目次

第 1 章

生成モデルとベイズ脳仮説

これまでの章では，知覚 (perception) のモデル，すなわち外界からの入力に対して，どのようにして神経回路網が意味のある出力を生成するのか，という問題を主に扱ってきた．ここで改めて知覚の基本的な定義を確認しておこう．知覚とは，外界からの刺激を感覚受容器によって受容し，それに意味を与える過程である．この「刺激に意味を与える」という個所を，より体系的に理解するために，「順問題」と「逆問題」という概念を導入しよう．

一般に，ある原因から結果を予測する問題は順問題 (forward problem) と呼ばれる．逆に，観測された結果からその原因を推定する問題は逆問題 (inverse problem) と呼ばれる．視覚を例にとって，順問題と逆問題について考えてみよう．たとえば，三次元の物体が光を反射し，それが二次元の網膜上にどのような像を結ぶか，という問いは順問題に分類される．これに対して，網膜上に投影された二次元像から，元の物体の三次元的な構造や大きさ，位置などを推定する課題が逆問題である[*1]．光学の分野では，それぞれの問題は順光学 (forward optics)，逆光学 (inverse optics) と呼ばれている．逆問題は多くの場合，不良設定問題 (ill-posed problem) となる．すなわち，解が存在しない，解が一意に定まらない，あるいはわずかな誤差に対して解が大きく変化するといった性質をもつ[*2]．例えば，先ほどの例であれば同じ 2 次元像を示す 3 次元物体は複数 (あるいは無数に) 存在する．そのため，逆問題を解くには，事前知識や

[*1] 他にも逆問題は数多く存在する．逆問題は様々な分野に現れるが，ここでは医学や神経科学に関連した例として，外部から脳の構造や機能を推定する問題を取り上げる．たとえば，医用画像解析では，コンピュータ断層撮影 (computed tomography; CT)，磁気共鳴画像法 (magnetic resonance imaging; MRI)，陽電子放射断層撮影 (positron emission tomography; PET) などにおいて，観測データから画像を再構成する必要がある．この再構成処理には，CT や PET では逆ラドン変換，MRI では逆フーリエ変換が用いられる．また，神経活動を非侵襲的に計測する手法として，脳波 (electroencephalography; EEG) や脳磁図 (magnetoencephalography; MEG) がある．これらにおける電流源推定 (source localization) も典型的な逆問題である．EEG や MEG における順問題は，脳内の神経電流源の位置・方向・強度から，頭皮上の電極 (EEG) や磁場センサ (MEG) によって観測される電位や磁場分布を予測することである．一方，逆問題は，実際に観測された電位や磁場データから，神経電流源の空間的位置と活動を推定することである．この逆問題は不良設定 (ill-posed) であるため，安定的に解くには，MRI から得られた頭部の構造データに基づいて構築された順モデル (forward model) が必要となる．

[*2] これに対して，良設定問題 (well-posed problem) とは，解が存在し，一意であり，かつ入力の変動に対して連続的に変化する (安定性をもつ) ような問題を指す．良設定問題では，入力データに小さなノイズや誤差が含まれていても，求められる解は大きく変わることなく，安定に計算することができる．

仮定 (制約条件，正則化) の導入などが必要となる．

こうした逆問題を踏まえ，知覚とは単なる入力情報の受動的な処理ではなく，感覚入力という結果から外界に存在する潜在的な原因を推定する逆推論 (abductive reasoning) の過程とみなす考えがある (**helmholtz1867**; **mumford1992computational**; **kawato1993forward**; **friston2003learning**)[*3]．この枠組みを推論的知覚 (perception as inference) と呼ぶ．推論的知覚は，外界の潜在的な原因から感覚入力が生成される過程を記述する確率的生成モデル (probabilistic generative model) に基づいて説明される．確率的生成モデルについて説明する前に，前提となるベイズ推論について次節で説明する．

1.1 確率的生成モデル

1.1.1 確率的生成モデルとベイズ推論

外界から感覚入力などを通じて観測データ x を得る状況を考えよう[*4] 。観測データが存在するということは、それを生成する確率分布[*5] $p_{\text{data}}(x)$ が存在する（すなわち $x \sim p_{\text{data}}(x)$ である）と仮定できる。この $p_{\text{data}}(\cdot)$ はしばしば真の確率分布と呼ばれるが、実際にそのような分布が存在する保証はなく、多くの場合は未知である。もし $p_{\text{data}}(\cdot)$ が既知であれば、任意のサンプル x をそこから直接生成（サンプリング）できるが、現実にはこれを直接知ることはできない。

このため、観測データがある確率的な生成過程に従って生じたと仮定し、その過程を記述する生成モデルを構築する。生成モデルは分布を明示的に表現するため、新たなデータの生成や欠損値の補完、潜在構造の抽出、外界の状態推定など、多様な推論を可能にする。ここではパラメータ θ をもつ条件付き確率密度関数 $p(X \mid \theta)$ を導入し、観測の背後にある生成過程を近似的に表す。このような確率分布 $p(X \mid \theta)$ を定めるモデルを確率的生成モデル（probabilistic generative model）と呼ぶ。また、このように有限個のパラメータ θ で分布形状を規定するモデルをパラメトリックモデル（parametric model）と呼ぶ[*6]。

例えば、正規分布 $p(x \mid \theta) = \mathcal{N}(x \mid \mu, \sigma^2)$（$\theta = \{\mu, \sigma\}$）はパラメトリックな確率的生成モデルの一例である。この場合、分布の形状（正規分布）はあらかじめ固定され、未知なのはパラメータ θ である。こ

[*3] Helmholtz は，知覚を単なる感覚の受容ではなく，感覚入力に意味を与え，対象として構成する過程であると捉えた．この過程には，観念の連合 (*Vorstellungsverbindungen*) が関与している．観念の連合とは，過去の経験によって形成された (必ずしも言語化を伴わない) 観念や知識が，現在の感覚入力と結び付けられる過程を指す．通常，推論とは意識的に行われるものと考えられているが，Helmholtz はこのような観念の連合を，意識されることなく行われる推論として捉え，無意識的推論 (*unbewusster Schluss*, unconscious inference) と表現した．なお，この脚注ではドイツ語を斜体で表記した．

[*4] 厳密には、確率変数は大文字 X、その実現値は小文字 x で表記して両者を区別すべきである。しかし、応用的な文脈では両者を混同しても問題となることは少ないため、本書では明確に区別しないこととする。特に、確率変数がスカラーの場合は大文字・小文字で容易に区別できるが、ベクトルや行列を扱う場合には表記が煩雑になり、かえって可読性を損ねることとなる。このため、本書では固定された観測値が必要な場合に限り x_{obs} などの記号を用いて区別し、その旨を明記する。

[*5] 本書では確率分布を確率密度関数の意味で用いる．

[*6] 有限個のパラメータで分布形状をあらかじめ規定せず、データ量に応じて表現可能な複雑さが変化するものをノンパラメトリックモデル（non-parametric model）と呼ぶ。代表例にはヒストグラムやカーネル法による密度推定、ガウス過程、分位点回帰などがある。特に分位点回帰は分布型強化学習への応用を通じて神経科学とも関連し、その詳細は第 9 章で述べる。

こで改めて強調しておくと、目標は真の分布 $p_{\text{data}}(\cdot)$ を近似できる生成モデルを構築することである。パラメトリックモデルの場合、この目標はモデルのパラメータを適切に推定することによって達成される。

パラメータ推定には、大きく分けて二つの方法がある。一つは、パラメータの最適な一点の値を求める点推定であり、もう一つはパラメータを確率変数として扱い、その不確実性を含めて推定する分布推定である。パラメータ推定には大きく二つの方法がある。一つは、パラメータの最適な一点の値を求める点推定であり、もう一つはパラメータを確率変数として扱い、その不確実性を含めて推定する分布推定である。分布推定には様々な方法があるが、ここではその代表例としてベイズ推論（Bayesian inference）を取り上げる。ベイズ推論では、観測前のパラメータ分布を事前分布（prior）$p(\theta)$、観測後の分布を事後分布（posterior）$p(\theta \mid X)$ と呼び、事前分布を事後分布へと更新する。この更新は、尤度（likelihood）とベイズの定理（Bayes' theorem）に基づいて行われる。

尤度

尤度に関してであるが，そもそも先ほど導入した $p(x \mid \theta)$ は観測データ x とパラメータ θ のどちらを変数とみなすかによって 2 通りの解釈がある．確率モデルとして解釈する場合，θ を固定し, x を確率変数として扱う．このとき，$p(x \mid \theta)$ は「θ が与えられたときに, どのようなデータ x がどの確率で得られるか」を表す．尤度として解釈する場合は，観測データを固定して, θ を変数として扱う．このとき，

$$L(\theta; x) := p(x \mid \theta) \tag{1.1}$$

を尤度関数 (likelihood function) と呼ぶ[*7]．尤度は, 仮定した θ の下でデータが得られる「尤もらしさ」を定量化する指標であり, 値が大きいほどデータをよく説明すると解釈できる．なお, 尤度関数は θ に関する確率分布ではないため, θ について積分しても必ずしも 1 にはならない．

ベイズの定理

事前分布と尤度が設定されれば，事後分布を次式で表されるベイズの定理によって求めることができる:

$$\underbrace{p(\theta \mid x)}_{\text{事後分布}} = \frac{\overbrace{p(x \mid \theta)}^{\text{尤度}}\ \overbrace{p(\theta)}^{\text{事前分布}}}{\underbrace{p(x)}_{\text{周辺尤度}}} \tag{1.2}$$

ここで

$$p(x) = \int p(x \mid \theta)\, p(\theta)\, \mathrm{d}\theta \tag{1.3}$$

は周辺尤度 (marginal likelihood) あるいは 証拠 (evidence)，正規化定数 (normalization constant) と呼ばれる．このベイズの定理は

$$p(\theta, x) = p(x \mid \theta)\, p(\theta) = p(\theta \mid x)\, p(x) \tag{1.4}$$

[*7] ここで，$L(\theta; x)$ は L は θ の関数であり，; の後の x は固定された引数であることを意味する．

が成り立つことから導かれる．ここで $p(\theta, x)$ は θ と x の同時確率分布 (joint probability distribution) である．

予測分布

事後分布 $p(\theta \mid x)$ は、観測データを得た後のパラメータ θ の確率分布であり、ベイズ推論はこの分布を更新する過程とみなせる。十分に更新された事後分布が得られれば、新たなデータの予測も可能となる。予測においては事後分布全体を平均化した事後予測分布 (posterior predictive distribution) を用いる．観測されたデータの実現値を x, 予測対象のデータを $\tilde{x}$ とすると,

$$p(\tilde{x} \mid x) = \int p(\tilde{x}, \theta \mid x)\, \mathrm{d}\theta \tag{1.5}$$

$$= \int p(\tilde{x} \mid \theta, x)\, p(\theta \mid x)\, \mathrm{d}\theta \tag{1.6}$$

$$= \int p(\tilde{x} \mid \theta)\, p(\theta \mid x)\, \mathrm{d}\theta \quad (\because \tilde{x} \perp\!\!\!\perp x \mid \theta) \tag{1.7}$$

となる．ここで最後の等式は,「θ が与えられた条件下で $\tilde{x}$ と x が独立」という条件付き独立性を用いたものである．ベイズ推論において, この事後予測分布 $p(\tilde{x} \mid x)$ が真の生成分布の推論結果となる．なお, データの観測をしていない場合の予測は周辺尤度と同様の形式

$$p(\tilde{x}) = \int p(\tilde{x} \mid \theta)\, p(\theta)\, \mathrm{d}\theta \tag{1.8}$$

で与えられ，これを事前予測分布 (prior predictive distribution) と呼ぶ.

1.1.2 ベイズ線形回帰

1.1.3 最尤推定と MAP 推定

1.1.4 潜在変数モデル

ここで，$p_\theta(\mathbf{x})$ は，観測変数 $\mathbf{x}$ に対する条件付き分布 $p(\mathbf{x} \mid \theta)$ の略記である．
確率的主成分分析

エネルギーベースモデル

1.1.5 階層ベイズモデル

1.1.6 スパース符号化モデル

生成モデルの学習における目的は，パラメータ θ を調整して，生成モデルが定める確率密度関数 $p_\theta(\mathbf{x})$ を，学習データが従う真の分布 $p_{\mathrm{data}}(\mathbf{x})$ に近づけることである．この「近づける」という操作には，両分

布間の差異を定量化する指標，すなわち確率分布間の距離 (あるいは不一致度) を定義する必要がある．ここではその尺度として，Kullback – Leibler ダイバージェンス (KL ダイバージェンス) を用いる：

$$D_{\mathrm{KL}}\left(p_{\mathrm{data}}(\mathbf{x}) \,\|\, p_\theta(\mathbf{x})\right) := \int p_{\mathrm{data}}(\mathbf{x}) \log \frac{p_{\mathrm{data}}(\mathbf{x})}{p_\theta(\mathbf{x})}\, d\mathbf{x} \tag{1.9}$$

この KL ダイバージェンスは，真の分布 $p_{\mathrm{data}}(\mathbf{x})$ を基準としたときに，モデル分布 $p_\theta(\mathbf{x})$ がどれだけ情報的に乖離しているかを測る指標である．すなわち，モデルが生成する分布が，実際のデータ分布からどの程度逸脱しているかを定量化するものである．この KL ダイバージェンスを展開すると，

$$\begin{aligned} D_{\mathrm{KL}}\left(p_{\mathrm{data}}(\mathbf{x}) \,\|\, p_\theta(\mathbf{x})\right) &= \int p_{\mathrm{data}}(\mathbf{x}) \log \frac{p_{\mathrm{data}}(\mathbf{x})}{p_\theta(\mathbf{x})}\, d\mathbf{x} && (1.10)\\ &= \int p_{\mathrm{data}}(\mathbf{x}) \log p_{\mathrm{data}}(\mathbf{x})\, d\mathbf{x} \; - \int p_{\mathrm{data}}(\mathbf{x}) \log p_\theta(\mathbf{x})\, d\mathbf{x} && (1.11)\\ &= \text{const.} - \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}\left[\log p_\theta(\mathbf{x})\right] && (1.12) \end{aligned}$$

となる．ここで第 1 項は θ に依存しない定数であるため，パラメータ θ を最適化する際には，第 2 項 (対数尤度の期待値) のみを考慮すればよい．したがって，最適なパラメータ θ^* は，

$$\theta^* = \arg\min_\theta D_{\mathrm{KL}}\left(p_{\mathrm{data}} \,\|\, p_\theta\right) = \arg\max_\theta \mathbb{E}_{\mathbf{x}\sim p_{\mathrm{data}}}\left[\log p_\theta(\mathbf{x})\right] \tag{1.13}$$

として求められる．しかし実際には，真の分布 $p_{\mathrm{data}}(\mathbf{x})$ の形は不明であり，観測されるのは有限個のデータ点 $\{\mathbf{x}_i\}_{i=1}^N$ のみである．そこで，真の分布の代替として，以下のような経験分布 (empirical distribution) $\hat{p}_{\mathrm{data}}(\mathbf{x})$ を用いる：

$$\hat{p}_{\mathrm{data}}(\mathbf{x}) := \frac{1}{N}\sum_{i=1}^{N} \delta(\mathbf{x} - \mathbf{x}_i) \tag{1.14}$$

ここで，$\delta(\cdot)$ は Dirac のデルタ関数であり，この経験分布 $\hat{p}_{\mathrm{data}}(\mathbf{x})$ は，観測された各データ点の位置にのみ確率を集中させるような離散的な点分布である．すなわち, サンプル $\{\mathbf{x}_i\}_{i=1}^N$ 以外の点では確率密度がゼロであり, 各 $\mathbf{x}_i$ に等しい重み $1/N$ を割り当てている．この近似を用いることで，最適化問題は次のように書き換えられる：

$$\theta^* \approx \arg\max_\theta \mathbb{E}_{\mathbf{x}\sim \hat{p}_{\mathrm{data}}}\left[\log p_\theta(\mathbf{x})\right] = \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(\mathbf{x}_i) \tag{1.15}$$

これは，観測されたデータに対する対数周辺尤度のサンプル平均を最大化する操作に対応し，最尤推定と一致する．

この最適化問題をさらに具体的に扱うためには，確率密度関数 $p_\theta(\mathbf{x})$ の形式を明示的に定める必要がある．そこで次に，この $p_\theta(\mathbf{x})$ をどのような構造のもとに構築するかを紹介する．