パターン認識と機械学習
第3章 線形回帰モデル

橋本龍二

January 6, 2021

1　線形基底関数モデル

入力変数 $\begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}$ が与えられた時、その線形結合を取ることで、目標変数 t の予測値 $y(\boldsymbol{x}, \boldsymbol{w})$ を

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + ... + w_D x_D$$

と構成する。(線形回帰)
次に、入力変数 $\boldsymbol{x}$ に対して非線形な基底関数 $\phi(\boldsymbol{x}), ..., \phi(\boldsymbol{x})$ を用いて、

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x})$$

なるモデルを構成する。$\phi_0(\boldsymbol{x}) = 1$ とすれば、

$$\begin{aligned} y(\boldsymbol{x}, \boldsymbol{w}) &= \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) \\ &= {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}) \end{aligned}$$

と書ける。ただし、$\boldsymbol{w} = \begin{pmatrix} w_0 \\ \vdots \\ w_{M-1} \end{pmatrix}$、$\boldsymbol{\phi} = \begin{pmatrix} \phi_0 \\ \vdots \\ \phi_{M-1} \end{pmatrix}$

1. ガウス基底関数

$$\phi_j(x) = \exp\{-\frac{(x - \mu_j)^2}{2s^2}\}$$

2. シグモイド基底関数

$$\phi_j(x) = \sigma(\frac{x-\mu_j}{s}), \quad \sigma(a) = \frac{1}{1+\exp(-a)}$$

以下、目標変数 t は一次元とする。

1.1 最尤推定と最小二乗法

条件付きガウスノイズ分布のもとでの線形モデルを考える。すなわち、

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon, \quad \epsilon \sim N(0, \frac{1}{\beta})$$

として、

$$p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) = N(t|y(\boldsymbol{x}, \boldsymbol{w}), \frac{1}{\beta})$$

を考える。N 個の入力データ集合 $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_N \end{pmatrix}$ とその目標値のデータ集合 $\boldsymbol{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix}$ を考える。$y(\boldsymbol{x}, \boldsymbol{w}) = {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x})$ とすると、尤度関数は

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) &= \Pi_{n=1}^N N(t_n|y(\boldsymbol{x}, \boldsymbol{w}), \frac{1}{\beta}) \\ &= \Pi_{n=1}^N N(t_n|\ {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}), \frac{1}{\beta}) \\ &= (2\pi)^{-\frac{N}{2}} \times \beta^{\frac{N}{2}} \times \exp[-\frac{\beta}{2}\sum_{n=1}^N \{t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2] \\ \log p(\boldsymbol{t}|\boldsymbol{x}, \boldsymbol{w}, \beta) &= \frac{N}{2}\log\beta - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\sum_{n=1}^N \{t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \end{aligned}$$

最尤推定によって $\boldsymbol{w}$ を求める。対数尤度の最大化は二乗和誤差関数の最小化である。以下、$\boldsymbol{x}$ を省略する。($\boldsymbol{x}$ の分布のモデル化は目指していない)
対数尤度関数を微分して 0 と置く。

$$\begin{aligned} 0 = \frac{\partial}{\partial \boldsymbol{w}} \log p(\boldsymbol{t}|\boldsymbol{w}, \beta) &= \beta \sum_{n=1}^N \{t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n)\}\ {}^T\boldsymbol{\phi}(\boldsymbol{x}_n) \\ \therefore 0 &= \sum_{n=1}^N t_n\ {}^T\boldsymbol{\phi}(\boldsymbol{x}_n) - {}^T\boldsymbol{w}\{\sum_{n=1}^N \boldsymbol{\phi}(\boldsymbol{x}_n)\ {}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\} \end{aligned}$$

ここで、

$$\begin{aligned}
\sum_{n=1}^{N} t_n \, {}^T\boldsymbol{\phi}(\boldsymbol{x}_n) &= \sum_{n=1}^{N} t_n \begin{pmatrix} \phi_0(\boldsymbol{x}_n) & \dots & \phi_{M-1}(\boldsymbol{x}_n) \end{pmatrix} \\
&= \begin{pmatrix} t_1 & \dots & t_N \end{pmatrix} \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_n) \end{pmatrix} \\
\sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n) \, {}^T\boldsymbol{\phi}(\boldsymbol{x}_n) &= \sum_{n=1}^{N} \begin{pmatrix} \phi_0(\boldsymbol{x}_n) \\ \vdots \\ \phi_{M-1}(\boldsymbol{x}_n) \end{pmatrix} \begin{pmatrix} \phi_0(\boldsymbol{x}_n) & \dots & \phi_{M-1}(\boldsymbol{x}_n) \end{pmatrix} \\
&= \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \dots & \phi_0(\boldsymbol{x}_N) \\ \vdots & \ddots & \vdots \\ \phi_{M-1}(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_n) \end{pmatrix}
\end{aligned}$$

であるから、

$$\Phi = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \dots & \phi_{M-1}(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \dots & \phi_{M-1}(\boldsymbol{x}_n) \end{pmatrix}$$

と置けば (計画行列)、

$$\begin{aligned}
0 &= {}^T\boldsymbol{t}\Phi - {}^T\boldsymbol{w} \, {}^T\Phi\Phi \\
\therefore {}^T\boldsymbol{w} &= {}^T\boldsymbol{t}\Phi({}^T\Phi\Phi)^{-1} \\
\therefore \boldsymbol{w}_{ML} &= {}^T\{ {}^T\boldsymbol{t}\Phi({}^T\Phi\Phi)^{-1}\} \\
&= \{ {}^T({}^T\Phi\Phi)\}^{-1} \, {}^T({}^T(\boldsymbol{t}\Phi)) \\
&= ({}^T\Phi\Phi)^{-1} \, {}^T\Phi\boldsymbol{t}
\end{aligned}$$

1.2 最小二乗法の幾何学

$\boldsymbol{t} = \begin{pmatrix} \boldsymbol{t}_1 \\ \vdots \\ \boldsymbol{t}_N \end{pmatrix}$ は N 次元数ベクトル空間の要素。Φ の j 番目の列ベクトルを $\boldsymbol{\varphi}_j$ と置く。

$$\boldsymbol{\varphi}_j = \begin{pmatrix} \phi_j(\boldsymbol{x}_1) \\ \vdots \\ \phi_j(\boldsymbol{x}_N) \end{pmatrix} \in \mathbb{R}^N, \quad j = 1, ..., M$$

$M < N$ ならば、M 本のベクトル $\boldsymbol{\varphi}_j$ は M 次元の線形部分空間 S を張る。次に、推定値 $y(\boldsymbol{x}_j, \boldsymbol{w})$ を並べたベクトルを $\boldsymbol{y}$ と置くと、

$$\begin{aligned}\boldsymbol{y} &= \begin{pmatrix} y(\boldsymbol{x}_1, \boldsymbol{w}) \\ \vdots \\ y(\boldsymbol{x}_j, \boldsymbol{w}) \end{pmatrix} \\ &= \begin{pmatrix} w_0\phi_0(\boldsymbol{x}_1) + ... + w_{M-1}\phi_{M-1}(\boldsymbol{x}_1) \\ \vdots \\ w_0\phi_0(\boldsymbol{x}_N) + ... + \boldsymbol{w}_{M-1}\phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix} \\ &= w_o\boldsymbol{\varphi}_0 + ... + \boldsymbol{w}_{M-1}\boldsymbol{\varphi}_{M-1}\end{aligned}$$

より、$\boldsymbol{y}$ はベクトル $\boldsymbol{\varphi}_j$ の任意の線形結合より、S 内の要素である。今、二乗誤差；

$$\sum_{n=1}^{N}\{t_n - y(\boldsymbol{x}_n, \boldsymbol{w})\}^2$$

とは $\boldsymbol{y}$ と $\boldsymbol{t}$ の二乗ユークリッド距離である。すなわち、二乗和誤差の最小化とは $\boldsymbol{t} \in \mathbb{R}^N$ の $\mathbb{R}^N$ の部分空間 $S = \mathbb{R}^M$ への正射影 $\boldsymbol{y}$ を選ぶことである。

1.3 逐次学習

データ点を一度に一つだけ用いてモデルのパラメータを順次学習していく方法を逐次学習という。パターン n が与えられた時、確率的勾配降下法ではパラメータベクトル $\boldsymbol{w}$ を

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta\nabla E_n$$

を用いて更新する。

1.4 正則化最小二乗法

誤差関数に正則化項を加えて、

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

を最小化することで、過学習を防ぐ。$q = 1$ のとき lasso、$q = 2$ のとき二次正則化項という。lasso では、いくつかの係数 w_j が 0 になり、疎な解が得られる。

2　バイアス-バリアンス分解

入力 $\boldsymbol{x}$ に対して t の値に対する特定の推定値 $y(\boldsymbol{x})$ を選ぶことで損失 $L(t, y(\boldsymbol{x}))$ を被るとしよう。期待損失 $E[L]$ は、

$$E[L] = \int\int L(t, y(\boldsymbol{x}))p(\boldsymbol{x}, t)d\boldsymbol{x}dt$$

今、$L(t, y(\boldsymbol{x})) = \{y(\boldsymbol{x}) - t\}^2$ とすれば、

$$\begin{aligned}\{y(\boldsymbol{x}) - t\}^2 &= \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}] + E[t|\boldsymbol{x}] - t\}^2 \\ &= \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 + 2\{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}(E[t|\boldsymbol{x}] - t) + (E[t|\boldsymbol{x}] - t)^2\end{aligned}$$

なので、期待二乗損失 $E[L]$ は

$$\begin{aligned}E[L] &= \int\int \{y(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t)d\boldsymbol{x}dt \\ &= \int \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 \{\int p(\boldsymbol{x}, t)dt\}d\boldsymbol{x} \\ &\quad +2\int \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}\{\int E[t|\boldsymbol{x}]\int p(\boldsymbol{x}, t)dt - \int tp(t|\boldsymbol{x})dt\}p(\boldsymbol{x})d\boldsymbol{x} \\ &\quad +\int\int (E[t|\boldsymbol{x}] - t)^2 p(t, \boldsymbol{x})dtd\boldsymbol{x} \\ &= \int \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 p(\boldsymbol{x})d\boldsymbol{x} \\ &\quad +2\int \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}\{E[t|\boldsymbol{x}]p(\boldsymbol{x}) - E[t|\boldsymbol{x}]p(\boldsymbol{x})\}d\boldsymbol{x} \\ &\quad +\int \{\int (t - E[t|\boldsymbol{x}])^2 p(t|\boldsymbol{x})dt\}p(\boldsymbol{x})d\boldsymbol{x} \\ &= \int \{y(\boldsymbol{x}) - E[t|\boldsymbol{x}]\}^2 p(\boldsymbol{x})d\boldsymbol{x} + \int var[t|\boldsymbol{x}]p(\boldsymbol{x})d\boldsymbol{x}\end{aligned}$$

の形で書くことができる。第二項は $y(\boldsymbol{x})$ に依存しないので、$h(\boldsymbol{x}) = E[t|\boldsymbol{x}]$ が期待二乗損失を最小化する回帰関数である。第二項は t の分布の分散を $\boldsymbol{x}$ に関して平均したもので、目標データが本質的に持つ変動 (ノイズ) である。

＜回帰の考え方＞
入力 $\boldsymbol{x}$ に対して t の推定値を選ぶ上で、期待損失を最小化するのは $E[t|\boldsymbol{x}]$ である。
$\rightarrow E[t|\boldsymbol{x}]$ を求めたい。
$\rightarrow p(t|\boldsymbol{x}) = N(t|y(\boldsymbol{x}, \boldsymbol{w}), \frac{1}{\beta})$ とすることで、$E[t|\boldsymbol{x}]$ を $y(\boldsymbol{x}, \boldsymbol{w})$ によってモデル

化
ここで、頻度主義的には $\boldsymbol{w}$ を点推定する。ベイズの立場では、$\boldsymbol{w}$ の事後分布を求める。

さて、今無限のデータ点を取れるのであれば推定値 $y(\boldsymbol{x})$ は理想の精度で $h(\boldsymbol{x})$ に近づけることができる。しかし、現実には有限のデータ集合 $\mathcal{D}$ を得ることしか出来ないので、それに対する予測関数 $y(\boldsymbol{x};\mathcal{D})$ を構成する。$y(\boldsymbol{x};\mathcal{D}) = h(\boldsymbol{x})$ は厳密には成り立たない。ではその誤差 $\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2$ はどのように変動するだろうか。

$$\begin{aligned}\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2 &= \{y(\boldsymbol{x};\mathcal{D}) - E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] + E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})\}^2 \\ &= \{y(\boldsymbol{x};\mathcal{D}) - E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2 + \{E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})\}^2 \\ &\quad +2\{y(\boldsymbol{x};\mathcal{D}) - E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}\{E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})\}\end{aligned}$$

これ全体に $\mathcal{D}$ に関して期待値を取る。$y(\boldsymbol{x};\mathcal{D})$ でない項は定数であるから、

$$E_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D}) - h(\boldsymbol{x})\}^2] = \{E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})\}^2 + E_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D}) - E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2]$$

右辺第一項は (バイアス)^2、第二項はバリアンスである。

- バイアス：$E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})$　　全てのデータ集合 $\mathcal{D}$ の取り方に関する予測値の平均が理想的な回帰関数 $h(\boldsymbol{x})$ からどのくらい離れているか
- バリアンス：$E_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D}) - E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2]$　　各のデータ集合 $\mathcal{D}$ に対する解 $y(\boldsymbol{x};\mathcal{D})$ の分散

以上より、期待損失は二乗バイアスとバリアンス、ノイズに分解できる。すなわち、

$$\begin{aligned}E[L] &= \int \{E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})] - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})d\boldsymbol{x} \\ &\quad + \int E_{\mathcal{D}}[\{y(\boldsymbol{x};\mathcal{D}) - E_{\mathcal{D}}[y(\boldsymbol{x};\mathcal{D})]\}^2]p(\boldsymbol{x})d\boldsymbol{x} \\ &\quad + \int var[t|\boldsymbol{x}]p(\boldsymbol{x})d\boldsymbol{x}\end{aligned}$$

バイアスとバリアンスはモデルの複雑さに依存する。単純なモデルではバリアンスは小さくなるがバイアスは大きくなる。複雑なモデルではバイアスは小さくなるがバリアンスは大きくなる。

3　ベイズ線形回帰

線形回帰モデルをベイズ的に取り扱うことによって、最尤推定の過学習を回避するとともに訓練データだけからモデルの複雑さを自動的に決定する。

3.1　パラメータの分布

$y(\boldsymbol{x}, \boldsymbol{w}) = {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x})$ とすると、尤度関数は

$$p(\boldsymbol{t}|\boldsymbol{w}) = \Pi_{n=1}^{N} N(t_n|\ {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}), \frac{1}{\beta})$$

と、$\boldsymbol{w}$ の二次関数の指数なので、共役事前分布は

$$p(\boldsymbol{w}) = N(\boldsymbol{w}|\boldsymbol{m}_0, S_0)$$

で与えるのが適切である。$\boldsymbol{a} = \begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{w} \end{pmatrix}$ とおく。

$$\begin{aligned}
\log p(\boldsymbol{a}) &= \sum_{n=1}^{N} {}^T(t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n))\beta(t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n)) \\
&\quad + {}^T(\boldsymbol{w} - \boldsymbol{m}_0)S_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0) \\
&= {}^T(\boldsymbol{t} - \Phi\boldsymbol{w})\beta(\boldsymbol{t} - \Phi\boldsymbol{w}) + {}^T(\boldsymbol{w} - \boldsymbol{m}_0)S_0^{-1}(\boldsymbol{w} - \boldsymbol{m}_0) \\
&= {}^T\boldsymbol{w}\ {}^T\Phi\beta\Phi\boldsymbol{w} + {}^T\boldsymbol{w}S_0^{-1}\boldsymbol{w} \\
&\quad - {}^T\boldsymbol{w}\ {}^T\Phi\beta\boldsymbol{t} - {}^T\boldsymbol{t}\beta\Phi\boldsymbol{w} + {}^T\boldsymbol{t}\beta\boldsymbol{t} \\
&\quad + {}^T\boldsymbol{w}S_0^{-1}m0 - {}^T\boldsymbol{m}_0S_0^{-1}\boldsymbol{w} \\
&= {}^T\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{t} \end{pmatrix}\begin{pmatrix} {}^T\Phi\beta\Phi + S_0^{-1} & -{}^T\Phi\beta \\ -\beta\Phi & \beta \end{pmatrix}\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{t} \end{pmatrix} - {}^T\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{t} \end{pmatrix}\begin{pmatrix} S_0^{-1}\boldsymbol{m}_0 \\ \boldsymbol{0} \end{pmatrix} - {}^T\begin{pmatrix} S_0^{-1}\boldsymbol{m}_0 \\ \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{w} \\ \boldsymbol{t} \end{pmatrix} \\
&= {}^T\boldsymbol{a}\Lambda\boldsymbol{a} - {}^T\boldsymbol{a}B - {}^TB\boldsymbol{a} \\
&= {}^T(\boldsymbol{a} - \Lambda^{-1}B)\Lambda(\boldsymbol{a} - \Lambda^{-1}B)
\end{aligned}$$

ただし、

$$\Lambda = \begin{pmatrix} {}^T\Phi\beta\Phi + S_0^{-1} & -{}^T\Phi\beta \\ -\beta\Phi & \beta \end{pmatrix}, \quad B = \begin{pmatrix} S_0^{-1}\boldsymbol{m}_0 \\ \boldsymbol{0} \end{pmatrix}$$

と置いた。よって、

$$\Sigma_{\boldsymbol{w}|\boldsymbol{t}} = (S_0^{-1} + {}^T\Phi\beta\Phi)^{-1}$$

条件付き平均を求める。

$$\begin{aligned}
\Lambda^{-1} &= \begin{pmatrix} S_0 & S_0\,{}^T\Phi \\ \Phi S_0 & \beta^{-1} + \Phi S_0\,{}^T\Phi \end{pmatrix} \\
\therefore \boldsymbol{\mu}_a &= \begin{pmatrix} S_0 & S_0\,{}^T\Phi \\ \Phi S_0 & \beta^{-1} + \Phi S_0\,{}^T\Phi \end{pmatrix} \begin{pmatrix} S_0^{-1}\boldsymbol{m}_0 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \boldsymbol{m}_0 \\ \Phi\boldsymbol{m}_0 \end{pmatrix}
\end{aligned}$$

より、

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}} &= \Sigma_{\boldsymbol{w}|\boldsymbol{t}}\{({}^T\Phi\beta\Phi + S_0^{-1})m_0 + {}^T\Phi\beta(\boldsymbol{t} - \Phi\boldsymbol{m}_0) \\
&= \Sigma_{\boldsymbol{w}|\boldsymbol{t}}(S_0^{-1}\boldsymbol{m}_0 + {}^T\Phi\beta\boldsymbol{t})
\end{aligned}$$

事前分布を $p(\boldsymbol{w}|\alpha) = N(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}I)$ と置けば、

$$\begin{aligned}
\Sigma_{\boldsymbol{w}|\boldsymbol{t}} &= (\alpha I + \beta\,{}^T\Phi\Phi)^{-1} \\
\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}} &= \beta\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\,{}^T\Phi\boldsymbol{t}
\end{aligned}$$

3.2 予測分布

訓練データが得られた時、新しい t の値を予測する。

$$\begin{aligned}
p(t|\boldsymbol{t}, \alpha, \beta) &= \int p(t, \boldsymbol{w}|\boldsymbol{t}, \alpha, \beta)d\boldsymbol{w} \\
&= \int p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta)p(t|\boldsymbol{w}, \beta)d\boldsymbol{w}
\end{aligned}$$

より、「$\boldsymbol{w}$ の分布」と「平均が $\boldsymbol{w}$ に依存する t の条件付き分布」が与えられたときの t の周辺分布なので、μ_t, Σ_t は、

$$\begin{aligned}
\mu_t &= {}^T\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}}\boldsymbol{\phi}(\boldsymbol{x}) \\
\Sigma_t &= \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\,{}^T\boldsymbol{\phi}(\boldsymbol{x})
\end{aligned}$$

である。予測分布の平均を t の推定値として採用する。

3.3 等価カーネル

事後平均 $\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}}$ を $\boldsymbol{w}$ の推定値として採用する。

$$\begin{aligned}\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}}) &= {}^T\boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}}\boldsymbol{\phi}(\boldsymbol{x}) \\ &= {}^T(\beta\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\,{}^T\Phi\boldsymbol{t})\boldsymbol{\phi}(\boldsymbol{x}) \\ &= \beta\,{}^T\boldsymbol{t}\Phi\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\boldsymbol{\phi}(\boldsymbol{x}) \\ &= \beta\,{}^T\boldsymbol{\phi}(\boldsymbol{x})\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\,{}^T\Phi\boldsymbol{t}\end{aligned}$$

(最終行について、y はスカラーなので全体に転置をとって良い)
${}^T\Phi\boldsymbol{t} = \sum_{n=1}^{N}\boldsymbol{\phi}(\boldsymbol{x}_n)t_n$　(「行列 × ベクトル」は列ベクトルの一次結合で表せる。) なので、

$$\begin{aligned}\boldsymbol{y}(\boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{w}|\boldsymbol{t}}) &= \sum_{n=1}^{N}\beta\,{}^T\boldsymbol{\phi}(\boldsymbol{x})\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\boldsymbol{\phi}(\boldsymbol{x}_n)t_n \\ &= \sum_{n=1}^{N}k(\boldsymbol{x}, \boldsymbol{x}_n)t_n\end{aligned}$$

($\boldsymbol{x}$ はテストデータのあるベクトル。$\boldsymbol{x}_n$ は N 個の訓練データの n 番目のベクトル)
このように、点 $\boldsymbol{x}$ に対する予測分布の平均は訓練データの目標変数 t_n の線形結合で与えられる。
ここで、二本の N 次元ベクトル $\boldsymbol{x}, \boldsymbol{x}'$ を引数にもつ関数

$$k(\boldsymbol{x}, \boldsymbol{x}') = \beta\,{}^T\boldsymbol{\phi}(\boldsymbol{x})\Sigma_{\boldsymbol{w}|\boldsymbol{t}}\boldsymbol{\phi}(\boldsymbol{x}')$$

を平滑化行列または等価カーネルという。等価カーネルは重みの役割を果たす。例えば p158 の図 3.11 では $x = 0$ としたときの x' の関数としての等価カーネルがプロットされている。$x' = 0$ 付近の重みが重要視されていることがわかる。ベイズ線形回帰では、訓練データの目標値の線形結合によって予測を行う。この際、訓練データのうち予測したいデータ点 $\boldsymbol{x}$ に近い点の目標データが重視される、ということ。
ガウス過程を用いることで、基底関数を用いずに、近傍点での相関を強く、より離れた点への相関は小さい局所的なカーネルを定義することで予測ができる。

4 ガウスモデル比較

データがモデル $\{\mathcal{M}_i\}$, $(i = 1, ..., L)$ のどれかにしたがって生成されているとする (モデルは $\mathcal{D}$ 上の確率分布)。訓練データ集合 $\mathcal{D}$ が与えられた時、モデ

ルの事後分布は次で与えられる。

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$$

事前分布 $p(\mathcal{M}_i)$ は大体分からないので、全てのモデルの事前確率は等しいとする。この時、モデルエビデンス $p(\mathcal{D}|\mathcal{M}_i)$ が重要な働きをする。モデルエビデンスは、

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}, \theta|\mathcal{M}_i)d\theta$$

と、モデル $\mathcal{M}_i$ の空間で θ を周辺化した尤度関数と見ることができるため、周辺尤度とも呼ばれる。また、モデル $\mathcal{M}_i$ の $\mathcal{M}_i$ の $\mathcal{M}_j$ に対するベイズ因子を

$$\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$$

と定義する。
一旦モデルの事後分布がわかると、予測分布は

$$\begin{aligned} p(t|\boldsymbol{x}, \mathcal{D}) &= \sum_{i=1}^{L} p(t, \mathcal{M}_i|\boldsymbol{x}, \mathcal{D}) \\ &= \sum_{i=1}^{L} p(\mathcal{M}_i|\boldsymbol{x}, \mathcal{D})p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) \end{aligned}$$

と書ける (混合分布)。

w を一次元とする。事後分布は $p(w|\mathcal{D}, \mathcal{M}_i) \propto p(w|\mathcal{M}_i)p(\mathcal{D}|w, \mathcal{M}_i)$ なので、周辺尤度は

$$\begin{aligned} p(\mathcal{D}|\mathcal{M}_i) &= \int p(\mathcal{D}, w|\mathcal{M}_i)dw \\ &= \int p(\mathcal{D}|w, \mathcal{M}_i)p(w|\mathcal{M}_i)dw \\ &\simeq \frac{1}{\Delta w_{prior}} \int p(\mathcal{D}|w, \mathcal{M}_i)dw \\ &\simeq \frac{1}{\Delta w_{prior}} \times p(\mathcal{D}|w_{MAP})\Delta w_{posterior} \\ \log p(\mathcal{D}|\mathcal{M}_i) &\simeq \log p(\mathcal{D}|w_{MAP}) + \log \frac{\Delta w_{posterior}}{\Delta w_{prior}} \end{aligned}$$

第一項は尤もな w によるデータへのフィッティング度を表し、第二項はモデルの複雑さに対する罰則項 (負) である。
では、パラメータ M 個を幾つにするかを選ぶ問題を考える。$w_{posterior}/w_{prior}$ が全てのパラメータで等しいとすると、

$$\log p(\mathcal{D}|\mathcal{M}_i) \simeq \log p(\mathcal{D}|w_{MAP}) + M \log(\frac{\Delta w_{posterior}}{\Delta w_{prior}})$$

となる。実際に、周辺尤度を最大化することで中くらいの複雑さのモデルが選ばれる。単純すぎるモデルでは $\mathcal{D}_0$ を生成する確率 $p(\mathcal{D}_0|\mathcal{M}_i)$ は低くなる (限られたデータ集合しか作り出せない) し、複雑すぎるモデルではありうるデータの組み合わせが多すぎてピンポイントで $\mathcal{D}_0$ を生成する確率はやはり小さくなる。このように、ベイズ線形回帰では (事前分布が単純、など一定の条件のもとで) パラメータの事後分布をパラメータで積分することで、解析的にモデル選択を行うことができる。

5 エビデンス近似

訓練データ $\boldsymbol{t}$ がえられたとき、新しい入力 $\boldsymbol{x}$ に対して目標変数 t の値を予測する。事前分布の精度 α とガウスノイズ β に事前分布を導入する。予測分布は、

$$\begin{aligned} p(t|\boldsymbol{t}) &= \int\int\int p(t, \boldsymbol{w}, \alpha, \beta|\boldsymbol{t}) d\boldsymbol{w} d\alpha d\beta \\ &= \int\int\int p(\alpha, \beta|\boldsymbol{t}) p(\boldsymbol{w}|\alpha, \beta, \boldsymbol{t}) p(t|\boldsymbol{w}, \beta, \boldsymbol{t}) d\boldsymbol{w} d\alpha d\beta \end{aligned}$$

5.1 エビデンス関数の評価

$p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{w}|\alpha) p(\boldsymbol{t}|\boldsymbol{w}, \beta)$ より、事後分布をパラメータで周辺化することで α, β の周辺尤度 $p(\boldsymbol{t}|\alpha, \beta)$ が次のように得られる。

$$\begin{aligned} p(\boldsymbol{t}|\alpha, \beta) &= \int p(\boldsymbol{t}, \boldsymbol{w}|\alpha, \beta) d\boldsymbol{w} \\ &= \int p(\boldsymbol{w}|\alpha) p(\boldsymbol{t}|\boldsymbol{w}, \beta) d\boldsymbol{w} \end{aligned}$$

ここで、

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{w},\beta) &= (\frac{\beta}{2\pi})^{N/2}\exp[-\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - {}^T\boldsymbol{w}\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2] \\ &= (\frac{\beta}{2\pi})^{N/2}\exp\{-\frac{\beta}{2}{}^T(\boldsymbol{t}-\Phi\boldsymbol{w})(\boldsymbol{t}-\Phi\boldsymbol{w})\} \\ p(\boldsymbol{w}|\alpha) &= \frac{\alpha^{M/2}}{(2\pi)^{M/2}}\exp(-\frac{1}{2}\alpha\,{}^T\boldsymbol{w}\boldsymbol{w}) \end{aligned}$$

より、

$$p(\boldsymbol{t}|\alpha,\beta) = (\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}\int\exp(-E[\boldsymbol{w}])d\boldsymbol{w}$$

ただし、

$$\begin{aligned} E[\boldsymbol{w}] &= \beta E_D[\boldsymbol{w}] + \alpha E_W[\boldsymbol{w}] \\ &= \frac{\beta}{2}{}^T(\boldsymbol{t}-\Phi\boldsymbol{w})(\boldsymbol{t}-\Phi\boldsymbol{w}) + \frac{\alpha}{2}{}^T\boldsymbol{w}\boldsymbol{w} \end{aligned}$$

$E[\boldsymbol{w}]$ を $\boldsymbol{w}$ について平方完成して積分を解くことを考える。

$$\begin{aligned} E[\boldsymbol{w}] &= \frac{\beta}{2}{}^T(\boldsymbol{t}-\Phi\boldsymbol{w})(\boldsymbol{t}-\Phi\boldsymbol{w}) + \frac{\alpha}{2}{}^T\boldsymbol{w}\boldsymbol{w} \\ &= \frac{\beta}{2}({}^T\boldsymbol{t}\boldsymbol{t} - {}^T\boldsymbol{t}\Phi\boldsymbol{w} - {}^T\boldsymbol{w}\,{}^T\Phi\boldsymbol{t} + {}^T\boldsymbol{w}\,{}^T\Phi\Phi\boldsymbol{w}) + \frac{\alpha}{2}{}^T\boldsymbol{w}\boldsymbol{w} \\ &= \frac{1}{2}{}^T\boldsymbol{w}(\beta\,{}^T\Phi\Phi + \alpha I)\boldsymbol{w} - \frac{\beta}{2}{}^T\boldsymbol{w}\,{}^T\Phi\boldsymbol{t} - \frac{\beta}{2}{}^T\boldsymbol{t}\Phi\boldsymbol{w} + \frac{\beta}{2}{}^T\boldsymbol{t}\boldsymbol{t} \\ &= \frac{1}{2}{}^T(\boldsymbol{w}-A^{-1}B)A(\boldsymbol{w}-A^{-1}B) + \frac{\beta}{2}{}^T\boldsymbol{t}\boldsymbol{t} - \frac{1}{2}{}^BA^{-1}B \end{aligned}$$

ただし、

$$A = \alpha I + \beta\,{}^T\Phi\Phi,\ B = \beta\,{}^T\boldsymbol{w}\,{}^T\Phi\boldsymbol{t}$$

とおいた。

$$\begin{aligned} \boldsymbol{m}_N &= A^{-1}B \\ &= \beta A^{-1}\,{}^T\Phi\boldsymbol{t} \end{aligned}$$

と置くと、$B = A\boldsymbol{m}_N$ なので、

$$\begin{aligned}
\frac{\beta}{2}\,{}^T\boldsymbol{t}\boldsymbol{t} - {}^TBA^{-1}B &= \frac{\beta}{2}\,{}^T\boldsymbol{t}\boldsymbol{t} - \frac{1}{2}\,{}^T\boldsymbol{m}_N\,{}^TA\boldsymbol{m}_N \\
&= \frac{\beta}{2}\,{}^T\boldsymbol{t}\boldsymbol{t} - \frac{1}{2}\,{}^T\boldsymbol{m}_N(\alpha I + \beta\,{}^T\Phi\Phi)\boldsymbol{m}_N \\
&= \frac{\beta}{2}\,{}^T(\boldsymbol{t} - \Phi\boldsymbol{m}_N)(\boldsymbol{t} - \Phi\boldsymbol{m}_N) + \frac{\alpha}{2}\,{}^T\boldsymbol{m}_N\boldsymbol{m}_N \\
&= E[\boldsymbol{m}_N]
\end{aligned}$$

より、

$$E[\boldsymbol{w}] = \frac{1}{2}\,{}^T(\boldsymbol{w} - \boldsymbol{m}_N)A(\boldsymbol{w} - \boldsymbol{m}_N) + E[\boldsymbol{m}_N]$$

したがって、

$$\begin{aligned}
\int \exp(-E[\boldsymbol{w}])d\boldsymbol{w} &= \exp(-E[\boldsymbol{m}_N]) \int \exp\{-\frac{1}{2}\,{}^T(\boldsymbol{w} - \boldsymbol{m}_N)A(\boldsymbol{w} - \boldsymbol{m}_N)\}d\boldsymbol{w} \\
&= \exp\{-E[\boldsymbol{m}_N]\}\frac{(2\pi)^{M/2}}{|A|^{1/2}}
\end{aligned}$$

となるので、周辺尤度は

$$\begin{aligned}
p(\boldsymbol{t}|\alpha,\beta) &= (\frac{\beta}{2\pi})^{N/2}(\frac{\alpha}{2\pi})^{M/2}\exp\{-E[\boldsymbol{m}_N]\}\frac{(2\pi)^{M/2}}{|A|^{1/2}} \\
\log p(\boldsymbol{t}|\alpha,\beta) &= \frac{M}{2}\log\alpha + \frac{N}{2}\log\beta - E[\boldsymbol{m}_N] - \frac{1}{2}\log|A| - \frac{N}{2}\log(2\pi)
\end{aligned}$$