数学的準備体操：数理的基盤と計算の実際

理化学研究所 脳科学総合研究センター
村田 昇

1　はじめに

脳の情報表現を定量的，あるいは定性的に議論するためには土台となる数理モデルを構築することは必要不可欠である．そのモデル化と解析に必要な数理的概念，特に

1) 神経回路の数理モデル

2) 情報理論の基礎

3) 統計的推測の基礎

に焦点を当てて概説し，演習を通じて理解を深めることを目的とする．

2　神経回路の数理モデル

神経細胞は多入力一出力の情報処理素子であるが，数学的な取り扱いを容易にし本質を把握するためには，簡潔でありながら注目する現象を説明するに十分なモデルを組み立てることが重要である．以下に基本的なモデルを列挙する．

2.1　神経細胞のモデル

神経細胞の働きの主な特徴としては

空間的加算　神経細胞の内部状態は入力信号の重ね合わせで決定する

時間的加算　入力信号の影響はある時間持続し，あとから入ってきた入力と相互に干渉する

閾値作用　出力は内部状態がある閾値を越えるまで発生しない

が挙げられる．多くの単純なモデルでは空間的加算と閾値作用に着目したモデル化が行われる．これ以外の特徴としては

不応期　閾値は定数ではなく出力に応じて短期的に変化する

疲労　閾値は出力に応じて長期的に変化する

可塑性　学習や記憶は素子間の結合荷重の変化によってもたらされる

などがある．単一神経細胞の微細なモデルでは不応期や疲労のモデル化が重要となる．

数学的な記述の上からは時間を離散とするか連続とするか，出力を離散とするか連続とするか，で大きく分けられ

離散時間–離散情報 時刻 $t = 1, 2, \ldots$ において，その時点での内部状態に基づいて出力を 0 または 1 に変化させるモデル．簡単な例としては，時刻 t における入力信号 $x_1, \ldots, x_n$ に対して内部状態を重み付き和

$$u = \sum_{i=1}^{n} w_i x_i \tag{1}$$

で計算し，出力 z を

$$z = 1(u - h) = \begin{cases} 1, & u > h \\ 0, & u \leq h \end{cases} \tag{2}$$

とするものである．

離散時間–連続情報 時刻 $t = 1, 2, \ldots$ において，その時点での内部状態に基づいて出力を適当な連続値に対応させるモデル．簡単な例は，上と同様にして内部状態を計算し，適当な非線形関数 f (例えば $f(x) = \tanh(x)$) を用いて

$$z = f(u - h) \tag{3}$$

とするものである．

連続時間–連続情報 時間 t における内部状態の変化は次々にやって来る入力の集積によると考え，これを微分方程式によって記述し，出力に適当な連続値を対応させるモデル．例えば

$$\tau \frac{du(t)}{dt} = -u(t) + \sum_{i=1}^{n} w_i x_i(t) - h \tag{4}$$

$$z(t) = f(u(t)) \tag{5}$$

といった方程式が使われる．

のいづれかが用いられる．離散時間モデルはコンピュータ上でそのまま実装できるが，連続時間モデルは微分方程式を数値的に解くことになる．

2.2 神経回路網のモデル

神経回路網は基本的に多重の階層構造を持つシステムであるが，数学的取り扱いにおいては

多層回路網 フィードフォワードネットワークとも呼ばれ下層から上層に向けて単一方向の結合を持つ．工学的にも広く応用されている多層パーセプトロンなどがこれにあたる．

再帰的回路網 リカレントネットワークとも呼ばれ，層内でフィードバック結合を持つ．典型的なものはホップフィールドネットワークと呼ばれるものである．

が基本的な構成要素となる．これらを組み合わせたモデルや，単一素子の性質を特殊化したものなど様々な変種はある．

2.3 文献

数理モデル考え方，および統計的な解析を扱った成書としては

[1] 甘利俊一. 神経回路網の数理 –脳の情報処理様式–. システム サイエンス シリーズ. 産業図書, 1978.

があり，古典的で重要な文献はこれに網羅されている．上で分類したモデルはこの第 2 章に，ここでは述べなかった学習則の分類については第 8 章に詳しく述べられている．

また工学的応用に供するモデルからコネクショニズムの思想的な解説までも含んだものとしては

[2] 麻生英樹. ニューラルネットワーク情報処理 –コネクショニズム入門，あるいは柔らかな記号に向けて–. 産業図書, 1988.

などがある．

3 情報理論の基礎

脳内の情報の流れを捉える上で情報の量という考え方が重要になる．以下では情報理論における基本的な概念を説明する．

3.1 エントロピー

確率変数 X の確率密度を $P_X(x)$ としたとき，X のエントロピーは

$$H(X) = -\int P_X(x) \log P_X(x) dx \tag{6}$$

と定義される．

エントロピーは直観的には確率変数のもつランダムさの目安になる．例えば確率変数 X が $0,1$ の二項分布で，$X = 0$ となる確率を p とすると，エントロピー関数は p を変数とする一変数関数

$$H(X) = -p \log p - (1-p) \log(1-p) \tag{7}$$

となる．この関数は上に凸な連続関数で，対称性から $p = 1/2$ において最大の値を取る．p が $1/2$ に近いときには X は 0 が出るのか 1 が出るのかほぼ五分五分で予測し難いが，p が 0 または 1 に近いときには X はほとんどの場合 0 または 1 が出ることになり，どちらが出るのか予測しやすいことを考えると，エントロピーという量は確率変数の不確定性，曖昧さの尺度になると考えられる．

多変数の場合の定義も自然に拡張でき，例えば 2 変数の場合

$$H(X_1, X_2) = -\int P_{X_1X_2}(x_1, x_2) \log P_{X_1X_2}(x_1, x_2) dx_1 dx_2 \tag{8}$$

となる．条件付き確率のエントロピーを平均したもの

$$\begin{aligned} H(X_2|X_1) &= \int H(X_2|x_1) P_{X_1}(x_1) dx_1 \\ &= -\int P_{X_1}(x_1) P_{X_2|X_1}(x_2|x_1) \log P_{X_2|X_1}(x_2|x_1) dx_2 dx_1 \end{aligned} \tag{9}$$

で条件付きエントロピーを定義すると，エントロピーの加法性と呼ばれる単純な関係

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1) \tag{10}$$

が成り立つ．$H(X_1, X_2)$ は確率変数 X_1, X_2 に関する曖昧さを表わし，$H(X_1)$ は確率変数 X_1 のみの曖昧さを表わしているので，その差にあたる条件付きエントロピー $H(X_2|X_1)$ は X_1 を知ったときに残る X_2 の曖昧さを表わすと考えられ，独立性・従属性の指標になることが理解る．

3.2 相互情報量

条件付きエントロピーをある意味で対称化した量として相互情報量が定義される．相互情報量は X_1 に関する曖昧さから X_2 を知ったあとに残る X_1 の曖昧さを引いたもの

$$\begin{aligned} I(X_1; X_2) &= H(X_1) - H(X_1|X_2) \\ &\quad (= H(X_2) - H(X_2|X_1)) \end{aligned} \tag{11}$$

として定義される．これは見方を変えれば X_2 が実質的に持っている X_1 に関する情報の量とも捉えられる．相互情報量を確率分布を使って書き下すと

$$\begin{aligned} I(X_1; X_2) &= H(X_1) + H(X_2) - H(X_1, X_2) \\ &= E_{P_{X_1X_2}}\left[\log \frac{P_{X_1X_2}(X_1, X_2)}{P_{X_1}(X_1) P_{X_2}(X_2)}\right] \end{aligned} \tag{12}$$

となる．これは結合分布 $P_{X_1X_2}(x_1, x_2)$ と各周辺分布の積 $P_{X_1}(x_1)P_{X_2}(x_2)$ の間の距離を Kullback-Leibler 情報量と呼ばれる量で測ったものであり，これからも X_1, X_2 の対称性がわかる．独立性の定義は

$$P_{X_1X_2}(x_1, x_2) = P_{X_1}(x_1) P_{X_2}(x_2) \tag{13}$$

のように結合分布が周辺分布の積で書けることであるから，X_1 と X_2 が独立なら相互情報量は 0 になる．

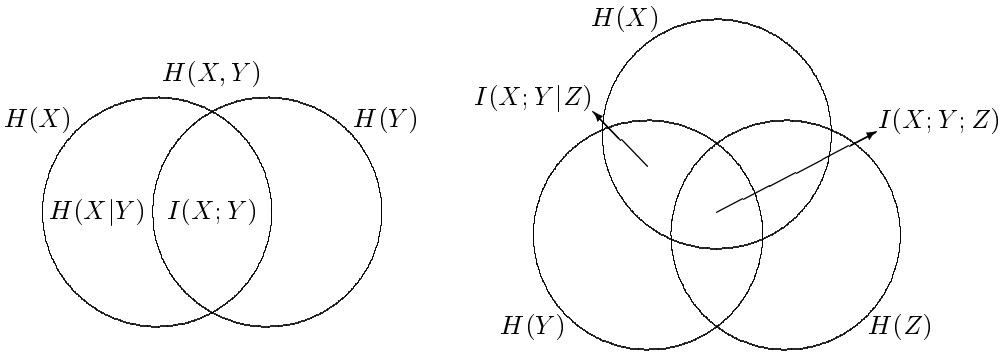

Figure 1: エントロピー・相互情報量の集合的解釈

3.3 集合的見方

エントロピーの加法性でみたように，これらの量は集合的な操作

$$
\begin{aligned}
H(X,Y) &= H(X) \cup H(Y) \\
H(X|Y) &= H(X,Y) - H(Y) = H(X) \cap \overline{H(Y)} \\
I(X,Y) &= H(X) - H(X|Y) = H(X) \cap H(Y)
\end{aligned}
$$

が成り立つ．このベン図から独立性と相互情報量の関係は容易に理解できる．

またこの見方によれば相互情報量は多変数の場合へも自然に拡張できる．例えば 3 変数 X,Y,Z の相互情報量は

$$
\begin{aligned}
& I(X;Y;Z) \\
=& I(X;Y) - I(X;Y|Z) \\
=& \{H(X) + H(Y) - H(X,Y)\} \\
& - \{H(X|Z) + H(Y|Z) - H(X,Y|Z)\} \\
=& \{H(X) + H(Y) - H(X,Y)\} \\
& - \{H(X,Z) - H(Z) + H(Y,Z) - H(Z) - H(X,Y,Z) + H(Z)\} \\
=& H(X) + H(Y) + H(Z) \\
& - H(X,Y) - H(Y,Z) - H(X,Z) + H(X,Y,Z)
\end{aligned}
$$

により定義される．

3.4 文献

情報理論に関する成書は多いが，体系的に捉えるのであれば

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. A Wiley-Interscience Publication, New York, 1991.

が定番である．和書では

[4] 韓太舜，小林欣吾． 情報と符号化の数理． 岩波講座 応用数学．岩波書店, 1994.

がよくまとまっている．

4 統計的推測の基礎

統計的な考え方の基本となるのは，観測している値をある確率分布にしたがう確率変数の実現値であると考え，観測される現象に内在する構造を確率によって記述することにある．確率的な揺らぎを含む観測値からどれだけの情報が引き出せるかを示すものとして Fisher 情報量が重要な働きをする．

4.1 Cramer-Rao の不等式

統計的推測の最も単純な設定は，母数 θ で表現される確率密度関数 $P_X(x;\theta)$ の族を想定し，その一つから発生する独立な確率変数 $X_1, X_2, \ldots, X_n$ を観測し，それを用いて分布の母数 θ を推定することである．この際 X も θ もベクトルであって構わない．観測値が適当な条件下で無限に集められるのであれば任意の精度で母数の推定を行うことができるが，有限個の観測値しか使えないところに推定の難しさがある．

n 個の独立な確率変数 $X_1, \ldots, X_n$ に基づく母数 θ の推定量を

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n) \tag{14}$$

と書くことにする．今考えている推定量は素性が良く，真の母数を θ_0 としたとき

$$E(\hat{\theta}) = E^{X_1,\ldots,X_n}(\hat{\theta}(X_1, \ldots, X_n)) = \theta_0 \tag{15}$$

が成り立っているとする．ただし $E^{X_1,\ldots,X_n}$ は $X_1, \ldots, X_n$ の分布に関して平均を取ることを意味する．このとき

$$I(\theta_0) = E^X\left(\left(\frac{d\log P_X(X;\theta_0)}{d\theta}\right)^2\right) \tag{16}$$

で定義される Fisher 情報量と推定量の分散

$$V = E^{X_1,\ldots,X_n}\left((\hat{\theta}(X_1, \ldots, X_n) - \theta_0)^2\right) \tag{17}$$

の間には

$$V \geq \frac{1}{nI(\theta_0)} \tag{18}$$

という関係が成り立つ．母数 $\boldsymbol{\theta}$ が (縦) ベクトルの場合には ij 成分を

$$I_{ij}(\boldsymbol{\theta}_0) = E^X\left(\frac{\partial\log P_X(X;\boldsymbol{\theta}_0)}{\partial\theta_i}\frac{\partial\log P_X(X;\boldsymbol{\theta}_0)}{\partial\theta_j}\right) \tag{19}$$

とする Fisher 情報行列および推定量の共分散行列

$$V = E^{X_1,\ldots,X_n}\left((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T\right) \tag{20}$$

を用いて

$$\boldsymbol{a}^T V \boldsymbol{a} \geq \frac{1}{n}\boldsymbol{a}^T I(\boldsymbol{\theta}_0)\boldsymbol{a},\ \forall \boldsymbol{a} \tag{21}$$

と表される．これは Cramer-Rao の不等式と呼ばれ，どんなに巧みな方法を用いても平均的な推定誤差は Fisher 情報量の逆数 (Fisher 情報行列の逆行列) より良くできないことを表している．

4.2 最尤推定

確率密度関数 $P_X(x,\theta)$ を母数 θ の関数とみたものを尤度関数と呼ぶ．観測値 $x_1,\ldots,x_n$ が与えられた場合尤度関数を最大にする母数の推定量を最尤推定量と呼ぶ．独立な n 変数の尤度関数は各々の密度関数の積

$$L(\theta) = P(x_1,\ldots,x_n,\theta) = P_X(x_1,\theta)\times\cdots\times P_X(x_n,\theta) \tag{22}$$

で表され，最尤推定量 $\hat{\theta}$ は

$$\max_{\theta} L(\theta) = L(\hat{\theta}) \tag{23}$$

で定義される．最尤推定量は

一致性 観測値の数 n が大きいとき，推定量 $\hat{\theta}$ が真値に確率収束する

漸近有効性 観測値の数 n が大きいとき，推定量の分散は Cramer-Rao の不等式の下限をほぼ達成する

という統計的に重要な性質を持っており，理論的に興味深い推定量である．

4.3 文献

現代的な統計の教科書は数多く出ているので，各自手持ちのものを参照されたし．理論的な基礎を追うのであれば

[5] William Feller. *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, Inc., 1957.

をお薦めする (邦訳も出ている)．