最適化理論

―最大事後確率推定と最尤推定―

今回学ぶこと

- **最大事後確率推定法** (MAP **推定法**)
- **最尤推定法**
- **応用：最尤推定を利用した確率分布学習**

最大事後確率推定法の考え方

ここでは、ベイズ推論の枠組からはいったん外れて、重要な推定則である

- **最大事後確率推定** (maximum a posteriori probability estimation; MAP estimation)
- **最尤推定** (maximum likelihood estimation; ML estimation)

について学ぶ。

これらの推定法は、観測できない確率変数の実現値、またはパラメータを値を予想 (推定) するタイプの推定手法であり、点推定手法のひとつである。

最大事後確率推定と最尤推定は、ベイズ推定とは違う考え方に基づく推定手法であるが、経験ベイズ法と呼ばれるベイズ推論手法ではハイパーパラメータの推定に最尤法の考え方を用いている。また、事前分布の推定において最尤法が用いられることもある。

説明のための問題設定

- **系は二つの離散型確率変数 X, Y を含む。**
- **$P_X(x)$ と $P_{Y|X}(y|x)$ が既知。**
- **Y の実現値 $y^* \in D(Y)$ が観測されている。**

MAP 推定則

最大事後確率推定則, MAP 推定則

X の推定値を $\hat{x}$ とするとき、

$$\hat{x} = \arg \max_{x \in D(X)} P_{X|Y}(x|y^*)$$

と $\hat{x}$ を決める推定法を MAP 推定法と呼ぶ。

考え方

- 事後確率分布の最大値 (モード) を X の推定値と考える。
- ベイズ的考え方では X の分布に注目するのに対して、X の推定値を一点定める考え方 (点推定) の考え方に基づく。

MAP 推定則の使いやすい形

多くの場合、前述の MAP 則の形をそのまま使うのではなく、下記の最後の等式の形で利用することが多い。

$$\begin{aligned}\hat{x} &= \arg \max_{x \in D(X)} P_{X|Y}(x|y^*) \\ &= \arg \max_{x \in D(X)} \frac{P_{Y|X}(y^*|x) P_X(x)}{P_Y(y^*)} \\ &= \arg \max_{x \in D(X)} P_{Y|X}(y^*|x) P_X(x)\end{aligned}$$

- **2 つ目の等式はベイズ則による。3 つ目の等式は分母の $P_Y(y^*)$ が変数 x に依存しないことから成り立つ。**
- **「尤度関数と事前確率分布の積を最大化する x を推定値とする」と解釈できる。**

MAP 推定則に関する問い

- $D(X) = D(Y) = \{0, 1\}$
- **事前確率分布**

$$P_X(0) = 0.3, P_X(1) = 0.7$$

- **条件付き確率分布**

$$\begin{aligned} P_{Y|X}(0|0) &= 0.5, P_{Y|X}(1|0) = 0.5 \\ P_{Y|X}(0|1) &= 0.25, P_{Y|X}(1|1) = 0.75 \end{aligned}$$

- **観測値** $y^* = 1$

以上の状況で MAP **推定を行え。**

解答

$$P_{Y|X}(1|x) \propto f(x) \stackrel{\triangle}{=} P_X(x)P_{Y|X}(1|x)$$

- $f(0)$ **を求める**:

$$f(0) = P_X(0)P_{Y|X}(1|0) = 0.3 \times 0.5 = 0.15$$

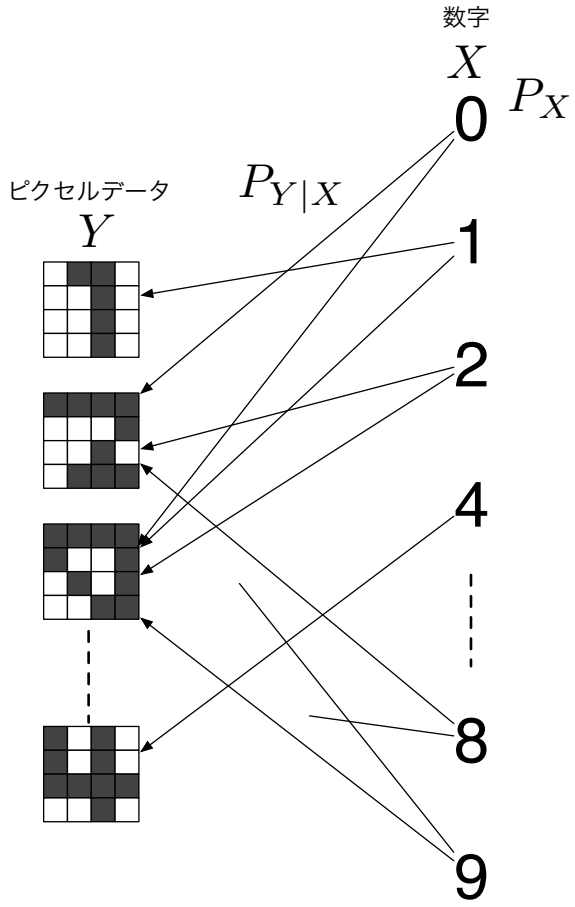
- $f(1)$ **を求める**:

$$f(1) = P_X(1)P_{Y|X}(1|1) = 0.7 \times 0.75 = 0.525$$

- $f(1) > f(0)$ **であるので、**$\hat{x} = 1$ **と推定する。**

MAP 推定の応用例：パターン認識

0 から 9 までの 10 個の数字に関する手書き文字認識システムを考える。書かれた数字は 16×16 白黒ピクセル (256 ビット) のデータして表現されるものとしよう。

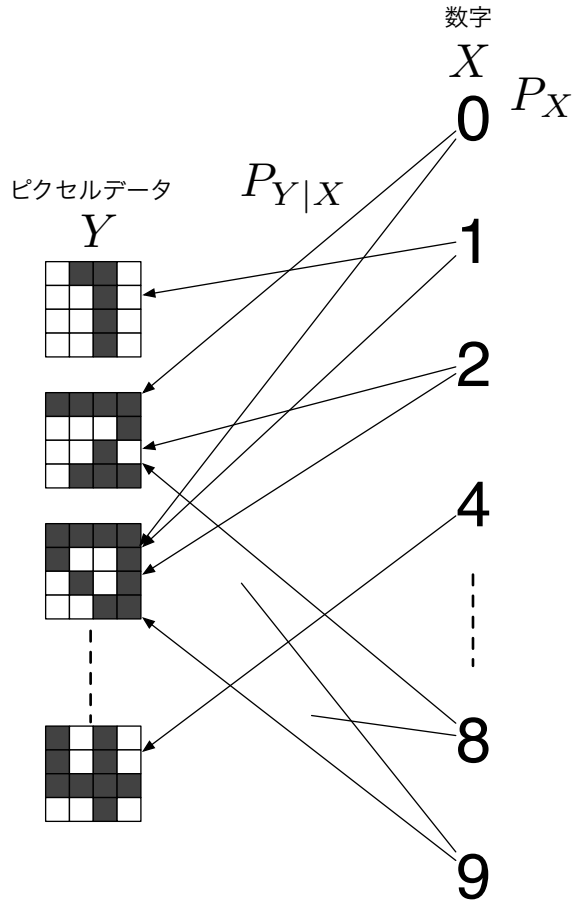


設定をもう少し詳しく

- 書いた数字を表す確率変数を X とする。
  $D(X) = \{0, 1, 2, \ldots, 9\}$
- ピクセルデータを Y とする。ここで、$D(Y)$ のサイズは 2^{256} である。
- 過去の大量の手書きデータから、十分に信用のおける $P_{Y|X}(y|x)$ が与えられているものとする。
- また、文字の生起確率 $P_X(x)$ も既知であるとする。
- 「手書き文字認識問題」として解きたい問題は、「観測された y^*(ピクセルデータ) から、書かれた数字を推定せよ」という問題である。

MAP 推定に基づく手書き認識

$$\hat{x} = \arg \max_{x \in D(X)} P_{Y|X}(y^*|x) P_X(x)$$



最尤推定則 (ML 推定則)

最尤推定則 (ML 推定則)

X の推定値を $\hat{x}$ とするとき、

$$\hat{x} = \arg \max_{x \in D(X)} P_{Y|X}(y^*|x)$$

と $\hat{x}$ を決める推定法を最尤推定法と呼ぶ。

考え方

- 最尤推定 = 事前分布が等確率分布の場合の MAP 推定
- 事前確率分布が必要ない。

対数尤度関数

観測値 y^* が与えられているとき、尤度関数の対数

$$L(x) \stackrel{\triangle}{=} \log P_{Y|X}(y^*|x)$$

を対数尤度関数と呼ぶ。ML 推定法は、対数尤度関数を最大化する x を推定値とする：

$$\hat{x} = \arg \max_{x \in D(X)} L(x)$$

最尤推定に基づく確率分布の学習

- 連続確率変数に基づく確率的な推論を行うシステムを構築する場合には、事後確率計算の前提となる同時確率密度関数・条件付確率密度関数を何らかの手段により与える必要がある。

- 推定したい確率密度関数を少数のパラメータ θ を持つ確率密度関数 $f(x|\theta)$ でモデル化し、適切にパラメータを最適化することにより確率密度関数を推定する手法をパラメトリック分布推定法と呼び、モデル関数 $f(x|\theta)$ をパラメトリックモデルと呼ぶ。

最尤推定に基づく確率分布の学習の例

未知の確率密度関数 $p_X(x)$ に従って生起した独立な観測値 a_1, a_2 がある。この p_X を次のパラメトリックモデルで近似したい。ここでモデルとなる確率密度関数を

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

とする (平均 θ、分散 1 のガウス分布)。最尤法を利用して、推定確率密度関数を求めよ。すなわち、$\hat{\theta}$ を求めよ。

ヒント：最尤推定に基づく確率分布の学習の例

サンプルが 2 つありそれぞれが独立であるので、2 次元分布

$$F(x_1, x_2|\theta) = \frac{1}{2\pi} \exp\left(-\frac{(x_1-\theta)^2 + (x_2-\theta)^2}{2}\right)$$

に対して ML 推定を行うことになる。

解答：最尤推定に基づく確率分布の学習の例

対数尤度関数を考える：

$$L(\theta) \stackrel{\triangle}{=} -\frac{(a_1-\theta)^2+(a_2-\theta)^2}{2}+C$$

$L(\theta)$ は 2 次関数であり上に凸であることから、その極値で最大値を取る。極値条件

$$\frac{L(\theta)}{d\theta}=(a_1-\theta)+(a_2-\theta)=0$$

を解くことで、最尤推定パラメータ

$$\hat{\theta}=\frac{a_1+a_2}{2}$$

を得る。

推定確率密度関数

最尤推定パラメータを最初のモデルに代入することにより、学習された分布 (推定確率密度関数)

$$f(x|\hat{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \frac{a_1+a_2}{2})^2}{2}\right)$$

が得られる。

最尤推定に基づく確率密度関数の学習

分布学習

未知の分布に従う独立な出力 $a_1, a_2, \ldots, a_T$ に対して、パラメトリック分布 $f(x|\theta)$ で分布学習を行う場合は、対数尤度関数の最大値を与える最尤パラメータ $\hat{\theta}$ を

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{T} \log f(x_i|\theta)$$

として求める。このとき、推定確率密度関数は $f(x|\hat{\theta})$ となる。

最尤推定と最適化

本講義で扱うほとんどすべての確率モデルにおいて、尤度関数がギブス分布

$$f(x|\theta) = \frac{1}{Z} \exp\left(-E(x,\theta)\right) \tag{1}$$

の形になる。ここで、Z は正規化定数である。関数 $E(x,\theta)$ はエネルギー関数と呼ばれる実数値関数であり、任意の x, θ について、$E(x,\theta) \geq 0$ である。

ギブス分布に対する最尤推定則

対数尤度関数の最大化を考えることにより、ただちに次の最尤推定則が得られる。

ギブス分布の尤度関数に対する最尤推定則

ここでは、$\theta = (\theta_1, \theta_2, \ldots, \theta_m) \in \mathbb{R}^m$ とする。尤度関数 (1) を仮定し、観測値を x^* とする。このとき最尤推定則は

$$\hat{\theta} = \arg\min_{\theta} E(x^*, \theta) \tag{2}$$

という形に書くことができる。

分布の学習 = 最適化

この最適化問題 (= 最尤推定、分布の学習) の代表的な解法は次の通り：

(1) **勾配ベクトル = 0 を解く** エネルギー関数が微分可能な関数であり制約条件が無い場合には、エネルギー関数の勾配ベクトル (グラディエント, gradient) を 0(極値条件) としそれを解く。すなわち、

$$\nabla E(x^*, \theta) = \begin{pmatrix} \frac{\partial E(x^*,\theta)}{\partial \theta_1} \\ \frac{\partial E(x^*,\theta)}{\partial \theta_2} \\ \vdots \\ \frac{\partial E(x^*,\theta)}{\partial \theta_m} \end{pmatrix} = 0 \qquad (3)$$

から得られる連立方程式を解くことにより、解析的に局所解が求まる

(2) **ラグランジュの未定定数決定法** 制約条件がある場合には、ラグランジュの未定定数決定法 (あるいは KKT 条件を解く) により、解析的に最適解を見つけることができる場合もある。

つづき

(3) **勾配法の利用** 数値的に最小化問題を解く手法の一つ。目的関数の勾配ベクトルの情報に基づいて最適解を探索する。

(4) **ニュートン法の利用** ニュートン法では、勾配ベクトル加えて目的関数のヘッセ行列 (2 次の微分) を利用することにより、勾配法よりも速い収束が達成される。

(5) **凸最適化手法の利用** 凸最適化問題 (目的関数・制約関数が凸関数)

(6) **他のヒューリスティック最適化手法の利用** 非線形計画法、シミュレーテッドアニーリング、遺伝的アルゴリズムなど。