(\Phi w, \sigma^2 I_n) \tag{7}$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} exp\{-\frac{1}{2\sigma^2}(y - \Phi w)^T (y - \Phi w)\} \tag{8}$$

式 (8) より，$P(y \mid w)$ の対数を w で偏微分し，$\hat{w}$ を求める．

$$\log P(y \mid w) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{(y - \Phi w)^T (y - \Phi w)}{2\sigma^2} \tag{9}$$

$$\frac{1}{P(y \mid w)} \frac{P(y \mid w)}{dw} = -(\Phi^T y + \Phi^T \Phi w) \tag{10}$$

$\frac{dP(y|w)}{dw} = 0$ のときを考えると．

$$\hat{w} = (\Phi^T \Phi)^{-1} \Phi^T y \tag{11}$$

従って，式 (11) より最尤推定による予測モデルは以下のようになる．

$$\hat{y} = \Phi \hat{w} = \Phi (\Phi^T \Phi)^{-1} y \tag{12}$$

これは，最小二乗法によって求められる予測モデルと同じである．

2.3　MAP 推定

最小二乗法，最尤推定に基づく方法では，モデルパラメータの数が多く，観測データの数が小さい時に過学習を起こしやすいという問題点がある．過学習が生じると，汎化性能が期待できないため，過学習を防ぐことが重要となる．MAP 推定は，w を確率変数として扱う．w の事前分布と観測データの尤度関数を以下のように導入する．α,β はハイパーパラメータとする．

$$P(w; \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1} I_n) \tag{13}$$

$$P(y \mid w; \beta) = \mathcal{N}(\Phi w, \beta^{-1} I_n) \tag{14}$$

MAP 推定では，w の事後分布 $P(w \mid y)$ を最大化する $\hat{w}$ を求める手法である．ベイズの定理より，

$$P(w \mid y) = \frac{P(y \mid w) P(w)}{P(y)} \tag{15}$$

ここで．$P(y|w)$ は尤度を，$P(w)$ は事前確率を表す．

$$P(w \mid y) = \frac{\frac{1}{(2\pi\beta^{-1})^{\frac{n}{2}}} exp\{-\frac{1}{2\beta^2}(y-\Phi w)^T(y-\Phi w)\}}{P(y)} \frac{\frac{1}{(2\pi\alpha^{-1})^{\frac{n}{2}}} exp\{-\frac{1}{2\alpha^2}w^T w\}}{P(y)} \tag{16}$$

$$P(w \mid y) = \frac{\frac{1}{(2\pi)^n(\alpha\beta)^{-\frac{n}{2}}}}{P(y)} \frac{exp\{-\frac{1}{2\beta^2}(y-\Phi w)^T(y-\Phi w) - \frac{1}{2\alpha^2}w^T w\}}{P(y)} \tag{17}$$

$P(w \mid y)$ を最大化する $\hat{w}$ は，$Z = -\frac{1}{2\beta^2}(y-\Phi w)^T(y-\Phi w) - \frac{1}{2\alpha^2}w^T w$ を最大化する $\hat{w}$ に等しい．

$$\frac{dZ}{dw} = -\frac{1}{2\beta^{-1}}(-\Phi^T + \Phi^T\Phi w) - \frac{1}{2\alpha^{-1}}w^T w \tag{18}$$

$\frac{dZ}{dw} = 0$ のときを考えると，

$$\hat{w} = (-\frac{\alpha}{\beta}I_n + \Phi^T\Phi)^{-1}\Phi^T y \tag{19}$$

従って，式 (19) より MAP 推定による予測モデルは以下のようになる．

$$\hat{y} = \Phi\hat{w} = \Phi(-\frac{\alpha}{\beta}I_n + \Phi^T\Phi)^{-1}\Phi^T y \tag{20}$$

2.4　ベイズ推定

最小二乗法，最尤推定，MAP 推定では，パラメータの推定値を一つの解として求めた．しかし，これではデータの予測にパラメータの不確かさを考慮することができない．事後分布をそのまま確率分布として取り扱うことで，パラメータ推定の不確かさを加味した予測分布を求める．MAP 推定では，$P(w \mid y)$ の最大化を考えたため，$P(D)$ については無視できたが，ベイズ推定では考える必要がある．同時確率 $P(y, w)$ から一方の確率変数を取り除き，周辺確率 $P(y)$ を求める．

$$P(y) = \int P(y \mid w)P(w)dw \tag{21}$$

$$P(w \mid y) = \frac{P(y \mid w)P(w)}{\int P(y \mid w)P(w)dw} \tag{22}$$

ここで，式 (13) と式 (14) を用いて，ガウス分布に対するベイズの定理より，

$$P(w \mid y) = \mathcal{N}(\mu_N, \Sigma_N) \tag{23}$$

参考文献 [1] P90 ガウス分布の周辺分布と条件付き分布を用いて計算すると

$$\mu_N = (\frac{\alpha}{\beta}I_n + \Phi^T\Phi)^{-1}\Phi^T y \tag{24}$$

$$\Sigma_N = (\alpha I_n + \beta\Phi^T\Phi)^{-1} \tag{25}$$

従って，以上よりベイズ推定による予測モデルは以下のようになる．

$$\hat{y} = \Phi\hat{w} = \Phi(-\frac{\alpha}{\beta}I_n + \Phi^T\Phi)^{-1}\Phi^T y \tag{26}$$

$$\Sigma_N = (\alpha I_n + \beta\Phi^T\Phi)^{-1} \tag{27}$$

3　回帰結果とその評価

$D = \{(x.train_i, y.train_i); i = 1, 2, ..., 15\}$ の訓練データに対して，最小二乗推定，最尤推定，MAP 推定，ベイズ推定を用いて予測モデルを作成した．それぞれのモデルで $D = \{(x.test_i, y.test_i); i = 1, 2, ..., 100\}$ のテストデータに対して予測分布を確認した．基底関数は以下のものを用いる．

$$f_j(x) = x^j, j = 0, 1, ...9 \tag{28}$$

ハイパーパラメータ $\alpha = 10, \beta = 10$ とする．

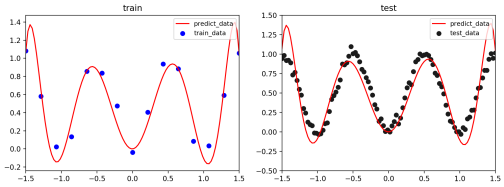

Fig. 1: 最小二乗推定・最尤推定

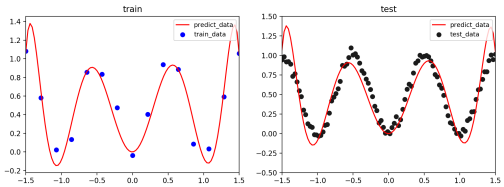

Fig. 2: MAP 推定

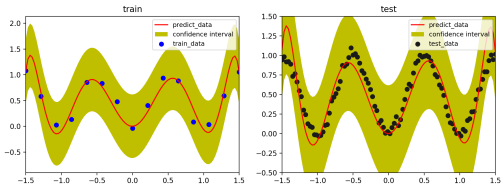

Fig. 3: ベイズ推定

決定係数 R^2 により点推定の評価を行う．決定係数とは回帰によって導いたモデルの当てはまりの良さを表現する値で、モデルによって予測した値

が実際の値とどの程度一致しているかを表現する評価指標である．決定係数 R^2 は実際のデータを (x_i, y_i)、回帰式から推定されたデータを $(x_i, \bar{y_i})$ として $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$ で求められる．0 から 1 の範囲で 1 に近づくほど良い値である．

手法	最小二乗推定・最尤推定	MAP 推定
R^2	0.76012297	0.769478635

Table 1: 決定係数 R^2

尤度関数の分散 β の値を変更する．

β	10	50	100	1000
R^2	0.76947	0.77227	0.769478	0.761849

Table 2: 決定係数 R^2

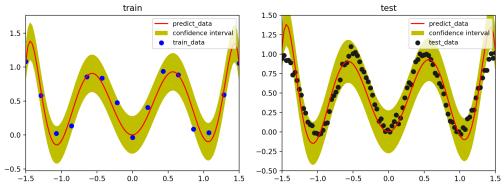

Fig. 4: $\beta = 50$ のときのベイズ推定

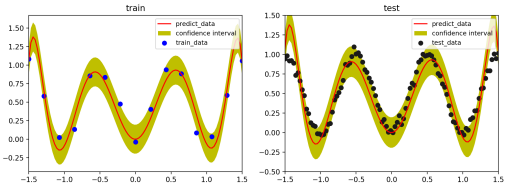

Fig. 5: $\beta = 100$ のときのベイズ推定

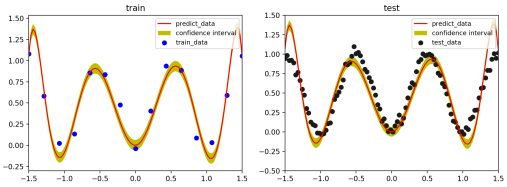

Fig. 6: $\beta = 1000$ のときのベイズ推定

4　まとめ

回帰問題に対して，最小二乗法，最尤推定，MAP 推定，ベイズ推定を適応して計算した．図 1-4 より，訓練データが存在しない部分の予測精度が下がることがわかった．

点推定である最小二乗法，最尤推定，MAP 推定については，表 2 より，今回の場合においては MAP 推定の方がより外れ値が減少し，すぐれたモデルを作成できることがわかる．

表 2，図 4-5 より，ベイズ推定，MAP 推定においては，尤度関数の分散を小さくするほど，誤差が小さくなるが，分散を小さくしすぎると過学習を起こして汎化性能が下がることがわかった．

参考文献

[1] C.M. ビショップ，“パターン認識と機械学習 上 ベイズ理論による統計的予測，” シュプリンガー・ジャパン，

[2] 須山 敦志，“ベイズ推論による機械学習入門，” 株式会社講談社サイエンティフィク