# ベイズ推定

Bayesian estimation

機械学習の目的は, データ $X$ からそれを説明するパラメータ $\theta$ を推定することである. しかし, $X$ は通常は有限で, 非常に少ないこともあり, $\theta$ の値を一意に決めるには不充分であることが多い. ベイズ推定は, このような場合でもパラメータ $\theta$ を確率分布として表現する方法であり, 18 世紀の英国の牧師 Thomas Bayes の発見にその起源を持つ. これにより, $\theta$ 自体がさらに確率分布に従う場合 (**階層ベイズ**) も, ベイズ推定では自然に扱うことができる.

## 1. 簡 単 な 例

たとえば, ある未知の確率 $q$ で表が出る ($=(1-q)$ の確率で裏が出る) コインを 4 回投げたところ, 結果が次のように, すべて表だったとしよう. このとき, $q$ の値はいくつだと推定すればよいのだろうか.

表 表 表 表

最尤推定に基づけば, この事象の確率は $p(X|q) = q^4(1-q)^0$ であり, これを最大にする $q$ の最尤推定値は $\hat{q}=1$ となる. すると, このコインは絶対に表が出ると考えることになるが, この結論はあまりに極端すぎるように思える.

そこで, たった 4 回の観測で $q$ を一意に決めたりせず, $q$ について分布を導入することにしてみよう. $q$ 自体が確率であるから, これは確率自体の確率分布となり, もっとも簡単なものとして, 次のベータ分布

$$p(q) = \mathrm{Be}(\alpha,\beta) \propto q^{\alpha-1}(1-q)^{\beta-1} \quad (1)$$

を使ってみる. 期待値は $E[q]=\alpha/(\alpha+\beta)$ であり, $\alpha=\beta=1$ のとき, $\mathrm{Be}(1,1)$ は $[0,1]$ の一様分布となる.

このとき, 上の観測 $X$ がわかった後での $q$ の分布 $p(q|X)$ は, ベイズの定理によると,

$$p(q|X) = \frac{p(q,X)}{p(X)} \propto p(q,X) = p(X|q)p(q) \quad (2)$$

であるから, $\mathrm{Be}(1,1)$ を事前分布とすれば

$$p(q|X) \propto p(X|q)\cdot p(q) \quad (3)$$

$$= q^4 \cdot q^{1-1}(1-q)^{1-1} = \mathrm{Be}(5,1) \quad (4)$$

となった. この分布は図 1 のようになり, 期待値は $E[q|X]=5/(5+1)=0.833$ である. 無事, 1 でない値が得られた!

一般に, パラメータ $\theta$ に**事前分布** $p(\theta)$ を置き, $\theta$ の下でのデータ $X$ の確率 (**尤度**) $p(X|\theta)$ から

$$p(\theta|X) \propto p(X|\theta)p(\theta) \quad (5)$$

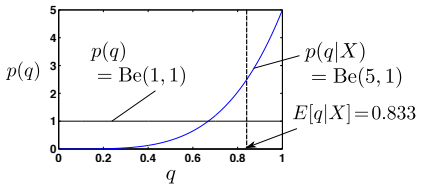



**図 1** コインの表が出る確率 $q$ のベイズ推定.

として $\theta$ の**事後分布**を求める方法を, ベイズ推定という. ベイズ推定は, 上の例のように最尤推定から得られる極端な解を緩和する効果があり, 特にデータ量が少ない時[*1)]に効果を発揮する.

さらに, ベイズ推定ではパラメータが確率変数であるため, 最初に述べたようにそれもさらに上位の確率分布から生成されたと考えること (階層ベイズ) により, 事前分布自体も学習する柔軟なモデリングが可能になる.

## 2. ベイズ統計のノンパラメトリック推定

上ではスカラー値のパラメータ $\theta$ の値を確率分布として表現する方法を示したが, それでは, $\theta$ が関数や分布の場合, ベイズ推定はどうすればよいのだろうか. この場合の $\theta$ の事前分布として機械学習で最も有名なものが, 連続の場合の**ガウス過程**と, 離散の場合の**ディリクレ過程**である. 以下, この 2 つについて解説する.

### 2.1 ガウス過程

ガウス過程 (Gaussian process, GP) とは, 「入力ベクトル $\mathbf{x}$ が似ていれば, 出力値 $y$ も似ている」ことを表すための回帰関数 (regressor) の確率モデルであり, 無限次元のガウス分布とも考えることができる.

GP では, 出力値 $y$ を, 入力 $\mathbf{x}$ に対する $H$ 個の基底関数 (=入力値の関数) $\phi_1(\mathbf{x}),\cdots,\phi_H(\mathbf{x})$ の線形結合

$$y = \mathbf{w}^T\phi(\mathbf{x}) = w_1\phi_1(\mathbf{x}) + \cdots + w_H\phi_H(\mathbf{x}) \quad (6)$$

でモデル化する. $n$ 個の入力 $\mathbf{x}^{(1)}\cdots\mathbf{x}^{(n)}$ と対応する出力 $y^{(1)}\cdots y^{(n)}$ について行列形式で書くと,

$$\underbrace{\begin{pmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} \phi_1(\mathbf{x}^{(1)})\cdots\phi_H(\mathbf{x}^{(1)}) \\ \ddots \\ \phi_1(\mathbf{x}^{(n)})\cdots\phi_H(\mathbf{x}^{(n)}) \end{pmatrix}}_{\mathbf{\Phi}} \underbrace{\begin{pmatrix} w_1 \\ \vdots \\ w_H \end{pmatrix}}_{\mathbf{w}} \quad (7)$$

すなわち, $\mathbf{y}=\mathbf{\Phi}\mathbf{w}$ である. いま, $\mathbf{w}$ がガウス分布

[*1)] データ全体が多くても, あるカテゴリに属するデータ (例えば, 関東地方で雪が降った日の積雪量) は非常に少ないことがあり, ベイズ推定はそのような場合にも有用である.

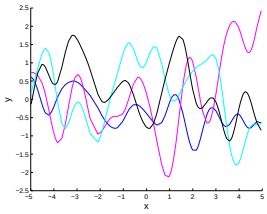


**図 2** ガウス過程からのサンプル (ガウスカーネル).

$\mathrm{N}(\mathbf{0}, \alpha^{-1}\mathbf{I})$ に従っているとすると, その線形変換である $\mathbf{y}$ もガウス分布に従い, 平均 $\mathbf{0}$, 分散

$$E[\mathbf{y}\mathbf{y}^T] = E[(\mathbf{\Phi}\mathbf{w})(\mathbf{\Phi}\mathbf{w})^T] = \mathbf{\Phi}E[\mathbf{w}\mathbf{w}^T]\mathbf{\Phi} \quad (8)$$

$$= \alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T \quad (9)$$

のガウス分布となる.

上の性質が任意の $\mathbf{y}$ について成り立つとき, $\mathbf{y}$ はガウス過程に従う, という. すなわち, $\alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T = \mathbf{K}$ とおくと,

$$\mathbf{y} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}) \quad (10)$$

と考えていることになる.

式 (10) は任意の次元の $\mathbf{y}$ について成り立つから, ガウス過程とは無限次元のガウス分布のことであり, (10) はそれをデータの存在する次元に関して周辺化したものだといえる. ガウス分布を任意の次元について周辺化しても, またガウス分布となることを思い出そう.

ここで, $\mathbf{K}$ の要素を $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ とすると,

$$k(\mathbf{x}_i, \mathbf{x}_j) = \alpha^{-1}\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \quad (11)$$

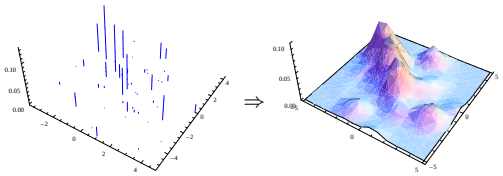
だけで GP が定まることに注意しよう. 式 (11) は $\mathbf{x}_i$ と $\mathbf{x}_j$ の「近さ」を与えるカーネル関数であり, 基底関数表示 $\phi(\mathbf{x})$ を陽に使わずに, カーネル関数 $k(\mathbf{x}_i, \mathbf{x}_j)$ だけで $y$ を求めることができる. この意味で, GP はベイズ的な (事後分布をもつ) **カーネルマシン**とも考えることができる.

カーネル関数として, ガウスカーネル $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-(\mathbf{x}_i - \mathbf{x}_j)^2/2)$ を用いた場合のガウス過程の出力の例を図 2 に示す. これは, 無限個の基底関数 $\phi(\mathbf{x})$ を考えたことに相当している.

ガウス過程は, 座標 $\mathbf{x}$ (典型的には, 時間や空間) 上のランダムな関数を与えると考えることができるため, 機械学習における多様な回帰問題のほか, 時系列解析や空間統計など, 様々な場所で使われている. ガウス過程について詳しくは, 成書[1] を参照されたい.

### 2.2 ディリクレ過程

これに対して, ディリクレ過程は離散分布の分布であり, 無限次元のディリクレ分布といってよい. ディリクレ分布とは, $K$ 次元の多項分布 $\mathbf{q} = (q_1, q_2, \cdots, q_K)$ の最も簡単な分布であり, 式 (1) のベータ分布の多次元版 (多変量ベータ分布) として,



DP からの無限個のクラスタ. 無限ガウス混合モデル.

**図 3** ディリクレ過程による無限ガウス混合モデル.

$$p(\mathbf{q}) = \mathrm{Dir}(\mathbf{q}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} q_k^{\alpha_k - 1} \quad (12)$$

で与えられる. パラメータは $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ である. ディリクレ分布の期待値は,

$$E[\mathbf{q}] = \bar{\boldsymbol{\alpha}} = (\alpha_1, \alpha_2, \cdots, \alpha_K)/\alpha \quad (13)$$

($\alpha = \sum_{k=1}^{K} \alpha_K$) であり, 実際にサンプルすると, この期待値を中心に, 集中度 $\alpha$ によって確率的にずれた分布が得られる.

ディリクレ過程 $\mathrm{DP}(\alpha, G_0)$ とはこの無限次元版であり, 上の $\bar{\boldsymbol{\alpha}}$ に相当する連続分布 $G_0$ に似た, 無限次元の離散分布 $G \sim \mathrm{DP}(\alpha, G_0)$ を作りだす.

実際には, 無限次元の $G$ 自体を直接扱うことは不可能なため, $G$ に従う離散データ $X_1, X_2, \cdots, X_n$ が与えられた時の $X_{n+1}$ の予測分布は

$$\begin{aligned} &p(X_{n+1}|X_1, \cdots, X_n) \\ &\quad = \int p(X_{n+1}|G)p(G|X_1, \cdots, X_n)dG \\ &\quad = \sum_{i=1}^{N} \frac{1}{\alpha+n}\delta(X_i) + \frac{\alpha}{\alpha+n}G_0(X_{n+1}) \end{aligned} \quad (14)$$

であること (中国料理店過程, CRP) を用いて, 逐次的に計算する. 詳しくは,[2] を見られたい.

ディリクレ過程はべき分布に従うクラスタリングを確率的に表現できるため, ディリクレ過程を事前分布としたベイズ推定では, 機械学習におけるクラスタ数, カテゴリ数, 単語種数, …などの上限を決めず, データに応じて適応的に学習することが可能になる. 図 3 に, 無限ガウス混合モデル (Infinite Gaussian Mixture Model) の例を示した. こうした性質から, ディリクレ過程やその拡張は, 統計的言語処理, 画像処理, バイオインフォマティクスなど, 多方面で現在適用が進んでいる.

**[持橋大地]**

### 参考文献

[1] Carl Edward Rasmussen and Christopher K. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[2] Nils Lid Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker. *Bayesian Nonparametrics*. Cambridge University Press, 2010.