ベイズモデル比較とエビデンス近似

PRML 輪読会 3 章後半

髙橋 優輝

2023.6.14

大阪大学大学院 情報科学研究科 情報数理学専攻 システム数理学講座 M1

目次

復習

核となる問題

PRML を通じて考える問題

- 入力
  - N 個の観測値 $\{\boldsymbol{x}_n\}(n = 1, \ldots, N)$（まとめて $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ と表すことも.）
  - 対応する目標値 $\{t_n\}(n = 1, \ldots, N)$（まとめて $\boldsymbol{t}$ と表すことも.）
- 出力
  - 与えられた入力変数 $\boldsymbol{x}$ から対応する目標変数 t を予測するような適当な関数 $y(\boldsymbol{x})$

線形回帰

y は次のように表されると仮定する.

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

ここで，w_i はモデルパラメータ，ϕ_i は基底関数である.

仮定（3 章全体）

目標変数 t が決定論的な関数 $y(\boldsymbol{x}, \boldsymbol{w})$ と加法性のガウスノイズの和で与えられる場合を考える．

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$

ただし，ϵ は期待値が 0 で精度（分散の逆数）が β のガウス確率変数である．

すなわち，次のように表すことができる．

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

β: ハイパーパラメータ

仮定（3.3.1 以降）

モデルパラメータ $\boldsymbol{w}$ の事前確率分布を簡単のため単一の精度パラメータ α で記述される期待値が 0 の等方的ガウス分布とする：

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

α：ハイパーパラメータ

3.4 ベイズモデル比較

復習：モデル選択 (1. 3 節)

多項式曲線フィッティング問題について考える．

最小二乗法で多項式曲線をあてはめた例において，最も良い汎化性能を持つ最適な次数の多項式があることを見た（1.1 節）（M は多項式の次数）：

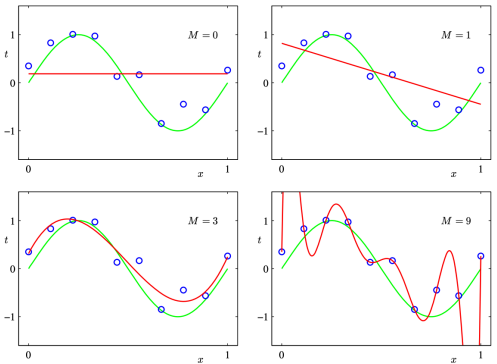
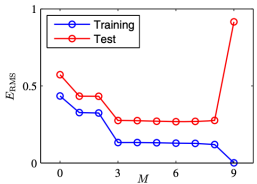


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

Figure 1.5 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.



モデルの良さをどう測るか？

復習：モデル選択の手法（1.3 節）

モデルの汎化性能をどう測るか？

- train, validation, test set にデータを分割する　⟶ データがもったいない
- cross-validation　⟶ 計算量：大

理想的には，1 回の訓練だけで複数のハイパーパラメータとモデルのタイプを比較したい．そこで，訓練データだけに依存し，過学習によるバイアスを持たない性能の尺度を見つけることが必要である．

このような文脈で，さまざまな情報量基準が提案されてきた：

- AIC
- BIC（4.4.1 節）

これらは，より複雑なモデルによる過学習を避けるための罰金項を足すことによって，最尤推定のバイアスを修正することを試みている．

3.4 節では，複雑さに罰金を課すのに自然で理にかなった方法として，ベイズ的なアプローチを採用する．

ベイズの立場からのモデル比較

ベイズの立場からのモデル比較では，モデルの選択に関する不確かさを表すために確率を用いる．

以下では，L 個のモデル $M_i (i = 1, \ldots, L)$ を比較する場合について考える．ここでは，データはこれらのモデルのどれかに従って生成されているが，そのどれかは分からないという問題設定を考える．

モデルの不確かさは事前確率分布 $p(\mathcal{M}_i)$ を通して表現する．そして訓練集合 $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{t})$ が与えられたとき，モデルの事後分布 $p(\mathcal{M}_i|\mathcal{D})$ を評価する．

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

ここで，$p(\mathcal{D}|\mathcal{M}_i)$ はモデルエビデンスと呼ばれ，データから見たモデルの好みを表す．また，簡単のため，すべてのモデルの事前確率が等しい場合を考える．

2 つのモデルエビデンスの比 $p(\mathcal{D}|\mathcal{M}_i)/p(\mathcal{D}|\mathcal{M}_j)$ はベイズ因子と呼ばれる．

予測分布

一旦モデルの事後分布が分かれば，予測分布は次式で与えられる．

$$
\begin{aligned}
p(t|\boldsymbol{x}, \mathcal{D}) &= \sum_{i=1}^{L} p(t, \mathcal{M}_i|\boldsymbol{x}, \mathcal{D}) \qquad (\because \text{加法定理}) \\
&= \sum_{i=1}^{L} p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i|\boldsymbol{x}, \mathcal{D}) \qquad (\because \text{乗法定理}) \\
&= \sum_{i=1}^{L} p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i|\mathcal{D}) \qquad (3.67)
\end{aligned}
$$

この予測分布は混合分布の一種である．混合分布では，全体の予測分布が個々のモデルの予測分布 $p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D})$ の事後確率 $p(\mathcal{M}_i|\mathcal{D})$ に関する重み付き平均で得られる．

モデル選択

一番尤もらしいモデルを 1 つ選ぶことを考える．これはモデル選択と呼ばれる．パラメータ $\boldsymbol{\theta}$ を持つモデルに対して，モデルエビデンスは次式で与えられる．

$$
\begin{aligned}
p(\mathcal{D}|\mathcal{M}_i) &= \int p(\mathcal{D}, \boldsymbol{\theta}|\mathcal{M}_i) d\boldsymbol{\theta} \qquad (\because \text{加法定理}) \\
&= \int p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M}_i) p(\boldsymbol{\theta}|\mathcal{M}_i) d\boldsymbol{\theta} \qquad (\because \text{乗法定理}) \qquad (3.68)
\end{aligned}
$$

標本化の観点から，モデルエビデンスは，パラメータがその事前分布からランダムにサンプリングされたモデルからデータ集合 $\mathcal{D}$ が生成される確率とみなすことができる．(11 章)

また，パラメータに関する積分を単純近似することにより，モデルエビデンスの別の解釈を得ることができる．

まず，パラメータが一つ（それを θ とする）しかないモデルについて考える．

以降，表記を簡単に保つため，モデル $\mathcal{M}_i$ への依存性を省略する．

モデルエビデンスの解釈

ここで，事後分布 $p(\mathcal{D}|\theta)$ が最頻値 θ_{MAP} の近傍で鋭く尖っているとき，その幅を $\Delta\theta_{posterior}$ で表せば，全体の積分は幅 $\Delta\theta_{posterior}$ と最大値の積で近似できる．

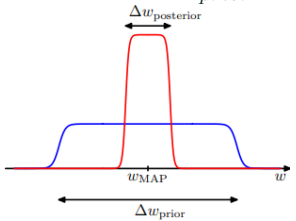
さらに，事前確率 $p(\theta)$ が平坦で幅 $\Delta\theta_{prior}$，つまり $p(\theta) = 1/\Delta\theta_{prior}$ のとき，次式が成り立つ．

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta \approx p(\mathcal{D}|\theta_{MAP})\frac{\Delta\theta_{posterior}}{\Delta\theta_{prior}} \tag{3.70}$$

対数を取れば式 (3.70) は次のようになる．

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\theta_{MAP}) + \ln\frac{\Delta\theta_{posterior}}{\Delta\theta_{prior}} \tag{3.71}$$

Figure 3.12 We can obtain a rough approximation to the model evidence if we assume that the posterior distribution over parameters is sharply peaked around its mode w_{MAP}.



w を θ に読み替える．

モデルエビデンスの解釈

ここでは，式 (3.71) の解釈について考える．

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\theta_{MAP}) + \ln \frac{\Delta\theta_{posterior}}{\Delta\theta_{prior}} \tag{3.71}$$

第一項は最も尤もらしいパラメータ値 θ によるデータへのフィッティング度に対応し，第二項はモデルの複雑さに基づいてペナルティを与えることに対応している．仮定より $\Delta\theta_{posterior} < \Delta\theta_{prior}$ なので，第二項は負であり，比 $\Delta\theta_{posterior}/\Delta\theta_{prior}$ が小さくなるにつれて，この項の絶対値は大きくなる．すなわち，モデルがデータに強くフィットするようにパラメータをうまく調整すれば，ペナルティは大きくなる．

モデルエビデンスの解釈

モデルが M 個のパラメータを含むとき，それぞれのパラメータに対して，順々に同様の近似を行うことができる．すべてのパラメータが同じ比 $\Delta\theta_{posterior}/\Delta\theta_{prior}$ を持つとき，次式が得られる．

$$\ln p(\mathcal{D}) \approx \ln p(\mathcal{D}|\boldsymbol{\theta}_{MAP}) + M \ln \frac{\Delta\theta_{posterior}}{\Delta\theta_{prior}} \tag{3.72}$$

モデルの複雑さを増したとき，モデルはデータにフィットしやすくなるため，第一項は通常増加するが，M により第二項は減少する．すなわち，エビデンスを最大にする最適なモデルは，これらの項をバランスよく小さくする．後に，ガウス近似に基づく，より洗練された事後分布の近似法を与える．（BIC,4.4.1 節）

モデルエビデンスの解釈

図を用いて，モデルエビデンスの最大化により，中間程度の複雑さのモデルが選ばれる理由を説明する．複雑さが単調増加の関係にある 3 つのモデルを $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ として，これらのモデルによって，どのようなデータ集合が生成されるかを考える．

まず，事前分布 $p(\boldsymbol{\theta})$ に従ってパラメータの値を選択し，これらのパラメータ値に対してデータを $p(\mathcal{D}|\boldsymbol{\theta})$ からサンプリングする．（目標変数にランダムノイズが加わることに注意する．）単純なモデル $\mathcal{M}_1$ は自由度が少ないため，こうして生成されるデータ集合は多様性に乏しく，その分布 $p(\mathcal{D})$ は横軸の比較的狭い範囲に集中する．$\mathcal{M}_2, \mathcal{M}_3$ についても同様に考えると，Figure 3.13 が得られる．

分布 $p(\mathcal{D})$ は正規化されるため，あるデータ集合 $\mathcal{D}_0$ に対しては，中間の複雑さを持つモデルのエビデンスが最大になることがある．

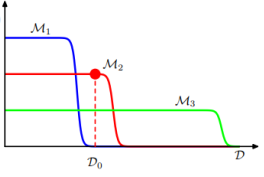


Figure 3.13 Schematic illustration of the distribution of data sets for three models of different complexity, in which $\mathcal{M}_1$ is the simplest and $\mathcal{M}_3$ is the most complex. Note that the distributions are normalized. In this example, for the particular observed data set $\mathcal{D}_0$, the model $\mathcal{M}_2$ with intermediate complexity has the largest evidence.

例）$\mathcal{M}_1$:1 次の多項式，$\mathcal{M}_2$:5 次の多項式，$\mathcal{M}_3$:9 次の多項式

ベイズ因子と KL ダイバージェンス

ベイズモデル比較の枠組みの中では，考えているモデルの集合の中にデータが生成される真の分布が含まれていると暗に仮定している．この仮定が正しければ，ベイズモデル比較によって平均的に正しいモデルが選ばれることを確認できる．

2 つのモデル $\mathcal{M}_1, \mathcal{M}_2$ について考える．ここで，$\mathcal{M}_1$ が正しいモデルと仮定する．このとき，ベイズ因子（の対数？）の真のデータ生成の分布の上での期待値を考えると，次の期待ベイズ因子が得られる．

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

これは KL ダイバージェンスの一例となっている．ここで，KL ダイバージェンスの性質より，式 (3.73) は $p(\mathcal{D}|\mathcal{M}_1) = p(\mathcal{D}|\mathcal{M}_2)$ のとき 0 となり，それ以外のとき正となる．

3.4 節のまとめと課題

3.4 節の議論で，ベイズの枠組みでは過学習の問題を回避できると共に訓練データだけに基づいてモデル比較を行えることが分かった．しかしながら，ベイズ的なアプローチではモデルの形に関する仮定を置く必要があり，これが正しくない場合は誤った結果を導くことがある．特に，モデルエビデンスは $\boldsymbol{\theta}$ に関する事前分布の様々な特性に強く依存する．

そのため，変則事前分布 (p.115) を考えたくなるが，正規化できないような分布に対してはモデルエビデンスを定義できない．この問題を避けるためには，例えば変則でない通常の事前分布を考え，適当な極限をとればよい（例えば，ガウス事前分布における分散の無限大の極限）．しかし，このとき，モデルエビデンスは 0 に収束する．その一方で，2 つのモデルのベイズ因子を先に考え，その後極限を取ることにより意味のある値を得ることができる場合もある．

したがって，実際の応用場面ではテスト用の独立なデータ集合をとっておき，それを用いて最終的なシステムの全体性能を評価するのが賢明だろう．（結局！？）

3.5 エビデンス近似

導入（3.5 エビデンス近似）

3.5 節では，3.4 節で一般的に話していたベイズモデル比較の理論を線形回帰の正則化パラメータの決定に適用する方法について述べる．

線形基底関数モデルを完全にベイズ的に取り扱うために，ハイパーパラメータ α, β に対しても事前分布を導入し，通常のパラメータ $\boldsymbol{w}$ だけでなく，α, β に関する予測分布について考える．しかしながら，$\boldsymbol{w}, \alpha, \beta$ について同時に考えるのは難しい（解析的に周辺化することが困難）ので，$\boldsymbol{w}$ だけに関して積分して得られた周辺尤度関数を最大にするように α, β の値を決めるという 2 段階の近似法について議論する．この枠組みは統計学の文献では，経験ベイズ，第二種の最尤推定，一般化最尤推定と呼ばれ，機械学習の文献ではエビデンス近似と呼ばれる．

α, β ってなんだったけ？（復習）

$\underline{\alpha}$(p.152)

モデルパラメータ $\boldsymbol{w}$ の事前確率分布を簡単のため単一の精度パラメータ α で記述される期待値が 0 の等方的ガウス分布とする：

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

$\underline{\beta}$(p.138)

目標変数 t が決定論的な関数 $y(\boldsymbol{x}, \boldsymbol{w})$ と加法性のガウスノイズの和で与えられる場合を考える．

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$$

ただし，ϵ は期待値が 0 で精度（分散の逆数）が β のガウス確率変数である．

すなわち，次のように表すことができる．

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

α, β に事前分布を導入したときの予測分布

ハイパーパラメータ α, β の事前分布と導入すれば，予測分布は同時分布を $\boldsymbol{w}, \alpha, \beta$ に関して周辺化することにより得られる．

$$
\begin{aligned}
p(t|\boldsymbol{t}) &= \iiint p(t, \boldsymbol{w}, \alpha, \beta|\boldsymbol{t}) d\boldsymbol{w} d\alpha d\beta (\because \text{加法定理}) \\
&= \iiint p(t|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta) p(\alpha, \beta|\boldsymbol{t}) d\boldsymbol{w} d\alpha d\beta \qquad (3.74)
\end{aligned}
$$

ここで，次式を用いた．

$$
\begin{aligned}
p(t, \boldsymbol{w}, \alpha, \beta|\boldsymbol{t}) &= p(t|\boldsymbol{t}, \boldsymbol{w}, \alpha, \beta) p(\boldsymbol{w}, \alpha, \beta|\boldsymbol{t}) \qquad (\because \text{乗法定理}) \\
&= p(t|\boldsymbol{t}, \boldsymbol{w}, \alpha, \beta) p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta) p(\alpha, \beta|\boldsymbol{t}) \qquad (\because \text{乗法定理}) \\
&= p(t|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta) p(\alpha, \beta|\boldsymbol{t}) \qquad (\because \text{依存しない変数を削除})
\end{aligned}
$$

3.3.2 節から引き続き，$\boldsymbol{X}$ を条件の部分から省いている．さらに，ここでは，表記を簡単に保つため，$\boldsymbol{x}$ も省いている．

α, β に事前分布を導入したときの予測分布

$$p(t|\boldsymbol{t}) = \iiint p(t|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\boldsymbol{t}, \alpha, \beta) p(\alpha, \beta|\boldsymbol{t}) d\boldsymbol{w} d\alpha d\beta \qquad (3.74, \text{再掲})$$

式 (3.74) について，事後分布 $p(\alpha, \beta|\boldsymbol{t})$ が $\alpha = \hat{\alpha}, \beta = \hat{\beta}$ の周りで鋭く尖っているとき，次のように変形できる．

$$p(t|\boldsymbol{t}) \approx \int p(t|\boldsymbol{w}, \hat{\beta}) p(\boldsymbol{w}|\boldsymbol{t}, \hat{\alpha}, \hat{\beta}) d\boldsymbol{w} \qquad (3.75)$$

$\hat{\alpha}, \hat{\beta}$ の求め方

ベイズの定理より，α, β の事後確率は，次のように与えられる．

$$p(\alpha, \beta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\alpha, \beta)p(\alpha, \beta) \tag{3.76}$$

事前分布 $p(\alpha, \beta)$ が平坦なとき，$\hat{\alpha}, \hat{\beta}$ は周辺尤度関数 $p(\boldsymbol{t}|\alpha, \beta)$ を最大化することで得られる．

以降，線形基底関数モデルに対して周辺尤度関数を計算してから，それを最大化することにする．これにより，クロスバリデーションを行うことなく，訓練データだけから，ハイパーパラメータの値を決定可能である．

比 α/β が式 (3.27) の正則化パラメータ λ と同様の働きをする (p.152, 式 (3.55)) ことに注意する．

$$E_W(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{M}\{t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 + \frac{\lambda}{2}\boldsymbol{w}^\top \boldsymbol{w} \tag{3.27}$$

3.5 エビデンス近似

3.5.1 エビデンス関数の評価

周辺尤度関数

周辺尤度関数 $p(\boldsymbol{t}|\alpha, \beta)$ は次の積分を計算することで得られる．

$$p(\boldsymbol{t}|\alpha, \beta) = \int p(\boldsymbol{t}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha)d\boldsymbol{w} \tag{3.77}$$

この積分は，直接評価することができるが（演習 3.16），ここではその代わりに指数関数の中身を直接平方完成し，ガウス関数の正規化係数の一般形を用いることによって積分を評価することにする．

周辺尤度関数の変形

$$\ln p(\boldsymbol{t}|\boldsymbol{w},\beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\boldsymbol{w}) \quad \text{(3.11，前回)}$$

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \quad \text{(3.12，前回)}$$

$$E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w} \quad \text{(3.25，前回)}$$

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}\boldsymbol{I}) \quad \text{(3.52，再掲)}$$

式 (3.11, 3.12, 3.52) を用いると，式 (3.77) は次のように変形できる．（演習 3.17）

$$\begin{aligned} p(\boldsymbol{t}|\alpha,\beta) &= \int p(\boldsymbol{t}|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha) \\ &= \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int e^{-E(\boldsymbol{w})}d\boldsymbol{w} \end{aligned} \quad \text{(3.78)}$$

ここで，　$E(\boldsymbol{w}) = \beta E_D(\boldsymbol{w}) + \alpha E_W(\boldsymbol{W}) = \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 + \frac{\alpha}{2}\boldsymbol{w}^\top\boldsymbol{w}$ (3.79)

周辺尤度関数の変形

$$E(\boldsymbol{w}) = \beta E_D(\boldsymbol{w}) + \alpha E_W(\boldsymbol{W}) = \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 + \frac{\alpha}{2}\boldsymbol{w}^\top \boldsymbol{w} \qquad (3.79，再掲)$$

ここで，式 (3.79) は，3.1.4 節で考えた正則化二乗和誤差関数[1] と
($\lambda = \alpha/\beta$とすれば) 定数倍の範囲で等しいことが分かる．

式 (3.79) を $\boldsymbol{w}$ に関して平方完成すれば，次式が得られる．(演習 3.18)

$$E(\boldsymbol{w}) = E(\boldsymbol{m}_N) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{m}_N)^\top \boldsymbol{A}(\boldsymbol{w} - \boldsymbol{m}_N) \qquad (3.80)$$

ただし，

$$\boldsymbol{A} \coloneqq \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \qquad (3.81)$$

$$E(\boldsymbol{m}_N) \coloneqq \frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + \frac{\alpha}{2}\boldsymbol{m}_N^\top \boldsymbol{m}_N \qquad (3.82)$$

$$\boldsymbol{m}_N \coloneqq \beta \boldsymbol{A}^{-1} \boldsymbol{\Phi}^\top \boldsymbol{t} \qquad (3.84)$$

[1] $E_W(\boldsymbol{w}) = 1/2 \sum_{n=1}^{M} \{t_n - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 + \lambda/2 \boldsymbol{w}^\top \boldsymbol{w}$

周辺尤度関数の変形

$$\boldsymbol{S}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \qquad \text{(3.54，前回)}$$

式 (3.54) と式 (3.81) を比較すると，$\boldsymbol{A} = \boldsymbol{S}_N^{-1}$ が成り立つことが分かり，このことから，式 (3.84) は事後分布 $p(\boldsymbol{w}|\boldsymbol{t})$ の平均であることが分かる．

上記の結果から，次式が成り立つ．（演習 3.19）

$$\int e^{-E(\boldsymbol{w})} d\boldsymbol{w} = e^{-E(\boldsymbol{m}_N)} (2\pi)^{M/2} |\boldsymbol{A}|^{-1/2} \qquad (3.85)$$

よって，周辺尤度関数の対数は，次のようになる．

$$\begin{aligned} \ln p(\boldsymbol{t}|\alpha, \beta) &= \ln\left(\left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int e^{-E(\boldsymbol{w})} d\boldsymbol{w} \right) \\ &= \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\boldsymbol{m}_N) - \frac{1}{2} \ln |\boldsymbol{A}| - \frac{N}{2} \ln(2\pi) \qquad (3.86) \end{aligned}$$

多項式回帰問題と周辺尤度関数

基底関数を $\phi_j(x) = x^j$ として，1.1 節で考えた多項式回帰問題について考える．

- 右図より，$M = 1$ から $M = 2$ に変化させたとき，残差誤差はほとんど変化していない．これは，真の関数の多項式展開において偶数次数の項を持たないことによる．
- 一方，左図より，$M = 1$ から $M = 2$ に変化させたとき，周辺尤度はモデルが複雑になったことにより，小さくなる（悪くなる）．
- 右図の汎化誤差は $M = 3$ から $M = 8$ でほとんど変化しておらず，モデルを選ぶのが難しいが，左図の周辺尤度では $M = 3$ で最大値を取っている事がわかる．これは，観測されたデータを説明できるモデルの中で $M = 3$ が最も単純なモデルだからである．

Figure 3.14 Plot of the model evidence versus the order M, for the polynomial regression model, showing that the evidence favours the model with $M = 3$.

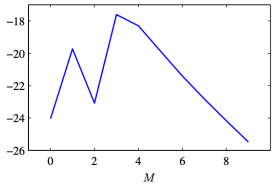
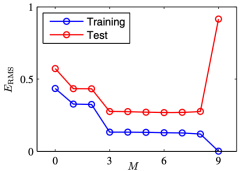


Figure 1.5 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.



$\alpha = 5 \times 10^{-3}$ に固定

$\sin x = \sum_{n=1}^{\infty} (-1)^n \frac{1}{(2n+1)!} x^{2n+1}$

$\ln p(\boldsymbol{t}|\alpha, \beta) = \frac{M}{2} \ln \alpha + \frac{N}{2} \ln \beta - E(\boldsymbol{m}_N) - \frac{1}{2} \ln |\boldsymbol{A}| - \frac{N}{2} \ln(2\pi)$

3.5 エビデンス近似

3.5.2 周辺尤度関数の最大化

周辺尤度関数（再掲）

周辺尤度関数の対数：

$$\begin{aligned}\ln p(\boldsymbol{t}|\alpha,\beta) &= \ln\left(\left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int e^{-E(\boldsymbol{w})}d\boldsymbol{w}\right)\\ &= \frac{M}{2}\ln\alpha+\frac{N}{2}\ln\beta-E(\boldsymbol{m}_N)-\frac{1}{2}\ln|\boldsymbol{A}|-\frac{N}{2}\ln(2\pi)\end{aligned} \tag{3.86}$$

ただし，

$$\boldsymbol{A} \coloneqq \alpha\boldsymbol{I}+\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi} \tag{3.81}$$

$$E(\boldsymbol{m}_N) \coloneqq \frac{\beta}{2}\|\boldsymbol{t}-\boldsymbol{\Phi}\boldsymbol{m}_N\|^2+\frac{\alpha}{2}\boldsymbol{m}_N^\top\boldsymbol{m}_N \tag{3.82}$$

$$\boldsymbol{m}_N \coloneqq \beta\boldsymbol{A}^{-1}\boldsymbol{\Phi}^\top\boldsymbol{t} \tag{3.84}$$

$\boldsymbol{m}_N$: $\boldsymbol{w}$ の事後分布のモード

周辺尤度関数の α に関する最大化

3.5.2 節では，周辺尤度関数を α, β に関して最大化することを考える．まずは，α について最大化する問題について考える．

$(\lambda_i, \boldsymbol{u}_i)$ を行列 $(\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})$ の固有対とする．すなわち，次式が成り立つとする．

$$(\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}) \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \qquad (i = 1, \ldots, M) \tag{3.87}$$

このとき，$\boldsymbol{A} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ より，$\boldsymbol{A}$ は固有値 $\alpha + \lambda_i$ を持つ．よって，周辺尤度関数の $\ln |\boldsymbol{A}|$ の項の α に関する導関数は次式で与えられる．

$$\frac{d}{d\alpha} \ln |A| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha} \tag{3.88}$$

周辺尤度関数の α に関する最大化

よって，周辺尤度関数の対数の α に関する停留点は以下の等式を満たす．

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\boldsymbol{m}_N^\top \boldsymbol{m}_N - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha} \tag{3.89}$$

この式を整理すると，周辺尤度を最大にする α の値は次式で満たすことが分かる．(演習 3.20)

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^\top \boldsymbol{m}_N} \tag{3.92}$$

ここで，
$$\gamma \coloneqq \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \tag{3.91}$$

ここで，γ は α に依存しており（γ の解釈を 3.5.3 節で与える．），また，$\boldsymbol{w}$ の事後分布のモード $\boldsymbol{m}_N$ も α に依存することから，式 (3.92) は α に関して陽に表せているわけではないことに注意する．

そこで，次ページで示す反復法で α の最大化を行う．

周辺尤度関数の α に関する最大化

Step 0 $\alpha \leftarrow$ (適当な値)，$\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ の固有値を計算;

Step 1 $\boldsymbol{m}_N \leftarrow \beta(\alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \boldsymbol{t}$;(式 (3.84))

$\gamma \leftarrow \sum_i \dfrac{\lambda_i}{\lambda_i + \alpha}$;(式 (3.91))

Step 2 $\alpha \leftarrow \dfrac{\gamma}{\boldsymbol{m}_N^\top \boldsymbol{m}_N}$;(式 (3.92))

Step 3 **if** 停止条件を満たす **then**

Step 1 に戻る;

else

このときの α を出力;

$\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ の固有値 λ_i について，この反復法の中で行列 $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ は不変なので，行列 $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ の固有値を前計算し，それらを β 倍すれば得られる．

上記の手順では，α の値を訓練データのみから決定できている．

周辺尤度関数の β に関する最大化

同様に，β に関しても（対数）周辺尤度関数の最大化問題について考える．

λ_i の定義より，固有値 λ_i は β に比例する．このことから，$d\lambda_i/d\beta = \lambda_i/\beta$ が成り立つので，次式が得られる．

$$\frac{d}{d\beta}\ln|\boldsymbol{A}| = \frac{d}{d\beta}\sum_i \ln(\lambda_i+\alpha) = \frac{1}{\beta}\frac{\lambda_i}{\lambda_i+\alpha} = \frac{\gamma}{\beta} \tag{3.93}$$

したがって，周辺尤度の停留点は，次式を満たす．（演習 3.22）

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_n \{t_n - \boldsymbol{m}_N^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 - \frac{\gamma}{2\beta} \tag{3.94}$$

$$\Longleftrightarrow \frac{1}{\beta} = \frac{1}{N-\gamma}\sum_n \{t_n - \boldsymbol{m}_N^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.95}$$

やはりこれも β に関して陽に表せていないので，α のときと同様の反復法で周辺尤度を最大化する β を得る．α, β を両方とも推定するのであれば，Step 2 で両方とも更新すれば良い．

3.5 エビデンス近似

3.5.3 有効パラメータ数

α, γ の解釈

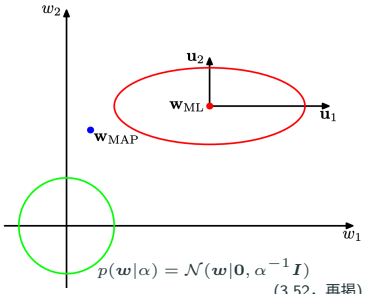
$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^\top \boldsymbol{m}_N} \qquad (3.92, \text{再掲})$$

より，ベイズ解 α の解釈を得ることができる．Figure 3.15 は $\boldsymbol{w}$ の尤度関数と事前分布の等高線，最尤解 $\boldsymbol{w}_{ML}$，MAP 解 $\boldsymbol{w}_{MAP}$，$\beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ の固有値 $\boldsymbol{u}_1, \boldsymbol{u_2}$ を表している．ただし，パラメータ空間の各軸は固有ベクトル $\boldsymbol{u}_i$ と重なるように回転してある．

$$\boldsymbol{w}_{ML} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{t} \qquad (3.15, \text{前回})$$

$$\boldsymbol{w}_{MAP} = \boldsymbol{m}_N \qquad (3.49, \text{前回})$$

Figure 3.15 Contours of the likelihood function (red) and the prior (green) in which the axes in parameter space have been rotated to align with the eigenvectors $\mathbf{u}_i$ of the Hessian. For $\alpha = 0$, the mode of the posterior is given by the maximum likelihood solution $\mathbf{w}_{\mathrm{ML}}$, whereas for nonzero α the mode is at $\mathbf{w}_{\mathrm{MAP}} = \mathbf{m}_N$. In the direction w_1 the eigenvalue λ_1, defined by (3.87), is small compared with α and so the quantity $\lambda_1/(\lambda_1 + \alpha)$ is close to zero, and the corresponding MAP value of w_1 is also close to zero. By contrast, in the direction w_2 the eigenvalue λ_2 is large compared with α and so the quantity $\lambda_2/(\lambda_2+\alpha)$ is close to unity, and the MAP value of w_2 is close to its maximum likelihood value.



$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha)$$

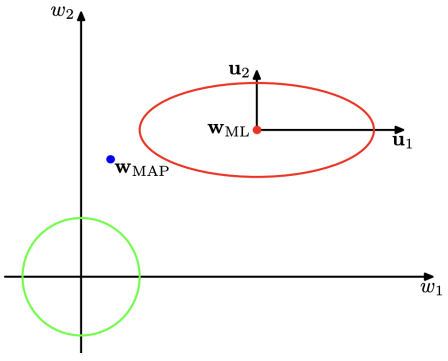
$$p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \qquad (3.10, \text{前回})$$

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \qquad (3.52, \text{再掲})$$

$$-\frac{\beta}{2}\sum_{n=1}^{N}(t_n - \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_n)) = -\frac{\beta}{2}\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 = -E_D(\boldsymbol{w}) \quad (\text{尤度関数の対数})$$

α, γ の解釈

尤度関数の等高線は，軸に沿った楕円となる．固有値 λ_i は尤度関数の曲率を表すため，この図では，$\lambda_1 < \lambda_2$ である．（曲率が小さいとき尤度関数の等高線は長く伸びるので．）



α, γ の解釈

$\alpha = 0$ のとき，事後分布について，

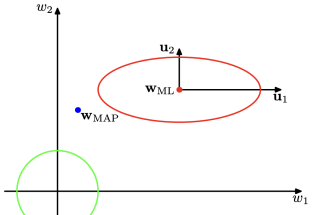
$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \beta)$$

となるので，$\boldsymbol{w}_{MAP} = \boldsymbol{w}_{ML}$ となる．

$\alpha \neq 0$ のとき，$w_{MAPi} = \gamma_i w_{MLi}$（ただし，$\gamma_i = \frac{\lambda_i}{\alpha + \lambda_i}$）である（要証明）．さらに，$\beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi} > 0$ であるので，$\lambda_i > 0, 0 < \gamma_i < 1$ が成り立つ．

よって，α より十分小さい λ_i に対応する w_{MAPi} は 0 に近い値をとり（図の w_1），α より十分大きい λ_i に対応する w_{MAPi} は最尤推定値に近い値をとる（図の w_2）．

このような理由から，$\gamma = \sum_i \frac{\lambda_i}{\alpha + \lambda_i} (= \boldsymbol{\gamma}^\top \mathbf{1})$ は有効なパラメータの数を表しており，そのようなパラメータは well-determined パラメータと呼ばれる $(0 \leq \gamma \leq M)$．



β の再推定の式 (3.95) の解釈

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n - \boldsymbol{m}_N^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \qquad \text{(3.95，再掲)}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}_{ML}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \qquad \text{(3.21，前回)}$$

上記の 2 つの式を比較すると，どちらも目標値とモデルによる推定値との差の二乗の平均の形で分散（精度の逆数）を表している．しかしながら，分母の数が最尤推定ではデータ点数 N であるのに対して，ベイズ推定では $N-\gamma$ である．

β の再推定の式 (3.95) の解釈

1 変数 x のガウス分布の分散の最尤推定値と，ベイズ推定値[2] はそれぞれ次のように与えられる．

$$\sigma_{ML}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{3.96}$$

$$\sigma_{MAP}^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{3.97}$$

ベイズの結果の分母の $N-1$ は自由度の 1 つを平均のフィッティングと最尤推定のバイアスを取り除くのに用いていることを考慮している．

[2] これは今後導出される？

β の再推定の式 (3.95) の解釈

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n - \boldsymbol{m}_N^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \qquad \text{(3.95，再掲)}$$

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}_{ML}^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \qquad \text{(3.21，前回)}$$

線形回帰モデルにおいては，目標分布の平均は M 個のパラメータを含む関数 $\boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$ で与えられる．しかしながら，すべてのパラメータがデータに調整されているわけでなく，データによって決まる有効パラメータの数は γ であり，残りの $M-\gamma$ 個のパラメータは事前分布によって，小さい値に設定される．このことは，分散のベイズ推定の結果 (式 (3.97)) に反映されており，それにより最尤推定の結果のバイアスを補正している．

β の再推定の式 (3.95) の解釈

1.1 節で用いた三角関数の例を 9 個の基底関数からなるガウス基底関数モデルによって近似する問題を考える．このモデルには，バイアス項を含めて $M = 10$ 個のパラメータがある．この例を用いて，エビデンスの枠組みによるハイパーパラメータの値の決定法を説明する．ここでは，単純のため β を真の値 11.1 に設定した．

$$\gamma = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha} \qquad (3.91,\ 再掲)$$

$$E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} \qquad (3.25,\ 再掲)$$

$$\arg\max_{\alpha} p(\boldsymbol{t}|\alpha, \beta) = \frac{\gamma}{\boldsymbol{m}_N^\top \boldsymbol{m}_N} \qquad (3.92,\ 再掲)$$

$$\ln p(\boldsymbol{t}|\alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\boldsymbol{m}_N) - \frac{1}{2}\ln|\boldsymbol{A}| - \frac{N}{2}\ln(2\pi) \qquad (3.86,\ 再掲)$$

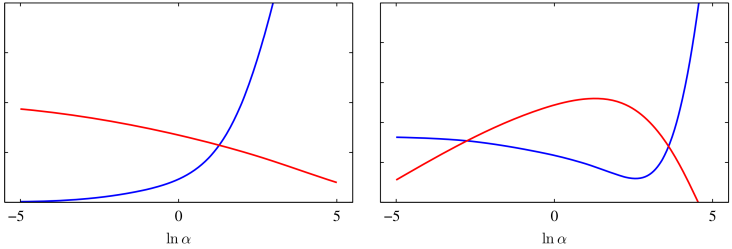
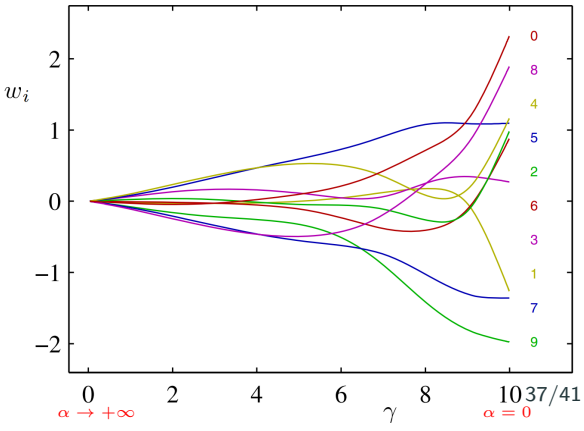


Figure 3.16 The left plot shows γ (red curve) and $2\alpha E_W(\mathbf{m}_N)$ (blue curve) versus $\ln\alpha$ for the sinusoidal synthetic data set. It is the intersection of these two curves that defines the optimum value for α given by the evidence procedure. The right plot shows the corresponding graph of log evidence $\ln p(\mathbf{t}|\alpha, \beta)$ versus $\ln\alpha$ (red curve) showing that the peak coincides with the crossing point of the curves in the left plot. Also shown is the test set error (blue curve) showing that the evidence maximum occurs close to the point of best generalization.

β の再推定の式 (3.95) の解釈

Figure 3.17 にそれぞれのパラメータの値を有効パラメータの数 γ の関数として表す．これより，ハイパーパラメータ α はパラメータ $\{w_i\}$ の大きさを制御していることが分かる．

Figure 3.17 Plot of the 10 parameters w_i from the Gaussian basis function model versus the effective number of parameters γ, in which the hyperparameter α is varied in the range $0 \leqslant \alpha \leqslant \infty$ causing γ to vary in the range $0 \leqslant \gamma \leqslant M$.



$$\gamma = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha}$$

$M >> N$ の状況

データ点数 N がパラメータ数 M と比べて十分に大きい場合について考える．このとき，どのパラメータ軸に対しても尤度関数は鋭く尖る．このとき，$\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$ の固有値は α より十分大きくなる．すなわち，$\lambda_i >> \alpha$ となり，$\gamma = M$ となる．このとき，α, β の再推定方程式は，それぞれ次のようになる．[3]

$$\alpha = \frac{\gamma}{\boldsymbol{m}_N^\top \boldsymbol{m}_N} \tag{3.92，再掲}$$

$$\longrightarrow \alpha = \frac{M}{2E_W(\boldsymbol{m}_N)} \tag{3.98}$$

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_n \{t_n - \boldsymbol{m}_N^\top \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{3.95，再掲}$$

$$\longrightarrow \beta = \frac{N}{2E_D(\boldsymbol{m}_N)} \tag{3.99}$$

[3] $E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top\boldsymbol{w}, E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^N \{t_n - \boldsymbol{w}^\top\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2$

$M >> N$ の状況

$$\alpha = \frac{M}{2E_W(\boldsymbol{m}_N)} \tag{3.98}$$

$$\beta = \frac{N}{2E_D(\boldsymbol{m}_N)} \tag{3.99}$$

このとき，$\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ の固有値を計算することなしに，α, β を求めることができる

また，このとき，$\boldsymbol{w}_{MAP} = \boldsymbol{w}_{ML}$ が成り立つ．

固定された基底関数の限界

線形回帰モデルの利点

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$$

本章を通じて，固定された非線形規定関数を線形結合したモデルを扱ってきた．このモデルは $\boldsymbol{w}$ に対しての線形性により，最小二乗問題の閉じた解が求まり（3.1.1 節），またベイズ推定の計算が簡単になるという利点があった (3.3 節以降)．さらに，基底関数を適切に選ぶことで，任意の $\boldsymbol{x}$ の非線形変換をモデル化することができた．

しかし，残念ながら，線形モデルには致命的な問題点がいくつかある．これらの問題点を克服するため，後の章では SVM や NN などのより複雑なモデルを扱うことになる．

線形回帰モデルの問題点

線形回帰モデルの問題点の 1 つは，訓練データ集合を観測する前に基底関数 ϕ_j を固定するという仮定から生じる．実際，このとき，入力空間の次元数 D に対して，指数的に基底関数の数を増やすことが必要である．

幸いにも，現実的なデータ集合には，この問題を軽減するための嬉しい性質がある．

- 入力変数同士に強い相関がある（12 章;PCA）
  - 局所的な基底関数を用いることにすれば，入力空間中のデータがある場所のみに基底関数を配置することができる．
  - RBF ネットワーク（6 章），SVM（7 章），関係ベクトルマシン（7 章）
- 目標変数は少数の次元に強く依存する（5 章;NN）．
  - NN はこの性質を活用する．