1

1.1

訓練集合として，N 個の観測地 x を並べた $\boldsymbol{m}x := (x_1, \cdots, x_n)^T$ とそれぞれ対応する観測値 t を並べた $\boldsymbol{m}t := (t_1, \cdots, t_N)^N$ が与えられたとする．そして目標データ集合 $\boldsymbol{m}t$ 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．

我々の目標は，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．

$$y(x, \boldsymbol{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

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

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

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

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{1.2}$$

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 (1.2) 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 (1.2) 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，

$$E_{RMS} = \sqrt{2E(\boldsymbol{w}^\star/N)} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}^\star) - t_n\}^2} \tag{1.3}$$

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．[4]

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

$$\tilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\|w\|^2 \tag{1.4}$$

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 (1.4) 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

確率の例を単純な例を使って導入する．赤と青の 2 つの箱があり，赤い箱にはリンゴ 2 個とオレンジ 6 個．青の箱にはリンゴ 3 個とオレンジ 1 個入っているとする．赤い箱を 40%, 青の箱を 60% で選び，箱の中の果物は分け隔てなく同じ確からしさで選ぶ.
この例ではどの箱を選ぶかを表すものが確率変数となり，今それを B で表すことにする．この変数は 2 つの可能な値，すなわち r(赤い箱) または b(青い箱) をとりうる．同様にどの果物かを表すのも確率変数であり，これを F で表すこれは a(リンゴ) か o(オレンジ) の値をとる．

事象の確率を，その事象が起きた回数と全試行回数の比で定義する．ただし，全試行回数が無

限に多くなった時の極限を考える．このとき

$$p(B=r)=\frac{4}{10},\quad p(B=b)=\frac{6}{10}$$

となる．事象の集合が互いに排反で，すべての可能な場合を含んでいれば，それらの事象の確率の総和は 1 になることが要請される．

このような状況で考えうる質問「リンゴを選び出す確率はいくらか」,「オレンジを選び出したとして，それが青い箱から取り出されたものである確率はいくつか」，あるいはパターン認識問題に関連したより込み入った質問にも答えるには，確率に関する 2 つの基本的な法則のみ知っていればよい．すなわち確率の加法定理と確率の乗法定理である．以下でこれらを導出することを考える．

2 つの確率変数 X,Y は任意の値 $x_i(i=1,\cdots,M), y_i(i=1,\cdots,L)$ をそれぞれとれるものとする．X,Y の両方についてサンプルを取り，全部で N 回の試行を行う．そのうち $X=x_i, Y=y_j$ となる試行の数を n_{ij} とする．また X が値 x_i を取る試行の数を c_i とし，同様に Y が y_j を取る試行の数を r_j とする．X が x_i，Y が y_j を取る確率を $p(X=x_i,Y=y_j)$ と書き，$X=x_i$ と $Y=y_j$ の同時確率と呼ぶ．この場合は

$$p(X=x_i,Y=y_j)=\frac{n_{i,j}}{N} \tag{1.5}$$

で与えられる．同様に X が c_i を取る確率は

$$p(X=x_i)=\frac{c_i}{N} \tag{1.6}$$

で与えられる．これは $c_i=\sum_j n_{ij}$ であることを用いれば

$$p(X=x_i)=\sum_{j=1}^{L} p(X=x_i,Y=y_j) \tag{1.7}$$

が成立する．これが確率の加法定理である．この場合，$p(X=x_i)$ を周辺確率と呼ぶこともある．

$X=x_i$ の事例だけを考え，その中での $Y=y_j$ の事例の比率を $p(Y=y_j \mid X=x_i)$ と書き，$X=x_i$ が与えられた下での $Y=y_j$ の条件付き確率と呼ぶ．この場合は

$$p(Y=y_j \mid X=x_i)=\frac{n_{ij}}{c_i} \tag{1.8}$$

となる．(1.5),(1.6),(1.8) 式より次の関係を得る．

$$\begin{aligned} p(X=x_i,Y=y_j) &= \frac{n_{ij}}{N} \\ &= \frac{n_{ij}}{c_i}\cdot\frac{c_i}{N} \\ &= p(Y=y_j \mid X=x_i)p(X=x_i) \end{aligned} \tag{1.9}$$

これは確率の乗法定理である．
まとめると，確率論の基本法則を以下のように書くことができる．

$$p(X) = \sum_{Y} p(X, Y) \tag{1.10}$$

$$p(X, Y) = p(Y \mid X)p(X) \tag{1.11}$$

乗法定理と対称性 $p(X, Y) = p(Y, X)$ より関係式

$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)} \tag{1.12}$$

を得る．これはベイズの定理と呼ばれ，パターン認識や機械学習において中心的な役割を果たす．
また，上式の分母は

$$p(X) = \sum_{Y} p(X \mid Y)p(Y) \tag{1.13}$$

と分子に現れる量を使って表すことができる．

果物の箱の例に戻る．赤，青の箱を選ぶ確率はそれぞれ

$$p(B = r) = 4/10 \tag{1.14}$$

$$p(B = b) = 6/10 \tag{1.15}$$

選んだ箱が青い箱であった場合にリンゴを選ぶ確率は，単に青い箱の中のリンゴの個数の比率で $3/4$ なので $p(F = a \mid B = b) = 3/4$ である．実際に箱の種類が与えられた下での果物の条件付き確率をすべて書き出すと

$$p(F = a \mid B = r) = 1/4 \tag{1.16}$$

$$p(F = o \mid B = r) = 3/4 \tag{1.17}$$

$$p(F = a \mid B = b) = 3/4 \tag{1.18}$$

$$p(F = o \mid B = b) = 1/4 \tag{1.19}$$

となり，これらの確率はすべて規格化されていることが確認できる．

$$p(F = a \mid B = r) + p(F = o \mid B = r) = 1 \tag{1.20}$$

$$p(F = a \mid B = b) + p(F = o \mid B = b) = 1 \tag{1.21}$$

確率の加法定理，乗法定理を用いることでリンゴを選ぶ確率は

$$\begin{aligned} p(F = a) &= p(F = a \mid B = r)p(B = r) + p(F = o \mid B = r)p(B = r) \\ &= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20} \end{aligned} \tag{1.22}$$

となり，また加法定理から $p(F=o)=1-\dfrac{11}{20}=\dfrac{9}{11}$ がいえる．一方果物が与えられたもとでの条件付確率はベイズの定理より

$$p(B=r\mid F=o)=\frac{p(F=o\mid B=r)p(B=r)}{p(F=o)}=\frac{3}{4}\times\frac{4}{10}\times\frac{20}{9}=\frac{2}{3} \tag{1.23}$$

となる．再び加法定理より $p(B=b\mid F=o)=1-\frac{2}{3}=\frac{1}{3}$ が言える．

ベイズの定理において，どの箱を選んだかに関する確率は事前確率と呼ばれ，一旦果物がオレンジであると判明した後に得られる $p(B\mid F)$ を事後確率と呼ぶ.
また, 2 つの変数の同時分布がその周辺分布の積に分解できるとき, すなわち $p(X,Y)=p(X)p(Y)$ となるとき，X と Y は独立であるという．このとき $p(Y\mid X)=P(Y)$ であることが乗法定理より得られる．果物の例では各箱に同じ比率でリンゴとオレンジが入っていれば $p(F\mid B)=p(F)$ となり，リンゴが選ばれる確率はどの箱が選ばれたかに独立になる．

1.2.1

連続な確率変数についても議論を進めていく．実数値をとる変数 x が区間 $(x,x+\delta x)$ に入る確率が $\delta x\to 0$ のとき，関数 $p(x)$ によって $p(x)\delta x$ で与えられるとき，$p(x)$ を x 上の確率密度関数と呼ぶ．このとき x が区間 (a,b) にある確率は

$$p(x\in(a,b))=\int_a^b p(x)dx \tag{1.24}$$

で与えられる．確率は非負で x は実数値上のどこかを値をとる必要があるため，確率密度は以下の 2 つの条件を満たす必要がある

$$p(x)\geq 0 \tag{1.25}$$

$$\int_{-\infty}^{\infty}p(x)dx=1 \tag{1.26}$$

変数に非線形な変換を施すと，確率変数はヤコビ行列により単純な関数とは異なる仕方で変換される．例えば，変数変換 $x=g(y)$ を考えると，関数 $f(x)$ は $\tilde{f}(y)=f(g(y))$ となる．確率密度 $p_x(x)$ に対応する，新しい変数 y に関する密度 $p_y(y)$ を考える．区間 $(x,x+\delta x)$ に入る観測値 x' に対応する観測値 y' は，δx が小さければ区間 $(y,y+\delta y)$ に入り，$p_x(x)\delta x\simeq p_y(y)\delta y$ となるから

$$\begin{aligned}p_y(y)&=p_x(x)\left|\frac{dx}{dy}\right|\\&=p_x(g(y))|g'(y)|\end{aligned} \tag{1.27}$$

となる．これより確率密度の最大値は変数の選び方に依存することがわかる．

x が区間 $(-\infty,z)$ に入る確率は累積分布関数

$$P(z)=\int_{-\infty}^{z}p(x)dx \tag{1.28}$$

で定義され，これは $P'(x) = p(x)$ を満たす．
いくつかの連続変数 $x_1, \cdots, x_D$ があるとき，これをまとめてベクトル $\mathbf{x}$ で表すと同時分布 $p(\mathbf{x}) = \mathbf{p}(\mathbf{x_1}, \cdots, \mathbf{x_D})$ を定義することができて，$\mathbf{x}$ が $\mathbf{x}$ を含む無限小の体積要素 $\delta\mathbf{x}$ に入る確率は $p(\mathbf{x})\delta\mathbf{x}$ で与えられる．この多変数確率密度は

$$p(\mathbf{x}) \geq 0 \tag{1.29}$$

$$\int p(\mathbf{x})\mathbf{dx} = 1 \tag{1.30}$$

を満たす必要がある．また x が離散変数の時は $p(x)$ は確率質量関数と呼ばれる．連続変数においても離散変数と同様に，確率の加法定理・乗法定理を用いることが可能で

$$p(x) = \int p(x, y)dy \tag{1.31}$$

$$p(x, y) = p(y \mid x)p(x) \tag{1.32}$$

の形をとる．

1.2.2

確率を含むもっとも重要な操作の 1 つは重み付きの平均を求めることである．関数 $f(x)$ の，確率分布 $p(x)$ の下での平均値を $f(x)$ の期待値と呼び，$\mathbb{E}[f]$ と書く．離散分布に対しては

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{1.33}$$

で与えられ，連続変数の場合の期待値は

$$\mathbb{E}[f] = \int p(x)f(x)dx \tag{1.34}$$

で与えられる．どちらの場合も確率分布や確率密度から得られた有限個の N 点を用いて，期待値はこれらの値の有限和

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{1.35}$$

で近似できる．この近似は大数の法則より，$N \to \infty$ で厳密になる．
多変数関数の期待値を考えることもあるが，この場合には，どの変数について平均を取るのかを示すのに添え字を使う．例えば，

$$\mathbb{E}_x[f(x, y)] \tag{1.36}$$

は関数 $f(x, y)$ の x の分布に関する平均を表す．$\mathbb{E}_x[f(x, y)]$ は y の関数になることに注意する．
条件付き分布についても条件付き期待値を考えることができ，

$$\mathbb{E}_x[f \mid y] = \sum_x p(x \mid y)f(x) \tag{1.37}$$

となり，連続変数についても同様に定義できる．

f の分散は

$$\mathrm{Var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{1.38}$$

と定義され，$f(x)$ がその平均のまわりでどれくらいばらつくかの尺度になる．上式を展開すると

$$\begin{aligned}\mathrm{Var}[f] &= \mathbb{E}\left[(f(x))^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\right] \\ &= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2\end{aligned} \tag{1.39}$$

と書くこともできる．特に確率変数 x 自体の分散を考えることができ，

$$\mathrm{Var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{1.40}$$

となる．2 つの確率変数 x, y の間の共分散は

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned} \tag{1.41}$$

と定義され，x と y が同時に変動する度合いを表している．x, y が独立なら $p(x, y) = p(x)p(y)$ となることより

$$\begin{aligned}\mathbb{E}[xy] &= \int xyp(x, y)dxdy \\ &= \int xp(x)dx \int yp(y)dy = \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

となり，共分散が 0 になる．
2 つの確率変数ベクトル $\boldsymbol{x}, \boldsymbol{y}$ に関して共分散は行列

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathbf{T}} - \mathbb{E}[\mathbf{y}^{\mathbf{T}}]\}] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathbf{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathbf{T}}]\end{aligned} \tag{1.42}$$

となる．この行列の各成分は

$$(\mathrm{cov}[\mathbf{x}, \mathbf{y}])_{i,j} = \mathbb{E}[x_i y_j] - \mathbb{E}[x_i]\mathbb{E}[y_j]$$

とかくことができ，特に $\mathbf{y} = \mathbf{x}$ とすると

$$(\mathrm{cov}[\mathbf{x}, \mathbf{x}])_{i,j} = \mathbb{E}[x_i x_j] - \mathbb{E}[x_i]\mathbb{E}[x_j]$$

となり，対角成分が分散を表す分散共分散行列となる．

1.2.3

これまでは確率をランダムな繰り返し試行の頻度とみなしてきたが，ここではより一般的なベイズ的な見方を導入する．そこでは確率は不確実性の度合いを与える．1.1 節の多項式曲線フィッティングの例を考える．観測される変数 t_n にのるノイズに頻度主義的な確率の概念を当てはめることは妥当であろう．しかしながら，我々はモデルパラメータ $\boldsymbol{w}$ の適切な選び方に関する不確実性を取り扱い，そして定量化したい．ベイズ的な観点を採用すれば，$\boldsymbol{w}$ といったモデルパラメータの他，モデルそのものの選択に関する不確実性を表すのに確率論の道具が使えることを見ていこう．

果物の箱の例では，果物の種類を観測することが，選ばれた箱が赤である確率を変える本質的な情報になっていた．そこでは観測されたデータで与えられた証拠を取り込むことで，事前確率を事後確率に変換できた．同様のアプローチは多項式曲線フィッティングにおけるパラメータの推論にも採用できる．データを観測する前にあらかじめ $\boldsymbol{w}$ に関する我々の仮説を事前確率分布 $p(\boldsymbol{w})$ の形で取り込んでおく．観測データ $\mathcal{D} = \{t_1, \cdots, t_N\}$ の効果は，後に見るように $p(\mathcal{D} \mid \boldsymbol{w})$ という条件付き確率で陽に表現される．ベイズの定理は

$$p(\boldsymbol{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{w})p(\boldsymbol{w})}{P(\mathcal{D})} \tag{1.43}$$

という形をとり，$\mathcal{D}$ を観測した事後に $\boldsymbol{w}$ に関する不確実性を事後分布 $p(\boldsymbol{w} \mid \mathcal{D})$ の形で評価することを可能にする．
右辺にある $p(\mathcal{D} \mid \boldsymbol{w})$ は尤度関数と呼ばれ，パラメータベクトル $\boldsymbol{w}$ を固定したときに観測されたデータがどれくらい起こりやすいかを表している．尤度は $\boldsymbol{w}$ に関する確率分布では無いことに注意する．
尤度の定義から，ベイズの定理は言葉でかけば

$$\text{事後確率} \propto \text{尤度} \times \text{事前確率} \tag{1.44}$$

となり，この式に現れる全ての値は $\boldsymbol{w}$ の関数とみなせる．(1.43) 式の分母は，左辺の事後分布が積分して 1 になることを保証する規格化定数である．実際 (1.43) 式の両辺を $\boldsymbol{w}$ で微分することで

$$\begin{aligned} \int p(\boldsymbol{w} \mid \mathcal{D}) d\boldsymbol{w} &= \int \frac{p(\mathcal{D} \mid \boldsymbol{w})p(\boldsymbol{w})}{P(\mathcal{D})} d\boldsymbol{w} \\ \Leftrightarrow \int p(\boldsymbol{w} \mid \mathcal{D})p(\mathcal{D}) d\boldsymbol{w} &= \int p(\mathcal{D} \mid \boldsymbol{w})p(\boldsymbol{w}) d\boldsymbol{w} \end{aligned} \tag{1.45}$$

となり，左辺は $p(\mathcal{D})$ を表している．ベイズと頻度主義両方の考え方で，尤度関数 $p(\mathcal{D} \mid \boldsymbol{w})$ は重要な役割を果たす．しかしながら，それをどう使うかは 2 つのアプローチで根本的に異なる．頻度主義的な設定では $\boldsymbol{w}$ は固定したパラメータと考えられ，その値は何らかの推定量として定められ，この推定の誤差範囲は可能なデータ集合 $\mathcal{D}$ の分布を考慮して得られる．一方ベイズ的な見方ではただ 1 つのデータ集合 $\mathcal{D}$ があって，パラメータに用いる不確実性は $\boldsymbol{w}$ の確率分布として表される．

頻度主義では最尤推定が広く用いられており，$\boldsymbol{w}$ は尤度関数を最大にする値である．これは観測されたデータ集合が，実際に観測される確率を最大にする $\boldsymbol{w}$ の値を選ぶことに相当する．機械学習の分野において，尤度の対数の符号を反転したものは誤差関数と呼ばれる．

頻度主義で誤差範囲を決める 1 つのアプローチはブートストラップと呼ばれているもので，そこではデータ集合 $\boldsymbol{X}$ から復元抽出を繰り返すことによって新たなデータ集合を作成し，異なるブートストラップデータ集合に対する予測の変動を見ることでパラメータ推定の統計的な精度を評価することができる．

ベイズ的な視点の利点は事前知識を自然に入れられる点にある．3 回のコイン投げで全て表が出た場合，最尤推定では表が出る確率は 1 になってしまうが，ベイズ的なアプローチを用いればそのように極端な結論を導くことが無い．ベイズアプローチに対する批判の 1 つに，事前分布の選び方によって結果が主観的になるというものがある．事前分布への依存を小さくするために無情報事前分布を用いることがあるが，これは異なるモデルを比較する際に困難が生じる．また，悪い事前分布を選べば高い確率で悪い結果が得られてしまう．これらの問題は頻度主義的な評価方法によってある程度防ぐことができ，交差確認といったテクニックがモデル選択などの問題には有効に働く．

1.2.4 ガウス分布

本書の多くで頻繁に用いられるについて述べる．
単一の実数値変数 x に対し，ガウス分布は

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.46}$$

と定義され，2 つのパラメータ，平均 μ と分散 σ^2 を持つ．分散の平方根 σ は標準偏差と呼ばれ，分散の逆数は $\beta = \frac{1}{\sigma^2}$ と書き，精度パラメータと呼ぶ．
(1.46) 式の形から，ガウス分布は

$$\mathcal{N}(x \mid \mu, \sigma^2) > 0 \tag{1.47}$$

を満たすことがわかる．また，規格化されていることは $I := \int \exp\left\{-\frac{x^2}{2\sigma^2}\right\} dx$ として

$$
\begin{aligned}
I^2 &= \int\int \exp\left\{-\frac{x^2}{2\sigma^2} - \frac{y^2}{2\sigma^2}\right\} dxdy \\
&= \int_0^{2\pi}\int_0^{\infty} \exp\left\{-\frac{r^2}{2\sigma^2}\right\} \cdot r dr d\theta \quad (\because x = r\sin\theta, y = r\cos\theta \text{ の変換}) \\
&= 2\pi \int_0^{\infty} r \exp\left\{-\frac{r^2}{2\sigma^2}\right\} dr \\
&= \pi \int_0^{\infty} \exp\left\{-\frac{u}{2\sigma^2}\right\} du (r^2 = u \text{ の変換}) \\
&= \pi \left[-2\sigma^2 \exp\left(-\frac{u}{2\sigma^2}\right)\right]_0^{\infty} \\
&= 2\pi\sigma^2 \\
\therefore I &= \sqrt{2\pi\sigma^2} \\
\therefore \int \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} dx &= 1
\end{aligned} \tag{1.48}
$$

となることより確認できる．また期待値は (1.48) 式の両辺を μ で微分することで

$$
\begin{aligned}
-\frac{1}{\sigma^2}\int \frac{1}{\sqrt{2\pi\sigma^2}}(x-\mu)\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx &= 0 \\
\mu\int \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx &= \int x \cdot \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\
\therefore \mu &= \int x \cdot \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \mathbb{E}[x]
\end{aligned} \tag{1.49}
$$

となり，分散は (1.48) 式の両辺を σ^2 で微分することで

$$
\begin{aligned}
\int \left\{-\frac{1}{2\sqrt{2\pi\sigma^2}}\cdot\frac{1}{\sigma^2}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \frac{1}{\sqrt{2\pi\sigma^2}}\cdot\frac{(x-\mu)^2}{2\sigma^4}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right\} dx &= 0 \\
\frac{1}{2\sigma^2} = \frac{1}{2\sigma^4}\int \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx & \\
\sigma^2 = \int \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx &
\end{aligned} \tag{1.50}
$$

となることがわかる．さらに分布の最大値を与えるモード (最頻値) を考えると，(1.46) 式を x で微分して 0 とおいた式を解いて

$$
\begin{aligned}
\frac{d}{dx}\mathcal{N}(x \mid \mu, \sigma^2) = -\frac{(x-\mu)}{\sigma^2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) &= 0 \\
\therefore x &= \mu
\end{aligned}
$$

となることより期待値と一致することがわかる．