1.1

訓練集合として，N 個の観測地 x を並べた $\boldsymbol{x} := (x_1, \cdots, x_n)^T$ とそれぞれ対応する観測値 t を並べた $\boldsymbol{t} := (t_1, \cdots, t_N)^N$ が与えられたとする．そして目標データ集合 $\boldsymbol{t}$ は，まず $\sin(2\pi x)$ の関数値を作成したのち，ガウス分布に従う小さなランダムノイズを加えて対応する t_n を作った．すなわち

$$t_n \sim sin(2\pi x_n) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$$

となるデータを作成している．このようにして生成されたデータは多くの現実データ集合の持つ性質をよく表している．すなわち，データはこれから学習しようとする規則性を保持してはいるが，それぞれの観測値はランダムノイズによって不正確なものになっている．このノイズは，放射性崩壊のように，本質的に確率的なランダムプロセスによる場合もあるが，多くはそれ自身は観測されない信号源の変動によるものである．

我々の木費用は，この訓練集合を利用して，新たな入力変数の値 $\hat{x}$ に対して $\hat{t}$ の値を予測することである．後で見るように，これは背後にある関数 $\sin(2\pi x)$ を暗に見つけようとすることとほぼ等価であるが，有限個のデータ集合から汎化しなければならない点で，本質的に難しい問題である．さらに，観測データはノイズが乗っており，与えられた $\hat{x}$ に対する $\hat{t}$ の値には不確実性がある．1.2 説では，そのような不確実性を厳密かつ定量的に評価する枠組みを与える．また 1.5 説で議論する決定理論は，確率論的な枠組みを利用して，適切な基準の下での最適な予測をすることを可能にする．

ただしここでは話を先に進めるために，曲線フィッティングに基づく単純なアプローチを，あまり形式ばらない形で考えよう．ここでは特に，以下のような多項式を使ってデータへのフィッティングを行うことにする．

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1}$$

ただし，M は多項式の次数で，x^j は x の j 乗を表す．多項式 $y(x, \boldsymbol{w})$ は x の非線形関数であるものの，係数 $\boldsymbol{w}$ の線形関数であることに注意する．すなわち，相異なる x_1, x_2 に対して

$$y(x_1, \boldsymbol{w}) + y(x_2, \boldsymbol{w}) = y(x_1 + x_2, \boldsymbol{w})$$

は常には成り立たないものの，相異なる $\boldsymbol{w}_1, \boldsymbol{w}_2$ に対しては

$$y(x, \boldsymbol{w}_1) + y(x, \boldsymbol{w}_2) = y(x, \boldsymbol{w}_1 + \boldsymbol{w}_2) \tag{2}$$

が成立する．多項式のように，未知のパラメータに関して線形であるような関数は非常に重要な性質を持つ．それらは線形モデルと呼ばれ，のちの章で詳細に議論する．

訓練データに多項式を当てはめることで係数の値を求めてみよう．これは $\boldsymbol{w}$ を任意に固定した時の関数 $y(x, \boldsymbol{w})$ の値と訓練集合のデータ点との間のずれを測る誤差関数の最小化で達成できる．誤差関数の選び方として，単純で広く用いられているのは，各データ点 x_n における予測値 $y(x_n, \boldsymbol{w})$ と対応する目標値 t_n との二乗和誤差

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{3}$$

となり，これを最小化することになる．のちにこの関数を選ぶ理由について議論するが，利点の 1 つは，これが微分可能な凸関数であることにある．また，上の関数が 0 となるのは $y(x, \boldsymbol{w}_n)$ が全訓練データを通るとき，かつその時に限ることに注意する．

このように $E(\boldsymbol{w})$ をできるだけ小さくするような $\boldsymbol{w}$ を選ぶことで曲線当てはめ問題を解くことができる．誤差関数は係数 $\boldsymbol{w}$ の二次関数だから, その係数に関する微分は $\boldsymbol{w}$ の要素に関して線形になり，誤差関数を最小にするただ 1 つの解 $\boldsymbol{w}^\star$ が $\varphi(x) := (x^0, x, x^2, \cdots, x^M)^T$ として

$$\boldsymbol{w}^\star = \left(\sum_{n=1}^{N}\varphi(x_n)\varphi(x_n)^T\right)^{-1}\left(\sum_{n=1}^{N}\varphi(x_n)^T t_n\right)$$

と閉じた形で求まる．実際 $y(x, \boldsymbol{w}) = \varphi(x)^T\boldsymbol{w}$ に注意して (3) 式を $\boldsymbol{w}$ で微分して 0 と置くと

$$\begin{aligned}
\nabla E(\boldsymbol{w}) &= \sum_{n=1}^{N}\varphi(x_n)\{\varphi(x)^T\boldsymbol{w} - t_n\} \\
&= \sum_{n=1}^{N}(\varphi(x_n)\varphi(x_n)^T w - \varphi(x_n)^T t_n) = 0 \\
\therefore \boldsymbol{w} &= (\varphi(x_n)\varphi(x_n)^T)^{-1}(\varphi(x_n)t_n)
\end{aligned}$$

が得られる．また，これより結果として得られる多項式は $y(x, \boldsymbol{w}^\star)$ となる．

あとは多項式の次数 M を選ぶ問題が残っているが，この問題はモデル比較 (モデル選択) と呼ぶ重要な概念の一例とみなすことができる．

例に見る通り $M = 0, 1$ の場合にはあまりデータへの当てはまりが良くない一方で，$M = 3$ の場合がこの中では $\sin(2\pi x)$ にもっともよく当てはまっているように見える．しかし $M = 9$ にした場合，訓練データには非常によく当てはまっている (実際にこの多項式は全データを通っている) ものの曲線は無茶苦茶に発振したようになっており，関数 $\sin(2\pi x)$ の表現として明らかに不適切である．このような学習は過学習として知られている．

我々の目標は新たなデータに対して正確な予測を行える高い汎化性能を達成することである．汎化性能が次数 M にどう依存するかを定量的に評価するため，100 個のデータ点からなる独立したテスト集合を，訓練データとまったく同じ方法で生成する．すると，選んだ M の各値について (3) 式で与えられる $E(\boldsymbol{w}^\star)$ の残渣が計算できるが，テスト集合についても $E(\boldsymbol{w}^\star)$ を評価できる．このとき，

$$\begin{aligned} E_{RMS} &= \sqrt{2E(\boldsymbol{w}^\star/N)} \\ &= \sqrt{\frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\boldsymbol{w}^\star)-t_n\}^2} \end{aligned} \tag{4}$$

で定義される平均二乗平方根誤差を用いると比較に便利なことがある．N で割ることによってサイズの異なるデータ集合を比較することができるようになり，平方根をとることによって E_{RMS} は目的変数 t と同じ尺度であることが保障される．テスト集合の誤差は新たな x を観測した時に t をどれだけよく予測できたかを表している．例の通り，M が小さいと誤差が大きく，これは $\sin(2\pi x)$ の振動をとらえることができないことを意味しており，$3 \leq M \leq 8$ では比較的誤差が小さく妥当な表現といえる．

$M = 9$ では訓練集合の誤差は 0 になる．しかし関数 $u(x,\boldsymbol{w}^\star)$ の無茶苦茶な発振が起きるので，テスト集合の誤差は非常に大きくなる．$M = 9$ 次の多項式が潜在的に $M = 3$ 次の多項式を含むことを考えると，このことはパラドックスのように思えることもある．
さらに新たなデータに対する最良な予測関数はデータを生成した関数 $\sin(2\pi x)$ であると考えることもできる．後にこのことを示す．関数 $\sin(2\pi x)$ の級数展開は区間 $(0, 1)$ ですべての次数の項を含むことより，M を増やせば増やすほど単調に良い結果が得られると期待してしまう．

いろいろな次数の多項式について得られた $\boldsymbol{w}^\star$ の値を検証してみると，M の増加に伴って係数の多くが大きな値をとるようになることがわかる．これにより訓練集合のデータ点にはピッタリ適合するものの，データ点とデータ点の間では大きな発振が起こってしまう．

次にモデルの次数は固定し，データ集合のサイズを変えてみた時の振る舞いについて，モデルの複雑さを固定した時，データ集合のサイズが大きくなるにつれて過学習の問題は深刻ではなくなってくることがわかる．別な言い方をすると，データ集合を大きくすればするほど，より複雑で柔軟なモデルをデータにあてはめられるようになる．大雑把な経験則としては，データ点の数はモデル中の適応パラメータの数の何倍かよりは小さくてはならない，と言われている．しかし，3 章で見るように，必ずしもパラメータの数がモデルの複雑さを測る最適な尺度というわけではない．

また入手できる訓練集合のサイズに応じてモデルのパラメータの数を制限する必要があ

るのは納得できない感もする．モデルの複雑さはデータ点の個数ではなく，解くべき問題の複雑さに応じて選ぶのがもっともに思える．最小二乗でモデルのパラメータを求めるアプローチが最尤推定の特別な場合に相当し，過学習の問題が最尤推定の持つ一般的性質として理解できることを後で示す．

過学習の問題を避けるにはベイズ的アプローチを採用すればよい．ベイズの観点からはモデルのパラメータ数がデータ点の数をはるかに超えても問題がないことが後にわかる．実際，ベイズモデルにおいては有効パラメータ数は自動的にデータ集合のサイズに適合する．

ベイズ的アプローチ以外で，複雑で柔軟なモデルを限られたサイズのデータ集合に対して使うことができるかを考える．過学習の現象を制御するためによく使われる手法として正則化がある．これは誤差関数に罰金項を付加することにより係数が大きな値になることを防ごうとするものである．そのようなもので最も単純な誤差関数は

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\|w\|^2 \tag{5}$$

である．ただし一般性を失うことなく $w_0 = 0$ とすることができ，正則化からは外すことも多い．この場合も誤差関数 (5) 式を最小にする解は

$$\boldsymbol{w}^\star = \left(\sum_{n=1}^{N}\varphi(x_n)\varphi(x_n)^T + \lambda I\right)^{-1}\left(\sum_{n=1}^{N}\varphi(x_n)^T t_n\right)$$

と閉じた形で求まる．このようなテクニックは統計学の分野で縮小推定と呼ばれており，特に 2 次の場合はリッジ回帰と呼ばれる．またニューラルネットワークの分野では荷重減衰として知られている．

同じデータ集合に対し，$M = 9$ を当てはめた結果を見ると，過学習が抑制されて背後の $\sin(2\pi x)$ にずっと近い表現が得られていることがわかる．しかしながら，λ を大きくし過ぎると当てはまりは再び悪くなる．

モデルの複雑さに関する議論は 1.3 節で詳しく議論するが，ここでは単に誤差関数を最小にするようなアプローチで実際の応用問題を解こうとする際には，モデルの複雑さを適切に決める方法を見つけなければならないということを注意しておく．得られたデータを係数決定のための訓練集合と，テストのための確認用集合の 2 つに分けるという単純な方法が思いつくが，この方法では貴重な訓練データを無駄にすることになることが多く，より洗練されたアプローチを探す必要がある．

1.2