PRML 輪講（3.3 節，3.4 節）

2009 年 6 月 17 日　担当　福嶋　誠

3.3. Bayesian Linear Regression (ベイズ線形回帰)

線形回帰モデルで最尤推定によりパラメータを決める際には，得られるデータ数によって実効的なモデルの複雑さ（基底関数の個数，正則化係数の値）を適切に選ばなければならない．

単純に尤度関数を最大化することによりパラメータを推定すると，モデルは常に複雑になり訓練データに対して過学習を引き起こす．1.3 節では過学習を避けるためにテストデータを用いてモデル選択を行ったが，この方法では貴重なデータが失われてしまう．そこで本節では線形回帰モデルをベイズ的枠組みで取り扱うことにより，訓練データのみから自動的にモデル複雑さを決定する方法を述べる．

3.3.1 Parameter distribution (パラメータの分布)

モデルパラメータ $\mathbf{w}$ の事前分布を導入することから線形回帰モデルをベイズ的に取り扱う．ここではノイズ精度 β を既知の定数として議論を進める．式 (3.10) で表される尤度関数の指数部分が $\mathbf{w}$ の 2 次形式であることから，対応する共役事前分布は以下のような正規分布で表される．

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0). \tag{3.48}$$

事後分布は尤度関数と事前分布の積で記述できるので，平方完成を用いて以下のように計算できる (Exercise 3.7 を参照).

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N). \tag{3.49}$$

ただし

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}) \tag{3.50}$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} \tag{3.51}$$

である．ここで事後分布は正規分布となるのでそのモード（つまり最大事後確率をとるパラメータ）は平均値と一致し $\mathbf{w}_{\mathrm{MAP}} = \mathbf{m}_N$ となる．仮に無限に広い事前分布 $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}/(\alpha \to 0)$ を考えると，事後分布の平均 $\mathbf{m}_N$ は式 (3.15) で得られる最尤推定量 $\mathbf{w}_{\mathrm{ML}}$ と一致する．そして $N = 0$ ならば事後分布は事前分布に一致する．さらにデータ点が逐次的に与えられる場合，任意の時点での事後分布が新たにデータ点が得られた際には事前分布として働く (Exercise 3.8 を参照).

これ以降は簡単のため，平均ゼロ，精度 α の等方的ガウス分布

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{3.52}$$

を考える．上の分布に対応する事後分布のパラメータはそれぞれ

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \tag{3.53}$$

$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} \tag{3.54}$$

と表される．

対数事後分布は対数尤度と対数事前分布の和から得られ，

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \mathrm{const} \tag{3.55}$$

となる．上式より事後分布の $\mathbf{w}$ についての最大化と，式 (3.27) で $\lambda = \alpha/\beta$ とおいたときの正則化二乗和誤差関数の最小化は等価であることが分かる．

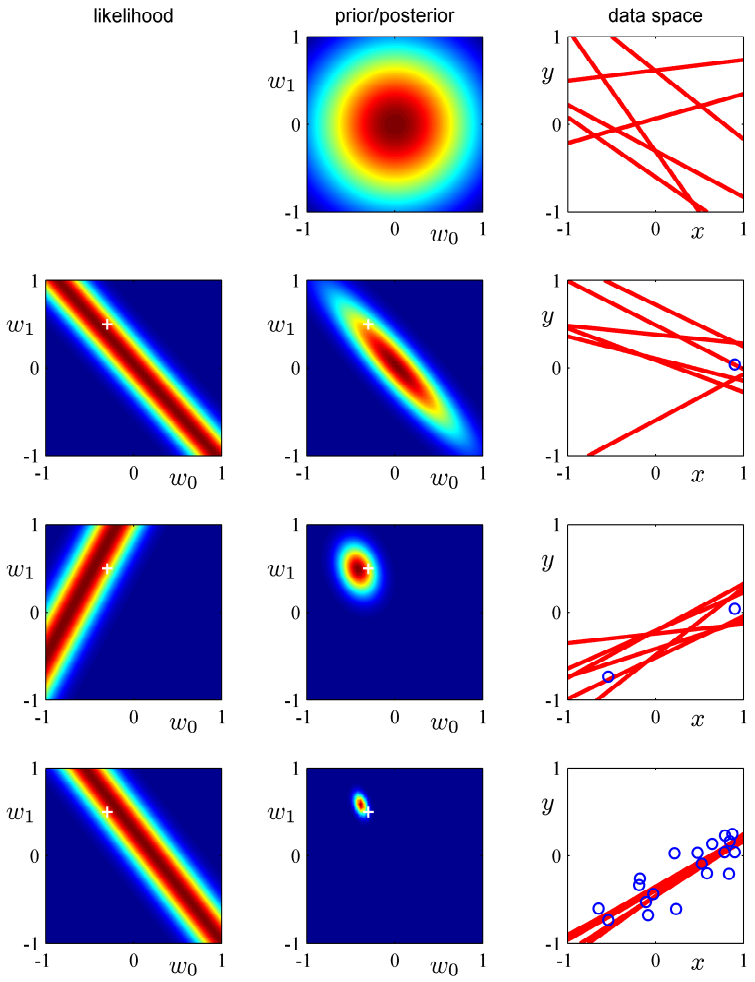

図 3.7: $y(x, \mathbf{w}) = w_0 + w_1 x$ に対する逐次ベイズ学習の例

直線フィッティングの例題を用いて，線形基底関数モデルにおけるベイズ学習と事後分布の逐次的な更新について説明する．ここでは１次元の入力変数 x と目的変数 t を考え，線形モデル $y(x, \mathbf{w}) = w_0 + w_1 x$ を用いる．まず x を一様分布 $\mathrm{U}(x|-1, 1)$ からランダムに選び，関数 $f(x, \mathbf{a}) = a_0 + a_1 x$ $(a_0 = -0.3, a_1 = 0.5)$ を評価した後，標準偏差 0.2 のノイズを加えることで人工訓練データを生成する．我々の目標は訓練データからモデルのパラメータ a_0，a_1 を復元すること，そしてデータセットのサイズと推定値との関係を示すことである．ここでノイズの分散は既知であるので，精度パラメータ β は $\beta = (1/0.2)^2 = 25$ と求められる．また $\alpha = 2.0$ と定める．

図 3.7 はデータ集合のサイズを大きくしていくと変化するベイズ学習の結果と，新しいデータ点が得られた際には現在の事後分布が事前分布を構成するというベイズの逐次性を示している．１行目にはまだ１つもデータ点が得られていない時点での事前分布とその事前分布から抽出したパラメータ $\mathbf{w}$ をもつ関数 $y(x, \mathbf{w})$ が６個プロットされている．

２行目はデータ点を一つ観測した時点でのプロットである．データ点 (x, t) は右の図で青い円印で表され，真のパラメータの値 $(a_0 = 0.3, a_1 = 0.5)$ は中央の図で白い＋印で表されている．左の図は得られたデータ点に対する尤度関数を $\mathbf{w}$ の関数で示している．尤度関数は右の図のグラフがデータ点の近傍を通らなければならないという柔らかな拘束条件を与えており，その度合いは精度 β の値で決まる．１列目の事前分布とこの尤度関数との積を正規化すると２列目中央の事後分布が得られ，この事後分布から抽出されたパラメータを持つ直線が右の図にプロットされる．

データ点が２個得られた際には３列目左の尤度関数が得られ，２列目の事後分布との積をとり正規化することで３列目中央の事後分布が求まる．右の図は事後分布より抽出されたパラメータをもつ直線をプロットしたもので，直線はデータ点の近くを通っていることが分かる．

最下段はデータ点が２０個得られたときの図であり，データ点が増えると事後分布のピークが鋭くなることが見てとれる．データ点を無限に増やすと，事後分布は真のパラメータ上のデルタ関数となる．

異なる形式のパラメータ事前分布を考えることもできる．例えばガウス事前分布を一般化して得られる事前分布

$$p(\mathbf{w}|\alpha) = \left[\frac{q}{2}\left(\frac{\alpha}{2}\right)^{1/q}\frac{1}{\Gamma(1/q)}\right]^M \exp\left(-\frac{\alpha}{2}\sum_{j=1}^{M}|w_j|^q\right) \tag{3.56}$$

などがある．$q = 2$ のときガウス事前分布と一致し，またそのときのみ尤度関数 (3.10) の共役事前分布となる．上式を $\mathbf{w}$ に関して最大化することは正則化誤差関数 (3.29) を最小化することと等価である．ガウス事前分布の際には事後分布のモードが平均値と一致するが，$q \neq 2$ のときは一般にそうはならない．

Exercise 3.7

平方完成を用いて式 (3.49) から (3.51) の導出を行う．事後分布 $p(\mathbf{w}|\mathbf{t})$ は以下のように書き下すことができる．

$$p(\mathbf{w}|\mathbf{t}) \propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) - \sum_{n=1}^{N}\frac{\beta}{2}\left(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2\right).$$

指数部分を取り出し -2 倍してから展開すると以下のように式変形できる．ここで簡単のため $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ とした．

$$\begin{aligned}
&(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0) + \sum_{n=1}^{N}\beta\left(t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2 \\
&= \mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 + \sum_{n=1}^{N}\left(\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n)^{\mathrm{T}} - 2\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_n t_n\right) + \text{const.} \\
&= \mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}}\right)\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n t_n\right) + \text{const..}
\end{aligned}$$

式 (3.49) と上式について係数比較を行うことにより以下のように $\mathbf{m}_N$，$\mathbf{S}_N^{-1}$ が求められる．

$$\begin{aligned}
\mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^{\mathrm{T}} \\
&= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} \\
\mathbf{m}_N &= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\sum_{n=1}^{N}\boldsymbol{\phi}_n t_n\right) \\
&= \mathbf{S}_N\left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right).
\end{aligned}$$

Exercise 3.8

基本的には前問と同様にして平方完成を用いて計算を行う．$N+1$ 個目のデータ点が得られた際の事後分布は,

$$p(\mathbf{w}|t_{N+1}, \mathbf{x}_{N+1}, \mathbf{m}_N, \mathbf{S}_N) \propto \exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) - \frac{\beta}{2}(t_{N+1} - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_{N+1}))^2\right)$$

となる．ここで指数部分を取り出し -2 倍してから展開すると以下のように式変形できる．また簡単のため $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n)$ としている．

$$\begin{aligned}
&(\mathbf{w}-\mathbf{m}_N)^{\mathrm{T}}\mathbf{S}_N^{-1}(\mathbf{w}-\mathbf{m}_N) + \beta(t_{N+1} - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^2 \\
&= \mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N + \beta(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1})^{\mathrm{T}} - 2\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}_{N+1}t_{N+1} + \text{const} \\
&= \mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}})\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta\boldsymbol{\phi}_{N+1}t_{N+1}) + \text{const.}
\end{aligned}$$

$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_{N+1}, \mathbf{S}_{N+1})$ と上式について係数比較を行うことにより以下のように $\mathbf{m}_N$，$\mathbf{S}_{N+1}^{-1}$ が求められる．

$$\begin{aligned}
\mathbf{S}_{N+1}^{-1} &= \mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}_{N+1}\boldsymbol{\phi}_{N+1}^{\mathrm{T}} \\
\mathbf{m}_{N+1} &= \mathbf{S}_{N+1}(\mathbf{S}_N^{-1}\mathbf{m}_N + \beta\boldsymbol{\phi}_{N+1}t_{N+1}).
\end{aligned}$$

3.3.2 Predictive distribution (予測分布)

ここからは新しい入力データ $\mathbf{x}$ に対する t を予測することを考える．そこで次式で定義される**予測分布**

$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)\,\mathrm{d}\mathbf{w} \tag{3.57}$$

を評価していく．式 (3.8) の条件付き分布と式 (3.49) の事後分布をガウス分布の周辺化の公式 (2.115) に代入することで，予測分布が以下のような形をとることがわかる (Exercise 3.10 を参照)．

$$p(t|\mathbf{x},\mathbf{t},\alpha,\beta) = \mathcal{N}(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}),\sigma_N^2(\mathbf{x})). \tag{3.58}$$

ただし，予測分布の分散 $\sigma_N^2(\mathbf{x})$ は

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) \tag{3.59}$$

と与えられる．上式の第 1 項はデータにのったノイズを表しており，第 2 項は $\mathbf{w}$ に関する不確かさを反映している．ノイズとパラメータの分布は独立なガウス分布であるため，予測分布の分散は加法的である．また新たにデータが観測されると，事後分布が鋭くなっていくので $\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2$ が成り立ち (Exercise 3.11 を参照)，$N \to \infty$ の極限では第 2 項が 0 となる．

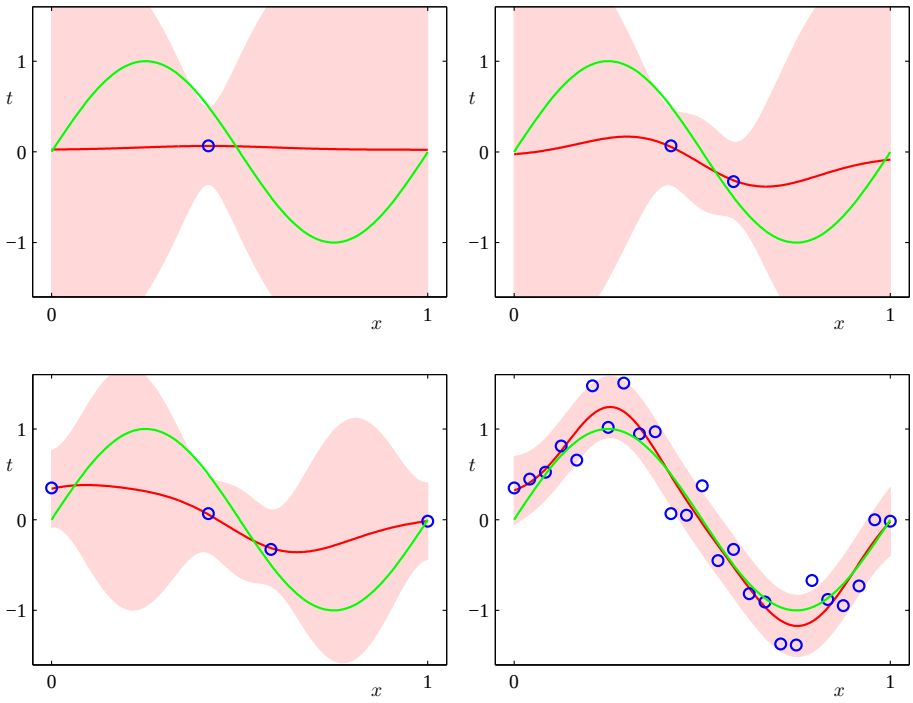

図 3.8: 三角関数データ集合に対する予測分布 (3.58) の例

ベイズ線形回帰モデルの予測分布の例を図 3.8 に示す．ここではガウス基底関数の線形結合からなるモデルを，さまざまなサイズの訓練データにあてはめたときの対応する事後分布を表示している．緑色の曲線は関数 $\sin(2\pi x)$ を表し，データ点はこの関数から生成しガウスノイズを加えることで得られる．訓練データは青い円印で表される ($N = 1$, $N = 2$, $N = 4$, $N = 25$)．また赤い曲線は予測分布の平均を示し，赤い領域は平均 $\pm$ 標準偏差の範囲を表す．予測の不確かさは x に依存し，データ点近傍ではその度合いが小さくなる．不確かさはデータ点を観測すればするほど減少する．

図 3.8 はある入力 x における予測分散のみを図示しているが，図 3.9 では異なる x に対する予測値同士の共分散を

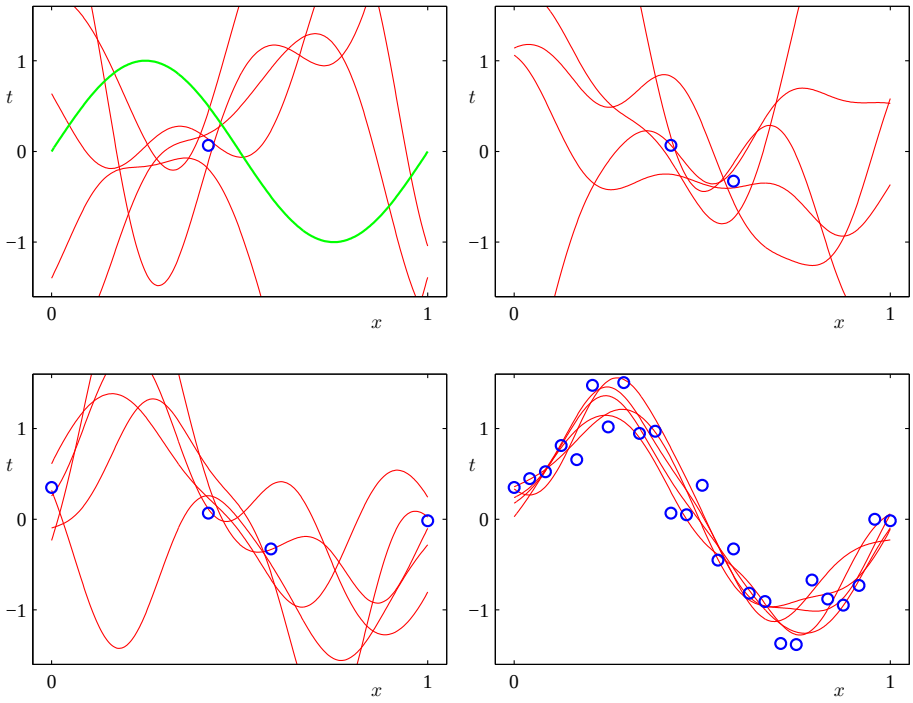

図 3.9: 図 3.8 に示した $\mathbf{w}$ の事後分布から得られた関数 $y(x, \mathbf{w})$ のプロット

調べるため，$\mathbf{w}$ を事後分布から選び対応する関数 $y(x, \mathbf{w})$ をプロットしている．

ガウス関数のような局所的な基底関数を選ぶと，基底関数の中心から離れれば離れるほど予測分散 (3.59) の第 2 項が小さくなる．これより基底関数の外側の領域を補完する場合には，推定の信頼度が非常に高くなるという問題が生ずる．

今まで β が既知として議論を進めてきたが，$\mathbf{w}$，β が未知の場合，共役事前分布 $p(\mathbf{w}, \beta)$ はガウス－ガンマ分布で与えられ，その予測分布は t 分布となる (Exercise 3.12，3.13 を参照)．

Exercise 3.10

予測分布は目標変数の条件付き分布 (3.8) を事後分布 (3.49) について, $\mathbf{w}$ で積分することにより得られるので,

$$\begin{aligned} p(t|\mathbf{t}, \alpha, \beta) &= \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta)\, \mathrm{d}\mathbf{w} \\ &= \int \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)\, \mathrm{d}\mathbf{w} \end{aligned}$$

と表される．ここで式 (2.115) のガウス分布の周辺分布の公式を用いることで，次式を導くことができる．

$$p(t|\mathbf{t}, \alpha, \beta) = \mathcal{N}\left(t|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}), \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x})\right).$$

Exercise 3.11

付録 C の行列の公式を用いて得られる式 (3.110) を用いて，$\sigma^2_{N+1}(\mathbf{x}) \leq \sigma^2_N(\mathbf{x})$ を証明する．

$$\begin{aligned}
\sigma^2_{N+1}(\mathbf{x}) &= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_{N+1}\boldsymbol{\phi}(\mathbf{x}) \\
&= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{S}_N^{-1} + \beta\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\right)^{-1}\boldsymbol{\phi}(\mathbf{x}) \\
&= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{S}_N - \frac{\beta\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}\right)\boldsymbol{\phi}(\mathbf{x}) \\
&= \beta^{-1} + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}) - \beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\frac{\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N}{1+\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}\boldsymbol{\phi}(\mathbf{x}) \\
&= \sigma^2_N - \beta\frac{\left(\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)\left(\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)}{1+\beta\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})} \\
&\leq \sigma^2_N(\mathbf{x}).
\end{aligned}$$

Exercise 3.12

尤度関数 (3.10) を変形すると，

$$\begin{aligned}
p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) &= \prod_{n=1}^{N}\mathcal{N}\left(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}\right) \\
&\propto \beta^{N/2}\exp\left\{-\frac{\beta}{2}\left(\sum_{n=1}^{N}t_n^2 - 2\sum_{n=1}^{N}\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n + \sum_{n=1}^{N}\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\boldsymbol{\phi}(\mathbf{x})_n^{\mathrm{T}}\mathbf{w}\right)\right\} \\
&\propto \beta^{N/2}\exp\left(-\frac{\beta}{2}\sum_{n=1}^{N}\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x})_n^{\mathrm{T}}\mathbf{w}\right)\exp\left(\beta\sum_{n=1}^{N}\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)t_n - \frac{\beta}{2}\sum_{n=1}^{N}t_n^2\right) \\
&\propto \beta^{N/2}\exp\left(-\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w}\right)\exp\left(\beta\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} - \beta\left(\frac{1}{2}\sum_{n=1}^{N}t_N^2\right)\right)
\end{aligned}$$

と計算できる．上式は式 (2.153) と同じ形式をしているので，共役事前分布はガウス－ガンマ分布になる．共役事前分布を展開すると，

$$\begin{aligned}
p(\mathbf{w},\beta) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\beta^{-1},\mathbf{S}_0)\mathrm{Gam}(\beta|a_0,b_0) \\
&\propto \beta^{M/2}\exp\left(-\frac{\beta}{2}(\mathbf{w}-\mathbf{m}_0)^{\mathrm{T}}\mathbf{S}_0^{-1}(\mathbf{w}-\mathbf{m}_0)\right)\beta^{\alpha-1}\exp(-b_0\beta) \\
&\propto \beta^{M/2}\exp\left(-\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w}\right)\exp\left(\beta\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \frac{\beta}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{m}_0 - b_0\beta\right)\beta^{a_0-1} \\
&\propto \beta^{M/2+a_0-1}\exp\left(-\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{w}\right)\exp\left(\beta\mathbf{w}^{\mathrm{T}}\mathbf{S}_0^{-1}\mathbf{m}_0 - \beta\left(\frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{m}_0 + b_0\right)\right)
\end{aligned}$$

となる．尤度関数と事前分布の積により事後分布が以下のように求まる．

$$p(\mathbf{w},\beta|\mathbf{t}) \propto \beta^{(N+M)/2+a_0-1}\exp\left(-\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}+\mathbf{S}_0^{-1}\right)\mathbf{w}\right)\exp\left(\beta\left(\mathbf{w}^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}+\mathbf{S}_0^{-1}\mathbf{m}_0\right) - \beta\left(\frac{1}{2}\sum_{n=1}^{N}t_n^2\right) + \frac{1}{2}\mathbf{m}_0^{\mathrm{T}}\mathbf{m}_0 + b_0\right)\right)$$

上式と以下の式

$$\begin{aligned}
p(\mathbf{w},\beta|\mathbf{t}) &= \mathcal{N}(\mathbf{w}|\mathbf{m}_N,\beta^{-1}\mathbf{S}_N)\mathrm{Gam}(\beta|a_N,b_N) \\
&\propto \beta^{M/2+a_N-1}\exp\left(-\frac{\beta}{2}\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{w}\right)\exp\left(\beta\mathbf{w}^{\mathrm{T}}\mathbf{S}_N^{-1}\mathbf{m}_N - \beta\left(\frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N + b_N\right)\right)
\end{aligned}$$

を係数比較することにより，$\mathbf{m}_N$，$\mathbf{S}_N$，a_N，b_N は次のように表すことができる．

$$
\begin{aligned}
\mathbf{S}_N &= (\mathbf{S}_0^{-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1} \\
\mathbf{m}_N &= \mathbf{S}_N(\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t} + \mathbf{S}_0^{-1}\mathbf{m}_0) \\
a_N &= a_0 + \frac{1}{2} \\
b_N &= b_0 + \frac{1}{2}\left(\sum_{n=1}^{N} t_n^2 + \mathbf{m}_0^{\mathrm{T}}\mathbf{m}_0 - \mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\right).
\end{aligned}
$$

Exercise 3.13

前問で求めた事後分布を用いて予測分布 $p(t|\mathbf{x}, \mathbf{t})$ を計算し，予測分布が t 分布になることを以下に示す．

$$
\begin{aligned}
p(t|\mathbf{x}, \mathbf{t}) &= \iint p(t|\mathbf{w}, \beta)p(\mathbf{w}, \beta|\mathbf{t})\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \iint \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\mathrm{Gam}(\beta|a_N, b_N)\,\mathrm{d}\mathbf{w}\,\mathrm{d}\beta \\
&= \int \left(\int \mathcal{N}(t|\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \beta^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m}_N, \beta^{-1}\mathbf{S}_N)\,\mathrm{d}\mathbf{w}\right)\mathrm{Gam}(\beta|a_N, b_N)\,\mathrm{d}\beta \\
&= \int \mathcal{N}\left(t|\mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}(1 + \boldsymbol{\phi}(\mathbf{x})\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))\right)\mathrm{Gam}(\beta|a_N, b_N)\,\mathrm{d}\beta \\
&= \int \frac{b_N^{a_N}e^{-b_N\beta}\beta^{a_N-1}}{\Gamma(a_N)}\left(\frac{\beta(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^{-1}}{2\pi}\right)^{1/2}\exp\left\{-\frac{\beta(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))^{-1}}{2}\left(t - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})\right)^2\right\}\mathrm{d}\beta \\
&= \frac{b_N^{a_N}}{\Gamma(a_N)}\left(\frac{1}{2\pi(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right)^{1/2}\int \beta^{\,a_N+1/2-1}\exp\left\{-\beta\left(b_N + \frac{1}{2}\frac{(t - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))^2}{1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}\right)\right\}\mathrm{d}\beta \\
&= \frac{b_N^{a_N}}{\Gamma(a_N)}\left(\frac{1}{2\pi(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right)^{1/2}\Gamma\left(a_N + \frac{1}{2}\right)\left(b_N + \frac{1}{2}\frac{(t - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))^2}{1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})}\right)^{-a_N-1/2} \\
&= \frac{\Gamma(a_N + 1/2)}{\Gamma(a_N)}\left(\frac{1}{\pi \cdot 2b_N(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right)^{1/2}\left(1 + \frac{(t - \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}))^2}{2b_N(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}))}\right)^{-a_N-1/2} \\
&= \frac{\Gamma(\nu/2 + 1/2)}{\Gamma(\nu/2)}\left(\frac{\lambda}{\pi\nu}\right)^{1/2}\left(1 + \frac{\lambda(t - \mu)^2}{\nu}\right)^{-\nu/2-1/2}.
\end{aligned}
$$

ただし，

$$
\begin{aligned}
\mu &= \mathbf{m}_N^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}) \\
\lambda &= \frac{a_N}{b_N}\left(1 + \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x})\right)^{-1} \\
\nu &= 2a_n
\end{aligned}
$$

である．

3.3.3 Eqivalent kernel (等価カーネル)

式 (3.3) に事後分布の平均 (3.53) を代入すると，予測分布の平均は次の形をとる．

$$y(\mathbf{x}, \mathbf{m}_N) = \mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) = \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$
$$= \sum_{n=1}^{N} \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}_n) t_n. \tag{3.60}$$

よって点 $\mathbf{x}$ での予測分布の平均は，訓練データの目標変数 t_n の線形結合で与えられるので，

$$y(\mathbf{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n \tag{3.61}$$

と書くことができる．ここで以下の関数

$$k(\mathbf{x}, \mathbf{x}') = \beta \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}} \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}') \tag{3.62}$$

は**平滑化行列**，または**等価カーネル**と呼ばれ，訓練データの目標値の線形結合で予測を行う (3.61) のような関数を**線形平滑器**と呼ぶ．等価カーネルは $\mathbf{S}_N$ を含むので入力値 $\mathbf{x}_n$ に依存することに注意する．

図 3.10 ではガウス基底関数の場合の等価カーネルが示されており，カーネル関数 $k(\mathbf{x}, \mathbf{x}')$ は異なる $\mathbf{x}$ に対して $\mathbf{x}'$ の関数としてプロットされている．カーネルは x の周りに局在しており，x における予測分布の平均は目標値の重み付き和で表される．ここで重みは x に近いほど大きくなり，これは遠くの情報より近くの情報をより強く重み付けすることに対応している．この局所性は局所的なガウス基底関数だけに備わっているわけではなく，局所的でない多項式やシグモイドの基底関数に対しても成り立つ (図 3.11)．

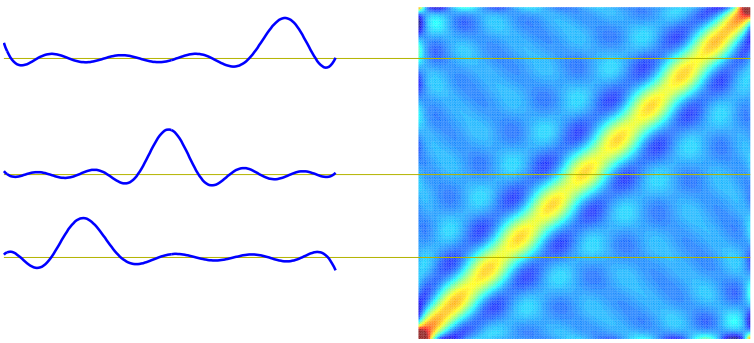

図 3.10: ガウス基底関数に対する等価カーネル $k(\mathbf{x}, \mathbf{x}')$

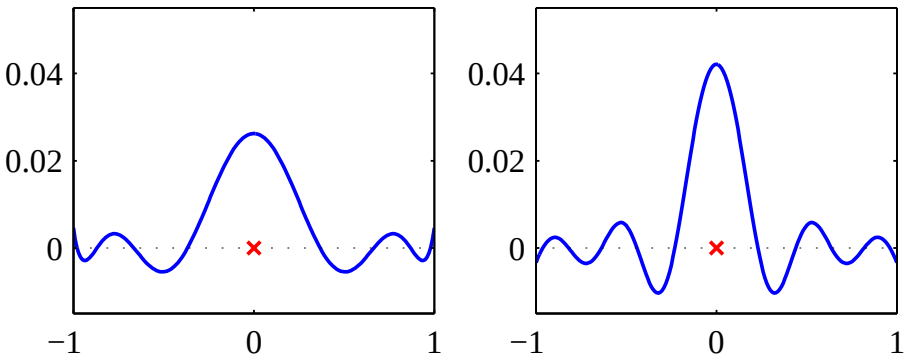

図 3.11: 等価カーネル $k(\mathbf{x}, \mathbf{x}')$ の例，左：多項式基底関数，右：シグモイド基底関数

$y(\mathbf{x})$ と $y(\mathbf{x}')$ の共分散を等価カーネルを用いて表すと

$$\begin{aligned}
\mathrm{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \mathrm{cov}[\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}, \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')] \\
&= \mathbb{E}\left[\left(\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w} - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w}]\right)\left(\boldsymbol{\phi}(\mathbf{x}'^{\mathrm{T}}\mathbf{w}) - \mathbb{E}[\boldsymbol{\phi}(\mathbf{x}')^{\mathrm{T}}\mathbf{w}]\right)^{\mathrm{T}}\right] \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbb{E}\left[(\mathbf{w} - \mathbb{E}[\mathbf{w}])(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}') - \mathbb{E}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')])\right] \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbb{E}\left[(\mathbf{w} - \mathbb{E}[\mathbf{w}])(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}') - \mathbb{E}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}')])\right] \\
&= \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\phi}(\mathbf{x}') = \beta^{-1}k(\mathbf{x}, \mathbf{x}')
\end{aligned} \tag{3.63}$$

となる．等価カーネルの形状より，近傍の予測平均同士の共分散の値は大きく (相関が大きい)，離れた予測平均の組では共分散が小さくなる (相関が小さい) ことがわかる．

基底関数の集合を固定すれば等価カーネルが暗に求まるので，線形回帰問題を扱う際，基底関数の集合の代わりに局所的なカーネル関数を直接定義することにより定式化することもできる．この定式化では観測された訓練集合が与えられたときに，カーネルを用いて新たな入力 $\mathbf{x}$ に対する予測値を求める．

等価カーネルは重みを定める役割を果たし，この重みに従って訓練集合の目標値の和をとり，新たな入力 $\mathbf{x}$ に対して予測を行う．これらの重みの和は，

$$\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1 \tag{3.64}$$

となり，すべての n について和をとると 1 となる (Exercise 3.14 を参照)．ただし，重みの和が 1 になるという制約条件を満たしていたとしても，カーネル関数自体は正の値でも負の値でもよいので，予測器は訓練集合の目標値の凸結合になるとは限らない．

また等価カーネル (3.62) は非線形関数 $\boldsymbol{\psi}(\mathbf{x})$ の内積で表されるという，通常のカーネル関数が満たすべき条件を満たす．

$$k(\mathbf{x}, \mathbf{z}) = \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{z}). \tag{3.65}$$

ただし，$\boldsymbol{\psi}(\mathbf{x}) = \beta^{1/2}\mathbf{S}_N^{1/2}\boldsymbol{\phi}(\mathbf{x})$ である．

Exercise 3.14

新しい基底関数として正規直交基底 $\boldsymbol{\psi}$ をとると，等価カーネルは $\alpha = 0$ のとき

$$\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= \beta\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\mathbf{S}_N\boldsymbol{\psi}(\mathbf{x}') \\
&= \beta\mathbf{x}^{\mathrm{T}}\left(0 \cdot \mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\psi}(\mathbf{x}') \\
&= \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\psi}(\mathbf{x}') \\
&= \frac{1}{N}\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}')
\end{aligned}$$

と求められる．n について等価カーネルの和をとると

$$\begin{aligned}
\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) &= \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_n) \\
&= \frac{1}{N}\left(\boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_1) + \cdots + \boldsymbol{\psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\psi}(\mathbf{x}_N)\right) \\
&= \frac{1}{N}(1 + \cdots + 1) = 1
\end{aligned}$$

となり，式 (3.116) が導かれた．

3.4 Bayesian Model Comparison (ベイズモデル比較)

1 章では，モデルを選択する際に生じる過学習の問題と，それを避けるために用いる交差確認が述べられているが，ここではベイズの立場からモデル選択の問題を取り扱う．

最尤推定による過学習は，モデルパラメータの値を点予測する代わりに周辺化することにより回避することができる．モデルは訓練データからのみで比較することができ，確認データは必要なくなる．得られるデータはすべて訓練用に用いることができ，また交差確認のように何度も訓練を行う必要もなくなる．訓練によってモデルパラメータの複数決めることも可能である (7 章)．

まず L 個のモデル $\{\mathcal{M}_i\}\,(i = 1, \ldots, L)$ を考える．ここでのモデルは観測されたデータ $\mathcal{D}$ 上の確率分布となっている．多項式フィッティングの問題では，分布は目標値 $\mathbf{t}$ の集合上に定義され，入力値の集合 $\mathbf{X}$ は既知である．データは仮定したモデル集合 $\mathcal{M}_i$ の中から生成されているが，そのどれかはわからない．モデルの不確かさは事前分布 $p(\mathcal{M}_i)$ によって表現され，訓練データ $\mathcal{D}$ が与えられると事後分布は

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i) \tag{3.66}$$

となる．事前分布 $p(\mathcal{M}_i)$ は各々のモデルに対する好み (preference) を表し，以下では簡単のため $p(\mathcal{M}_i)$ がすべての i について等しいとする．$p(\mathcal{D}|\mathcal{M}_i)$ は**モデルエビデンス**と呼ばれ，モデル $\mathcal{M}_i$ の下でのデータ $\mathcal{D}$ の起こりやすさを表している．モデルエビデンスは後でみるように，モデルの空間上で尤度関数をパラメータに関して積分したものであるので，周辺尤度とも呼ばれる．2 個のモデルに対するモデルエビデンスの比

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

は**ベイズ因子**として知られている．

事後分布が求まれば，確率の加法，乗法定理を用いて予測分布

$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D}) \tag{3.67}$$

が得られる．これは混合分布の一例となっており，全体の予測分布は個々のモデルの予測分布 $p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})$ を事後分布確率 $p(\mathcal{M}_i|\mathcal{D})$ に関して重み付け平均することにより求められる．モデル平均をとる代わりに，一番もっともらしいモデルを一つ選ぶ方法はモデル選択と呼ばれる．

パラメータ $\mathbf{w}$ を持つモデルに関しては，モデルエビデンスは以下の式で与えられる．

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}. \tag{3.68}$$

サンプリング (11 章) の観点では，周辺尤度は事前分布 $p(\mathbf{w}|\mathcal{M}_i)$ からパラメータをサンプリングし，データ $\mathcal{D}$ を生成する確率と見なすことができる．また周辺尤度は事後分布を計算する際に用いるベイズの定理

$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)} \tag{3.69}$$

の分母に表れる正規化項に等しい．

周辺尤度のパラメータに関する積分を近似することによって，モデルエビデンスの解釈をすることができる．はじめにモデルがパラメータを 1 つもつ場合を考える．事後分布はモデル $\mathcal{M}_i$ の依存性を省略すると $p(\mathcal{D}|w)p(w)$ に比例する．ここで事後分布が w_{MAP} で鋭くピークを持ち，その幅が $\Delta w_{\mathrm{posterior}}$ の場合を考えると，分布の積分はピークの高さと幅の積で近似できる (図 3.12)．さらに事前分布が平坦でその幅が $\Delta w_{\mathrm{prior}}$ であると仮定すれば $p(w) = 1/\Delta w_{\mathrm{prior}}$ となるので，

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\,\mathrm{d}w \simeq p(\mathcal{D}|w_{\mathrm{MAP}})\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \tag{3.70}$$

が成り立つ．両辺について対数をとれば以下の式が得られる．

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right). \tag{3.71}$$

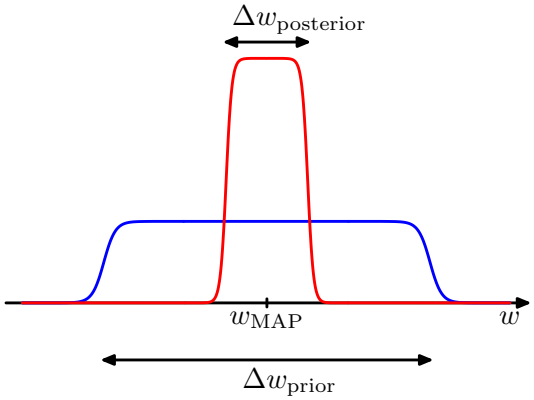

図 3.12: パラメータが w_{MAP} の近傍で尖っているときの事後分布の近似

第 1 項はパラメータが w_{MAP} のときのデータへのフィティング度を表し，第 2 項はモデル複雑さに対するペナルティ項を表す．$\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$ は常に負であり，モデルがデータに強くフィットするようにパラメータでは $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$ は小さくなり対数周辺尤度へのペナルティは大きくなる．

モデルがパラメータを M 個もつときには同様の近似をそれぞれのパラメータについて行い，またすべてのパラメータについて $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$ が等しいと仮定すると，

$$\begin{aligned}\ln p(\mathcal{D}) &= \int p(\mathcal{D}|\mathbf{w})p(w_1)\cdots p(w_M)\,\mathrm{d}w_1\cdots\mathrm{d}w_M \\ &\simeq \ln p(\mathcal{D}|\mathbf{w}_{\text{MAP}}) + M\ln\left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}\right) \qquad (3.72)\end{aligned}$$

が得られる．上式よりパラメータの数が増えるとそれとともにペナルティ項も線形に増加することがわかる．モデルを複雑にするとデータに対するあてはまりがよくなり通常第 1 項は増加するが，第 2 項は減少する．したがって周辺尤度最大化の観点からモデル選択を行うと，適切な複雑さを備えたモデルが選択されることになる．

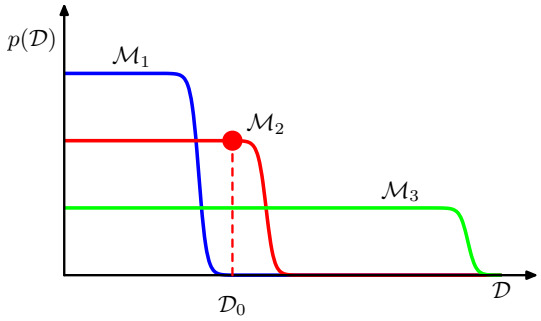

図 3.13: 複雑さの異なるモデルに対するデータの分布

図 3.13 を用いてなぜ周辺尤度の最大化により中間程度の複雑さをもつモデルが選択されるのかについて説明する．図の横軸は存在しうるデータセットの空間を 1 次元で示しており，軸上の 1 点が特定のデータセットに対応する．ここでモデル複雑さが単調増加の関係である 3 つのモデル $\mathcal{M}_1$，$\mathcal{M}_2$，$\mathcal{M}_3$ を考える．データ集合を得る際には，まずパラメータが事前分布 $p(\mathbf{w})$ から選ばれ，選ばれたパラメータを用いて $p(\mathcal{D}|\mathbf{w})$ からサンプリングを行う．

モデル $\mathcal{M}_1$ のような単純なモデル (例えば 1 次多項式) は多様性に乏しいデータしか生成できず，図 3.13 の横軸上の狭い領域に集中する．逆にモデル $\mathcal{M}_3$ のような複雑なモデル (例えば 9 次多項式) は非常に多様なデータを生成することができる．しかし，$p(\mathcal{D}|\mathcal{M}_i)$ は正規化されるので，あるデータに対する周辺尤度 $p(\mathcal{D})$ は小さくなってしまう．図 3.13 ではデータセット $\mathcal{D}_0$ に対して中間程度の複雑度を持つモデル $\mathcal{M}_2$ のエビデンスが最大となっている．単純すぎるモデルではデータにうまくフィットせず，複雑すぎるモデルでは予測分布があまりに広く分布するため個々のデータのどれかに割り当てられる確率が低くなる．

考えているモデル集合の中に真のモデルが含まれているという仮定の下で，ベイズモデル比較によりエビデンスを最大化することで平均的に正しいモデルが選ばれることを示す．モデル $\mathcal{M}_1$，$\mathcal{M}_2$ を考え $\mathcal{M}_1$ が正しいモデルであると仮定する．ベイズ因子を真のデータ生成の分布で平均すると、期待ベイズ因子が以下の形で表される．

$$\int p(\mathcal{D}|\mathcal{M}_1) \ln \frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} \mathrm{d}\mathcal{D}. \tag{3.73}$$

上式は KL-ダイバージェンスの例となっており，2 つの分布が一致するときに 0 となり，それ以外では正となる．このことより平均的には正しいモデルのベイズ因子の方が大きくなることがわかる．

ベイズモデル比較では過学習を避けることができ、訓練データのみを用いて比較を行うことができる．しかしながら他のアプローチと同様にモデルの形に関する仮定をおく必要があり，もしそれが間違っていたら誤った結果を導くことになる. モデルエビデンスは事前分布の特性に強く依存し，変則事前分布に対しては分布が正規化できないためエビデンスを定義できない．このとき変則でない事前分布の極限 (ガウス事前分布では分散無限大の極限) を考えることによりエビデンスを扱うことができるが．図 3.12，式 (3.70) よりエビデンスが 0 に収束してしまう．先に 2 つのモデルエビデンスの比を考え，その後極限を取れば意味のある値が得られるかもしれない.

したがって，実際の応用場面では独立なデータセットを分けておいて，そのデータを訓練終了後のシステムの評価に用いる方がよい.