パターン認識と機械学習
第 1 章 序論

鈴木拓己

2024 年 11 月 20 日

目次

1 **序論** 2
1.1 例：多項式フィッティング . . . 2
1.2 確率論 . . . 4
1.2.1 確率密度 . . . 5
1.2.2 期待値と分散 . . . 6
1.2.3 ベイズ確率 . . . 7
1.2.4 ガウス分布 . . . 9
1.2.5 曲線フィッティング再訪 . . . 12
1.2.6 ベイズ曲線フィッティング . . . 13
1.3 モデル選択 . . . 14
1.4 次元の呪い . . . 15
1.5 決定理論 . . . 16
1.5.1 誤識別率の最小化 . . . 16
1.5.2 期待損失の最小化 . . . 16
1.5.3 棄却オプション . . . 16
1.5.4 推論と決定 . . . 16
1.5.5 回帰のための損失関数 . . . 16
1.6 情報理論 . . . 16
1.6.1 相対エントロピーと相互情報量 . . . 16
1.7 演習問題 . . . 16

概要

本章では，種々の機械学習アルゴリズムの柱になる基本的な概念のうち，最も重要なもののいくつかを導入し，簡単な例を用いて説明する．また，本章の後半では，実世界のパターン認識に適用できる，より洗練されたモデルを提示する．さらに，本書全体で必要な 3 つの重要なツールである，確率論・決定理論・情報理論についての導入を行う．

Notation

基本的な記法は前書きで書かれているものに合わせているが，一部異なる記法を用いる場合がある．

1　序論

1.1　例：多項式フィッティング

まず最初に単純な回帰問題から始める．実数値の入力変数 $x \in \mathbb{R}$ を観測し，それを用いて実数値の目標変数 $t \in \mathbb{R}$ を予測したいとする．

訓練集合として，N 個の観測値 x を並べた $\mathbf{x} = (x_1, \ldots, x_N)^\top$ と，それぞれに対応する観測値 t を並べた $\mathbf{t} = (t_1, \ldots, t_N)^\top$ が与えられたとする．このとき，目標は，訓練集合を利用して，新たな入力変数 $\hat{x}$ に対して目標変数 $\hat{t}$ を予測することである．これは，観測データの背後にある関数 $y(x)$ を暗に見つけようとするの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 (uncertainty) がある．1.2 節で議論する確率論はそのような不確実性を厳密かつ定量的に表現する枠組みを与える．また，1.5 節で議論する決定理論は，確率論的な枠組みを利用して，適切な規準の下で最適な予測をすることを可能にする．

ここでは，話を先に進めるために，曲線フィッティング (曲線あてはめ：curve fitting) に基づく単純なアプローチを考える．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** (order) で，x^j は x の j 乗を表す．多項式の係数 $w_0, \ldots, w_M$ をまとめて $\boldsymbol{w}$ と書く．多項式 $y(x, \boldsymbol{x})$ は x の非線形関数であるが，係数 $\boldsymbol{w}$ の線形関数であることに注意する．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ち，**線形モデル** (linear model) と呼ばれ，3 章と 4 章で詳細に議論する．

訓練データに多項式をあてはめることで係数の値を求めてみよう．これは，$\boldsymbol{w}$ を任意に固定したときの関数 $y(x, \boldsymbol{w})$ 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** (error function) 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ものは，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 (sum-of-squares error) である．式で書けば，

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

となり，これを最小化することになる．後で見るように，二乗和誤差関数の最大化は，観測データ x_n のノイズがガウス分布に従うと仮定した下での尤度関数の最大化と等価になる．ここでは単に，二乗和誤差関数は非負であり，その値が 0 になることと $y(x, \boldsymbol{w})$ が全訓練データ点を通ることは同値であることに注意すればよい．

このように，$E(\boldsymbol{w})$ をできるだけ小さくするような $\boldsymbol{w}$ を選ぶことで曲線フィッティング問題を解くことができる．誤差関数は係数 $\boldsymbol{w}$ の 2 次関数であるため，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，通常，誤差関数を最小にする一意的な解を持つ．その解 $\boldsymbol{w}^*$ は閉形式で求まり (**演習 1.1**)，結果として得られる多項式は $y(x, \boldsymbol{w}^*)$ 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** (model conparison) あるいは**モデル選択** (model selection) と呼ぶ重要な概念の一例とみなすことができる．

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にある．汎化性能が M にどう依存するかを定量的に評価するために，訓練集合の一部から用意したテスト集合を用いれば，選んだ M の各値について，訓練データに対して (1.2) で与えられる $E(\boldsymbol{w}^*)$ の残差が計算できるが，テスト集合についても $E(\boldsymbol{w}^*)$ が評価できる．このとき，

$$E_{\mathrm{RMS}} = \sqrt{2E(\boldsymbol{w}^*)/N} \tag{1.3}$$

で定義される平均二乗平方根誤差 (root-mean-square error, RMS error) を用いると便利なことがある．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 (単位) であることが保証される．いろいろな M に対する訓練集合とテスト集合の RMS 誤差を測ることにより，妥当な M の値を評価することができる．

ある次数の多項式がそれより低い次数のすべての多項式を含むことを考えると，M の値は大きければ大きいほど良い結果が得られるように思えるが，一般に M の値が大きすぎるとデータを生成する関数に対する曲線のあてはまりが悪くなり，このような振る舞いは**過学習** (over-fitting) として知られている．これは，直観的には，M が大きく自由度の高い多項式モデルは目的値のランダムノイズに引きずられてしまうと解釈すればよい．

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ときの振る舞いを考える．一般に，モデルの複雑さを固定したとき，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なくなる．別の言い方をすると，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 (例えば 5 とか 10) より小さくてはいけない，と言われている．しかしながら，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しなければならないのは納得できない感じがする．モデルの複雑さはむしろ解くべき問題の複雑さに応じて選ぶのがもっともであるように思える．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** (maximum likelihood) (1.2.5 節で議論する) 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** (Bayesian) アプローチ (3.4 節) 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** (effective number of parameters) は自動的にデータ集合のサイズに適合する．

ここでは，これまで説明したアプローチに沿って，実際にどうやって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もう少し考える．過学習の現象を制御するためによく使われるテクニックに**正則化** (regularization) がある．これは，誤差関数 (1.2) に罰則項 (penalty term) 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．そのような罰則項のうち最も単純なものは，係数の L2-ノルムをとったもので，誤差関数は

$$\widetilde{E}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2 \tag{1.4}$$

で与えられる．ここで，$\|\boldsymbol{w}\|_2^2 = \boldsymbol{w}^{\top}\boldsymbol{w} = w_0^2 + w_1^2 + \cdots + w_M^2$ であり，係数 λ は正則化項と二乗誤差の和の項との相対的な重要度を調節している．ただし，係数 w_0 は正則化から外すことも多い．というのは，w_0 は目的変数の原点の選び方に依存しているためであり，正則化に入れるとしてもそれ専用の正則化係数を掛けたりする (詳細な議論については 5.5.1 節で扱う)．この場合も誤差関数 (1.4) を最小にする解は閉形式で求めることができる (**演習 1.2**)．このようなテクニックは，係数の値を小さくするという意味で，統計学の分野で**縮小推定** (shrinkage) と呼ばれている．特に，2 次の正則化の場合は**リッジ回帰** (ridge regression) と呼ばれ，ニューラルネットワークの文脈では**荷重減衰** (weighted decay) として知られている．

正則化項が汎化誤差に与える影響は，訓練集合とテスト集合の両方に対する RMS 誤差 (1.3) の値を $\ln\lambda$ に対してプロットしてみればよい．これを見ることで，λ がモデルの実質的な複雑さを制御し，過学習の度合いを決定していることがわかる．

モデルの複雑さの問題は重要で，1.3 で詳しく議論する．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に注意する．上記の議論から，得られたデータを，係数 $\boldsymbol{w}$ を決めるために使われる訓練集合と，

それとは別の**検証用集合** (validation set) に分けるという単純な方法が思いつく．検証用集合は**ホールドアウト集合** (hold-out set) とも呼ばれ，モデルの複雑さ (M または λ) を最適化するのに使われる．ただし，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いので，より洗練されたアプローチを探す必要がある (1.3 節).

ここまでは，直観に依拠した多項式曲線フィッティングの議論を扱ってきたが，ここからは，確率論的な議論の枠組みを用いて，パターン認識における問題を解くためのより原理的なアプローチを扱う．

1.2 確率論

パターン認識の分野の鍵となる概念は不確実性である．これは計測ノイズやデータ集合のサイズが有限であることによって起きる．確率論 (probability theory) は不確実性に関する定量化と操作に関して一貫した枠組みを与え，パターン認識の基礎の中心をになっている．また，1.5 節で議論する決定理論と組み合わせることにより，与えられた情報が不完全で曖昧なものであっても，そのすべての情報の下で最適な予測をすることが可能となる．

始めに，ある事象の確率を，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ときの極限を考える．パターン認識問題に関連した確率の重要な法則は**確率の加法定理** (sum rule of probability) と**確率の乗法定理** (product rule of probability) である．これらの法則を導出するため，2 つの離散型確率変数 X, Y を考える．ただし，X の標本空間を $\mathcal{X} = \{x_i \in \mathbb{R} \mid i = 1, \ldots, M\}$ とし，Y の標本空間を $\mathcal{Y} = \{y_j \in \mathbb{R} \mid j = 1, \ldots, L\}$ とする．X と Y の両方についてサンプルをとり，合計 N 回の試行を行う．そのうち，$(X, Y) = (x_i, y_j)$ となる試行の数を n_{ij} とする．また，$X = x_i$ となる試行の数を c_i とし，$Y = y_j$ となる試行の数を r_j とする．$(X, Y) = (x_i, y_j)$ となる確率を $p(X = x_i, Y = y_j)$ と書き，$(X, Y) = (x_i, y_j)$ の**同時確率** (joint probability) と呼ぶ．これは，

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

で与えられる．ここで，暗に $N \to \infty$ という極限を考えている．同様に，$X = x_i$ となる確率を $p(X = x_i)$ と書き，

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

で与えられる．いま，$c_i = \sum_j n_{ij}$ であるから，(1.5) と (1.6) より

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

が成り立ち，これが確率の**加法定理** (sum rule) である．$p(X = x_i)$ は他の変数 (ここでは Y) についての周辺化であるから，**周辺確率** (marginal probability) と呼ぶ.

$X = x_i$ が与えられたときの $Y = y_j$ の事象の比率を $p(Y = y_j \mid X = x_i)$ 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** (conditional probability) と呼ぶ．この定義より，

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

となる．ここで，(1.5)，(1.6) および (1.8) から

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

が成り立つ．これは確率の**乗法定理** (product rule) である．以下，文脈上明らかなときは確率変数 X 上の確率分布を単に $p(X)$ と書き，$X = x_i$ をとる確率を $p(x_i)$ と書く．この記法を用いると，確率論の 2 つの基本

法則を以下のように書くことができる．

$$\textbf{(加法定理)} \qquad p(X) = \sum_Y p(X, Y), \tag{1.10}$$

$$\textbf{(乗法定理)} \qquad p(X, Y) = p(Y \mid X)p(X). \tag{1.11}$$

特に，乗法定理および対称性 $p(X, Y) = p(Y, X)$ から，

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

を得る．これは**ベイズの定理** (Bayes' theorem) 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，加法定理を用いれば，ベイズの定理の分母は分子に現れる量を用いて

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

と表すことができる．これにより，ベイズの定理の分母は，(1.12) 式の左辺の条件付き確率を Y について和をとったものが 1 になることを保証するための規格化 (正規化) 定数とみなすことができる．ベイズの定理における $p(Y)$ を Y の**事前確率** (prior probability)，$p(Y \mid X)$ を X が与えられた下での Y の**事後確率** (posterior probability) と呼ぶ．

2 つの確率変数の同時分布が $p(X, Y) = p(X)p(Y)$ と周辺確率の積に分解できるとき，X と Y は**独立** (independent) であるという．確率変数 X, Y が独立であるとき，乗法定理から $p(Y \mid X) = p(Y)$ を得るので，X が与えられた下での Y の条件付き確率は X の値に独立になる．

1.2.1 確率密度

連続的な事象集合に対する確率を考える．実数値をとる変数 $x \in \mathbb{R}$ が区間 $(x, x + \delta x)$ に入る確率が，$\delta x \to 0$ のとき $p(x)\delta x$ で与えられるとき，$p(x)$ を x 上の**確率密度** (probability density) と呼ぶ．このとき，x が区間 (a, b) にある確率は

$$p(x \in (a, b)) = \int_a^b p(x)dx \tag{1.14}$$

で与えられる．確率は非負で x は実数値上のどこかの値をとらなければならないため，確率密度は

$$p(x) \geq 0, \tag{1.15}$$

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{1.16}$$

を満たす必要がある．変数に非線形な変換を施すと，確率密度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 $\widetilde{f}(y) = f(g(y))$ となる．ここで，確率密度 $p_x(x)$ に対応する，新たな変数 y に関する密度 $p_y(y)$ を考える．ただし，ここでは添え字によって異なる密度 $p_x(x), p_y(y)$ を表す．g に連続性を仮定すれば，区間 $(x, x + \delta x)$ に入る観測値は δx が十分小さければ区間 $(y, y + \delta y)$ に入り，$p_x(x)\delta x \simeq p_y(y)\delta y$ となるので，

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right| = p_x(g(y))|g'(y)| \tag{1.17}$$

となる．この性質から，確率密度の最大値は変数の選び方に依存するということがわかる (**演習 1.4**).

x が区間 $(-\infty, x)$ に入る確率は**累積分布関数** (cumulative distribution function)

$$P(x) = \int_{-\infty}^{x} p(x)dx \tag{1.18}$$

で定義され，微分積分学の基本定理より $P'(x) = p(x)$ を満たす．

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

$$p(\boldsymbol{x}) \geq 0, \tag{1.19}$$

$$\int p(\boldsymbol{x})d\boldsymbol{x} = \int_{\mathcal{X}_D} \cdots \int_{\mathcal{X}_1} p(x_1, \ldots, x_D)dx_1 \cdots dx_D = 1 \tag{1.20}$$

を満たす必要がある．ここで，第 2 式の積分は $\boldsymbol{x}$ の定義域全体にわたってとる．離散変数と連続変数が組み合わせになっている場合も同時確率分布を考えることができる．x が離散変数の場合には，$p(x)$ を**確率質量関数** (probability mass function) と呼ぶ.

確率の加法・乗法定理およびベイズの定理は確率密度や離散変数と連続変数の組み合わせに対しても同様に適用可能であり，x, y を 2 つの実変数として，その形は

$$p(x) = \int p(x, y)dy, \tag{1.21}$$

$$p(x, y) = p(y \mid x)p(x) \tag{1.22}$$

をとる．連続変数の加法・乗法定理を厳密に示すには測度論が必要となるが，ここでは測度論的確率論は扱わない．しかし，厳密でない言い方をすれば，各連続変数を幅 Δ の区間に分けて，その上の離散確率分布を考えることにより理解できる．$\Delta \to 0$ という極限をとると，和は積分になり所望の結果が得られる．

1.2.2　期待値と分散

確率を含む最も重要な操作の 1 つは関数の重み付きの平均を求めることである．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** (expectation) と呼び,$\mathbb{E}[f]$ と書く．離散分布に対しては,

$$\mathbb{E}[f] = \sum_x p(x)f(x) \tag{1.23}$$

で与えられ，連続変数の場合は対応する確率密度を用いて

$$\mathbb{E}[f] = \int p(x)f(x)dx \tag{1.24}$$

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，期待値はこれらの点の有限和によって

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n) \tag{1.25}$$

で近似される．この結果は，11 章でサンプリング法について議論する際に頻繁に用いることになる．(1.25) の近似は $N \to \infty$ の極限で厳密になる．

多変数関数の期待値を考える際には，どの変数について平均をとるかを示すのに添字を用いる．例えば,

$$\mathbb{E}_x[f(x, y)] \tag{1.26}$$

は関数 $f(x, y)$ の x の分布に関する平均を表す．このとき，$\mathbb{E}_x[f(x, y)]$ は y の関数で表される確率変数である．

条件付き分布についても**条件付き期待値** (conditional expectation) を考えることができ,

$$\mathbb{E}_x[f \mid y] = \sum_x p(x \mid y)f(x) \tag{1.27}$$

となり，連続変数に対しても同様に定義される．$f(x)$ の**分散** (variance) は

$$\mathbb{V}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \tag{1.28}$$

で定義され，$f(x)$ がその平均値 $\mathbb{E}[f(x)]$ の周りでどの程度ばらつくかの尺度となる．期待値の中身の 2 乗を展開すると，期待値の線形性から，分散は $f(x)$ と $f(x)^2$ の期待値を用いて

$$\mathbb{V}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}\left[f(x)^2 - 2f(x)\mathbb{E}[f(x)] + \mathbb{E}[f(x)]^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \tag{1.29}$$

と書くこともできる (**演習 1.5**)．特に，確率変数 x 自身の分散を考えることができ，

$$\mathbb{V}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 \tag{1.30}$$

となる．2 つの確率変数 x と y の**共分散** (covariance) は

$$\begin{aligned}\mathrm{Cov}[x, y] &= \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy - x\mathbb{E}[y] - y\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y]] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned} \tag{1.31}$$

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 と y が独立ならば共分散は 0 になる (**演習 1.6**)．

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

$$\begin{aligned}\mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{y}^\top - \mathbb{E}[\boldsymbol{y}^\top])] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[\boldsymbol{x}\boldsymbol{y}^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}^\top] \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\begin{pmatrix} x_1 - \mathbb{E}[x_1] \\ \vdots \\ x_N - \mathbb{E}[x_N] \end{pmatrix} \begin{pmatrix} y_1 - \mathbb{E}[y_1] & \cdots & y_N - \mathbb{E}[y_N] \end{pmatrix}\right] \\ &= \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\left[\begin{pmatrix} (x_1 - \mathbb{E}[x_1])(y_1 - \mathbb{E}[y_N]) & \cdots & (x_1 - \mathbb{E}[x_1])(y_N - \mathbb{E}[y_N]) \\ \vdots & & \vdots \\ (x_N - \mathbb{E}[x_N])(y_1 - \mathbb{E}[y_N]) & \cdots & (x_N - \mathbb{E}[x_N])(y_N - \mathbb{E}[y_N]) \end{pmatrix}\right]\end{aligned} \tag{1.32}$$

となる．ベクトル $\boldsymbol{x}$ の成分間の共分散を表すのには，より単純に $\mathrm{Cov}[\boldsymbol{x}] \coloneqq \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}]$ と書く．

1.2.3 ベイズ確率

本章ではここまで確率をランダムな繰り返し試行の頻度とみなしてきた．これを**古典的確率** (classical probability) あるいは**頻度主義的** (frequentist) な確率解釈と呼ぶ．ここではより一般的な**ベイズ的** (Bayesian) な見方を導入する．そこでは，確率は不確実性の度合いを考える．

不確かな事象，例えば月がかつて太陽を回る軌道上にあったかどうかとか，北極の氷が今世紀末には消えるかどうかといったことを考える．これらの事象は確立の概念を定義するために果物の箱を使って行ったような，たくさんの繰り返し観測ができる事象ではない．しかしながら，一般にどれぐらい速く極地の氷が溶けるかといったことに関して我々は何らかの知見を持っている．今，何か新たな証拠，例えば地球観測衛星が集めた新たな形の診断情報が得られれば，氷を失う速さに関する意見を修正するかもしれない．こうした問題に対する評価は，どれくらいまで温室ガスの放出を減らす努力をすべきかなど，とるべき行動に影響を与える．したがって，そのような状況に対しては，不確実性を定量的に表現し，新たな証拠に照らしてそれを正しく修正し，その結果として最適な行動や決定を下したくなる．これらはすべて，エレガントで非常に一般的なベイズ的な確率の解釈によって実現できる．

不確実性を表現するのに確率を使うことは場当たり的なものではなく，合理的で一貫した推論を行う際の常識というものを考えれば，必然的なものであることがわかる．例えば，Cox (1946) は，信念の度合いを数値で表そうとする際，信念に関する常識の性質を公理の単純な集合で表すと，信念の度合いを操作する法則の集合が一意的に導かれ，それが確率の加法・乗法定理と等価であることを示した．これは，確率論がブール論理を不確実性を含む場合に拡張したものとみなせることの，最初の厳密な証明である (Jaynes, 2003)．他にも多くの研究者が，不確実性の尺度が満たすべき性質や公理を様々に提案している (Ramsey, 1931; Good, 1950; Savage, 1961; deFinetti, 1970; Lindley, 1982)．いずれの場合も結果的に得られる数値的な量は厳密に確率の加法・乗法定理に従う．従って，これらの量を (ベイズ) 確率とみなすのは自然である．

パターン認識の分野でも，確率のより一般的な概念を導入することが有用である．1.1 節の多項式曲線フィッティングの例を考えよう．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あ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ほか，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．

ここでベイズの定理が新たな重要性を獲得することを見る．果物の箱の例を思い出してみると，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．この例ではベイズの定理により，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．後で詳しくみるように，同様のアプローチは多項式曲線フィッティングの例において $\boldsymbol{w}$ などのパラメータに関する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 $\mathcal{D} = \{t_1, \ldots, t_N\}$ の効果は，後で 1.2.5 節で見るように，$p(\mathcal{D} \mid \boldsymbol{w})$ という条件付き確率で陽に表現される．ベイズの定理は

$$p(\boldsymbol{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{w})p(\boldsymbol{w})}{p(\mathcal{D})} \tag{1.33}$$

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．

ベイズの定理の右辺にある $p(\mathcal{D} \mid \boldsymbol{w})$ という量はデータ集合 $\mathcal{D}$ に対する評価であって，パラメータベクトル $\boldsymbol{w}$ の関数とみなせる．これを**尤度関数** (likelihood function) と呼ぶ．これは，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集合がどれぐらい起こりやすいかを表している．尤度は $\boldsymbol{w}$ の確率分布ではなく，$\boldsymbol{w}$ に関する積分は 1 になるとは限らないことに注意する．

尤度の定義から，ベイズの定理は言葉で書けば

$$\text{事後確率} \propto \text{尤度} \times \text{事前確率} \tag{1.34}$$

となる．この式に現れるすべての値は $\boldsymbol{w}$ の関数とみなせる．(1.33) の分母は，左辺の事後分布が厳密に確率密度になっており，積分すると 1 になることを保証する規格化定数である．実際，(1.33) の両辺を $\boldsymbol{w}$ で積分すると，ベイズの定理の分母を事前分布と尤度関数で表すことができ，

$$p(\mathcal{D}) = \int p(\mathcal{D} \mid \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} \tag{1.35}$$

となる．

ベイズと頻度主義の両方のパラダイムで，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」をとして定められ，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 (つまり実際観測された) データ集合 $\mathcal{D}$ があって，パラメータに関する不確実性は $\boldsymbol{w}$ の確率分布として表される．

頻度主義で広く用いられている推定量は**最尤推定** (maximum likelihood) で，$\boldsymbol{w}$ は尤度関数 $p(\mathcal{D} \mid \boldsymbol{w})$ を最大にする値である．これは観測されたデータ集合の確率を最大にする $\boldsymbol{w}$ の値を選ぶことに相当する．機械学習の分野では，尤度関数の対数の符号を反転したものは**誤差関数** (error function) と呼ばれる．対数のマイナスは単調減少関数だから，尤度の最大化は誤差の最小化と等価である．

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** (bootstrap) と呼ばれているもので，そこでは複数のデータ集合を次のように作る．まず，もととなるデータ集合が N 個のデータ点 $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ からなるとする．$\mathbf{X}$ からランダムに N 点を復元抽出することによって，新たなデータ集合 $\mathbf{X}_{\mathrm{B}}$ を作ることができる．$\mathbf{X}$ のいくつかの点は $\mathbf{X}_{\mathrm{B}}$ に複数回出現するが，一方 $\mathbf{X}_{\mathrm{B}}$ に入っていない点も存在することになる．この試行を L 回繰り返すことにより L 組のデータ集合を作ることができ，それぞれのサイズは N で，もと

のデータ集合 $\boldsymbol{X}$ からサンプリングして得られたものになる．パラメータ推定の統計的な精度は，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によって評価できる[*1]．

ベイズ的な視点の利点の 1 つは事前知識を自然に入れられることである．例えば，公平に見えるコインを 3 回投げて毎回表が出たとしよう．古典的な最尤推定では表が出る確率は 1 になってしまう (2.1 節)．これは未来永劫表が出ることを意味している！逆にベイズ的アプローチでは妥当な事前分布を使えばそれほど極端な結論を導くことはない．

頻度主義かベイズ主義かといったパラダイムの利点について多くの論争がなされてきた．頻度主義独自の視点とかベイズ主義独自の視点はないことが問題を困難にしてきた[*2]．例えば，ベイズアプローチによくある批判として，事前分布が何らかの事前の信念というよりは数学的な便宜によって選ばれることが多いというものがある．事前分布の選び方によって結果が主観的になることも人によっては難点と思えるだろう．事前分布への依存を小さくしたいときに，いわゆる**無情報事前分布** (noninformative prior) (2.4.3 節) を使うことがある．しかしながら，これは異なるモデルを比較する際には困難が生じる．また，実際，ベイズ的な方法は，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 (1.3 節) といったテクニックがモデルの選択などの問題には有効に働く．

ベイズ的な手法が近年実用面で非常に重要になってきているのを考慮し，本書ではベイズ的な視点を強調する一方で，必要に応じて，有用な頻度主義的な概念も議論する．

ベイズの枠組みの源は 18 世紀にまでさかのぼるものの，ベイズ法の実際への応用は長い間非常に限られたものであった．というのも，ベイズ法を完全に実行するには特に全パラメータ空間での周辺化 (和または積分) を必要としたからである．これらの操作は後から見るように，予測を行ったり，異なるモデルを比較したりするために必要なものである．マルコフ連鎖モンテカルロ法 (11 章で議論する) のようなサンプリング法の開発や，計算機の速度やメモリ量の大幅な進歩により，ベイズ法が極めて広い範囲の問題に実用的に使えるようになった．モンテカルロ法は非常に柔軟で，さまざまなモデルに適用可能である．しかしながら，計算量は非常に大きいのでこれまでは主に小さいスケールの問題に用いられてきた．

さらに最近になって，変分ベイズ法や EP 法 (期待値伝播法) といった，非常に能率的な決定論的近似法 (10 章で議論する) が開発された．これらの手法はサンプリング法が使えない場合，その代替的手法として使われ，ベイズ法を大規模な応用に適用することを可能にした．

1.2.4 ガウス分布

2 章全体でいろいろな確率分布やその重要な性質について述べる．ここでは，連続変数の確率分布の中で最も重要な**正規分布** (normal distribution) または**ガウス分布** (Gaussian distribution) と呼ばれる分布を導入する．本章の残りの部分を始め，本書の多くの場所でガウス分布を頻繁に用いる．

実数値変数 $x \in \mathbb{R}$ に対し，ガウス分布は

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1.36}$$

で定義され，2 つのパラメータ，**平均** (mean) μ および**分散** (vaeriance) σ^2 をもつ．分散の平方根 σ は**標準偏差** (standard deviation) と呼ばれ，分散の逆数は $\beta = 1/\sigma^2$ と書き，**精度パラメータ** (precision parameter) と呼ぶ．

[*1] データ集合を生成した真の分布がわかっていれば，そこからデータ集合を生成しパラメータを推定するというプロセスを何度も繰り返すことによって，パラメータの推定値の分布を得ることができる．しかし一般に真の分布はわからないので，ブートストラップ法ではその矜持として，与えられたデータ集合を真の分布とみなし，そこからブートストラップデータ集合を生成することによって，パラメータの推定値がどのように分布するかを評価するのである..

[*2] 頻度主義とベイズ主義の論争のポイントを一言で言えば「どこまで主観性を認めるか」という哲学的問題となる．したがって，さまざまな立場の人がおり，単純に頻度主義対ベイズ主義という構図ではないというのがこの文の意味するところである．本書に関しては，この後にも書かれているように，実用性や有用性を重視し，数学的に矛盾がなければどちらのパラダイムも採用するという立場である．

(1.36) の形から，ガウス分布は

$$\mathcal{N}(x \mid \mu, \sigma^2) > 0 \tag{1.37}$$

を満たすことがわかる．また，ガウス分布が規格化されていることは簡単に示すことができ，

$$\int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) dx = 1 \tag{1.38}$$

となる (**演習 1.7**)．従って，(1.36) は確率密度の満たすべき 2 つの要件を満たしている．

ガウス分布の下で x の関数の期待値はすぐに計算することができ，特に x の平均値は，

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) x dx = \mu \tag{1.39}$$

で与えられる．パラメータ μ はこの分布の下での x の平均値になるので平均と呼ばれる．同様に，2 次のモーメントは

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2 \tag{1.40}$$

となる (**演習 1.8**)．(1.39) と (1.40) より，x の分散は

$$\mathbb{V}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \tag{1.41}$$

で与えられるので，σ^2 を分散パラメータと呼ぶ．分布の最大値を与える x はモード (最頻値) である．ガウス分布に関しては，モードは平均に一致する (**演習 1.9**)．

D 次元ベクトルの連続変数 $\boldsymbol{x}$ に対して定義されるガウス分布は，

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right\} \tag{1.42}$$

で与えられ，D 次元ベクトル $\boldsymbol{\mu}$ は平均，$D \times D$ 行列 $\boldsymbol{\Sigma}$ は共分散と呼ばれ，$|\boldsymbol{\Sigma}|$ は $\boldsymbol{\Sigma}$ の行列式を表す．本章では，多変量ガウス分布も多少用いるが，その詳細な性質は 2.3 節で述べる．

ここで，スカラー変数 x の N 個の観測値からなるデータ集合 $\mathbf{x} = (x_1, \ldots, x_N)^\top$ があるとする．未知の平均 μ と分散 σ^2 をもつガウス分布から独立に生成された観測値があったとき，これらのパラメータの値をデータ集合から定めることを考える．データ点が同じ分布から独立に生成されるとき，**独立同分布** (independent identically distributed) であるといい，i.i.d. と略す．2 つの独立な事象の同時確率はそれぞれの事象の周辺確率の積で与えられるので，データ集合 $\mathbf{x}$ が i.i.d. であることから，μ と σ^2 が与えられたとき，データ集合の確率は

$$p(\mathbf{x} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n \mid \mu, \sigma^2) \tag{1.43}$$

という形に書ける．μ と σ^2 の関数とみなすと，これはガウス分布に対する尤度関数である．

観測されたデータ集合を使って確率分布のパラメータを決める普通の方法は，尤度関数を最大にするようなパラメータの値を求めることである．この規準は，確率論に関する今までの議論からすると，奇妙に思えるかもしれない．なぜなら，パラメータが与えられた下でのデータの確率でなく，むしろデータが与えられた下でのパラメータの確率を最大化する法が自然に見えるからである．実際にこの 2 つの規準は関連しており，曲線フィッティングの問題を使って後ほど議論する (1.2.5 節)．

ここでは，尤度関数 (1.43) を最大化することによって，ガウス分布の未知のパラメータ μ と σ^2 を決めることにする．実際には，以下の観点から，尤度関数の対数 (対数尤度関数) を最小化する法が便利である．

- 対数関数は単調増加関数なので，尤度関数の最大化は対数尤度関数の最小化と等価
- 尤度関数における積を和に変換することができ，微分計算が容易

- 数値計算をする際，尤度関数における小さな確率値の積は計算機の数値精度のアンダーフローを容易に引き起こすが，これが確率値の対数の和に変換されることで，この問題を解決することができる

(1.36) と (1.43) から，対数尤度関数は

$$\begin{aligned}\ln p(\mathbf{x} \mid \mu, \sigma^2) &= \ln \left(\prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x_n - \mu)^2 \right\} \right) \\ &= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)\end{aligned} \tag{1.44}$$

という形になる．(1.44) を μ と σ^2 に関して最大化することで，最尤推定の解が得られる．(1.44) を μ に関して最大化すると，

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{1.45}$$

で与えられ，これは**標本平均** (sample mean)，すなわち観測値 $\{x_n\}$ の平均に等しい．同様に，(1.44) を σ^2 に関して最大化すると，分散に対する最尤解

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 \tag{1.46}$$

を得る (**演習 1.11**)．これは標本平均 μ_{ML} に関する**標本分散** (sample variance) である．μ と σ^2 に関して (1.44) の同時最大化を行うとき，ガウス分布の場合には μ と σ^2 は分離して解けるので，まず (1.45) を評価し，この結果を使って (1.46) を評価することができる．

本章の後半およびその後の章では，最尤アプローチの重大な限界について述べる．ここでは，1 変数ガウス分布の最尤パラメータの設定に関してその問題を取り扱う．最尤アプローチでは特に分布の分散が系統的に過小評価されている．これは**バイアス** (bias) と呼ばれる現象の例であり，多項式曲線のフィッティングにおける過学習の問題に関連している (1.1 節)．まず，最尤解 $\mu_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}}$ はデータ集合の値 $x_1, \ldots, x_N$ の関数であることに注意する．これらの量の，パラメータ μ, σ^2 を持つガウス分布に従うデータ集合に関する期待値を考える．簡単な計算で

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu, \tag{1.47}$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N} \right) \sigma^2 \tag{1.48}$$

となり，最尤推定の平均は正しい平均になるが，真の分散は $(N-1)/N$ 倍過小評価されることが示される (**演習 1.12**)．(1.48) から，

$$\widetilde{\sigma}^2 = \frac{N}{N-1} \sigma^2_{\mathrm{ML}} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 \tag{1.49}$$

は分散パラメータの不偏推定量になる．10.1.3 節では，この結果がベイズアプローチによってどのように自動的に得られるかがわかる．

最尤解のバイアスはデータ点の数 N が増えればあまり重大ではなくなり，$N \to \infty$ の極限では分散の最尤解はデータを生成した分布の真の分散に一致することに注意する．実際には N が小さいという理由以外では，バイアスは深刻な問題にはならないことがはっきり示されている．しかしながら本書を通して，多くのパラメータを持つより複雑なモデルを扱うので，最尤推定に伴うバイアスの問題ははるかに厳しいものとなる．実際，最尤推定のバイアスの問題は多項式曲線フィッティングの例で前にみたように，過学習の問題の根本にあることがわかる．

1.2.5　曲線フィッティング再訪

1.1 節では，多項式曲線のフィッティング問題が誤差最小化としてどう表現することができるかを見てきた．ここでは曲線フィッティングの例に戻って，それを確率的な観点から眺め，誤差関数と正則化に関する洞察を得るとともに，完全なベイズ的取り扱いに進むことにする．

曲線フィッティング問題の目標は，N 個の入力値で構成される訓練データの集合 $\mathbf{x} = (x_1, \ldots, x_N)^\top$ とそれに対応する目標値 $\mathbf{t} = (t_1, \ldots, t_N)^\top$ に基づいて，与えられた新たな入力値 x に対する目的変数 t の予測ができるようにすることである．目標変数の値に関する不確実性は確率分布を使って表すことができる．そのために，与えられた x に対し，対応する t は，平均が (1.1) で与えられる多項式曲線 $y(x, \boldsymbol{w})$ に等しいガウス分布に従うものとする．すなわち，

$$p(t \mid x, \boldsymbol{w}, \beta) = \mathcal{N}(t \mid y(x, \boldsymbol{w}), \beta^{-1}) \tag{1.50}$$

となる．ここで事象以降の記号と合わせるために，分布の逆分散に相当する精度パラメータ β を定義した．

これで訓練データ $\{\mathbf{x}, \mathbf{t}\}$ を使って未知のパラメータ $\boldsymbol{w}, \beta$ を求めるのに最尤推定を使うことができる．データ (1.50) の分布から独立に生成されたものと仮定すれば，尤度関数は

$$p(\mathbf{t} \mid \mathbf{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid y(x_n, \boldsymbol{w}), \beta^{-1}) \tag{1.51}$$

で与えられる．この尤度関数 $p(\mathbf{t} \mid \mathbf{x}, \boldsymbol{w}, \beta)$ に関する対数尤度関数は，

$$\ln p(\mathbf{t} \mid \mathbf{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \tag{1.52}$$

という形で得られる．まず，最尤解によって定まる多項式の係数 $\boldsymbol{w}_{\mathrm{ML}}$ を考える．これは (1.52) を $\boldsymbol{w}$ について最大化することによって求められるが，$\boldsymbol{w}$ に依存しない (1.52) の右辺第 2 項および第 3 項を無視し，同じく $\boldsymbol{w}$ に依存しない $\beta/2$ を $1/2$ に置き換え，対数尤度関数の最大化の代わりに負の対数尤度関数の最小化を考えることで，$\boldsymbol{w}$ を求めるという観点からは (1.2) で定義される**二乗和誤差** (sum-of-squares error) の最小化と等価であることがわかる．したがって，二乗和誤差関数はノイズがガウス分布に従うという仮定の下で尤度の最大化の結果としてみなせる．

条件付きガウス分布の精度パラメータ β を決めるのにも最尤推定を使うことができる．(1.52) を β について最大化すると，

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}_{\mathrm{ML}}) - t_n\}^2 \tag{1.53}$$

を得る．この場合も，パラメータベクトル $\boldsymbol{w}_{\mathrm{ML}}$ を最初に決めて，そこから上の式の平均を計算することによって，精度パラメータ β_{ML} を求めることができる．これは単純なガウス分布の場合 (1.2.4 節) と同様である．

パラメータ $\boldsymbol{w}, \beta$ が決まれば，x の新たな値に対する予測ができる．確率モデルで定式化したので，それらは単なる点予測値ではなく，**予測分布** (predictive distribution) という形で t の確率分布を与えることができる．それは (1.50) 式を最尤パラメータで置き換えれば，

$$p(t \mid x, \boldsymbol{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}(t \mid y(x, \boldsymbol{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}) \tag{1.54}$$

という形で与えられる．

さて，多項式の係数 $\boldsymbol{w}$ に関する事前分布を導入し，よりベイズ的なアプローチに進むことにする．簡単のため，$\boldsymbol{w}$ に関する事前分布として，

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\boldsymbol{w}^\top \boldsymbol{w}\right\} \tag{1.55}$$

という形のガウス分布を考える．ここで α は分布の精度パラメータであり，$M+1$ は M 次多項式に対するベクトル $\boldsymbol{w}$ の要素数である．α のようにモデルパラメータの分布を制御するパラメータを**ハイパーパラメータ** (hyperparameter) と呼ぶ．ベイズの定理から，$\boldsymbol{w}$ の事後分布は事前分布と尤度関数との積に比例し，

$$p(\boldsymbol{w} \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t} \mid \mathbf{x}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \alpha) \tag{1.56}$$

となる．これで与えられたデータに基づいて最も確からしい $\boldsymbol{w}$ の値を見つける．言い換えれば，事後分布を最大化する $\boldsymbol{w}$ を決めることができる．このテクニックを**最大事後確率推定** (maximum posterior) あるいは単に **MAP** 推定と呼ぶ[*3]．(1.56) の対数を符号反転し，(1.52) および (1.55) における $\boldsymbol{w}$ に依存する項と組み合わせて，事後確率の最大値は，

$$\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{\alpha}{2} \boldsymbol{w}^\top \boldsymbol{w} = \beta \left(\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{1}{2} \frac{\alpha}{\beta} \|\boldsymbol{w}\|_2^2 \right) \tag{1.57}$$

の最小値として与えられることがわかる．したがって，事後分布の最大化は，前に (1.4) の形で用いた正則化された二乗和誤差の最小化と等価であることがわかる．なお，正則化パラメータは $\lambda = \alpha/\beta$ で与えられる．

1.2.6 ベイズ曲線フィッティング

事前分布 $p(\boldsymbol{w} \mid \alpha)$ を組み込んだものの，今のところ $\boldsymbol{w}$ の点推定を行なっているだけで，これはまだベイズ的な扱いとは言えない．完全なベイズアプローチでは，確率の加法・乗法定理を矛盾なく適用して，$\boldsymbol{w}$ のすべての値に関して積分する必要があることをこの後すぐに示す．そのような周辺化はパターン認識のベイズ手法の根幹になる．

曲線フィッティングの問題では，訓練データとして $\mathbf{x}$ と $\mathbf{t}$ が与えられ，新たなテスト点 x に対する値 t を予測することが目標である．したがって，予測分布 $p(t \mid x, \mathbf{x}, \mathbf{t})$ を評価してみたい．ここで，パラメータ α, β は固定されており事前にわかっているとする (後の章で，こうしたパラメータがベイズの枠組みでデータからどのように推論できるかを議論する)．

ベイズ的な扱いというのは，確率の加法・乗法定理を矛盾なく適用することに他ならない．すなわち，予測分布は，

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid x, \boldsymbol{w}) p(\boldsymbol{w} \mid \mathbf{x}, \mathbf{t}) d\boldsymbol{w} \tag{1.58}$$

という形で書ける．ここで，$p(t \mid x, \boldsymbol{w})$ は (1.50) で与えられ，α と β への依存は単純化のため省略して書いた．また，$p(\boldsymbol{w} \mid \mathbf{x}, \mathbf{t})$ はパラメータの事後分布で，(1.56) の右辺を規格化することにより求められる．3.3 節で示すように，曲線フィッティングの例のような問題では，この事後分布はガウス分布となり，解析的に求めることができる．同様に (1.58) の積分も解析的に解け，予測分布はガウス分布

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t \mid m(x), s^2(x)\right) \tag{1.59}$$

の形で与えられる．平均と分散は

$$m(x) = \beta \boldsymbol{\phi}(x)^\top \mathbf{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n, \tag{1.60}$$

$$s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^\top \mathbf{S} \boldsymbol{\phi}(x) \tag{1.61}$$

となり，行列 $\mathbf{S}$ は

$$\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^\top \tag{1.62}$$

[*3] MAP は **M**aximum **a** **p**osteriori の頭文字をとったものである．

で与えられる．ただし，$\mathbf{I}$ は単位行列で，$\phi_i(x) = x^i$ $(i = 0, \ldots, M)$ に対して $\boldsymbol{\phi}(x) = (\phi_0(x), \ldots, \phi_M(x))^\top$ である．

(1.59) における予測分布の分散や平均は x に依存していることがわかる．(1.61) の第 1 項は，t の予測値の目標変数のノイズによる不確実性を表しており，最尤推定の予測分布 (1.54) で β_{ML}^{-1} としたものに対応する．しかしながら，第 2 項はベイズ的な扱いによって出てきたもので，パラメータ $\boldsymbol{w}$ に対する不確実性である．

1.3 モデル選択

最小二乗法で多項式曲線をあてはめた例において，最も良い汎化を示した最適な次数の多項式があることを見た．多項式の次数はモデルの自由パラメータの数を制御し，したがってモデルの複雑さを支配する．正則化した最小二乗法では，正則化係数 λ もモデルの実質的な複雑さを制御しており，一方混合分布やニューラルネットワークといったより複雑なモデルにおいては，複雑さを支配する複数のパラメータがあり得る．実際の応用ではそのようなパラメータの値を決めなければならないが，その主な目的は通常，新たなデータに対して最も良い予測をすることである．さらに，与えられたモデル内の複雑さパラメータの適切な値を決めるのと合わせて，異なる型のモデルも考慮して，それぞれの応用ごとに最も良いモデルを見つけたい．

すでに見たように，最尤アプローチでは過学習の問題があるので，訓練集合に対する性能というのは未知データの予測性能の良い指標ではない．データが十分にあるときの単純なアプローチは，手持ちのデータの一部を使っていろいろなモデルを学習するか，あるいは 1 つのモデルの複雑さパラメータの値を変えるかした後，独立なデータでそれらを比較し，最も予測性能の良いものを選ぶというものである．この比較用のデータは**検証用集合** (validation set) と呼ばれる．限られたサイズのデータ集合を使ってモデルの設計を何度も繰り返すと検証用集合にも過学習してしまうことがあるので，3 番目の**テスト集合** (test set) を別に用意しておいて，選んだモデルの性能を最終的に評価する必要がある．

しかしながら，多くの応用では訓練とテストに使えるデータは限られており，良いモデルを作るためには得られたデータはできるだけたくさん訓練に使いたい．一方，検証用集合が小さいと予測性能の推定の誤差が大きくなる．このジレンマを解くために，**交差検証** (cross-validation) という方法がある．S 分割交差検証では，得られたデータのうち $(S-1)/S$ の割合部分を訓練に使いつつ，全データを性能の評価に使うことができる．データが特に少ないときには，データ点数を N としたときに $S = N$ と考えるのが妥当であり，これを **LOO 法** (1 個抜き法; leave-one-out method) と呼ぶ．

交差検証の大きな欠点の 1 つは，訓練を行わなければならない回数が S に比例して大きくなることであり，1 回の訓練自体に大きな計算量が必要な場合には問題である．さらに，交差検証のように性能を測るのに別のデータを使うテクニックでは，単独のモデルでも複数の複雑さパラメータを持つ場合に問題が出てくる (例えば複数の正則化パラメータを持つ場合)．そのようなパラメータの組み合わせをいろいろ試すには，最悪の場合，パラメータの数に対して指数関数的に訓練回数が増える可能性があるため，明らかにより良いアプローチが必要となる．理想的には，訓練データだけに依存し，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できるものが望ましい．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となる．

歴史的には，さまざまな「情報量規準」(information criterion) と呼ばれるものが提案されてきた．これは，より複雑なモデルによる過学習を避ける罰則項を足すことによって最尤推定のバイアスを修正しようというものである．例えば**赤池情報量規準** (Akaike information criterion) あるいは AIC (Akaike, 1974) は

$$\ln p(\mathcal{D} \mid \boldsymbol{w}_{\mathrm{ML}}) - M \tag{1.63}$$

という値が最大になるモデルを選ぶ．ここで，$p(\mathcal{D} \mid \boldsymbol{w}_{\mathrm{ML}})$ は最尤推定を行った場合の対数尤度で，M はモデルの中の可変パラメータの数である．これの変種に**ベイズ情報量規準** (Bayesian information criterion) あるいは **BIC** というものがあり，4.4.1 節で議論する．しかしながら，こうした規準はモデルパラメータの不確実性は考慮しておらず，実際には過度に単純なモデルを選ぶ傾向にある．そこで，3.4 節では複雑さに罰則を

与えるのに自然で理にかなった方法として，完全なベイズアプローチを採用する．

1.4 次元の呪い

多項式曲線フィッティングの例では 1 つだけの入力変数 x を考えた．しかしながら，パターン認識の実際の応用では多くの入力変数を持つ高次元空間を扱う場合がある．以下で議論するように，高次元空間の取り扱いは非常に難しい課題であり，パターン認識テクニックの設計に重要な影響を与える要因である．

高次元の問題について洞察を得るために，多項式曲線フィッティングの例 (1.1 節) に戻って，このアプローチを複数個の入力変数の場合に拡張した場合にどうなるかを見てみよう．入力変数が D 個あるとき，3 次までの多項式は一般に

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k \tag{1.64}$$

という形をとる．D が増えると独立な係数 (変数 x 内の置換対称性によってすべての係数が独立というわけではない) の数は D^3 に比例する．実際，データの複雑な依存関係を捉えるためには，より高次の多項式が必要となることもある．M 次の多項式では係数の数は D^M のように増える (**演習 1.16**)．これは指数的な増加ではなく，べき乗の増加であるものの，これでもまだ手に負えないほど速い増加であり，実際に使うには限界がある．

3 次元空間での生活経験を通じて形成された我々の幾何的直感は，高次元空間を考えるときには誤りに陥りやすい．簡単な例として，D 次元空間の半径 $r = 1$ の球を考え，$r = 1 - \epsilon$ と $r = 1$ の間にある体積の割合がどれだけになるかを考える．この割合を評価するために，まず，D 次元の半径 r の球の体積を考えると，それは r^D のスケールになることから

$$V_D(r) = K_D r^D \tag{1.65}$$

と書ける．ここで，定数 K_D は D のみに依存する (**演習 1.18**)．したがって，求める比は

$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = \frac{K_D - K_D(1-\epsilon)^D}{K_D} = 1 - (1-\epsilon)^D \tag{1.66}$$

で与えられる．これをいろいろな D の値に対して ϵ の関数としてプロットすると，大きな D では，この比が小さな ϵ に対しても 1 に近いことがわかる．したがって，高次元では球のほとんどの体積は表面に近い薄皮に集中している．

さらなる例として，パターン認識に直接関連する高次元ガウス分布の振る舞いを考える．直交座標を極座標に変換し，角度方向について積分すると，原点からの半径 r の関数として密度 $p(r)$ が得られる (**演習 1.20**)．したがって，$p(r)\delta r$ は半径 r の位置の暑さ δr の薄皮中の確率質量である．この分布の値をいろいろな D の値に対してプロットすると，D が大きくなるにつれて，ガウス分布の確率質量はある特定の半径における薄皮に集中することがわかる．

大きい次元の空間に伴う困難のことを**次元の呪い** (curse of dimensionality) (Bellman, 1961) と呼ぶことがある．本書では各種手法を図的に解説するのが特に容易なので入力次元が 1 次元または 2 次元での例示を多用する．しかしながら，低次元での直感が大きい次元に一般化できるとは限らないことに注意する必要がある．

次元の呪いはパターン認識の応用において重要な問題となることは確かではあるものの，高次元空間に有効な手法がないわけではない．この理由は 2 つある．1 つ目は，実データは多くの場合，実質的には低い次元の領域に入っており，さらに，特に目標変数の重要な変化が生じる方向というのは限定されることが多いことである．2 つ目は，実データが (少なくとも局所的には) 一般的に滑らかな性質を持っており，そのため大体において入力空間上での小さな変化は目標変数に関して小さい変化しか与えないので，局所的な内挿のような手法で入力変数の新たな値に対する予測をすることができることである．うまく働くパターン認識テクニックはこれらの性質の 1 つまたは両方を備えている．例えば，工場でベルトコンベアの上の 2 次元形状の物体をキャプ

チャした画像から，その向きを決めるという問題を考える．各画像は次元がピクセル数で決まる高次元空間上の点である．物体は画像中で異なった位置と方向を持ち得るので，画像の間には 3 つの自由度があり，画像の集合は高次元空間に埋め込まれた 3 次元の**多様体** (manifold) の上にある．物体の位置や方向とピクセル強度の間には複雑な関係があるので，この多様体は高度に非線形となる．画像を入力とし，物体の位置とは無関係に，その方向だけを出力するモデルを学習することが目標ならば，多様体の中の 1 自由度だけが重要な変量となる．

1.5　決定理論

1.5.1　誤識別率の最小化
1.5.2　期待損失の最小化
1.5.3　棄却オプション
1.5.4　推論と決定
1.5.5　回帰のための損失関数

1.6　情報理論

1.6.1　相対エントロピーと相互情報量

1.7　演習問題

1.1 **(基本)**　www 関数 $y(x, \boldsymbol{w})$ が多項式 (1.1) で与えられたときの (1.2) の二乗和誤差関数を考える．この誤差関数を最小にする係数 $\boldsymbol{w} = \{w_i\}$ は以下の線形方程式の解として与えられることを示せ．

$$\sum_{j=0}^{M} A_{ij} w_j = T_i. \tag{1.67}$$

ただし，

$$A_{ij} = \sum_{n=1}^{N} (x_n)^{i+j}, \qquad T_i = \sum_{n=1}^{N} (x_n)^i t_n. \tag{1.68}$$

ここで，下付き添え字の i や j は成分を表し，$(x)^i$ は x の i 乗を表す．

(解答)　(1.1) を (1.2) に代入すると，

$$E(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \left(\sum_{j=0}^{M} w_j (x_n)^j - t_n \right)^2 \tag{1.69}$$

となり，

$$\begin{aligned} \frac{\partial E(\boldsymbol{w})}{\partial w_i} &= \frac{1}{2} \sum_{n=1}^{N} \cdot 2 \left(\sum_{j=0}^{M} w_j (x_n)^j - t_n \right) \cdot (x_n)^i \\ &= \sum_{n=1}^{N} \sum_{j=0}^{M} w_j (x_n)^{i+j} - \sum_{n=1}^{N} (x_n)^i t_n \\ &= \sum_{j=0}^{M} \left(\sum_{n=1}^{N} (x_n)^{i+j} \right) w_j - \sum_{n=1}^{N} (x_n)^i t_n \\ &= \sum_{j=0}^{M} A_{ij} w_j - T_i \end{aligned} \tag{1.70}$$

を得る．これより，誤差関数を最小にする $\boldsymbol{w} = \{w_i\}$ は，線形方程式系

$$\frac{\partial E(\boldsymbol{w})}{\partial w_i} = 0 \iff \sum_{j=0}^{M} A_{ij} w_j = T_i \tag{1.71}$$

の解として与えられる. □

1.2 **(基本)** 正則化された二乗和誤差関数 (1.4) を最小にする係数 w_i が満たす, (1.67) に類似した線形方程式系を書き下せ.

(解答) 正則化された二乗和誤差関数 (1.4) は

$$\widetilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \frac{\lambda}{2}\|w\|_2^2 \tag{1.72}$$

ゆえ, 演習 1.1 の結果より

$$\begin{aligned}\frac{\partial \widetilde{E}(\boldsymbol{w})}{\partial w_i} &= \frac{\partial E(\boldsymbol{w})}{\partial w_i} + \frac{\partial}{\partial w_i}\frac{\lambda}{2}\|w\|_2^2 \\ &= \sum_{j=0}^{M} A_{ij}w_j - T_i + \lambda w_i \\ &= \sum_{j=0}^{M}(A_{ij} + I_{ij}\lambda)w_j - T_i\end{aligned} \tag{1.73}$$

を得る. ただし,

$$I_{ij} = \begin{cases}1, & i = j, \\ 0, & \text{otherwise}\end{cases} \tag{1.74}$$

である. これより, 正則化された二乗和誤差関数 (1.4) を最小にする係数 w_i が満たす線形方程式系は,

$$\frac{\partial \widetilde{E}(\boldsymbol{w})}{\partial w_i} = 0 \iff \sum_{j=0}^{M}(A_{ij} + I_{ij}\lambda)w_j = T_i \tag{1.75}$$

で与えられる.

1.3 **(標準)** 3 個の色分けされた箱 r (赤) , b (青) , g (緑) を考える. 箱 r には 3 個のりんご, 4 個のオレンジ, 3 個のライムが入っており, 箱 b には 1 個のりんご, 1 個のオレンジ, 0 個のライムが入っており, 箱 g には 3 個のりんご, 3 個のオレンジ, 4 個のライムが入っている. 箱を $p(r) = 0.2, p(b) = 0.2, p(g) = 0.6$ という確率でランダムに選び, 果物を箱から 1 個取り出す (箱の中のものは等確率で選ばれるとする) とき, りんごを選び出す確率を求めよ. また, 選んだ果物がオレンジであったとき, それが緑の箱から取り出されたものである確率はいくらか?

(解答) 省略.

1.4 **(標準)** www 連続変数 x 上で定義された確率密度 $p_x(x)$ を考える. $x = g(y)$ により非線形変換を施すと密度は (1.17) の変換を受ける. (1.17) を微分して, y に関する密度を最大にする位置 $\hat{y}$ と x に関する密度を最大にする位置 $\hat{x}$ とが, ヤコビ因子の影響により一般には単純な $\hat{x} = g(\hat{y})$ という関係にないことを示せ. これは確率密度の最大値が, (通常の関数と異なり) 変数の選択に依存することを示している. 線形変換の場合には最大値の位置が変数自身と同じ変換を受けることを確かめよ.

(解答)

1.5 **(基本)** (1.28) の定義を使って $\mathbb{V}[f(x)]$ が (1.29) を満たすことを示せ.

(解答) 期待値の線形性より明らか. □

1.6 **(基本)** 2 つの変数 x, y が独立なら, その共分散は 0 になることを示せ.

(解答) 独立性の定義より

$$\begin{aligned}\mathrm{Cov}[x, y] &= \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\ &= \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{x,y}(x, y)d\mu_x(x)d\mu_y(y) - \int_{\mathcal{X}} p_x(x)d\mu_x(x)\int_{\mathcal{Y}} p_y(y)d\mu_y(y) \\ &= 0 \qquad (\because p_{x,y}(x, y) = p_x(x)p_y(y))\end{aligned} \tag{1.76}$$

を得る. ただし. $d\mu_x(x)$ は x が離散型確率変数の場合は計数測度を表し, x が連続型確率変数の場合はルベーグ測度を表す. □

1.7 **(標準)** www この演習問題では，1 変数ガウス分布に関する規格化条件 (1.38) を証明する．このために，積分

$$I = \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2\right)dx \tag{1.77}$$

を考え，その 2 乗を

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right)dxdy \tag{1.78}$$

の形で書いて評価する．直交座標系 (x, y) から極座標 (r, θ) に変換し，$u = r^2$ を代入する．θ と u に関する積分を実行し，両辺の平方根をとることにより，

$$I = (2\pi\sigma^2)^{1/2} \tag{1.79}$$

が得られることを示せ．最後にこの結果からガウス分布 $\mathcal{N}(x \mid \mu, \sigma^2)$ が規格化されていることを示せ．

(解答) 直交座標 (x, y) に対して極座標変換 $x = r\cos\theta, y = r\sin\theta$ $(r \geq 0, 0 \leq \theta < 2\pi)$ を施すと，そのヤコビアンは

$$|J| = \begin{vmatrix} \partial x/\partial r & \partial x/\partial\theta \\ \partial y/\partial r & \partial y/\partial\theta \end{vmatrix} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r \tag{1.80}$$

となるので，$dxdy = rdrd\theta$ ゆえ

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}x^2 - \frac{1}{2\sigma^2}y^2\right)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}(x^2+y^2)\right)dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} r\exp\left(-\frac{1}{2\sigma^2}r^2\right)drd\theta = \int_0^{2\pi}\left[-\sigma^2\exp\left(-\frac{1}{2\sigma^2}r^2\right)\right]_{r=0}^{r=\infty} d\theta \\ &= 2\pi\sigma^2 \end{aligned} \tag{1.81}$$

となるので，両辺の平方根をとることにより (1.79) を得る．ガウス分布 $\mathcal{N}(x \mid \mu, \sigma^2)$ が規格化されていることは明らか． □

1.8 **(標準)** www 変数変換を使って 1 変数ガウス分布 (1.36) が (1.39) を満たすことを確かめよ．次に規格化条件

$$\int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)dx = 1 \tag{1.82}$$

の両辺を σ^2 に関して微分し，ガウス分布が (1.40) を満たすことを確かめよ．最後に，(1.41) が成り立つことを示せ．

(解答) まず，(1.26) より

$$\begin{aligned} \mathbb{E}[x] &= \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)xdx = \frac{1}{(2\pi\sigma)^{1/2}}\int_{-\infty}^{\infty} x\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}dx \\ &= \frac{1}{(2\pi\sigma)^{1/2}}\int_{-\infty}^{\infty} (y+\mu)\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}dy \qquad (y \coloneqq x-\mu) \\ &= \frac{1}{(2\pi\sigma)^{1/2}}\left\{\int_{-\infty}^{\infty} y\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}dy + \mu\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2\sigma^2}y^2\right\}dy\right\} \\ &= \frac{1}{(2\pi\sigma)^{1/2}}\cdot\mu(2\pi\sigma^2)^{1/2} \qquad \left(\because y\exp\left\{-\frac{1}{2\sigma^2}y^2\right\} \text{ は奇関数}\right) \\ &= \mu \end{aligned} \tag{1.83}$$

となり，(1.39) を得る．次に，ガウス分布が (1.40) を満たすことを確かめる．ガウス分布の規格化条件の両辺を σ^2 に関して微分すると，

$$\frac{\partial}{\partial\sigma^2}\int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2)dx = 0 \tag{1.84}$$

であり，ガウス分布 $\mathcal{N}(x \mid \mu, \sigma^2)$ は $(x, \sigma^2) \in \mathbb{R}^2$ において連続 $(*)$ であることに注意すると，

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) dx &= \int_{-\infty}^{\infty} \frac{\partial}{\partial \sigma^2} \mathcal{N}(x \mid \mu, \sigma^2) dx \qquad (\because \text{条件 } (*)) \\
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial \sigma^2} \left\{ \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \right\} dx \\
&= \int_{-\infty}^{\infty} \left\{ -\frac{1}{2}(2\pi\sigma^2)^{-3/2} \cdot 2\pi \cdot \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \right. \\
&\qquad \left. + \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \cdot \left(\frac{1}{2\sigma^4}(x-\mu)^2 \right) \right\} dx \\
&= -\frac{\pi}{(2\pi\sigma^2)^{3/2}} \int_{-\infty}^{\infty} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} dx \\
&\qquad + \frac{1}{2\sigma^4} \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{(2\pi\sigma^2)} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} dx \\
&= -\frac{\pi}{(2\pi\sigma^2)^{3/2}} \cdot (2\pi\sigma^2)^{1/2} \\
&\qquad + \frac{1}{2\sigma^4} \left(\mathbb{E}[x^2] - 2\mu\mathbb{E}[x] + \mu^2 \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu, \sigma^2) dx \right) \\
&= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\mathbb{E}[x^2] - \mu^2 \right) \qquad (\because (1.82))
\end{aligned} \tag{1.85}$$

ゆえ，

$$-\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \left(\mathbb{E}[x^2] - \mu^2 \right) = 0 \iff \mathbb{E}[x^2] = \mu^2 + \sigma^2 \tag{1.86}$$

となり，(1.40) が成り立つ．(1.41) については，(1.39) および (1.40) より明らか． □

1.9 **(基本)** www ガウス分布 (1.36) のモード (つまり分布が最大となる x の値) が μ で与えられることを示せ．同様に，多変量ガウス分布 (1.42) のモードは $\boldsymbol{\mu}$ で与えられることを示せ．

(解答) まず，ガウス分布 (1.36) については

$$\begin{aligned}
\frac{\partial}{\partial x} \mathcal{N}(x \mid \mu, \sigma^2) &= \frac{\partial}{\partial x} \left\{ \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} \right\} \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}} \left(-\frac{x-\mu}{\sigma^2} \right) \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}
\end{aligned} \tag{1.87}$$

ゆえ，$x \leq \mu$ で $\mathcal{N}(x \mid \mu, \sigma^2)$ は増加，$x \geq 0$ で $\mathcal{N}(x \mid \mu, \sigma^2)$ は減少するので，モードは μ で与えられる．次に，多変量ガウス分布 (1.42) のモードが $\boldsymbol{\mu}$ で与えられることを示す．多変量ガウス分布 (1.32) を $\boldsymbol{x}$ で微分すると，共分散 Σ は対称行列ゆえ，

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{x}} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) &= \frac{\partial}{\partial \boldsymbol{x}} \left\{ \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} \right\} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} \cdot \left(-\frac{1}{2} \cdot 2\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right) \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right\} \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})
\end{aligned} \tag{1.88}$$

となるので，モードは $\boldsymbol{\mu}$ で与えられる． □

1.10 **(基本)** www 2 変数 x, z が統計的に独立であるとする．それらの和の平均と分散が

$$\mathbb{E}[x+z] = \mathbb{E}[x] + \mathbb{E}[z] \tag{1.89}$$

$$\mathbb{V}[x+z] = \mathbb{V}[x] + \mathbb{V}[z] \tag{1.90}$$

を満たすことを示せ．

(解答)　(1.89) については期待値の線形性より明らか．(1.90) については

$$\begin{aligned}\mathbb{V}[x+z] &= \mathbb{E}[(x+z)^2] - \mathbb{E}[x+z]^2 \\ &= \mathbb{E}[x^2] + 2\mathbb{E}[xz] + \mathbb{E}[z^2] - (\mathbb{E}[x]^2 + 2\mathbb{E}[x]\mathbb{E}[z] + \mathbb{E}[z]^2) \\ &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 + \mathbb{E}[z^2] - \mathbb{E}[z]^2 \qquad (\because x, z \text{ が独立} \Rightarrow \mathbb{E}[xz] = \mathbb{E}[x]\mathbb{E}[z]) \\ &= \mathbb{V}[x] + \mathbb{V}[z]\end{aligned} \tag{1.91}$$

ゆえ成り立つ．　□

1.11 **(基本)**　対数尤度関数 (1.44) の μ と σ^2 に関する微分を 0 とおいて，(1.45) と (1.46) を確かめよ．

(解答)　まず，(1.45) については，

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x} \mid \mu, \sigma^2) = \frac{\partial}{\partial \mu} \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right\} = \frac{1}{\sigma^2} \sum_{n=1}^{N} (x_n - \mu) \tag{1.92}$$

より，

$$\frac{\partial}{\partial \mu} \ln p(\mathbf{x} \mid \mu, \sigma^2) \bigg|_{\mu=\mu_{\mathrm{ML}}} = 0 \iff \mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{1.93}$$

ゆえ成り立つ．次に，(1.46) については，

$$\begin{aligned}\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{x} \mid \mu_{\mathrm{ML}}, \sigma^2) &= \frac{\partial}{\partial \sigma^2} \left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \right\} \\ &= \frac{1}{2\sigma^4} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 - \frac{N}{2\sigma^2}\end{aligned} \tag{1.94}$$

より，

$$\frac{\partial}{\partial \sigma^2} \ln p(\mathbf{x} \mid \mu_{\mathrm{ML}}, \sigma^2) \bigg|_{\sigma^2=\sigma^2_{\mathrm{ML}}} = 0 \iff \sigma^2_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (x_n - \mu_{\mathrm{ML}})^2 \tag{1.95}$$

ゆえ成り立つ．　□

1.12 **(標準)**　www (1.39) と (1.40) を使って

$$\mathbb{E}[x_n x_m] = \mu^2 + I_{nm}\sigma^2 \tag{1.96}$$

を示せ．ただし，x_n と x_m は平均 μ，分散 σ^2 のガウス分布から生成されたデータ点を表し，I_{nm} は $n = m$ のとき $I_{nm} = 1$ でそれ以外は $I_{nm} = 0$ であるとする．これから，(1.47) と (1.48) を証明せよ．

(解答)　$n = m$ のときは，演習 1.8 の結果より，

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n^2] = \mu^2 + \sigma^2 \tag{1.97}$$

となり，(1.96) が成り立つ．$n \neq m$ のときは，$x_n, x_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(x \mid \mu, \sigma^2)$ より，

$$\mathbb{E}[x_n x_m] = \mathbb{E}[x_n]\mathbb{E}[x_m] = \mu^2 \tag{1.98}$$

となり，$n = m$ のときと同様に (1.96) が成り立つ．これを用いて (1.47) と (1.48) を示す．まず，(1.47) は

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^{N} x_n\right] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[x_n] = \frac{1}{N} \cdot N\mu = \mu \tag{1.99}$$

ゆえ成り立つ．次に，(1.48) は，

$$
\begin{aligned}
\mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2\right] = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbb{E}[x_n^2]-2\mathbb{E}[x_n\mu_{\mathrm{ML}}]+\mathbb{E}[\mu^2_{\mathrm{ML}}]\right)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2-2\mathbb{E}\left[x_n\frac{1}{N}\sum_{m=1}^{N}x_m\right]+\mathbb{E}\left[\frac{1}{N^2}\left(\sum_{n=1}^{N}x_n\right)^2\right]\right)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2-\frac{2}{N}\mathbb{E}\left[x_n^2+\sum_{m\neq n}x_nx_m\right]+\frac{1}{N^2}\mathbb{E}\left[\sum_{n=1}^{N}\sum_{m=1}^{N}x_nx_m\right]\right)\\
&= \frac{1}{N}\sum_{n=1}^{N}\Bigg(\mu^2+\sigma^2-\frac{2}{N}\left(\mathbb{E}[x_n^2]+(N-1)\mathbb{E}[x_nx_m]\right)\\
&\qquad\qquad +\frac{1}{N^2}\left(N\mathbb{E}[x_n^2]+N(N-1)\mathbb{E}[x_nx_m]\right)\Bigg) \qquad (n\neq m)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2-\frac{2}{N}\left(\mu^2+\sigma^2+(N-1)\mu^2\right)+\frac{1}{N}\left(\mu^2+\sigma^2+(N-1)\mu^2\right)\right)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\mu^2+\sigma^2-2\mu^2-\frac{2}{N}\sigma^2+\mu^2+\frac{1}{N}\sigma^2\right)\\
&= \frac{1}{N}\sum_{n=1}^{N}\left(\frac{N-1}{N}\right)\sigma^2 = \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned}
\tag{1.100}
$$

ゆえ成り立つ．□

参考文献

[1] Christopher.M.Bishop 著. 元田浩／栗田多喜夫／樋口知之／松本裕治／村田昇 監訳. **パターン認識と機械学習 上 – ベイズ理論による統計的予測**. 丸善出版, 2012.

[2] 久保川達也. **現代数理統計学の基礎**. 共立出版, 2017.