データ科学のための基礎数理 章末問題 (略解と解説)

第 1 章

1. $f(\boldsymbol{x}) = 7x_1^2 + 8x_1x_2 + 5x_2^2$ であるので，

$$f(\boldsymbol{x}) = (x_1, x_2) \begin{pmatrix} 7 & 4 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

である。行列 $\boldsymbol{A}$ を

$$\boldsymbol{A} = \begin{pmatrix} 7 & 4 \\ 4 & 5 \end{pmatrix}$$

とおくと，

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$$

と書くこともできる。

2. 勾配ベクトル $\nabla f(\boldsymbol{x})$ は，

$$\nabla f(\boldsymbol{x}) = 2\boldsymbol{A}\boldsymbol{x}$$

で与えられるので，点 $\boldsymbol{x}_0 = (1, 1)^T$ を代入すれば，

$$\nabla f(\boldsymbol{x}_0) = 2(11, 9)^T = (22, 18)^T$$

が得られる。

3.

$$\nabla f(\boldsymbol{x}) = \boldsymbol{b} + \boldsymbol{A}\boldsymbol{x}$$

4.

$$\frac{\partial f(\boldsymbol{x})}{\partial x_1} = 3x_1 + 3x_2x_3$$
$$\frac{\partial f(\boldsymbol{x})}{\partial x_2} = 3x_1x_3 + 6x_2$$
$$\frac{\partial f(\boldsymbol{x})}{\partial x_3} = 3x_1x_2 - 4x_3$$

より，点 $\boldsymbol{x}_0 = (1, 1, 1)^T$ における勾配ベクトルは

$$\nabla f(\boldsymbol{x}_0) = (6, 9, -1)^T$$

で与えられる。

5.

$$\begin{vmatrix} 1-\lambda & 1 \\ 1 & -1-\lambda \end{vmatrix} = \lambda^2 - 2 = 0$$

より，$\lambda = \pm\sqrt{2}$。

固有値 $\lambda_1 = \sqrt{2}$ のときの固有ベクトル $\boldsymbol{u}_1$ は

$$\boldsymbol{u}_1 = \left(\frac{\sqrt{2}+1}{\sqrt{4+2\sqrt{2}}}, \frac{1}{\sqrt{4+2\sqrt{2}}} \right)^T$$

固有値 $\lambda_2 = -\sqrt{2}$ のときの固有ベクトル $\boldsymbol{u}_2$ は

$$\boldsymbol{u}_2 = \left(\frac{-\sqrt{2}+1}{\sqrt{4-2\sqrt{2}}}, \frac{1}{\sqrt{4-2\sqrt{2}}} \right)^T$$

で与えられる。

6.

$$\begin{vmatrix} -\lambda & 1 & 1 \\ 1 & -\lambda & 1 \\ 1 & 1 & -\lambda \end{vmatrix} = (\lambda+1)^2(\lambda-2) = 0$$

より，$\lambda = -1$(重解), 2。固有値 $\lambda_1 = \sqrt{2}$ のときの固有ベクトル $\boldsymbol{u}_1$ は

$$\boldsymbol{u}_1 = \left(\frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)^T$$

固有値 $\lambda_2 = -1$ のときの固有ベクトルは，2 つの方向に取ることができ，

$$\boldsymbol{u}_2 = \left(\frac{-1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}}\right)^T$$

$$\boldsymbol{u}_3 = \left(\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0\right)^T$$

となる。

7.

$$\begin{vmatrix} 1-\lambda & -2 \\ -2 & 1-\lambda \end{vmatrix} = (\lambda-3)(\lambda+1) = 0$$

より，$\lambda = 3, -1$。固有値 $\lambda_1 = 3$ のときの固有ベクトル $\boldsymbol{u}_1$ は

$$\boldsymbol{u}_1 = \left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$$

固有値 $\lambda_2 = -1$ のときの固有ベクトル $\boldsymbol{u}_1$ は

$$\boldsymbol{u}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$$

となる。

8. 行列 $\boldsymbol{A}$ の対角成分が 0 であるので，2 次関数は

$$f(\boldsymbol{x}) = 5x_1^2 + 2x_2^2$$

のような標準形で与えられる。

9. 行列 $\boldsymbol{A}$ の固有値は $\lambda = 6, 1$，固有値 $\lambda_1 = 6$ に対応する固有ベクトルは

$$\boldsymbol{u}_1 = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)^T$$

固有値 $\lambda_2 = 1$ に対応する固有ベクトルは

$$\boldsymbol{u}_2 = \left(-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right)^T$$

で与えれる。これは，$\tilde{x}_1 = \boldsymbol{u}_1^T\boldsymbol{x}$，$\tilde{x}_2 = \boldsymbol{u}_2^T\boldsymbol{x}$ を用いて，

$$f(\boldsymbol{x}) = 6\tilde{x}_1^2 + \tilde{x}_2^2$$

のような標準形に変換できることを示している。

10. 最もシンプルな例は，

$$\boldsymbol{A} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

のようなものである。このとき，2 次関数は $f(\boldsymbol{x}) = x_1^2 - x_2^2$ となるので，明らかに最大値も最小値も持たない。これは，この関数が x_1 軸方向には下に凸であるので $x_1 = 0$ が最小値であるのに対し，x_2 軸方向には上に凸であるので $x_2 = 0$ が最大値になっているためである（このような点を鞍点と言う）。

対角成分が 0 ではないような 2 次関数で，このような形状になっているものを考えて見よ。

11. ラグランジュの未定乗数法により，

$$F(x,y) = xy + \lambda(2x + 2x - L)$$

を x, y で微分し，

$$\frac{\partial F(x,y)}{\partial x} = y + 2\lambda = 0$$

$$\frac{\partial F(x,y)}{\partial y} = x + 2\lambda = 0$$

を解けばよい。これらより $x = y$ が得られ，$2x + 2x - L = 0$ とから

$$x = y = L/4$$

が得られる。

12. ラグランジュの未定乗数法により，

$$F(p_1, p_2, \cdots, p_M) = -\sum_{j=1}^{M} p_j \log_2 p_j + \lambda(p_1 + p_2 + \cdots + p_M - 1)$$

を p_1,p_2,$\cdots$,p_M で微分すればよい。$j = 1, 2, \cdots, M$ の全てに対して

$$\frac{\partial F(p_1, p_2, \cdots, p_M)}{\partial p_j} = -\log_2 p_j - 1 + \lambda = 0$$

であることから，$p_1 = p_2 = \cdots = p_M$ となり，$p_1 + p_2 + \cdots + p_M - 1 = 0$ とから，

$$p_1 = p_2 = \cdots = p_M = \frac{1}{M}$$

が得られる。

13. ラグランジュの未定乗数法により，

$$F(x_1, x_2) = 2x_1 + x_2 + \lambda(x_1^2 + x_2^2 - 4)$$

を x_1 と x_2 で微分して 0 と置き，$x_1^2 + x_2^2 - 4 = 0$ と合わせて解けばよい。

14. (x_1, x_2, x_3) 空間上で，直方体の中心を原点 $\mathbf{0} = (0,0,0)^T$ とし，各辺が x_1 軸，x_2 軸，x_3 軸に並行になるように取っても一般性を失わない。$x_1 > 0$，$x_2 > 0$，$x_3 > 0$ の領域の頂点を $\boldsymbol{x} = (x_1, x_2, x_3)^T$ とおくと，この直方体の体積は

$$f(x_1, x_2, x_3) = 8x_1x_2x_3$$

で与えられることがわかる。また，この頂点が半径 r の球に内接していることから

$$x_1^2 + x_2^2 + x_3^2 = r^2$$

が成り立つ必要がある。ラグランジュの未定乗数法により，

$$F(x_1, x_2, x_3) = 8x_1x_2x_3 + \lambda(x_1^2 + x_2^2 + x_3^2 - r^2)$$

を x_1,x_2,x_3 で微分して 0 と置き，$x_1^2 + x_2^2 + x_3^2 = r^2$ と合わせて解けばよい。

第 2 章

1. 式 (2.1) の関数

$$f(\boldsymbol{x}) = c + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$$

を $\boldsymbol{x}$ で微分して $\mathbf{0}$ と置くことにより，

$$\boldsymbol{b} + \boldsymbol{A}\boldsymbol{x} = \mathbf{0}$$

が得られる。$\boldsymbol{A}$ が逆行列を持てば，その極値は

$$\boldsymbol{x} = -\boldsymbol{A}^{-1}\boldsymbol{b}$$

で与えられることがわかる。

2. 非線形関数 $f_{\boldsymbol{\theta}}(x)$ が $\boldsymbol{\theta}$ に関して微分可能であるとき，

$$E = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - f_{\boldsymbol{\theta}}(x_i)\right)^2$$

も $\boldsymbol{\theta}$ に関して微分可能となる。具体的には，

$$\frac{\partial E}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n}\left(y_i - f_{\boldsymbol{\theta}}(x_i)\right)\frac{\partial f_{\boldsymbol{\theta}}(x_i)}{\partial \boldsymbol{\theta}}$$

が得られる。この式を $\mathbf{0}$ と置いて解くことが難しい場合であっても，勾配法を用いてこの二乗誤差を極小化する $\boldsymbol{\theta}$ を探索することが可能である。

3. 勾配法は，ある適当な初期値から開始し，その勾配の情報を用いて探索を続ける方法である。関数の最大化の問題であれば，その関数の勾配によって関数の値が増大する方向へ解を更新する。逆に，関数の最小化問題であれば，その関数の勾配によって関数の値が減少する方向へ解を更新する。関数の極値では，関数の値がそれ以上増大しない，もしくは減少しないため，勾配法によって解が更新されなくなった時点で極値が得られていることが期待できる。

   ただし，極値でない地点でも勾配がゼロベクトルになっている場合も有り得るため，確実に極値を求められる保証はないが，現実問題では適当な改良手法によって

4. 最急降下法は，適当な初期値から開始して，目的関数 $f(\boldsymbol{x})$ の勾配ベクトル $\nabla f(\boldsymbol{x})$ を求め，最も目的関数の値が減少する方向（最急降下方向）$-\nabla f(\boldsymbol{x})$ に $\boldsymbol{x}$ を更新しながら極小解を探索する方法である。一方，確率的勾配降下法は，目的関数が

$$f(\boldsymbol{x}) = \sum_{i=1}^{n} f_i(\boldsymbol{x})$$

   のように異なる関数 $f_i(\boldsymbol{x})$ の和で与えられる場合に，最急降下方向ではなく，ある選ばれた i に対する $-\nabla f_i(\boldsymbol{x})$ の方向に $\boldsymbol{x}$ を更新しながら極小解を探索する方法である。

5. 勾配降下法における直線探索とは，探索方向が決まった後に，その方向にどれだけ更新をするのかを探索的に求める手続きである。

6. ステップ幅を固定する方法のメリットは，1 回あたりの更新にかかる計算量が少なくて済むことである。一方で，ステップ幅を適切に設定する必要があるが，探索の初期段階と最終段階で同じステップ幅が用いられるため，

7. 勾配降下法やニュートン法を適用するためには，最適化対象の関数の勾配ベクトルが計算できることが必要である。これは，最適化対象の関数が微分可能であることを意味する。とくに，ニュートン法では二次微分も必要である。

第 3 章

1. $\mathcal{B}$ は Ω 上の σ-集合族であるので，σ-集合族の 3 条件を満たす。すなわち，3 条件の (3) より，$A_1, A_2, \cdots \in \mathcal{B}$ ならば $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ である。

   一方，3 条件の (2) より，$B_i \in \mathcal{B}$ ならば $\bar{B}_i \in \mathcal{B}$ である。すなわち，$\cup_{i=1}^{\infty} B_i \in \mathcal{B}$ であることから，$\overline{\cup_{i=1}^{\infty} B_i} \in \mathcal{B}$ であり，

$$\overline{\cup_{i=1}^{\infty} B_i} = \cap_{i=1}^{\infty} \bar{B}_i$$

   より，$\cap_{i=1}^{\infty} \bar{B}_i \in \mathcal{B}$ も成り立つ。$A_i = \bar{B}_i$ と置けば題意が成り立つ。

2. $\mathcal{B}$ は Ω 上の σ-集合族であるので，σ-集合族の 3 条件より，$A_1, A_2, \cdots \in \mathcal{B}$ ならば

$$\cup_{k=n}^{\infty} A_k \in \mathcal{B}$$

が成り立つ。また，先の【 1 】の結果から，

$$\cap_{k=n}^{\infty} A_k \in \mathcal{B}$$

も成り立つ。$B_n = \cup_{k=n}^{\infty} A_k$，$C_n = \cap_{k=n}^{\infty} A_k$ と置けば，$B_1, B_2, \cdots \in \mathcal{B}$，$C_1, C_2, \cdots \in \mathcal{B}$ であることから，

$$\cup_{n=1}^{\infty} B_n \in \mathcal{B}$$

$$\cap_{n=1}^{\infty} C_k \in \mathcal{B}$$

も成り立つ。

3. 確率の 3 公理が成り立つとき，任意の事象 $A \in \mathcal{B}$ に対して $0 \leq P(A) \leq 1$ が成り立つ。一方，確率の 3 公理 (3) より，互いに排反な事象 A と B に対して，

$$P(A \cup B) = P(A) + P(B)$$

が成り立つ。$A = \Omega$，$B = \phi$ と置けば，これらの事象も排反であるため，

$$P(\Omega \cup \phi) = P(\Omega) + P(\phi)$$

となる。一方，確率の 3 公理 (2) より $P(\Omega) = 1$ であることから，

$$P(\Omega \cup \phi) = 1 + P(\phi)$$

となり，任意の事象 $A \in \mathcal{B}$ に対して $0 \leq P(A) \leq 1$ であることから

$$P(\phi) = 0$$

でなければならない。

4. サイコロを 1 つ振ったときに出る目が k の事象を A_k とすると，$A_1, A_2, \cdots, A_6$ の 6 通りの基本事象が定義でき，公正なサイコロであれば $P(A_k) = 1/6$ である。同様に，もう 1 つの公正なサイコロを 1 つ振ったときに出る目が k の事象を B_k とすると，$P(B_k) = 1/6$ である（$k = 1, 2, \cdots, 6$）。公正なサイコロを 2 つ振ったときに出る目の和が 7 となるのは，$A_1 \cap B_6$，$A_2 \cap B_5$，$A_3 \cap B_4$，$A_4 \cap B_3$，$A_5 \cap B_2$，$A_6 \cap B_1$ の 6 通りの場合である。

一方，A_k と B_j は互いに排反であることから

$$P(A_k \cap B_j) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

であり，$A_1 \cap B_6$，$A_2 \cap B_5$，$\cdots$，$A_6 \cap B_1$ も互いに排反であることから，公正なサイコロを 2 つ振ったときに出る目の和が 7 となる事象 N_7 の確率は

$$P(N_7) = P(A_1 \cap B_6) + P(A_2 \cap B_5) + \cdots + P(A_6 \cap B_1) = \frac{6}{36} = \frac{1}{6}$$

となる。

5. いま，各サイコロの出る目が，それぞれ i, j, k となる象を A_i，B_j，C_k とすると（$i, j, k = 1, 2, \cdots, 6$），

$$P(A_i \cap B_j \cap C_k) = \frac{1}{6^3} = \frac{1}{216}$$

である。

公正なサイコロを 3 つ振ったとき，それらの目の最大値が 4 となる確率を求めるためには，そのような場合の組み合わせを全て列挙し，場合の数を数え上げればよい。例えば，$A_1 \cap B_1 \cap C_4$ や $A_2 \cap B_4 \cap C_3$ などの場合に，それらの目の最大値が 4 となる。しかし，これらの事象を全て列挙して数え上げることはなかなか大変な作業である。

そこで次のように考えてみる。サイコロを振って出る目が 1 ～ 4 のいずれかである場合の組み合わせは，$A_1 \cap B_1 \cap C_1$ から $A_4 \cap B_4 \cap C_4$ までの 4^3 通りある。これらの中には出る目に 5 や 6 が含まれていないので，出る目の最大値は 4 以下である。これらの中には，出る目の最大値が 1 や 2，3 の場合も含まれるが，そのような 1 ～ 3 のいずれかである場合の組み合わせは，$A_1 \cap B_1 \cap C_1$ から $A_3 \cap B_3 \cap C_3$

までの 3^3 通りある。すなわち，サイコロを振って出る目が $1 \sim 4$ のいずれかであり，かつ 4 が 1 つ以上出ている組み合わせの数は $4^3 - 3^3$ 通りである。従って，求める確率は

$$\frac{4^3 - 3^3}{6^3} = \frac{64 - 27}{216} = \frac{37}{216}$$

となる。

6. サイコロを 1 つ振ったときに出る目が k の事象を A_k とすると“出る目が奇数である”という事象 $A_{\text{奇数}}$ は $A_{\text{奇数}} = A_1 \cup A_3 \cup A_5$ である。従って，求める条件付確率は

$$P(A_1|A_{\text{奇数}}) = \frac{P(A_1 \cap A_{\text{奇数}})}{P(A_{\text{奇数}})} = \frac{1/6}{1/2} = \frac{1}{3}$$

となる。

7. 3 人兄弟の性別は 2^3 通りの組み合わせがあるが，そのうち少なくとも 1 人が女の子である事象は，全体から全員男の子である場合を除いて $2^3 - 1$ 通りの組み合わせがある。このうち，2 人以上，女の子であるのは“女女女”,“女女男”,“女男女”,“男女女”の 4 通りなので，求める条件付確率は

$$P(\text{"2 人以上が女の子"}|\text{"少なくとも 1 人が女の子"}) = \frac{4}{2^3 - 1} = \frac{4}{7}$$

となる。

8. 3 つのコップのうち，赤玉が入っているコップは 1 つであるので，そのコップを言い当てる確率は 1/3 である。

9. 最初に自分が指定したコップの中に赤玉が入っている事象を A，それ以外の 2 つのコップに赤玉が入っている事象をそれぞれ B，C とする。$P(A) = P(B) = P(C) = 1/3$ である。また，最初に自分が指定したコップの中に赤玉が入っている確率は $P(A) = 1/3$ であり，それ以外のコップに赤玉が入っている確率は $P(B \cup C) = 2/3$ である。

   いま，相手がそれ以外の 2 つのコップのうち，赤玉が入っていない方のコップを開示したとする。これは，$\bar{B}$ または $\bar{C}$ のいずれかが生起していることを知ることを意味しているが，これらは必ずいずれかは起こるので $P(\bar{B} \cup \bar{C}) = 1$ である。従って，求める条件付確率は，

$$\begin{aligned} P(A|\bar{B} \cup \bar{C}) &= \frac{P(A \cap (\bar{B} \cup \bar{C}))}{P(\bar{B} \cup \bar{C})} \\ &= \frac{P(A)}{P(\bar{B} \cup \bar{C})} \\ &= \frac{1}{3} \end{aligned}$$

$$\begin{aligned} P(B \cup C|\bar{B} \cup \bar{C}) &= \frac{P\left((B \cup C) \cap (\bar{B} \cup \bar{C})\right)}{P\left(\bar{B} \cup \bar{C}\right)} \\ &= \frac{P(B \cup C)}{P(\bar{B} \cup \bar{C})} \\ &= \frac{2}{3} \end{aligned}$$

   となる。すなわち，最初に自分が指定したコップではなく，相手が外れを開示したもとで，その残った他方のコップに選び直した方が赤玉の確率は 2/3 と高くなることがわかる。

   これは，相手が 1 つの外れコップを開示したあとは，コップが 2 つになるため，どちらを選んでもあたりの確率は変わらないという直感を持つ人が多い問題で，確率論でよく知られるパラドックスの 1 つである。モンティ・ホール問題という名称で知られているので，興味のある読者は調べてみるとよい。

10. ”対象の病気に罹患している”という事象を A，”対象の病気に罹患していない”という事象を $\bar{A}$ とする。罹患率がは 0.0001 であるので，

$$P(A) = 0.0001, \qquad P(\bar{A}) = 0.9999$$

である。いま，”検査で陽性になる”という事象を B，”検査で陰性になる”という事象を $\bar{B}$ とすれば，この検査の誤り率は 0.01 であることから，

$$P(B|A) = 0.99, \qquad P(\bar{B}|A) = 0.01$$

$$P(B|\bar{A}) = 0.01, \qquad P(\bar{B}|\bar{A}) = 0.99$$

である。

ベイズの定理より，

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\bar{A})P(\bar{A})} \\ &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \\ &= \frac{0.000099}{0.000099 + 0.009999} = 0.009804 \end{aligned}$$

が得られる。

この結果から，この検査で陽性となっても，実際にこの被験者がこの病気に罹患している確率は 0.01 未満と非常に低いことが分かる。この検査の結果が正しい確率は 0.99 と高いにも関わらず，この検査で陽性となった被験者が実際に病気に罹患している確率が低いという事実は，多くの人の直感に反する結果であり，これも確率のパラドックスとしてよく知られている。ちなみに，検査の誤り率が 0.001，すなわち，検査の正解率が 99.9% という高精度な検査であったとしても，陽性となった被験者が実際に病気に罹患している確率は 0.0908 程度であり，やはり直感と比べて非常に低い値となる。

第 4 章

1. 二項定理とは，

$$(a+b)^n = \sum_{x=0}^{n} \frac{n!}{x!(n-x)!} a^{n-x} b^x$$

という公式である。$a = 1-\theta$，$b = \theta$ と置けば，$(a+b)^n = 1^n = 1$ となることから，

$$\sum_{x=0}^{n} \frac{n!}{x!(n-x)!} \theta^x (1-\theta)^{n-x} = 1$$

が得られる。

2. 二項確率分布は，

$$P(x) = \frac{n!}{(n-x)!x!} \theta^x (1-\theta)^{n-x} \qquad 0 \leq x \leq n$$

で与えられるので，平均は

$$\begin{aligned} E[X] &= \sum_{x=0}^{n} x \frac{n!}{(n-x)!x!} \theta^x (1-\theta)^{n-x} \\ &= \sum_{x=1}^{n} x \frac{n!}{(n-x)!x!} \theta^x (1-\theta)^{n-x} \\ &= \sum_{x=1}^{n} \frac{n!}{(n-x)!(x-1)!} \theta^x (1-\theta)^{n-x} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{x=1}^{n} \frac{n(n-1)!}{(n-x)!(x-1)!}\theta\theta^{x-1}(1-\theta)^{n-x} \\
&= n\theta\sum_{x=1}^{n} \frac{(n-1)!}{(n-1-(x-1))!(x-1)!}\theta^{x-1}(1-\theta)^{n-x} \\
&= n\theta\sum_{\tilde{x}=0}^{n-1} \frac{(n-1)!}{(n-1-\tilde{x})!\tilde{x}!}\theta^{\tilde{x}}(1-\theta)^{n-1-\tilde{x}} \\
&= n\theta
\end{aligned}
$$

と計算することができる。

一方，分散 $V[X]$ は $V[X] = E[X^2] - \{E[X]\}^2$ で与えられるので，$E[X^2]$ を計算してみる。

$$
\begin{aligned}
E[X^2] &= \sum_{x=0}^{n} x^2 \frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x} \\
&= \sum_{x=0}^{n} x(x-1)\frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x} + \sum_{x=0}^{n} x\frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x} \\
&= n(n-1)\sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!(x-2)!}\theta^x(1-\theta)^{n-x} + E[X] \\
&= n(n-1)\theta^2\sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!(x-2)!}\theta^{x-2}(1-\theta)^{n-x} + n\theta \\
&= n(n-1)\theta^2\sum_{\tilde{x}=0}^{n-2} \frac{(n-2)!}{(n-2-\tilde{x})!\tilde{x}!}\theta^{\tilde{x}}(1-\theta)^{n-2-\tilde{x}} + n\theta \\
&= n(n-1)\theta^2 + n\theta
\end{aligned}
$$

より，

$$
\begin{aligned}
V[X] &= E[X^2] - \{E[X]\}^2 \\
&= n(n-1)\theta^2 + n\theta - (n\theta)^2 \\
&= n\theta(1-\theta)
\end{aligned}
$$

が得られる。

3. 分散の定義 $V[X] = E[(X - E[X])^2]$ より，

$$
\begin{aligned}
V[aX+b] &= E[(aX + b - E[aX+b])^2] \\
&= E[(aX + b - aE[X] - b)^2] \\
&= E[(aX - aE[X])^2] \\
&= E[a^2(X - E[X])^2] \\
&= a^2E[(X - E[X])^2] = a^2V[X]
\end{aligned}
$$

4. 分散の定義 $V[X] = E[(X - E[X])^2]$ より，

$$
\begin{aligned}
V[X_1 - X_2] &= E[(X_1 - X_2 - E[X_1 - X_2])^2] \\
&= E[(X_1 - X_2 - E[X_1] + E[X_2])^2] \\
&= E[(X_1 - E[X_1] - (X_2 - E[X_2]))^2] \\
&= E[(X_1 - E[X_1])^2] - 2E[(X_1 - E[X_1])(X_2 - E[X_2])] + E[(X_2 - E[X_2])^2] \\
&= V[X_1] - 2\cdot 0 + V[X_2] = V[X_1] + V[X_2]
\end{aligned}
$$

が得られる。ただし，X_1 と X_2 が互いに独立であることから

$$E[(X_1 - E[X_1])(X_2 - E[X_2])] = 0$$

となることを用いた。

5. 区間 $[a,b]$ 上の一様分布の確率密度関数は，

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases}$$

で与えられるので，その平均値と分散は

$$\begin{aligned} E[X] &= \int_a^b x \frac{1}{b-a} dx \\ &= \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b \\ &= \frac{1}{b-a}(\frac{b^2}{2} - \frac{a^2}{2}) \\ &= \frac{1}{b-a}\frac{(b-a)(b+a)}{2} \\ &= \frac{a+b}{2} \end{aligned}$$

$$\begin{aligned} V[X] &= \int_a^b (x - E[X])^2 \frac{1}{b-a} dx \\ &= \frac{1}{b-a}\int_a^b (x - \frac{a+b}{2})^2 dx \\ &= \frac{1}{b-a}\left[\frac{\left(x - \frac{a+b}{2}\right)^3}{3}\right]_a^b \\ &= \frac{1}{b-a}\left(\frac{\left(b - \frac{a+b}{2}\right)^3}{3} - \frac{\left(a - \frac{a+b}{2}\right)^3}{3}\right) \\ &= \frac{1}{3(b-a)}\left\{\left(\frac{b-a}{2}\right)^3 - \left(\frac{a-b}{2}\right)^3\right\} \\ &= \frac{1}{3(b-a)}\frac{(b-a)^3}{4} = \frac{(b-a)^2}{12} \end{aligned}$$

6. 先の問題と同様の展開により，

$$E[X] = \frac{1}{2}$$

$$V[X] = \frac{1}{12}$$

で与えられる。

7.

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

に対し，I^2 を計算すると，

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2} - \frac{(y-\mu)^2}{2\sigma^2}\right\} dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} dxdy \end{aligned}$$

ここで，極座標変換 $x = r\sin\theta$， $y = r\cos\theta$ を用いると，$x^2 + y^2 = r^2$ であり，ヤコビアンが

$$\begin{vmatrix} r\cos\theta & \sin\theta \\ -r\sin\theta & \cos\theta \end{vmatrix} = r$$

となるので，

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left\{-\frac{x^2+y^2}{2\sigma^2}\right\} dxdy \\ &= \int_0^{\infty}\int_0^{2\pi} r\exp\left\{-\frac{r^2}{2\sigma^2}\right\} d\theta dr \\ &= \int_0^{2\pi} d\theta \int_0^{\infty} re^{-r^2/2\sigma^2} dr \\ &= 2\pi\left[-\sigma^2 e^{-\frac{r^2}{2\sigma^2}}\right]_0^{\infty} = 2\pi\sigma^2 \end{aligned}$$

したがって，

$$I = \sqrt{2\pi\sigma^2}$$

が得られる。

8. $X = a + bZ$ が従う分布は，平均 a，分散 b^2 の正規分布 $N(a, b^2)$ である。

9. $Y = 2X_1 - 3X_2$ が従う分布は，平均 $2\mu_1 - 3\mu_2$，分散 $4\sigma_1^2 + 9\sigma_2^2$ の正規分布 $N\left(2\mu_1 - 3\mu_2, 4\sigma_1^2 + 9\sigma_2^2\right)$ である。

第 5 章

1. 不偏推定量は，推定量の期待値が真のパラメータと一致している推定量をいう。一方，最尤推定量は尤度関数を最大化するパラメータで与えられる推定量である。

2. 確率 θ で 1 を，確率 $1-\theta$ で 0 を出力するモデルから，1 を y 個，0 を $n-y$ 個含むような 1 つの系列 x^n の出てくる確率は

$$P(x^n) = \theta^y(1-\theta)^{n-y}$$

である。もし，1 が y 回，0 が $n-y$ 回出現する確率 $P(y)$ を考えるならば，そのような系列の本数が ${}_nC_y$ 本あるので，

$$P(y) = {}_nC_y\theta^y(1-\theta)^{n-y}$$

となる（これがいわゆる二項分布の確率）。これらを θ の関数としてみたものを尤度関数という。尤度関数の増減表は，θ で微分して極値を求めればよい。

$$\begin{aligned} \frac{\partial l(\theta|x^n)}{\partial\theta} &= y\theta^{y-1}(1-\theta)^{n-y} - (n-y)\theta^y(1-\theta)^{n-y-1} \\ &= \left\{y(1-\theta) - (n-y)\theta\right\}\theta^{y-1}(1-\theta)^{n-y-1} \\ &= (y-n\theta)\,\theta^{y-1}(1-\theta)^{n-y-1} = 0 \end{aligned}$$

より，$y \geq 2$，$n-y \geq 2$ のとき $\theta = 0, 1, y/n$ で極値を取る。

θ の最尤推定量 $\hat{\theta}_M$ は尤度関数を最大化する推定量であるので，$\hat{\theta}_M = y/n$ で与えられる。一方，θ の不偏推定量 $\hat{\theta}_U$ とは，期待値が真の θ となっている，すなわち，

$$E\left[\hat{\theta}_U\right] = \theta$$

となるような推定量のことである。いま，1 回の試行で 1 が出る確率は θ であるので，n 回の試行における 1 の出現回数 k の期待値は $n\theta$ となる。従って，

$$\hat{\theta}_U = \frac{k}{n}$$

とすれば，明らかに，

$$E[\hat{\theta}_U] = E\left[\frac{k}{n}\right] = \frac{E[k]}{n} = \frac{n\theta}{n} = \theta$$

となるので，この $\hat{\theta}_U = k/n$ が θ の不偏推定量となる（結果として，最尤推定量と同じ）。

(1) $l(\theta) = \theta^7(1-\theta)^3$

(2) $\hat{\theta}_M = 7/10$

(3) $\hat{\theta}_U = 7/10$

(4) 増減表は，尤度関数を θ で微分して極値を求めればよい。

(注意) 尤度関数

$$l(\theta|x^n) = \theta^y(1-\theta)^{n-y}$$

は，データ x^n が与えられたもとでの θ の関数であることに注意しよう。この関数の概形は必ず描けるようにしておく必要がある。$l(\theta|x^n)$ を θ で微分して増減表を描けばよい。ただし，$y = 1$ や $n - y = 1$ の場合と，$y \geq 2$ かつ $n - y \geq 2$ の場合は分布の極値の個数が異なるので注意。

また，この尤度関数 $l(\theta|x^n)$ に対して，$f(\theta) \propto l(\theta|x^n)$ となるような θ の確率密度関数は，どのような形で与えられるだろうか。後の問題で，ベータ分布という分布を考える際に，その基準化項が

$$\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

で与えられることが出てくるが，この式を用いると上の疑問に答えることができる。この式から，容易に

$$\int_0^1 \theta^{y}(1-\theta)^{n-y}d\theta = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}$$

であることが分かる。従って，$f(\theta) \propto l(\theta|x^n)$ となるような θ の確率密度関数は，

$$f(\theta) = \frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)}\theta^y(1-\theta)^{n-y}$$

で与えられる。この分布の平均値は $\frac{y+1}{n+2}$，モードは $\frac{y}{n}$ で与えられる。モードは尤度関数を最大化する点と等しいので $\theta = \frac{y}{n}$ となっていることに注意しよう。一方，この確率密度関数の平均値は $E[\theta] = \frac{y+1}{n+2}$ となっているが，このように n 回中，y 回観測されたときに，

$$\hat{\theta} = \frac{y+1}{n+2}$$

とする推定量はしばしばラプラス推定量と呼ばれ，現実問題においてしばしば精度の高い推定量として用いられることがある。

3. 正規分布 $N(\mu, \sigma^2)$ に従う独立な確率変数 $X_1, X_2, \cdots, X_n$ については，定義より

$$E[X_i] = \mu$$

$$E[(X_i - \mu)^2] = \sigma^2$$

が成り立っている。また，n 個のサンプルの平均値 $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ について

$$E\left[\bar{X}\right] = \mu$$

も容易に導かれれる。また，その分散は

$$E\left[(\bar{X}-\mu)^2\right]=\frac{1}{n}\sigma^2$$

で与えられることも容易に導かれる。

(1) いま，問題のサンプル平均 $\bar{X}$ を真の平均値 μ で置き換えた場合を試しに考えてみると，

$$\begin{aligned}E\left[\sum_{i=1}^{n}(X_i-\mu)^2\right] &= \sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]\\ &= \sum_{i=1}^{n}\sigma^2\\ &= n\sigma^2\end{aligned}$$

が成り立つ。

(2) これは，真の平均値 μ をサンプルの平均 $\bar{X}$ で置き換えた場合の期待値である。計算してみると，

$$\begin{aligned}E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] &= E\left[\sum_{i=1}^{n}\left(X_i-\mu-(\bar{X}-\mu)\right)^2\right]\\ &= E\left[\sum_{i=1}^{n}\left\{\left(X_i-\mu\right)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2\right\}\right]\\ &= \sum_{i=1}^{n}E\left[\left(X_i-\mu\right)^2\right]-2\sum_{i=1}^{n}E\left[(X_i-\mu)(\bar{X}-\mu)\right]+E\sum_{i=1}^{n}\left[(\bar{X}-\mu)^2\right]\\ &= n\sigma^2-2\sum_{i=1}^{n}E\left[(X_i-\mu)(\bar{X}-\mu)\right]+n\cdot\frac{1}{n}\sigma^2\\ &= (n+1)\sigma^2-2\sum_{i=1}^{n}E\left[(X_i-\mu)(\bar{X}-\mu)\right]\\ &= (n+1)\sigma^2-2\sum_{i=1}^{n}E\left[(X_i-\mu)\left(\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right)\right]\\ &= (n+1)\sigma^2-2\sum_{i=1}^{n}E\left[(X_i-\mu)\frac{(X_1-\mu)+(X_2-\mu)+\cdots+(X_n-\mu)}{n}\right]\\ &= (n+1)\sigma^2-\frac{2}{n}\sum_{i=1}^{n}E\left[(X_i-\mu)^2\right]\\ &= (n+1)\sigma^2-\frac{2}{n}n\sigma^2\\ &= (n+1)\sigma^2-2\sigma^2\\ &= (n-1)\sigma^2\end{aligned}$$

が成り立つ。ただし，$i\neq j$ に対し X_i と X_j は独立であることから，$E[(X_i-\mu)(X_j-\mu)]=0$ であることを用いている。

(3) 最尤推定量は，尤度関数を最大化する推定量で与えられる。よって，

$$\begin{aligned}\hat{\mu}_M &= \frac{1}{n}\sum_{i=1}^{n}X_i\\ \hat{\sigma}^2_M &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\end{aligned}$$

となる。

(4) 不偏推定量は，期待値が母数と一致しているものを指す。不偏推定量は一意ではないが，最良な不偏推定量は以下で与えられる。

$$\begin{aligned}\hat{\mu}_U &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{\sigma}_U^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\end{aligned}$$

で与えられる。

4. 正規分布の密度関数は

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

で与えられるので，n 個のデータ $x_1, x_2, \cdots, x_n$ が与えられた時の最尤関数は，

$$\begin{aligned}L(\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\end{aligned}$$

となる。対数尤度関数は，

$$\begin{aligned}L(\mu, \sigma^2) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\end{aligned}$$

となるので，これを微分して 0 とおくことで，μ と σ^2 の最尤推定量が得られる。
簡単な計算により，μ と σ^2 の最尤推定量 $\hat{\mu}$，$\hat{\sigma}^2$ はそれぞれ

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

で与えれることは容易に確かめられる。
以上により，5 つのデータ

5.0, 8.0, 10.0, 12.0, 15.0

が得られたとき，これらが正規分布からのサンプルであると仮定した場合の平均値，分散の最尤推定量は，

$$\hat{\mu} = \frac{1}{5}(5.0 + 8.0 + 10.0 + 12.0 + 15.0) = 10.0$$

$$\hat{\sigma}^2 = \frac{1}{5}(5^2 + 2^2 + 0^2 + 2^2 + 5^2) = 11.6$$

で与えられる。

(注) 一般の統計で不偏分散が使われるのは，平均値もデータから推定しなければならない場合である。平均値はすでに分かっている状況で，分散を推定したいなら，既知の μ を用いて計算すればよいので，$n-1$ で割る必要はない。ちなみに，この $n-1$ は実質的なサンプル数のようなものを表わす量として自由度と呼ばれている。

5. $X_1, X_2, \cdots, X_n$ が独立に $N(\mu, \sigma^2)$ に従うとき，

$$\begin{aligned} E[X_i] &= \mu \\ V[X_i] &= E[(X_i - \mu)^2] = \sigma^2 \\ V[\alpha_i X_i] &= E[(\alpha_i X_i - \alpha\mu)^2] = \alpha_i^2 \sigma^2 \end{aligned}$$

である。

(1) すでに示したように，μ の最尤推定量 $\hat{\mu}_M$ は

$$\hat{\mu}_M = \frac{1}{n} \sum_{i=1}^{n} x_i$$

で与えれる。

不偏推定量としては，推定量の期待値が真のパラメータと一致していれば不偏であるので，多くの不偏推定量を作ることができる。例えば，母平均 μ の推定量として，

$$\hat{\mu}_U = X_1 + X_2 - X_3$$

や

$$\hat{\mu}_U = 3X_1 - X_2 - X_3$$

のような式を作っても，これらの期待値は μ になるので，いずれも不偏推定量であることに注意（ただし，バラツキが大きいことから良い推定量ではなく，現実には使われない)。一般に，$\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$ となる n 個のスカラを用いて，

$$\bar{X}_\alpha = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n$$

という推定量を構成すると，これは不偏推定量となる（次の問題参照)。

(2) σ^2 の最尤推定量 $\hat{\sigma}_M^2$ は

$$\hat{\sigma}_M^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

で与えれる。

σ^2 の不偏推定量 $\hat{\sigma}_U^2$ は，平均値の場合と同様に，例えば $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$ となる n 個のスカラを用いて，

$$\bar{X}_\alpha = \alpha_1 (x_1 - \mu)^2 + \alpha_2 (x_2 - \mu)^2 + \cdots + \alpha_n (x_n - \mu)^2$$

という推定量を構成すると，これは不偏推定量となる。

6. (1) $\bar{X}_\alpha$ の期待値が μ と一致すれば不偏推定量である。

$$\begin{aligned} E[\bar{X}_\alpha] &= E[\alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n] \\ &= \alpha_1 E[X_1] + \alpha_2 E[X_2] + \cdots + \alpha_n E[X_n] \\ &= \alpha_1 \mu + \alpha_2 \mu + \cdots + \alpha_n \mu \\ &= (\alpha_1 + \alpha_2 + \cdots + \alpha_n)\mu \\ &= \mu \end{aligned}$$

であることから，不偏推定量であることが示された。

(2) $\bar{X}_\alpha$ の分散は次式で与えられる。

$$\begin{aligned} V[\bar{X}_\alpha] &= V[\alpha_1 X_1] + V[\alpha_2 X_2] + \cdots + V[\alpha_n X_n] \\ &= \alpha_1^2 \sigma^2 + \alpha_2^2 \sigma^2 + \cdots + \alpha_n^2 \sigma^2 \\ &= (\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2)\sigma^2 \end{aligned}$$

従って，これを $\alpha_1 + \alpha_2 + \cdots + \alpha_n = 1$ の制約のもとで最小化すればよい。
あとはラグランジュの未定乗数法を用い，

$$L = (\alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2)\sigma^2 - \lambda(\alpha_1 + \alpha_2 + \cdots + \alpha_n - 1)$$

を，各 α_i で微分して 0 とおき，連立方程式を解けばよい。以下，計算のみであるため省略する（各自で計算してみよ)。

7. 既知の平均 μ, 分散 σ^2 で与えられる正規分布からデータ x が得られる場合を考えている。正規分布は，平均値，モード，メディアン共に μ で与えられる。

   (1) 0-1 損失，絶対誤差損失，二乗誤差損失に関するリスク関数を最小化する x の予測値 $\hat{x}$ は，それぞれ分布のモード，メディアン，平均値で与えられる。従って，正規分布の場合は全て μ である。

   (2) 上で示した $\hat{x} = \mu$ で予測した場合，平均の二乗予測誤差は

   $$E[(X - \hat{x})^2] = E[(X - \mu)^2] = \sigma^2$$

   で与えられる。

8. この確率分布のモードは $\arg\max_x P(x) = 2$，メディアンは $2 < x < 4$ の間の任意の値，平均値は ∞ である。

   従って，0-1 損失，絶対誤差損失，二乗誤差損失のそれぞれに対するリスク関数を最小化する x の予測値 $\hat{x}$ は，それぞれ

   $$\hat{x}_{0-1} = 2$$

   $$\hat{x}_{ab} = 3$$

   $$\hat{x}_{sq} = \infty$$

   などとなる。

   なお，$\hat{x}_{ab}$ は 3 でなくても $2 < \hat{x}_{ab} < 4$ であれば正解ではあるが，区間の中央が 3 であるので 3 とするのが最も好ましい。

9. 平均は，平均値の定義式

   $$\int_0^1 \theta f(\theta|\alpha,\beta)d\theta = \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha}(1-\theta)^{\beta-1}d\theta$$

   から計算を進めればよい。計算する過程において，

   $$\int_0^1 \theta^{\alpha}(1-\theta)^{\beta-1}d\theta = \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}$$

   と $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ の関係を使えば導くことができる。モードはベータ分布の密度関数 $f(\theta|\alpha,\beta)$ を θ で微分して 0 とおけばよい。

   (1) 実際に計算すれば，

   $$\begin{aligned}
   \int_0^1 \theta f(\theta|\alpha,\beta)d\theta &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int_0^1 \theta^{\alpha}(1-\theta)^{\beta-1}d\theta \\
   &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)} \\
   &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)} \\
   &= \frac{\alpha}{\alpha+\beta}
   \end{aligned}$$

   (2) ベータ分布の密度関数 $f(\theta|\alpha,\beta)$ を θ で微分して 0 とおくと，

   $$\begin{aligned}
   \frac{\partial f(\theta|\alpha,\beta)}{\partial\theta} &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left\{(\alpha-1)\theta^{\alpha-2}(1-\theta)^{\beta-1} - (\beta-1)\theta^{\alpha-1}(1-\theta)^{\beta-2}\right\} \\
   &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-2}(1-\theta)^{\beta-2}\left\{(\alpha-1) - (\alpha+\beta-2)\theta\right\} = 0
   \end{aligned}$$

   となるので，モードは

   $$\theta = \frac{\alpha-1}{\alpha+\beta-2}$$

   となる。

10. (1) $\alpha = \beta = 1$ のときは $(0, 1)$ 上の一様分布になる。

    (2)

    (3)

    (4)

11. この問題では，事前分布であるベータ分布のパラメータ α, β に特定の数値が与えられておらず，予測分布のモードは，これらの条件によって値が変わる点に注意。

    (1) 1 が 3 回，0 が 7 回生起しているとき，事後確率密度もベータ分布となり，

    $$f(\theta|x^{10}) = \frac{\Gamma(\alpha+\beta+10)}{\Gamma(\alpha+3)\Gamma(\beta+7)}\theta^{\alpha+2}(1-\theta)^{\beta+6}$$

    となる。

    (2) 二乗誤差損失 $L_{sq}(\hat{\theta}(x^{10}), \theta) = \left(\hat{\theta}(x^{10}) - \theta\right)^2$ を取った場合のベイズ最適な推定値 $\hat{\theta}_{sq}$ は，この事後確率密度の平均値で与えられるので，

    $$\hat{\theta}_{sq} = \frac{\alpha+3}{\alpha+\beta+10}$$

    となる（平均二乗誤差を最小化する推定値は，確率分布の平均値で与えられるため）。

    (3) 次のデータ $x = x_{11}$ の予測分布 $p(x|x^{10})$ は，

    $$p(x|x^{10}) = \begin{cases} \frac{\beta+7}{\alpha+\beta+10} & (x = 0 \text{ のとき}) \\ \frac{\alpha+3}{\alpha+\beta+10} & (x = 1 \text{ のとき}) \end{cases}$$

    で与えられる（0 と 1 においてのみ，確率を持つ分布であるので各自，概形を描いてみよ）。

    次のデータ $x = x_{11}$ の予測として，二乗誤差損失 $L(x_{11}, \hat{x}(x^{10}))) = (x_{11} - \hat{x}(x^{10}))^2$ をとった場合のベイズ最適な予測値は，予測分布 $p(x|x^{10})$ の平均値となる。従って，この場合のベイズ最適な予測は，

    $$\hat{x}_{sq} = \frac{\alpha+3}{\alpha+\beta+10}$$

    となる。

    一方，次のデータ $x = x_{11}$ の予測として，0-1 損失を考えた場合，モード (最頻値) がベイズ最適な予測値となる。すなわち，

    $$\hat{x}_{0-1} = \begin{cases} 0 & (\beta+4 > \alpha \text{のとき}) \\ 1 & (\beta+4 \leq \alpha \text{のとき}) \end{cases}$$

    のようになる。

    (4) $\alpha = \beta = 1$ を代入すればよい。二乗誤差損失をとった場合のベイズ最適な予測値は

    $$\hat{x}_{sq} = \frac{4}{12} = \frac{1}{3}$$

    となる。

    一方，0-1 損失をとった場合のベイズ最適な予測値は

    $$\hat{x}_{0-1} = 0$$

    となる。

12. 事前分布として一様分布が仮定されている。この一様分布は，実はベータ分布において $\alpha = \beta = 1$ とした特殊な場合であることに気づいていれば，先の問題とほぼ同じ展開になる。

    (1) 1 が 6 回，0 が 4 回出現しているので尤度は，

    $$\theta^6(1-\theta)^4$$

よって事後確率密度は，

$$f(\theta|x^{10}) \propto \theta^6(1-\theta)^4$$

であり，基準化項は

$$\int_0^1 \theta^6(1-\theta)^4 d\theta = \frac{\Gamma(7)\Gamma(5)}{\Gamma(12)}$$

となるので，事後確率密度は，

$$f(\theta|x^{10}) = \frac{\Gamma(12)}{\Gamma(7)\Gamma(5)}\theta^6(1-\theta)^4$$

となる。

(解説) これがやはりベータ分布になっていることに注意！ $\alpha = 5, \beta = 7$ とすればベータ分布の形になっていることが分かるであろう。一様分布はベータ分布に含まれる特殊系と考えられるので，事後密度はやはりベータ分布になって当然である。また，事後確率

$$f(\theta|x^{10}) = \frac{\Gamma(12)}{\Gamma(7)\Gamma(5)}\theta^6(1-\theta)^4$$

の平均値は 7/12，モードは 6/10 で与えられる。

(2) 次のデータ $x = x_{11}$ の予測分布 $p(x|x^{10})$ は，

$$p(x|x^{10}) = \begin{cases} \frac{5}{12} & (x = 0 \text{ のとき}) \\ \frac{7}{12} & (x = 1 \text{ のとき}) \end{cases}$$

で与えられる。よって，二乗誤差損失を適用した場合のベイズ最適な予測は，この予測分布の平均値になるので，

$$\hat{x}_{sq} = \frac{7}{12}$$

となる。

(3) 0-1 損失を適用した場合のベイズ最適な予測は，予測分布のモードになるので，

$$\hat{x}_{0-1} = 1$$

となる。

13. 自然共役事前分布は，共役事前分布とも呼ばれる。x が 0, 1 の 2 元の場合と同じように考えればよい。

独立に確率 θ_1 で 1 を，確率 θ_2 で 2 を，確率 $1-\theta_1-\theta_2$ で 3 が生起する確率分布モデルを考えたとき，長さ n の系列 x^n 中の 1 の出現回数 n_1, 2 の出現回数 n_2, 3 の出現回数 n_3 を用いて尤度関数は，

$$(\theta_1)^{n_1}(\theta_2)^{n_2}(1-\theta_1-\theta_2)^{n_3}$$

で与えられる。事後確率密度は，

$$\begin{aligned} f(\theta_1,\theta_2|x^n) &\propto p(x^n|\theta)f(\theta_1,\theta_2) \\ &= (\theta_1)^{n_1}(\theta_2)^{n_2}(1-\theta_1-\theta_2)^{n_3}f(\theta_1,\theta_2) \end{aligned}$$

の形で与えられるので，ベータ分布の時と同じように，事後確率密度 $f(\theta_1,\theta_2|x^n)$ と 事前確率密度 $f(\theta_1,\theta_2)$ が同じ関数の形であるためには，

$$f(\theta_1,\theta_2) \propto (\theta_1)^{\alpha-1}(\theta_2)^{\beta-1}(1-\theta_1-\theta_2)^{\gamma-1}$$

のような形にすると，

$$f(\theta_1,\theta_2|x^n) \propto (\theta_1)^{n_1+\alpha-1}(\theta_2)^{n_2+\beta-1}(1-\theta_1-\theta_2)^{n_3+\gamma-1}$$

のようになって共役事前分布の性質をみたすことが分かる。

さらに基準化項を求めると，

$$\int\int (\theta_1)^{\alpha-1}(\theta_2)^{\beta-1}(1-\theta_1-\theta_2)^{\gamma-1}d\theta_1 d\theta_2 = \frac{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}{\Gamma(\alpha+\beta+\gamma)}$$

のようになるので，

$$f(\theta_1,\theta_2) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)}(\theta_1)^{\alpha-1}(\theta_2)^{\beta-1}(1-\theta_1-\theta_2)^{\gamma-1}$$

が得られる。これはディリクレ分布と呼ばれる確率分布である。

14. 正規分布 $N(\mu,1^2)$ の確率密度関数は，

$$f(x|\mu) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x-\mu)^2}{2}\right\}$$

で与えられる。
観測データ $x^n = x_1x_2\cdots x_n$ が与えられたときの事後確率密度関数は，

$$\begin{aligned} f(\mu|x^n) &\propto \prod_{j=1}^{n} f(x_j|\mu)f(\mu) \\ &= \prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x_j-\mu)^2}{2}\right\}f(\mu) \\ &= \frac{1}{\sqrt{2\pi}^n}\prod_{j=1}^{n}\exp\left\{-\frac{(x_j-\mu)^2}{2}\right\}f(\mu) \end{aligned}$$

のように与えられるので，$f(\mu)\propto\exp\left\{-\frac{(\mu-\eta)^2}{2\tau^2}\right\}$ のような形式であれば，$f(\mu|x^n)$ も同型の確率密度関数になることがわかる。
実際，

$$f(\mu) = \frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{(\mu-\eta)^2}{2\tau^2}\right\}$$

とし，$n\bar{x} = \sum_{j=1}^{n} x_j$ であることを利用して，μ に比例する項を展開すれば

$$\begin{aligned} f(\mu|x^n) &\propto \prod_{j=1}^{n} f(x_j|\mu)f(\mu) \\ &= \prod_{j=1}^{n}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(x_j-\mu)^2}{2}\right\}\frac{1}{\sqrt{2\pi}\tau}\exp\left\{-\frac{(\mu-\eta)^2}{2\tau^2}\right\} \\ &\propto \exp\left\{-\sum_{j=1}^{n}\frac{(x_j-\mu)^2}{2}-\frac{(\mu-\eta)^2}{2\tau^2}\right\} \\ &= \exp\left\{-\sum_{j=1}^{n}\frac{x_j^2}{2}+\mu\sum_{j=1}^{n}x_j-\frac{n\mu^2}{2}-\frac{\mu^2-2\eta\mu+\eta^2}{2\tau^2}\right\} \\ &\propto \exp\left\{n\bar{x}\mu-\frac{n}{2}\mu^2-\frac{1}{2\tau^2}\mu^2-2\frac{\eta}{2\tau^2}\mu\right\} \\ &\propto \exp\left\{-\left(\frac{n}{2}+\frac{1}{2\tau^2}\right)\mu^2-2\left(\frac{n}{2}\bar{x}+\frac{\eta}{2\tau^2}\right)\mu\right\} \\ &= \exp\left\{-\left(\frac{n\tau^2+1}{2\tau^2}\right)\left(\mu^2-2\frac{n\tau^2\bar{x}+\eta}{n\tau^2+1}\mu\right)\right\} \end{aligned}$$

が得られる。これは，μ に関する正規分布の関数になっていることがわかる。

結局，μ の事後分布も正規分布となり，

$$\mu \sim N\left(\frac{n\tau^2\bar{x}+\eta}{n\tau^2+1}, \frac{\tau^2}{n\tau^2+1}\right)$$

という正規分布に従うことがわかる。

以上により，正規分布の平均値パラメータに対する共役事前分布はやはり正規分布であることがわかる。

第 6 章

1. 階層モデル族とは，複数のパラメトリックモデル族の集合で表され，かつ，それらの間に階層関係があるモデル族を指す。

   例えば，説明変数の集合 $\{X_1, X_2, \cdots, X_R\}$ と目的変数 Y の線形回帰モデル（重回帰モデル）を考えると，実際の解析において，どの説明変数が効いているかは未知であり，説明変数の選択が行われる。この場合，説明変数の部分集合と目的変数の線形回帰モデルを一つのモデル m と考え，各モデルを

   $$\begin{aligned}
   m_0 &: \quad Y = \alpha_0 + \epsilon, \\
   m_1 &: \quad Y = \alpha_0 + \alpha_1 X_1 + \epsilon, \\
   m_2 &: \quad Y = \alpha_0 + \alpha_1 X_2 + \epsilon, \\
   \cdots\cdots & \qquad \cdots\cdots\cdots\cdots \\
   m_{2^R-1} &: \quad Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_R X_R + \epsilon,
   \end{aligned}$$

   と表せば，これらのモデルには入れ子構造が存在するので，階層モデル族となる。

   例えば，モデル m_1 や m_2 において $\alpha_1 = 0$ とすればモデル m_0 と同様になるので，モデル m_0 はモデル m_1 や m_2 の中に完全に含まれることになる。

2. 階層モデル族の例は，実際のデータ解析で用いられる多くのモデルで確認することができる。例えば，次のような階層モデル族の例が挙げられる。これらについて，実際にモデル m_0,m_1,m_2,$\cdots$ の例を作ってみよ。

   (1) 時間 t と共に得られる時系列データ $X_1, X_2, \cdots$ に対して $X_t = \beta_1 X_{t-1} + \beta_2 X_{t-2} + \cdots + \beta_K X_{t-K} + \epsilon$ という関係を仮定したモデルを自己回帰モデルという（$E[X_t] = 0$ が仮定されている）。この自己回帰モデルについても，重回帰モデルと同様に入れ子構造を確認することができる。

   (2) 目的変数 Y と説明変数 X の間に $Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \cdots + \alpha_K X^K + \epsilon$ という関係を仮定したモデルを多項式回帰モデルという。この多項式回帰モデルについても，重回帰モデルと同様に入れ子構造を確認することができる。

   (3) 機械学習でよく用いられる階層ニューラルネットワークモデルにおいて，中間層のユニット数（ニューロン数）を変化させると，中間層のユニット数が少ないモデルは，多いモデルの一部のウェイトパラメータを 0 とすると表現できるため，入れ子構造があり，階層モデル族の一種である。

   (4) 機械学習でよく用いられる決定木モデルにおいて，木の最大深さを K と制限すると，これは 1 つの確率モデルとなる。木の最大深さを増やすと，例えば木の最大深さ K のモデルは，木の最大深さが $K+1$,$K+2$,$\cdots$ のモデルに含まれるので，入れ子構造があり，階層モデル族であることがわかる。

3. 階層モデル族とは，モデルのパラメータ数が増えるとその表現能力が豊かになっていくモデルであり，複雑なモデルがよりシンプルなモデルを包含するという特徴を持つ。

4. 一般に，パラメータ数が大き過ぎると，限られたサンプル数（学習データ数）からたくたんのパラメータを推定しなければならないため，推定精度が悪化してしまう。

5. モデルがシンプル過ぎると，データに存在する統計的性質を表現できなくなってしまう。

6. モデル m_1 は男女区別しないので，θ の最尤推定量は

$$\hat{\theta} = \frac{60}{100} = \frac{3}{5}$$

一方，モデル m_2 は男女別に比率を推定するモデルなので，

$$\hat{\theta}_M = \frac{30}{60} = \frac{1}{2}$$

$$\hat{\theta}_F = \frac{30}{40} = \frac{3}{4}$$

7. モデル m_1 のパラメータ数は 1，モデル m_2 のパラメータ数は 2 である。AIC 基準の定義により，

$$\begin{aligned} AIC(m_1) &= \log \left(\frac{3}{5}\right)^{60} \left(1-\frac{3}{5}\right)^{40} - 1 \\ &= 60(\log 3 - \log 5) + 40(\log 2 - \log 5) - 1 \\ &= 60\log 3 + 40\log 2 - 100\log 5 - 1 \\ &= 60 \times 1.099 + 40 \times 0.693 - 100 \times 1.609 - 1 = -68.24 \end{aligned}$$

$$\begin{aligned} AIC(m_2) &= \log \left(\frac{1}{2}\right)^{30} \left(1-\frac{1}{2}\right)^{30} \left(\frac{3}{4}\right)^{30} \left(1-\frac{3}{4}\right)^{10} - 2 \\ &= -60\log 2 + 30\log 3 - 40\log 4 - 2 \\ &= -60\log 2 + 30\log 3 - 80\log 2 - 2 \\ &= 30\log 3 - 140\log 2 - 2 \\ &= 30 \times 1.099 - 140 \times 0.693 - 2 = -66.05 \end{aligned}$$

従って，AIC 基準では，モデル $m2$ の方が値が大きいため，モデル m_2 の方が適切であると選択される。

8. BIC 基準，MDL 基準は，ペナルティ項が $\frac{k}{2}\log n$ に変更されるだけ。

$$\begin{aligned} BIC(m_1) &= 60\log 3 + 40\log 2 - 100\log 5 - \frac{1}{2}\log 100 \\ &= 60 \times 1.099 + 40 \times 0.693 - 100 \times 1.609 - \frac{1}{2} \times 4.605 = -69.54 \end{aligned}$$

$$\begin{aligned} BIC(m_2) &= 30\log 3 - 140\log 2 - \frac{2}{2} \times \log 100 \\ &= 30 \times 1.099 - 140 \times 0.693 - 4.605 = -68.655 \end{aligned}$$

やはり，BIC 基準でも，モデル $m2$ の方が値が大きいため，モデル m_2 の方が適切であると選択される。

9. 説明変数は X_1, X_2, X_3 の 3 つであるので，これらのどれをモデルに取り込むかによって $2^3 = 8$ 通りのモデルが構成できる。全て列挙すると以下のようになる。

$$\begin{aligned} m_0 &: \quad Y = \alpha_0 + \varepsilon, \\ m_1 &: \quad Y = \alpha_0 + \alpha_1 X_1 + \varepsilon, \\ m_2 &: \quad Y = \alpha_0 + \alpha_2 X_2 + \varepsilon, \\ m_3 &: \quad Y = \alpha_0 + \alpha_3 X_3 + \varepsilon, \\ m_4 &: \quad Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon, \\ m_5 &: \quad Y = \alpha_0 + \alpha_1 X_1 + \alpha_3 X_3 + \varepsilon, \\ m_6 &: \quad Y = \alpha_0 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon, \\ m_7 &: \quad Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \varepsilon, \end{aligned}$$

10. モデル m_7 において $\alpha_1 = 0$ とするとモデル m_6 になり，$\alpha_2 = 0$ とするとモデル m_5 になり，$\alpha_3 = 0$ とするとモデル m_4 になる。従って，モデル m_4,m_5,m_6 はモデル m_7 に含まれる。

    さらに，モデル m_6 において，$\alpha_2 = 0$ とするとモデル m_3 になり，$\alpha_3 = 0$ とするとモデル m_2 になる。モデル m_5 において，$\alpha_1 = 0$ とするとモデル m_3 になり，$\alpha_3 = 0$ とするとモデル m_1 になる。モデル m_4 において，$\alpha_1 = 0$ とするとモデル m_2 になり，$\alpha_2 = 0$ とするとモデル m_1 になる。従って，モデル m_1 はモデル m_4,m_5 に含まれ，モデル m_2 はモデル m_4,m_6 に含まれ，モデル m_3 はモデル m_5,m_6 に含まれる。

    同様に，モデル m_0 はモデルモデル m_1,m_2,m_3 に含まれる。上記の関係性から，モデル m_0 はモデルモデル m_4,m_5,m_6,m_7 に含まれることもわかる。

    以上が，この重回帰モデルの入れ子構造がであり，これらの順序は半順序関係であることも分かる。

11. 重回帰モデルは典型的な半階層構造を持つモデルクラスの一例である。

    (1) 重回帰モデル m_2 においてパラメータ $\beta_2 = 0$ とおくと，単回帰モデル m_1 と同等になってしまう。従って，重回帰モデル m_2 は単回帰モデル m_1 を完全に含む確率モデルとなっており，これをモデルの階層構造（入れ子構造）と言う。

    (2) 原理的には，重回帰モデル m_2 は単回帰モデル m_1 を完全に含むので，重回帰モデル m_2 を用意しておけば，モデルの表現能力の面からすれば十分である。

    もし，学習データが無限個あるような理想状態を考えれば，得られた無限個のデータから，未知の回帰係数である β_0，β_1，β_2 は完全に正しい値が推定できると考えられる。この場合には，重回帰モデル m_2 を用いれば，全ての β_0，β_1，β_2 に対して推定が可能であり，例えデータが単回帰モデル m_1 から発生していたとしても，β_2 の推定量が $\hat{\beta}_2 = 0$ となるだけである。

    逆に，単回帰モデル m_1 を用いた場合には，$\beta_2 \neq 0$ である重回帰モデル m_2 からデータが発生している場合には，例えデータが無限個あったとしても，この線形構造を正しく推定することはできない。

    (3) 有限個の学習データの場合，単回帰モデル m_1 から発生したデータに対しては，同じ単回帰モデル m_1 を用いて推定する方が精度が良い。

    同じ学習データから単回帰モデル m_1 は β_0，β_1, σ^2 の 3 つのパラメータを推定するが，重回帰モデル m_2 では β_0，β_1, β_2, σ^2 の 4 つのパラメータを推定しなければならない。確かに，重回帰モデル m_2 は単回帰モデル m_1 を含むので，モデルの表現能力としては m_2 で十分であるが，逆に推定すべきパラメータ数が増えてしまうので，推定精度が落ちてしまう。

    一般に“統計”では，モデルが固定であれば「学習データ数 n が多くなればなるほど，推定量は真の値に近づく」ことがいえる。データ数の方が固定であるならば「より未知パラメータ数の少ない，シンプルなモデルの方が推定精度が高い」と言える。

    (4) 逆に，重回帰モデル m_2 からデータが発生している場合には，$\beta_2 \neq 0$ のケースを考える必要がある。

    この場合，$|\beta_2|$ がある程度大きく，かつデータ数 n が十分大きければ，重回帰モデル m_2 を用いて推定する方が良くなると考えられる。逆に，$\beta_2 \approx 0$ であり，データ数 n もあまり大きくない場合には，単回帰モデル m_1 で推定した方が，推定するパラメータ数が少なくて済むため，精度が高くなる場合がある。このような場合には，AIC や BIC のようなモデル選択基準を用いて，適切なモデルを選択する必要がある。

    (注) 例えば，真のモデルが

$$\begin{aligned} Y &= 100 + 100X_1 + 0.01X_2 + \epsilon \\ \epsilon &\sim N(0, 10^2) \end{aligned}$$

    であり，これに従って，10 個のデータが得られていた場合を考えてみよう。得られる 10 個のデータには，標準偏差が 10 の誤差が乗っているので，これを同じ重回帰モデル

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

    を使って推定しようとしても β_2 の推定値が $\hat{\beta}_2 \approx 0.01$ となることは，ほぼ期待できない。この場合，たった 10 個しかないデータから β_2 を推定しようとするよりも，単回帰モデル

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

$$\epsilon \quad \sim \quad N(0, \sigma^2)$$

を仮定して，β_0 や β_1 を推定した方が，推定する未知パラメータの数が減るので，推定精度が向上することになる。

どちらのモデルが良いかについては「データの統計構造の複雑さ」と「データ数」の兼ね合いで決まる事に注意しよう。データ数が膨大なのであれば，複雑なモデルも推定できる。上の例でも，データ数が数億個という単位であれば，β_2 の推定値はほぼ正しく $\hat{\beta}_2 \approx 0.01$ という値が得られるはずである。その場合には，単回帰よりも，重回帰モデルを採用した方が良いであろう。

12.

第 7 章

1. 統計的決定では，決定を行う前にデータを観測することができ，その結果に基づいて決定を行う。これは，決定関数 $d(x)$ はデータの空間 $\mathcal{X}$ から決定の空間 $\mathcal{A}$ への写像を表す関数であるということもできる。合理的な決定を行うためには，決定の良し悪しを測る何らかの規準が必要であり，これを損失関数という。さらに，確率的な不確実性が存在する状況下では，この損失関数を平均化した平均損失を最小化する必要があり，これをリスク関数という。

   リスク関数を最小化できる状況では問題は単純であるが，一般に統計的決定が対象とする問題は，真の確率的構造が未知である状況が想定されており，それゆえにデータを観測して推測を行うことになる。この未知である真の確率的構造をパラメータ θ で定義した場合に，適切な決定を行うことが問題となる。

   ベイズリスクとは，リスク関数が真の状況（パラメータ）θ に依存するときに，このパラメータ θ に対して事前確率分布 $p(\theta)$ を仮定し，この事前確率分布でリスク関数を平均化したものとして定義される。一般にベイズ最適な決定は，このベイズリスクを最小化する決定で求めることができる。

2. 統計的決定では，問題設定から対象の確率的構造を決めるパラメータ θ が未知である状況で決定関数 $d(X)$ を選ばなければならない。その際，意思決定者側にとって最悪の状況を想定し，その最悪のリスクを最小化する決定をミニマックス決定という。

   ベイズ決定が，パラメータ θ に対して事前確率分布 $p(\theta)$ を仮定し，この事前確率分布でリスク関数を平均化したものとして定義されるのに対し，ミニマックス決定では最悪の θ に対するリスクを最小化する決定となっており，平均的なパフォーマンスよりもその最悪の状態におけるリスクを最小化しようとする消極的戦略，あるいはリスク回避型戦略であると言える。

3. 一般に，$0-1$ 損失関数を平均化したリスクが誤り率となる。従って，つぎのような $0-1$ 損失を設定すればよい。

損失関数 $L(a, y)$

$L(a, y)$	$y_1 = 1$ (赤球)	$y_2 = 0$ (白球)
$a_1 = 1$ (赤球)	0	1
$a_2 = 0$ (白球)	1	0

4. 例 7.1 では，表 7.1 に示したような偏りのある損失を仮定した場合のベイズ決定を導いたが，この問題では $0-1$ 損失の場合について同様にベイズ決定を導くことが求められている。

リスク関数 $R(\theta, d)$

$R(d, \theta)$	$d_1(X)$	$d_2(X)$	$d_3(X)$	$d_4(X)$
$\theta = \theta_1$ $(= 1/4)$				
$\theta = \theta_2$ $(= 3/4)$				

ベイズリスク $BR(d)$

	$d_1(X)$	$d_2(X)$	$d_3(X)$	$d_4(X)$
$BR(d)$				

5. 決定理論におけるベイズ最適解とミニマックス解の関係性については，7.2 節で詳しく説明した通りである。

   ベイズ最適解はリスクを平均化した図形がリスクセット上を通過するという条件で，ベイズリスクを最小化するため，この図形とリスクセットが接する点で最適解を持つ。一方，ミニマックス決定は，最大リスクを最小化しようとするため，ミニマックス解はリスクの空間上において，各軸の値が等しくなる直線とリスクセットが交わる点の最小値で与えられる。これらの関係性は「最悪の事前分布」を用いて関連付けることが可能であり,「最悪の事前分布」を用いて計算したベイズリスクを最小化するベイズ最適解がミニマックス解と等価になるという性質を有している。

6. この問題は第 3 章の【10】で示した条件付確率を求める問題に対して，意思決定の誤りの重大さを損失関数で定義したうえでベイズ決定を導いてみるという問題になっている。いま，損失関数として次のような関数を仮定してみる。

損失関数 $L(a, y)$

$L(a, y)$	$y_1 = 1$ (罹患)	$y_2 = 0$ (正常)
$a_1 = 1$ (罹患と診断)	0	1
$a_2 = 0$ (正常と診断)	500	0

   これは，実際には罹患していない正常な状況（$y_2 = 0$）で $a_1 = 1$ (罹患と診断) と決定してもあまり大きな問題とはならないが，逆に病気に罹患している状況 $y_1 = 1$ (罹患) にも関わらず $a_2 = 0$ (正常と診断) と決定してしまうとより重大な問題であると考えられるため，このような意思決定の重篤さを反映した損失関数となっている。

   "対象の病気に罹患している"という事象を $y_1 = 1$，"対象の病気に罹患していない"という事象を $y_2 = 0$ とする。罹患率がは 0.0001 であるので，

$$P(y_1 = 1) = 0.0001, \qquad P(y_2 = 0) = 0.9999$$

   である。いま，"検査で陽性になる"という事象を $x_1 = 1$，"検査で陰性になる"という事象を $x_2 = 0$ とすれば，この検査の誤り率は 0.01 であることから，

$$P(x_1|y_1) = 0.99, \qquad P(x_0|y_1) = 0.01$$

$$P(x_1|y_2) = 0.01, \qquad P(x_0|y_2) = 0.99$$

   である。

   第 3 章と同様の展開により，

$$\begin{aligned} P(y_1|x_1) &= \frac{P(x_1|y_1)P(y_1)}{P(x_1)} \\ &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \\ &= \frac{0.000099}{0.000099 + 0.009999} = 0.009804 \end{aligned}$$

$$\begin{aligned} P(y_2|x_1) &= \frac{P(x_1|y_2)P(y_2)}{P(x_1)} \\ &= \frac{0.01 \times 0.9999}{0.99 \times 0.0001 + 0.01 \times 0.9999} \\ &= \frac{0.009999}{0.000099 + 0.009999} = 0.990196 \end{aligned}$$

   が得られる。

   $a_1 = 1$ (罹患と診断) と決定した場合のリスク関数，並びに $a_2 = 0$ (正常と診断) と決定した場合の条件付リスク関数を求めると次のようになる。

$$\begin{aligned} R(a_1) &= L(a_1, y_1)P(y_1|x_1) + L(a_1, y_2)P(y_2|x_1) \\ &= 0 \times 0.009804 + 1 \times 0.990196 = 0.990196 \end{aligned}$$

$$
\begin{aligned}
R(a_2) &= L(a_2, y_1)P(y_1|x_1) + L(a_2, y_2)P(y_2|x_1) \\
&= 500 \times 0.009804 + 0 \times 0.990196 = 4.902
\end{aligned}
$$

となり，$a_1 = 1$ (罹患と診断) と決定した場合のリスク関数の方が小さくなるので「この被験者は病気に罹患している」と診断する方の決定の方が合理的である。

なお，先の損失関数の定義では $L(a_2, y_1) = 500$ と定義したが，これを $L(a_2, y_1) = 100$ とすると微妙に a_2(正常と診断) の方がリスクが小さくなる（色々な損失関数で試してみよ)。

第 8 章

1. に自然界に存在する音や画像，時間，温度などの連続情報であるアナログ情報は実数データをそのまま扱うため，データの伝送中，もしくはデータの読み取り中にノイズが混入した場合，それを除去することが困難である。すなわち，アナログ情報は連続量を扱える半面，情報の減衰が生じたり，ノイズに弱く，長期間情報を保存すると劣化が生じるなどの問題があった。

   これに対し，0, 1, 2, $\cdots$ と数が数えられる離散量で表現されたデジタル情報では，デジタル情報は少々のノイズであれば除去可能であり，より信頼性の高い情報処理が可能となる。また，アナログ情報と比べて，情報の劣化が少なく，情報を記憶するうえでもメリットが大きい。そのため，現在の情報処理技術では，アナログ情報からデジタル化の方向へ発展してきている。

2. コンピュータを実装するためのハードウェア技術を考慮すると，情報を 2 値で表現することの有用性は非常に大きい。これは，離散値を値に持つ素子の中でも最も状態数が少ない 2 値の素子を用いて，情報が表現できるためである。2 値の情報であれば，電圧のプラスとマイナス，電流のオンとオフ，磁場の向き（正方向と逆方向）といった方法で，2 つの状態を表現することが可能であり，比較的安価な素子を用いて実装可能である。

   “ON”, “OFF” や電圧の “高”, “低” といった 2 つの状態を持つ素子を使った情報の表示は，状態を “1” と “0” に対応して表現することが可能であり，2 進数はこのような情報機器や記憶装置による情報表現に適した数字であることから，その重要性は極めて高い。

3. 10 進数→ 2 進数は，2 で剰余を取り続けて，逆から並べればよい.

   (1) $(20)_{10} = (10100)_2$
   (2) $(200)_{10} = (11001000)_2$
   (3) $(777)_{10} = (1100001001)_2$
   (3) $(1024)_{10} = (10000000000)_2$

   これらが正しい事は，$2^7 + 2^6 + 2^3 = 200$ と検算してみれば確かめられる。なお，$2^{10} = 1024$ は便利な知識であるので覚えておくとよい。

4. (1) $(101)_2 = (2^2 + 2^0)_{10} = (5)_{10}$
   (2) $(1000)_2 = (8)_{10}$
   (3) $(1111)_2 = (15)_{10}$
   (4) $(101011)_2 = (43)_{10}$

5. 2 進数を 16 進数への変換は，$2^4 = 16$ であることを利用すれば，表を用いて 4 桁ずつ簡単に変換できる。

   (1) $(1011)_2 = (B)_{16}$
   (2) $(11111011)_2 = (FB)_{16}$
   (3) $(101011111011)_2 = (AFB)_{16}$

6. 16 進数を 2 進数への変換は，1 桁ずつ 4 ビット 2 進数に変換して連接すればよい。

   (1) $(1A0F)_{16} = (0001\ 1010\ 0000\ 1111)_2$
   (2) $(5BCA\ 12F7)_{16} = (0101\ 1011\ 1100\ 1010\ 0001\ 0010\ 1111\ 0111)_2$
   (3) $(89AB\ CDEF)_{16} = (1000\ 1001\ 1010\ 1011\ 1100\ 1101\ 1110\ 1111)_2$

7. (1) $(1A0F)_{16} = (6671)_{10}$
   (2) $(5BCA\ 12F7)_{16} = (1539969783)_{10}$
   (3) $(89AB\ CDEF)_{16} = (2309737967)_{10}$

8. 1024 種類の記号を区別して 2 進数表現するためには $\log_2 1024 = 10$ 桁が必要である。

9. $1,000,000$ 種類の記号を区別して 2 進数表現するためには $\log_2 1,000,000 = 19.93$ であることから，20 桁が必要である。

10. r 桁二進数 X に対する 2 の補数 Y は，二進数の加算で

$$(X)_2 + (Y)_2 = (\underbrace{1000\cdots0}_{r+1\text{ 桁}})_2 \tag{1}$$

となるような Y の下 r 桁の二進数で定義される。このような Y は，r 桁の二進数演算において，Y を $-X$ と定義することが可能になる。その結果，減算を可算で表現することができ，またある数の負の数を自然に定義し，これらの和が 0 となるような体系を作ることができる。

11. ある数の 2 の補数を直接表から求めるのは桁が大きくなると大変面倒な計算になるため，1 の補数を活用すると計算が大変シンプルになる。r 桁二進数 X に対し，二進数の加算で，

$$(X)_2 + (Y)_2 = (\underbrace{1111\cdots1}_{r\text{ 桁}})_2 \tag{2}$$

となるような Y を X の 1 の補数といい，これに二進数で 1 を加えると X の 2 の補数 Y を得ることができる。

12. 1 の補数は 0 と 1 を反転させればよい．両者を足すと全 1 になることから，1 の補数と呼ばれる．
   (1) $(0011\ 1100)_2$ の 1 の補数は，$(1100\ 0011)_2$
   (2) $(0110\ 1010)_2$ の 1 の補数は，$(1001\ 0101)_2$
   (3) $(0000\ 0000)_2$ の 1 の補数は，$(1111\ 1111)_2$

13. 2 の補数は，1 の補数に 1 を足し，9 ビット目の桁は切り捨てる．
   (1) $(0011\ 1100)_2$ の 2 の補数は，$(1100\ 0100)_2$
   (2) $(0110\ 1010)_2$ の 2 の補数は，$(1001\ 0110)_2$
   (3) $(0000\ 0000)_2$ の 2 の補数は，$(0000\ 0000)_2$

14. シフトは，2 値で表現されたビットパターンを必要なビット数だけ，右または左にずらす操作をいい，このような演算をシフト演算という。

   基本的な方法である算術シフトでは，右に r だけシフト（右シフト）すると (10 進数で) 2^{-r} 倍，逆に左に r だけシフト（左シフト）すると 2^r 倍の乗算をしたことと等価である。ただし，先頭ビットを符号ビットとして，これはシフトをせず，残りのビットをシフトさせ，シフトでは溢れたビットは単に消える。一方，シフトによって空いたビット部分には，左シフトの場合は全てゼロが入り，右シフトでは符号ビットと同じ値が入るものとする。

   このようなシフト演算を利用することで，2 進数の乗算や除算を大変効率的に計算を行うことができ，コンピュータ上で表現された数に対して演算を行うための処理方法として大変優れた方法となっている。

15. 乗算は左シフト演算で簡単に出来る。
   (1) $(0000\ 0011)_2 \times (0000\ 0100)_2 = (0000\ 1100)_2$
   (2) $(0010\ 1010)_2 \times (0000\ 0110)_2 = (1010\ 1000)_2 + (0101\ 0100)_2 = (1111\ 1100)_2$
   (3) は省略。各自でやってみること。

16. 除算は，普通の割り算をやればよい。その際，8 ビットの 2 進表示での割り算なので，少数以下は計算せず，余りは切り捨てる。

    (1) $(11)_2 \div (100)_2 = (0)_2$ 余り $(11)_2$。よって答えは，$(0000\ 0000)_2$

    (2) $(101010)_2 \div (110)_2 = (111)_2$ 余り $(0)_2$．割り切れる。よって答えは，$(0000\ 0111)_2$

    (3) は省略。各自でやってみること。

17. (1) $(0.1011)_2 = 2^{-1} + 2^{-3} + 2^{-4} = 0.6875$

    (2) $(0.111)_2 = 2^{-1} + 2^{-2} + 2^{-3} = 0.875$

    (3) は省略。各自でやってみること。

18. (1) $(0.1)_{10} = (0001100110011\cdots)_2$

    (2) $(0.34375)_{10} = (0.01011)_2$

    (3) は省略．各自でやってみること。

19. 整数部と小数部に分けてそれぞれ 10 進数に変換してから加えるとよい．小数部分については【17】で求めているので活用するとよい。

    (1) $(1.1011)_2 = 2^0 + 2^{-1} + 2^{-3} + 2^{-4} = 1.6875$

    (2) $(11.111)_2 = 2^1 + 2^0 + 2^{-1} + 2^{-2} + 2^{-3} = 3.875$

    (3) は省略。各自でやってみること。

20. 整数部と小数部に分けてそれぞれ 2 進数に変換してから加えるとよい．小数部分については【18】で求めているので活用するとよい。

    (1) $(10.1)_{10} = (1010.0001100110011\cdots)_2$

    (2) $(20.34375)_{10} = (10100.01011)_2$

    (3) は省略。各自でやってみること。

第 9 章

1. 9.2 節に詳しい説明があるため省略。

2. 9.3 節に詳しい説明があるため省略。

3. M 個のものから 1 つを特定する情報を知った時に得られる自己情報量は，特に確率などが定められていなければ $\log_2 M$(ビット) で与えられる (Hartrey の情報量)。これは，いずれの事象の同等に確からしく，確率は一様である場合の Shannon の自己情報量を同じである。

   (1) は $\log_2 8 = 3$ (ビット)，(2) は $\log_2 32 = 5$ (ビット)，(3) は $\log_2 256 = 8$ (ビット) となる。

4. 系列 $\boldsymbol{x}$ の自己情報量を $I(\boldsymbol{x})$ と表すことにすると，各系列の自己情報量は次のように計算される。

   (1) $I(11111111) = -8\log_2 \frac{3}{4} = 16 - 8\log_2 3$

   (2) $I(11011011) = -6\log_2 \frac{3}{4} - 2\log_2 \frac{1}{4} = 16 - 6\log_2 3$

   (3) $I(01010101) = -4\log_2 \frac{3}{4} - 4\log_2 \frac{1}{4} = 16 - 4\log_2 3$

   (4) $I(00100100) = -2\log_2 \frac{3}{4} - 6\log_2 \frac{1}{4} = 16 - 2\log_2 3$

   (4) $I(00000000) = -8\log_2 \frac{1}{4} = 16$

5. エントロピーの式を用いて計算するのみ。対数の底は 2 で計算すれば，きれいに計算できる。

$$H(X) = -\frac{1}{4}\log_2 \frac{1}{4} - \frac{1}{4}\log_2 \frac{1}{4} - \frac{1}{8}\log_2 \frac{1}{8} - \frac{1}{4}\log_2 \frac{1}{4} - \frac{1}{8}\log_2 \frac{1}{8} = \frac{9}{4} \quad (\text{ビット})$$

6. 例題 9.4 を参考に，同様の計算を行なえばよい。

$$H(X) = -\frac{1}{4}\log_2\frac{1}{4} - \frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} = \frac{3}{2}$$

$$H(Y) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$\begin{aligned} H(X,Y) &= -\sum_{x=0}^{2}\sum_{y=0}^{1} P(x,y)\log P(x,y) \\ &= -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - 0\log 0 - 0\log 0 - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} = 2 \end{aligned}$$

$$\begin{aligned} H(X|Y) &= \sum_{y=0}^{1} P(y)\left\{-\sum_{x=0}^{2} P(x|y)\log P(x|y)\right\} \\ &= \frac{1}{2}\left\{-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} - 0\log 0\right\} + \frac{1}{2}\left\{-0\log 0 - \frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right\} \\ &= 1 \end{aligned}$$

$$\begin{aligned} H(Y|X) &= -\sum_{x=0}^{2} P(x)\left\{\sum_{y=0}^{1} P(y|x)\log P(y|x)\right\} \\ &= \frac{1}{4}\{-1\log 1 - 0\log 0\} + \frac{1}{2}\left\{-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right\} + \frac{1}{4}\{-1\log 1 - 0\log 0\} \\ &= \frac{1}{2} \end{aligned}$$

7. コインを 4 回投げる場合は，情報源 X と Y は次のようになる。

$$X = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1/16 & 4/16 & 6/16 & 4/16 & 1/16 \end{pmatrix}$$

$$Y = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix}$$

となる。従って，

$$\begin{aligned} H(X) &= -\frac{1}{16}\log_2\frac{1}{16} - \frac{4}{16}\log_2\frac{4}{16} - \frac{6}{16}\log_2\frac{6}{16} - \frac{4}{16}\log_2\frac{4}{16} - \frac{1}{16}\log_2\frac{1}{16} \\ &= 3 - \frac{3}{4}\log_2 3 \end{aligned}$$

$$H(Y) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$$

$$\begin{aligned} H(X,Y) &= -\sum_{x=0}^{4}\sum_{y=0}^{1} P(x,y)\log P(x,y) \\ &= -\frac{1}{16}\log\frac{1}{16} - \frac{3}{16}\log\frac{3}{16} - \frac{3}{16}\log\frac{3}{16} - \frac{1}{16}\log\frac{1}{16} - 0\log 0 \\ &\qquad - 0\log 0 - \frac{1}{16}\log\frac{1}{16} - \frac{3}{16}\log\frac{3}{16} - \frac{3}{16}\log\frac{3}{16} - \frac{1}{16}\log\frac{1}{16} \\ &= 4 - \frac{3}{4}\log_2 3 \end{aligned}$$

$$H(X|Y) = \sum_{y=0}^{1} P(y)\left\{-\sum_{x=0}^{4} P(x|y)\log P(x|y)\right\}$$

$$
\begin{aligned}
&= \frac{1}{2}\left\{-\frac{1}{8}\log\frac{1}{8}-\frac{3}{8}\log\frac{3}{8}-\frac{3}{8}\log\frac{3}{8}-\frac{1}{8}\log\frac{1}{8}-0\log 0\right\} \\
&\quad+\frac{1}{2}\left\{-0\log 0-\frac{1}{8}\log\frac{1}{8}-\frac{3}{8}\log\frac{3}{8}-\frac{3}{8}\log\frac{3}{8}-\frac{1}{8}\log\frac{1}{8}\right\} \\
&= \frac{3}{4}-\frac{3}{4}\log_2 3
\end{aligned}
$$

$$
\begin{aligned}
H(Y|X) &= -\sum_{x=0}^{4}P(x)\left\{\sum_{y=0}^{1}P(y|x)\log P(y|x)\right\} \\
&= \frac{1}{16}\{-1\log 1-0\log 0\}+\frac{4}{16}\left\{-\frac{1}{4}\log\frac{1}{4}-\frac{3}{4}\log\frac{3}{4}\right\}+\frac{6}{16}\left\{-\frac{1}{2}\log\frac{1}{2}-\frac{1}{2}\log\frac{1}{2}\right\} \\
&\quad+\frac{4}{16}\left\{-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}\right\}+\frac{1}{16}\{-1\log 1-0\log 0\} \\
&= \frac{35}{8}-\frac{3}{2}\log_2 3
\end{aligned}
$$

8. 問題より，$P(X=0)=p_0$，$P(X=1)=1-p_0$，

$$
\begin{array}{ll}
P(Y=1|X=1)=1-\varepsilon, & P(Y=0|X=1)=\varepsilon \\
P(Y=1|X=0)=\varepsilon, & P(Y=0|X=0)=1-\varepsilon
\end{array}
$$

である。

(1)

$$
\begin{aligned}
H(X) &= -p_0\log p_0-(1-p_0)\log(1-p_0) \\
H(Y) &= -\left(p_0(1-\varepsilon)+(1-p_0)\varepsilon\right)\log\left(p_0(1-\varepsilon)+(1-p_0)\varepsilon\right) \\
&\quad-\left(p_0\varepsilon+(1-p_0)(1-\varepsilon)\right)\log\left(p_0\varepsilon+(1-p_0)(1-\varepsilon)\right)
\end{aligned}
$$

(2)

$$
H(Y|X) = -\varepsilon\log\varepsilon-(1-\varepsilon)\log(1-\varepsilon)
$$

(3)

$$
\begin{aligned}
I(Y;X) &= H(Y)-H(Y|X) \\
&= \left(p_0(1-\varepsilon)+(1-p_0)\varepsilon\right)\log\left(p_0(1-\varepsilon)+(1-p_0)\varepsilon\right) \\
&\quad-\left(p_0\varepsilon+(1-p_0)(1-\varepsilon)\right)\log\left(p_0\varepsilon+(1-p_0)(1-\varepsilon)\right) \\
&\quad+\varepsilon\log\varepsilon+(1-\varepsilon)\log(1-\varepsilon)
\end{aligned}
$$

9. 定理 9.1 のエントロピーの性質については，いずれもシンプルな計算によって導くことができる。

(1) エントロピーの式から，これが $0<p_i<1$ の範囲で，p_i で微分可能であり，連続な関数であることは明らか。

(2) 最大値はラグランジュの未定乗数法で解くことができる。

(3) 実際にエントロピーの式に代入して示せばよい。

(4) 実際にエントロピーの式に代入して示せばよい。

(5) エントロピーの式から明らか。

(6) エントロピーの式から明らか。

(7) 実際にエントロピーの式に代入して示せばよい。

10. シャノンの補助定理とは，$p_1 + p_2 + \cdots + p_M = 1$, $q_1 + q_2 + \cdots + q_M = 1$，かつ $p_i \geq 0$, $q_i \geq 0$ をみたす任意の確率ベクトル $\mathbf{p} = (p_1, p_2, \cdots, p_M)$ と $\mathbf{q} = (q_1, q_2, \cdots, q_M)$ に対し，

$$H(\mathbf{p}) = -\sum_{i=1}^{M} p_i \log p_i \leq -\sum_{i=1}^{M} p_i \log q_i$$

であり，等号は全ての i に対して $p_i = q_i$ となる場合に成り立ち，かつその時に限るというものである．これを示すためには，

$$-\sum_{i=1}^{M} p_i \log\left(\frac{p_i}{q_i}\right) \leq 0$$

を示せばよい (この対数の底は何でもよい)。

対数の底を a とし，$x > 0$ において，$\log_e x \leq x - 1$ (等号は $x = 1$ のとき) が成り立つことを利用すると，

$$\begin{aligned}
-\sum_{i=1}^{M} p_i \log_a\left(\frac{p_i}{q_i}\right) &= \sum_{i=1}^{M} p_i \log_a\left(\frac{q_i}{p_i}\right) \\
&\leq \sum_{i=1}^{M} p_i \left(\frac{q_i}{p_i} - 1\right) \Big/ \log_e a \\
&= \sum_{i=1}^{M} (q_i - p_i) \Big/ \log_e a \\
&= \left(\sum_{i=1}^{M} q_i - \sum_{i=1}^{M} p_i\right) \Big/ \log_e a \\
&= (1-1)/\log_e a = 0
\end{aligned}$$

以上により，証明が完結した。

11. 9.4 節に詳しい説明があるため省略。

12. KL 情報量の式

$$D(\mathbf{p}\|\mathbf{q}) = \sum_{i=1}^{M} p_i \log_2\left(\frac{p_i}{q_i}\right)$$

を用いて計算すればよい。

(1) は，

$$\begin{aligned}
D(\mathbf{p}\|\mathbf{q}) &= 0.5 \log\left(\frac{0.5}{0.25}\right) + 0.25 \log\left(\frac{0.25}{0.125}\right) + 0.125 \log\left(\frac{0.125}{0.5}\right) + 0.125 \log\left(\frac{0.125}{0.125}\right) \\
&= \frac{1}{2} \cdot \log 2 + \frac{1}{4} \cdot \log 2 + \frac{1}{8} \cdot \log \frac{1}{4} \\
&= \frac{1}{2} = 0.5 \ (\text{ビット})
\end{aligned}$$

となる。(2),(3) も同様に計算できる。

(2) $D(\mathbf{p}\|\mathbf{q}) = 0.625$

(3) $D(\mathbf{p}\|\mathbf{q}) = 0.125$

13. KL 情報量を計算するとそれぞれ次のようになる。

(1) $D(\mathbf{p}\|\mathbf{q}_1) = 0.5$

(2) $D(\mathbf{p}\|\mathbf{q}_2) = 0.125$

(3) $D(\mathbf{p}\|\mathbf{q}_3) = 0.1851$

従って，KL 情報量の意味で $\mathbf{p}$ から最も近い分布は (2) の $\mathbf{q}_2$ である。一方，$\mathbf{p}$ から最も遠い分布は (1) の $\mathbf{q}_1$ である。

第 10 章

1. 例えば，袋の中に「表の出る確率が 2/3，裏の出る確率が 1/3 のコイン A」と「表の出る確率が 1/3，裏の出る確率が 2/3 のコイン B」の 2 枚が入っており，この中から無作為に 1 枚を取り出す。この取り出したコインを繰り返し投げて，表が出たら $X_t = 1$，裏が出たら $X_t = 0$ を記録するという行為を繰り返す試行を考える。このとき，最初にどちらのコインが選ばれるかは確率 1/2 であり，コイン A が選ばれたときに 1 が出る確率は 2/3，コイン B が選ばれたときに 1 が出る確率は 1/3 であることから，各時点 t における期待値 $E[X_t]$（1 が生起する確率と等価）は $E[X_t] = 1/2 \cdot 1/3 + 1/2 \cdot 2/3 = 1/2$ で等しいため，この試行によって得られるデータを生成する情報源は定常情報源である。

   しかし，最初に選ばれたコインが繰り返し投げられるため

   - 最初にコイン A が選ばれた場合は，観測系列の平均 $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ は 2/3 に収束する
   - 最初にコイン B が選ばれた場合は，観測系列の平均 $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ は 1/3 に収束する

   となり，系列で平均を取った時間平均が先の平均値 1/2 には収束しない。従って，個の情報源は定常情報源ではあるが，エルゴード情報源ではない。

   同様の情報源は，他にも「偶数の生起確率が高いサイコロ」と「奇数の生起確率が高いサイコロ」が半々ずつ入っており，この中から無作為に 1 つを取り出す。取り出したサイコロを繰り返し振って出た目を観測するといった試行によっても作ることができる。

2. (1) 3 元単純マルコフ情報源なので，状態数は 3 である．状態を s_1,s_2,s_3 とし，これらの定常確率をそれぞれ $q_1 = p(s_1)$, $q_2 = p(s_2)$, $q_3 = p(s_3)$ とすれば，(q_1, q_2, q_3) は遷移確率行列 Q によって遷移しても確率が変わらないという意味で定常であるので，

   $$(q_1, q_2, q_3) = (q_1, q_2, q_3)Q$$

   の関係が成り立つ．すなわち，

   $$\begin{aligned} q_1 &= 0.5q_1 + 0.25q_2 \\ q_2 &= 0.5q_1 + 0.5q_2 + 0.5q_3 \\ q_3 &= 0.25q_2 + 0.5q_3 \end{aligned}$$

   であり，これを解いて，$q_1^* = q_3^* = 1/4$, $q_2^* = 1/2$ を得る．したがって，定常分布 (q_1^*, q_2^*, q_3^*) は、$(q_1^*, q_2^*, q_3^*) = (1/4, 1/2, 1/4)$ で与えられる。

   (2) s_1 のもとでの条件付エントロピーは，

   $$H(X|s_1) = -0.5\log_2 0.5 - 0.5\log_2 0.5 - 0\log_2 0 = 1$$

   同様に，$H(X|s_2) = 3/2$，$H(X|s_3) = 1$ であるので，このマルコフ情報源のエントロピーは，

   $$\begin{aligned} H(X) &= p(s_1)H(X|s_1) + p(s_2)H(X|s_2) + p(s_3)H(X|s_3) \\ &= \frac{1}{4} \cdot 1 + \frac{1}{2} \cdot \frac{3}{2} + \frac{1}{4} \cdot 1 \\ &= \frac{5}{4} \end{aligned}$$

3. 等長符号 C_1 は，a〜f に対して，0000, 0001, 0010, $\cdots$ といった同じ長さの符号の例を一つ示せばよい。

   一意復号不可能な符号 C_2 は，異なる記号に対して同じ符号語が対応しているような場合，

   $$\begin{aligned} a &\rightarrow 0000 \\ b &\rightarrow 0000 \\ &\cdots \end{aligned}$$

   符号語の連接が他の符号語になっている場合

   $$\begin{aligned} a &\rightarrow 0100 \\ b &\rightarrow 0 \end{aligned}$$

$$
\begin{aligned}
c &\rightarrow 100 \\
&\cdots
\end{aligned}
$$

など，様々なパターンがあるので，そのうちの一例を示せばよい。

また，一意復号可能な符号 C_3 は様々なパターンがある。一般に，瞬時に一意復号可能な瞬時符号は，符号の木を作ったときに，葉ノードにだけ符号語が割り当てられているようなものを作れば比較的簡単に作成できる。

4. ハフマン符号の作り方は 10.4 節を参照のこと。平均符号長が最小のハフマン符号を構成すると次のようになる。

$$
\begin{aligned}
a &\rightarrow 0 \\
b &\rightarrow 10 \\
c &\rightarrow 110 \\
d &\rightarrow 1110 \\
e &\rightarrow 11110 \\
f &\rightarrow 11111
\end{aligned}
$$

この符号が平均符号長の意味では最適である。ただし，0 と 1 が反転してもよい。平均符号長は

$$L = 1 \cdot 1/2 + 2 \cdot 1/4 + 3 \cdot 1/8 + 4 \cdot 1/16 + 5 \cdot 1/32 + 5 \cdot 1/32 = 1.9375 \text{（ビット）}$$

となる。

5. 平均符号長が最小のハフマン符号を構成すると次のようになる。ただし，これは一例であり，全く平均符号長が等しい最適な符号構成が複数存在する（それが何故か考えて見よ）。

$$
\begin{aligned}
a &\rightarrow 10 \\
b &\rightarrow 1111 \\
c &\rightarrow 00 \\
d &\rightarrow 110 \\
e &\rightarrow 1110 \\
f &\rightarrow 01
\end{aligned}
$$

また，平均符号長は

$$L = 2 \cdot 0.30 + 4 \cdot 0.10 + 2 \cdot 0.25 + 3 \cdot 0.13 + 4 \cdot 0.02 + 2 \cdot 0.2 = 2.37 \text{（ビット）}$$

となる。

6.

$$
\begin{aligned}
H(X^3) &= -\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p_i p_j p_k \log p_i p_j p_k \\
&= -\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p_i p_j p_k (\log p_i + \log p_j + \log p_k) \\
&= -\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p_i p_j p_k \log p_i - \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p_i p_j p_k \log p_j - \sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{k=1}^{M} p_i p_j p_k \log p_k \\
&= -\sum_{j=1}^{M} p_j \sum_{k=1}^{M} p_k \sum_{k=1}^{M} p_i \log p_i - \sum_{i=1}^{M} p_i \sum_{k=1}^{M} p_k \sum_{j=1}^{M} p_j \log p_j - \sum_{i=1}^{M} p_i \sum_{j=1}^{M} p_j \sum_{k=1}^{M} p_k \log p_k \\
&= -\sum_{i=1}^{M} p_i \log p_i - \sum_{j=1}^{M} p_j \log p_j - \sum_{k=1}^{M} p_k \log p_k \\
&= 3H(X)
\end{aligned}
$$

7. エントロピー $H(X)$ を持つ情報源から生起する情報源系列を一意復号が可能な符号で符号化する際，その平均符号長の下限が $H(X)$ で与えられることが知られている。また，データ長 n を十分長く取るとき，平均符号長がエントロピー $H(X)$ にいくらでも近くなるような符号化を構成可能であることも知られている。このことから，エントロピーは情報源符号化の限界を与える量として重要である。

   情報源符号化の目的は，情報源から出てくる記号列をできるだけ短い符号に変換し，情報を圧縮することである。その評価は，情報源シンボル 1 記号あたりの平均符号長であり，この平均符号長が最小となる符号が優れていることになる。情報源符号化の目的が圧縮である以上，符号化アルゴリズムの性能を測るための規準として，圧縮の限界を示すことが大変重要であり，その限界がエントロピー $H(X)$ で与えられるということになる。

8. 十分長い系列を観測すると，実際に生起するのはある統計的特徴を有した系列のみになる。このような情報源から実際に生起する系列を標準系列，もしくは代表的系列という。

   いま，記憶のない情報源

$$X = \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_M \\ p_1 & p_2 & p_3 & \cdots & p_M \end{pmatrix}$$

   から発生する長さ n の出力記号列 (系列) $x^n = x_1 x_2 \cdots x_n$ と，この系列中に含まれる記号 a_i の個数を n_i とする。$n \to \infty$ のとき，

$$\frac{n_i}{n} \to p_i$$

   となりたつので，十分大きな n については，

$$n_i \sim np_i$$

   と考えることができる。

   長さ n の記号列は M^n 通りの記号列が存在するが，その中で実際に出てくるのは，$n_i \sim np_i$ となっているような記号列だけということになる。標準系列は，このような実際に出てくる $n_i \sim np_i$ となっているような系列の総称である。

9. 一つの標準系列 x^n の確率 $P(x^n)$ を考えてみると，$n_i \sim np_i$ を使って，

$$\begin{aligned} P(x^n) &= \prod_{i=1}^{M} p_i^{n_i} \\ &\sim \prod_{i=1}^{M} p_i^{np_i} \\ &= \prod_{i=1}^{M} \left(2^{\log_2 p_i}\right)^{np_i} \\ &= \prod_{i=1}^{M} 2^{np_i \log_2 p_i} \\ &= 2^{n\sum_{i=1}^{M} p_i \log_2 p_i} \\ &= 2^{-nH(X)} \end{aligned}$$

   と書き下すことができる。$P(x^n) = 2^{-nH(X)}$ は，もはや x^n の中身によって変化しない n と $H(X)$ によって決められる値であり，情報源 X から十分大きな長さ n を持つ標準系列は，どれも等しく

$$P(x^n) = 2^{-nH(X)}$$

   の生起確率を持つという事になる。

   また，標準系列以外が生起する確率は 0（出てこない）と見なすことができるため，標準系列の個数は，$2^{nH(X)}$ 個である．

10. 確率 0.7 で 1 を，確率 0.3 で 0 を出力する定常無記憶情報源であるから，各系列

(1) $P(1111111111) = 0.7^{10}$, $P(1101011011) = 0.7^7 0.3^3$ であることから，明らかに $P(1111111111) > P(1101011011)$ である。

(2) (1) と同様に確率を計算してみる。1 が 100 回続く長さ 100 の系列の生起確率は 0.7^{100} であり，1 を 70 個，0 を 30 個を含む長さ 100 の 1 本の系列の生起確率は $0.7^{70}0.3^{30}$ となる。すなわち，1 が 100 回続く長さ 100 の系列の生起確率の方が，1 を 70 個，0 を 30 個を含む長さ 100 の 1 本の系列の生起確率よりも大きい。

(3) (1) と (2) で示したように，1 本の系列の生起確率の大小で比べた場合，より確率の大きい 1 が多く含まれる系列の生起確率が高くなる。しかし，1 本 1 本の系列の生起確率は系列長 n が長くなると，いずれも 0 に近づいていくものであり，2^n 本存在する系列の 1 本 1 本が生起する確率はほぼ 0 になってしまう。

一方，実際に長い系列を観測すると，全体の 7 割程度が 1，3 割程度が 0 となることが統計的に知られている事実である。これは，1 本 1 本の系列の生起確率は系列長 n が長くなると 0 に近づく反面，全体の 7 割程度が 1，3 割程度が 0 となるような系列の本数が ${}_nC_{0.7n}$ 程度で増えていくことから，系列全体のうち，全体の 7 割程度が 1，3 割程度が 0 となるような系列全体の確率がほぼ 1 になることを意味している。

11. 0 と 1 を使った 2 元符号ではなく，q 元符号を用いた場合の平均符号長の下限は，対数の底を q とした場合のエントロピー

$$H(X) = \sum_{i=1}^{M} p_i \log_q p_i$$

で与えられる。

12. q 元の場合の必要十分条件は，

$$q^{-l_1} + q^{-l_2} + q^{-l_3} + \cdots + q^{-l_M} \leq 1$$

で与えられる。証明は 2 元の場合とほぼ同様．1 つのノードから 2 つの枝が分岐する二分木ではなく，1 つのノードから q 本の枝が分岐する q 分木を考えて，同様の手順で示せばよい。

第 11 章

1. 一般的な式展開は，11.3 節に記載の通りである。

(1) $P(Y) = P(Y|X=0)P(X=0) + P(Y|X=1)P(X=1)$ より y の分布を求め，$H(Y)$ を計算すればよい。$H(Y|X)$，$I(X;Y)$ も定義式より計算できる。

(2) 誤り率を p としたときの一般式は，$C = \max_{P(X)} I(X;Y) = \max_{P(X)} H(Y) - H(Y|X)$ を計算し，

$$C = 1 + (1-p)\log_2(1-p) + p\log_2 p \quad [\text{ビット／記号}]$$

を得る。この p に $p = 0.2$ を代入すればよい。

(3) 11.3 節に計算例があるため省略。上の式に $p = 1.0$，$p = 0.5$ を代入すればよい。

(4) 通信路の誤り確率が 1.0 であるとき，必ず誤るため，$Y = 0$ を受け取れば $X = 1$，$Y = 1$ を受け取れば $X = 0$ と確実に送信された記号を特定できる。従って，このとき伝達される情報量 C_2 は誤り率 0 の時と同じになる。

誤り率 0.5 のとき，$P(X=0|Y=1) = P(X=1|Y=1) = P(X=0|Y=0) = P(X=1|Y=0)$ となってしまい，Y を受け取っても，X に関する情報は何も得られず，$C_3 = 0$ となる．誤り率 0.2 の状態は，これらの中間にあって，0.8 の確率で正しい記号が送られてくるので，その分の情報量を受け取ることができる。従って，$C_2 > C_1 > C_3$ となる。

2. もともとの 1 というシンボルを繰り返すのが，繰り返し符号である。

(1) 1111

(2) 111111

3. 偶重み符号は，1 の和が偶数になるように，最後に 1 ビットを追加した符号である。
   (1) 10100
   (2) 01001

4. (7,4,3) ハミング符号の符号語を $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ とするとき，

$$x_1 + x_2 + x_3 + x_5 = 0$$
$$x_2 + x_3 + x_4 + x_6 = 0$$
$$x_1 + x_2 + x_4 + x_7 = 0$$

   の関係式を持つ。これは，x_5, x_6, x_7 を，

$$x_5 = x_1 + x_2 + x_3$$
$$x_6 = x_2 + x_3 + x_4$$
$$x_7 = x_1 + x_2 + x_4$$

   のように生成することと等価である。

5. ハミング符号の最小距離（符号語間の距離の最小値）は 3 であるため，1 ビットの誤りが生じても，最も距離の近い符号語に復号すれば元の符号語に戻る。いま，受信系列を $\mathbf{y}$，誤り系列を $\mathbf{e}$, パリティ検査行列を $\boldsymbol{H}$ とすると，シンドロームは，

$$\mathbf{s} = \mathbf{y}\boldsymbol{H} = (\mathbf{x} + \mathbf{e})\boldsymbol{H} = \mathbf{x}\boldsymbol{H} + \mathbf{e}\boldsymbol{H} = \mathbf{e}\boldsymbol{H}$$

   となる。先の問題【４】で示した関係式を意味するパリティ検査行列は，

$$\boldsymbol{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}^T$$

   で与えられるので，

$$\mathbf{e}_1 = (1,0,0,0,0,0,0)$$
$$\mathbf{e}_2 = (0,1,0,0,0,0,0)$$
$$\mathbf{e}_3 = (0,0,1,0,0,0,0)$$
$$\mathbf{e}_4 = (0,0,0,1,0,0,0)$$
$$\mathbf{e}_5 = (0,0,0,0,1,0,0)$$
$$\mathbf{e}_6 = (0,0,0,0,0,1,0)$$
$$\mathbf{e}_7 = (0,0,0,0,0,0,1)$$

   という 7 パターンある 1 ビット誤りに対して，シンドローム $\mathbf{s}$ は，

$$\mathbf{e}_1\boldsymbol{H} = (1,0,1)$$
$$\mathbf{e}_2\boldsymbol{H} = (1,1,1)$$
$$\mathbf{e}_3\boldsymbol{H} = (1,1,0)$$
$$\mathbf{e}_4\boldsymbol{H} = (0,1,1)$$
$$\mathbf{e}_5\boldsymbol{H} = (1,0,0)$$
$$\mathbf{e}_6\boldsymbol{H} = (0,1,0)$$
$$\mathbf{e}_7\boldsymbol{H} = (0,0,1)$$

   と互いに異なるパターンになるため，これによって，どの 1 ビット誤りが生じたかを特定することができる。1 ビット誤りが特定できれば，その箇所の信号を反転させれば誤りを訂正することができる。

6. (1) 一例を示すと，

$$\boldsymbol{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}^T \tag{3}$$

あるいは，

$$\boldsymbol{H} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}^T \tag{4}$$

などでも正解である。

(2) 例えば，式 (3) の検査行列が意味するのは，1 が立っている箇所のビットを足し合わせると 0 になるという制約を符号語に課しているということである。従って，$\boldsymbol{H}$ が式 (3) であれば答えは，

$$x_1 + x_2 + x_3 + x_5 = 0$$
$$x_2 + x_3 + x_4 + x_6 = 0$$
$$x_1 + x_2 + x_4 + x_7 = 0$$

この正解は勿論，H によって変わる．

(3) (3) 式に対応する生成行列として

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

(4) 式に対応する生成行列として

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

などとなる．$\boldsymbol{GH} = \boldsymbol{0}$ が成り立っていることが分かる。(1) で示した $\boldsymbol{H}$ に対して $\boldsymbol{G}$ がきちんと出来ていればよい。

(解説) 生成行列 $\boldsymbol{G}$ は，全零の符号語を除く k 個 (ここでは $k = 4$) のそれぞれ異なる符号語を取り出し，それぞれを $\boldsymbol{G}$ の行ベクトルとすればよい。答えは一つではないので，$\boldsymbol{H}$ に対応する正しい生成行列を 1 つ示せばよい。
ただし，こうして得られた G に対して，行列の基本操作を行なっても本質的に同じ符号が得られる。したがって行列の基本操作を行い以下の構造の生成行列 $\boldsymbol{G}'$ を用いた方が扱いやすい。

$$\boldsymbol{G}' = [\boldsymbol{I}_k, \boldsymbol{P}^T]$$

ここで $\boldsymbol{I}_k$ は $k \times k$ の単位行列であり，$\boldsymbol{P}$ は $k \times (n-k)$ のパリティビット付加行列と呼ばれ，

$$\boldsymbol{H}' = [\boldsymbol{P}, \boldsymbol{I}_{n-k}]^T$$

という形のパリティ検査表列の最初の k 列の転置 $\boldsymbol{P}^T$ によって得られる．

(4) $\mathbf{x}_1 = \mathbf{w}_1\boldsymbol{G}$，$\mathbf{x}_2 = \mathbf{w}_2\boldsymbol{G}$ で計算すればよい。生成行列 $\boldsymbol{G}$ の作り方は一意ではないので，(2) の解答と連動してちきんと合っていればよい。
パリティ検査行列を (3) 式とした場合には，

$$\mathbf{x}_1 = (1, 1, 0, 0, 0, 1, 0)$$

$$\mathbf{x}_2 = (1, 0, 1, 0, 0, 1, 1)$$

(5) $\mathbf{y}_1 = \mathbf{x}_1 + \mathbf{e}$ で計算する．
パリティ検査行列を (3) 式とした場合には，

$$\mathbf{y}_1 = \mathbf{x}_1 + \mathbf{e} = (1, 1, 1, 0, 0, 1, 0)$$

(6) $\mathbf{s} = \mathbf{y}_1\boldsymbol{H}$ を計算し，誤り位置を特定すればよい．

$$\mathbf{s} = \mathbf{y}_1\boldsymbol{H} = (1, 1, 0)$$

より，これはパリティ検査行列 H の 3 行目であるので，3 ビット目が誤りである．

$$\mathbf{y}_1 = (1, 1, 1, 0, 0, 1, 0)$$

の 3 ビット目が誤りであると判定し，

$$\hat{\mathbf{x}}_1 = (1, 1, 0, 0, 0, 1, 0)$$

と復号することができる．

7. この符号化は，生成行列 $\boldsymbol{G}$ の形を見れば明らかであるように，最初の 4 ビットは情報記号列そのままで，最後に総和を追加するもので，符号長 5 の偶重み符号である．

   (1) 符号語は $(1, 0, 1, 0, 0)$

   (2) 2^4 個ある符号語のうち，どの二つを取ってきても，それらのハミング距離は 2 以上である．また，$(0, 0, 0, 0, 0)$ と $(1, 0, 0, 0, 1)$ などのように距離が 2 であるものも存在するので，最小距離は 2 である．

8. 符号の最小距離 d が $t_2 = d - 2t_1 - 1 > 0$ を満たすものとする。いま，限界距離復号法により t_1 個以下の誤りは訂正される。すなわち，符号語からハミング距離が t_1 以下の領域の受信系列はその符号語に復号される。t_1 以上の誤りについては復号されないが，$t_2 = d - 2t_1 - 1 > 0$ であるとき，$t_1 + t_2$ 個の誤りが起こっても，他の符号語への復号領域（その符号語からの距離が t_1 以内の範囲）に入らないため，誤りが発生していることは知ることができる。すなわち，$t_1 + t_2$ 個の誤りが混入していることを検出することができる。

第 12 章

1. パターン分類は，すでに所属するクラスが分かっている学習データを用いて学習し（教師あり学習），新たなパターンがどのクラスに属するのかを予測する問題をいう。過去のデータとして，パターンとクラスの組からなる学習データが与えられ，この学習データを学習させることで識別器を構成する。

   クラスタリング問題とは，与えられた複数のパターンをそれらの類似度や距離によって似たもの同士を自動グルーピングする問題をいう。予めクラスが与えられている訳ではないが，特徴ベクトルの類似性によってグループ（クラスタ）が作られることになる。

2. パターン分類問題の例としては，教師ありデータを作成して，教師あり学習によって各パターンを分類する分類器を構築するタイプの「画像分類」,「文字分類」,「文書分類」,「良品・不良品の分類」などが挙げられる。具体的には,「果物を撮影した画像データから，A 級品とそれ以外を自動選別する分類器の構築」,「人間の顔画像から本人を自動識別する分類器の構築」といった応用技術が考えられる。

   一方，クラスタリング問題の例としては，対象となるデータ間の類似性や距離を用いて，似たもの同士をグルーピングするタイプの「画像クラスタリング」,「文書クラスタリング」,「ログデータのクラスタリング」,「購買履歴データによる顧客クラスタリング」などが挙げられる。具体的には,「学生のレポートを類似性によって自動クラスタリング」,「EC サイトに投稿された商品レビューの自動クラスタリング」などの応用技術が考えられる。

3. これは識別関数法の原理について理解を得るための問題で，特徴空間が 1 次元であることに注意。

   $g_1(x) > g_2(x)$ である領域はクラス c_1 に判別し，$g_1(x) < g_2(x)$ である領域はクラス c_2 に判別される（$g_1(x) = g_2(x)$ の場合はランダム決定とすればよい）。$g_1(x) = x^2 - 9$，$g_2(x) = -x^2 + 9$ であることから，これらの関数は $x = -3$,$x = 3$ で交差する。これらの大小関係から，

   $$\begin{aligned} x < -3 \ \text{ or } \ 3 < x \quad &\Rightarrow c_1\text{に判別される領域} \\ -3 < x < 3 \quad &\Rightarrow c_2\text{に判別される領域} \end{aligned}$$

   となる。

4. 「線形分離不可能」とは，平面で 2 つのクラスを分類できないことを指す。従って，2 つのクラスが重なっていたり，直線できれいに分離できないような例を図示すればよい。

5. クラス c から発生したパターンを間違ってクラス c' に属していると判断した時の損失として，損失関数 $l(c,c')$ を考えると，平均損失

$$L(c|\boldsymbol{x}) = \sum_{c' \in \mathcal{C}} l(c,c') P(c'|\boldsymbol{x})$$

を考えることができる．これを最小にする c を $\boldsymbol{x}$ が属するクラスであると決定する方法は，平均損失を最小化するという意味で最適であり，ベイズ最適な決定と呼ばれている。

ここで，最も基本的な損失関数として，0-1 損失

$$l(c,c') = \begin{cases} 0\ , & c = c' \\ 1\ , & c \neq c' \end{cases}$$

を仮定する。このとき，入力パターン $\boldsymbol{x}$ に対する損失の期待値は，

$$\begin{aligned} L(c|\boldsymbol{x}) &= \sum_{c' \in \mathcal{C}} l(c,c') P(c'|\boldsymbol{x}) \\ &= 1 - P(c|\boldsymbol{x}) \end{aligned}$$

となる。この平均損失を最小化する $\hat{c}$ は

$$\begin{aligned} \hat{c} &= \arg\max_{c \in \mathcal{C}} \{\log P(c|\boldsymbol{x})\} \\ &= \arg\max_{c \in \mathcal{C}} \{\log p(\boldsymbol{x}|c) + \log P(c)\} \end{aligned}$$

で与えられる。この識別を事後確率最大のベイズ識別という。

6. 同じ事前確率 1/2 を持つ 2 つのクラス c_1, c_2 から，共分散行列が同じで，平均ベクトルがそれぞれ μ_1, μ_2 の多次元正規分布によって特徴ベクトルが生起する場合，識別境界の関数は $\boldsymbol{x}$ の 1 次関数の形になる。

7. それぞれのクラスの代表ベクトルが $m_1 = (6,2)^T$, $m_2 = (3,1)^T$ で与えられるとき，識別境界は，それらの点を結んだ線分の垂直二等分線で与えられる。これは全ての点が，代表ベクトルと近い方のクラスに分類されるためである。

   従って，この場合は，垂直二等分線は，2 点の中点 $(4.5, 1.5)$ を通り，傾きは -3 で与えられる事から，$x_2 = -3x_1 + 15$ が求める識別境界の図形になる。これを x_1-x_2 平面に図示すればよい。

8. あるクラスの代表ベクトルが複数ある場合，そのどちらかに近ければ，そのクラスに識別される。

   従って，全ての点は，$m_1^1 = (1,2)^T$, $m_2^1 = (5,2)^T$，$m_1^2 = (3,0)^T$ の 3 つの代表ベクトルと比較され，そのうち最も近い代表ベクトルが m_1^1 か m_2^1 であればクラス c_1 に，m_1^2 であれば c_2 に判別される。そのため，識別境界線は，$m_1^1 = (1,2)^T$ と $m_1^2 = (3,0)^T$ を結ぶ線分の垂直二等分線と $m_2^1 = (5,2)^T$ と $m_1^2 = (3,0)^T$ を結ぶ線分の垂直二等分線からなる区分線形関数となる。

   まず，$m_1^1 = (1,2)^T$, $m_2^1 = (5,2)^T$，$m_1^2 = (3,0)^T$ の 3 点を図示し，$m_1^1 = (1,2)^T$ と $m_1^2 = (3,0)^T$ を結ぶ線分の垂直二等分線と $m_2^1 = (5,2)^T$ と $m_1^2 = (3,0)^T$ を結ぶ線分の垂直二等分線はどんな直線であるのかを描いてみればよい。

9. 特徴量同士に相関がある場合，d 次元空間上に各データが均等に分布しておらず，ある方向に統計的な関係性を持って分布していることになる。このとき，これらの相関を持つデータは，情報量の大きな損失なしに，より低次元の空間上に写像して表現することができ，多変量解析における主成分分析がこのような次元縮約を目的とした分析となっている。

   このような特徴量同士に相関があるデータに対して，ユークリッド距離によってテンプレートマッチング法を行う場合，d 次元空間上でユークリッド距離を測るため，この距離の大小がデータ同士の類似の度合いを正しく表現できていない可能性がある。例えば，強い相関を持っている方向に対してはデータが分布しているため，ユークリッド距離が遠くてもパターン同士は類似性が高いが，データが分布していない方向に向かってはユークリッド距離が近くてもパターン同士の類似性が極めて低いという現象が起こり得る。

   ユークリッド距離は，データの相関を考慮してデータ同士の類似性を測るような距離にはなっていないので注意が必要である。

10. 各クラスに複数の代表ベクトル（テンプレート）を持たせるマルチテンプレートマッチングでは，各データを最も近いテンプレートのクラスに分類する。従って，クラス c への識別領域は，その各テンプレートからの距離が他のクラスの全てのテンプレートからの距離よりも近い領域となり，識別境界は必然的に区分線形な関数となる。そのため，線形識別が不可能なデータに対しても，各クラスに十分な数の代表ベクトル（テンプレート）を持たせることにより，適切な分類が可能になる。

    ただし，複雑な非線形の分類問題に対してマルチテンプレートマッチングを適用しようとすると，非常に多くの代表ベクトル（テンプレート）が必要となってしまうので，メモリ量や計算量の観点から現実的ではなくなる点には注意が必要である。

11. 分類誤り率を最小化する事後確率最大のベイズ識別では，$p(\boldsymbol{x}|c)$，もしくは $p(c|\boldsymbol{x})$ が与えられれば，ベイズ最適な識別が可能となる。そのため，$p(\boldsymbol{x}|c)$，もしくは $p(c|\boldsymbol{x})$ を近似するようなモデルを何らかの方法で得ることが目標となる。すなわち，実際のパターン認識では，$p(\boldsymbol{x}|c)$ か $p(c|\boldsymbol{x})$ のどちらかの条件付確率をモデル化することで，分類ルールを与えることができる。

    クラス c のもとでの特徴ベクトル $\boldsymbol{x}$ の条件付確率 $p(\boldsymbol{x}|c)$ をモデル化したものを生成モデル，一方，特徴ベクトル $\boldsymbol{x}$ のもとでクラス c の条件付確率 $p(c|\boldsymbol{x})$ をモデル化したものを識別モデルという。

    これらは，対象問題の性質によって，モデル化のし易い方法を適切に選択することが肝要である。また，生成モデル，識別モデルという分類は，あくまでどちらの方向で条件付確率をモデル化するかという観点で見たときの違いであり，生成モデルや識別モデルの具体的なモデルや手法としては，非常に多くの提案がなされている。

12. 実際のパターン認識では，パラメータを持つ何らかの確率モデルを使って特徴パターンの分布を近似できることも多い。このような場合，それぞれクラス c に依存して異なるパラメトリックな確率モデル $p(\boldsymbol{x}|c,\boldsymbol{\theta}_c)$ で表現することができる。最も単純な例としては，特徴量の次元 d が比較的小さく，特徴パターンがある点を中心に分布しており，多次元正規分布によって表現できる場合である。また，d が高次元であっても，各単語の出現頻度データによる文書分類や購買履歴データによる顧客の分類などのように，多次元であるがスパースなデータ（0 が多く生起するデータ）では，ナイーブベイズ法などのように，確率モデル $p(\boldsymbol{x}|c,\boldsymbol{\theta}_c)$ を比較的シンプルなモデルで近似できることが多い。これらの生成モデルによるモデル化の方がよい性能を示すケースにおいては，生成モデルに基づく方法を採用すべきである。

    一方，近年のパターン認識が対象とする問題では，特徴ベクトルの次元 d は一般に非常に高次元であることが多く，適切な生成モデルの仮定が困難な問題も多い。これに対し，識別モデル $p(c|\boldsymbol{x})$ で分類するカテゴリ $c_1, c_2, \cdots, c_M$ の数 M は，多くの場合，特徴ベクトルの空間の次元 d よりはかなり小さい（$M << d$）ことが一般的であるので，複雑な識別境界も精度よく推定できる可能性がある。このような場合には，識別モデルを適用する方が適切である。