演習問題 (2) 離散型確率変数の事後確率計算

1　事後確率の計算: 2 変数の場合

1.1　同時分布からの計算

確率変数 X, Y からなる系が与えられているものとし、その同時分布が既知であるものとする。Y の実現値 $y \in D(Y)$ が観測されたとき、未知の X に関して事後確率 (posterior probability)$P_{X|Y}(x|y)$ を計算したいものとしよう。

表 1: 確率変数 X, Y の同時分布 $P_{XY}(x, y)$

$y \backslash x$	0	1
0	0.1	0.2
1	0.4	0.3

例として、表 1 に示される同時分布を持つ確率変数 X, Y を考える。確率変数 X, Y の間の確率的相関はこの同時分布により定まる。この確率的相関を利用することにより、推論が可能となる。いま、$y = 1$ が観測されたときの X の事後確率を考えてみよう。条件付確率の定義より直ちに

$$P_{X|Y}(0|1) = \frac{P_{XY}(0,1)}{P_Y(1)} = 0.4/0.7 = 4/7 \quad (1)$$

$$P_{X|Y}(1|1) = \frac{P_{XY}(1,1)}{P_Y(1)} = 0.3/0.7 = 3/7 \quad (2)$$

を得る。観測がない場合の我々の X に関する知識は、周辺確率 $P_X(0) = P_X(1) = 0.5$ というものであった。$y = 1$ の観測から、X が 0 である確率が 1 である確率よりも高いという情報を得られたことになる。

> 問 1：確率変数 X_1, X_2 の同時分布は表 2 に示されている。$X_1 = 1$ を観測した場合の X_2 に関する事後確率を求めよ。

表 2 をみると $X_1 = 0, X_2 = 0$ の組が非常に起こりやすいことが見て取れる。すなわち、$X_1 = 0$ で

表 2: 確率変数 X_1, X_2 の同時分布

$X_2 \backslash X_1$	0	1
0	0.8	0.1
1	0.1	0

あれば、未観測の X_2 も 0 である可能性は非常に高い。事後確率を計算することにより、このような確率的な相関に基づいた推論が可能となる。

1.2　ベイズ則に基づく計算

確率変数 W は道が濡れているかどうかを示す確率変数であり、$D(W) = \{0, 1\}$ である。一方、確率変数 S は道で滑ったかどうかを表す確率変数であり、$D(S) = \{0, 1\}$ である。

次の問題を考えてみよう。

> 「道で人が滑っているのを観測した。道が濡れている確率は？」

この問題は、観測値 $S = 1$ のもとで事後確率 $P_{W|S}(w|1)$ を計算する問題と考えることができる。ここで、$P_W(0) = 0.7, P_W(1) = 0.3$ である。また、条件付確率 $P_{S|W}(s|w)$ は表 3 に与えられている。

表 3: $P_{S|W}(s|w)$ の値

$S \backslash W$	0	1
0	0.9	0.5
1	0.1	0.5

> 問 2：事後確率 $P_{W|S}(w|1)$ を求めよ。

この問題は、典型的な「確率に基づく推論問題」である。問題の構造を改めて整理してみよう。

上記の問題の構造

- 原因を表す確率変数 X について、確率分布 $P_X(x)$ は既知である (上の例では、$X = W$)。
- 確率変数 X の実現値は観測されておらず未知。
- 結果を表す確率変数 Y と X との間の**確率的因果関係**が $P_{Y|X}(y|x)$ でモデル化されている。$P_{Y|X}(y|x)$ は既知 (上の例では、$Y = S$)。
- 確率変数 Y の実現値 $y \in D(Y)$ が観測されている。
- X に関する事後分布 $P_{X|Y}(x|y)$ を求めよ。

この計算で重要な役割を果たすのがベイズ則である。

$$P_{X|Y}(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{P_Y(y)} \tag{3}$$

$$= \frac{P_X(x)P_{Y|X}(y|x)}{\sum_{x \in D(X)} P_X(x)P_{Y|X}(y|x)} \tag{4}$$

右辺に注目するとこの問題設定において全て既知の量であることに注意したい。

ベイズ則に登場する量の呼び名

- $P_X(x)$: 事前確率 (prior probability) 変数 x に関する 1 変数関数。
- $P_Y(y)$: 規格化定数、エビデンス (y は定まっているので定数になる)
- $P_{Y|X}(y|x)$ 尤度関数 (likelihood function) 変数 x に関する 1 変数関数。
- $P_{X|Y}(x|y)$: 事後確率 (posterior probability) 変数 x に関する 1 変数関数。

特に注意する点として、$P_{Y|X}(y|x)$ は確率ではない点に注意が必要である。すなわち、$P_{Y|X}(y|x)$ を x について和をとっても「足して 1 ルール」は一般には成立しない。

事前確率 (事前分布) は対象とする系に関する初期の知識であるとみることができる。観測値 y により、事後確率を計算することにより「観測値 y」の影響を織り込んだ系の知識として $P_{X|Y}(x|y)$ が新たに得られるものと考えられる。この事後確率に基づいて、行動の選択など様々な処理を行うことができる。また、得られた事後確率は、今度は事前確率して利用されることが多い (知識のアップデート)。すなわち、確率を用いた推論システムにおいては、確率という形で知識を蓄え、推論 (事後確率計算) を行い知識を更新していくということがよく行われる。

なお、ベイズ則は、

$$P_{X|Y}(x|y) = \frac{P_X(x)P_{Y|X}(y|x)}{P_Y(y)} = \frac{P_{XY}(x,y)}{P_Y(y)} \tag{5}$$

となるため、ベイズ則による事後確率の計算と前節で考えた同時分布に基づく事後確率の計算は本質的には差異はないことに注意しておきたい。

2 事後確率の計算: 多変数の場合

前節は、系に含まれる確率変数が 2 個の場合の事後確率の計算について議論をした。本節では、系に含まれる確率変数が n 個 ($n > 2$) の場合の事後確率の計算について議論する。多変数の場合も 2 変数の場合と同様に「周辺化」が事後確率計算の鍵となる。

問 3：3 つ確率変数 X_1, X_2, X_3 を含む系を考える。表 4 に同時分布 $P_{X_1X_2X_3}(x_1, x_2, x_3)$ を示す。いま、$X_3 = 0$ という観測値を得たものとしよう。次の問に答えよ。

- $P_{X_1X_2|X_3}(x_1, x_2|0)$ を求めよ。ただし、$x_1 \in \{0, 1\}$, $x_2 \in \{0, 1\}$ である。
- $P_{X_1|X_3}(x_1|0)$ を求めよ。ただし、$x_1 \in \{0, 1\}$ である。

(条件付確率の定義と周辺化則を適用)

表 4: 3 変数の同時分布 $P_{X_1X_2X_3}(x_1, x_2, x_3)$

$x_1x_2x_3$	$P_{X_1X_2X_3}(x_1, x_2, x_3)$
000	0.2
001	0.2
010	0.1
011	0.1
100	0.1
101	0.1
110	0.1
111	0.1

多変数の場合も、同時分布により表される「確率的な相関」を利用して推論が可能となる。詳細に入る前に多変数の場合の確率的な推論の枠組みを説明する。

多変数の確率変数に関する確率的な推論

- 系の確率的振る舞いは、同時分布 (同時確率) で完全に記述される。
- いくつかの確率変数については、その実現値が「観測」される。
- 「確率的な推論」=「観測されていない確率変数について、その事後確率を計算すること」
- 事後確率の計算は、適切な周辺化操作に基づく。

事後確率計算の問題をより一般的な形で議論しよう。確率変数 $X_1, \ldots, X_n$ を含む系を考える。ここでは、これらの確率変数の同時分布は既知であるものと仮定する。また確率変数の集合を $U = (X_1, \ldots, X_n)$ とする。確率変数の部分集合 $V = \{X_{i_1}, X_{i_2}, \ldots, X_{i_v}\} \subset U$ について、観測値 $(x^*_{i_1}, x^*_{i_2}, \ldots, x^*_{i_s})$ が得られているものとする。すなわち、V は可観測な確率変数を表している。観測されていない確率変数の集合を $H = \{X_{j_1}, X_{j_2}, \ldots, X_{j_t}\} \subset U \backslash V$ とするとき、特に興味深い問題は、確率変数の集合 H について事後分布

$$P_{H|V}(x_{j_1}, \ldots, x_{j_t} | x^*_{i_1}, x^*_{i_2}, \ldots, x^*_{i_s})$$

を計算することである。ここで、この事後分布は、$(x_{j_1}, \ldots, x_{j_t})$ の t 変数関数であり、t 個の自由度を持っていることに注意したい。次の例で示すようにこの事後確率計算は周辺化計算により「原理的には」計算できる。

次の問題を考える。

- $U = \{X_1, X_2, X_3, X_4\}$
- $V = X_3, X_4$, 観測値 (x^*_3, x^*_4)
- $H = \{X_1, X_2\}$

事後確率計算 $P_{X_1X_2|X_3X_3}(x_1, x_2 | x^*_3 x^*_4)$ の計算は次の通りである。

$$\begin{aligned} & P_{X_1X_2|X_3X_4}(x_1, x_2 | x^*_3 x^*_4) \\ = \; & \frac{P_{X_1X_2X_3X_4}(x_1, x_2, x^*_3 x^*_4)}{P_{X_3X_3}(x^*_3 x^*_4)} \\ = \; & \frac{P_{X_1X_2X_3X_4}(x_1, x_2, x^*_3 x^*_4)}{\sum_{x_1 \in D(X_1)} \sum_{x_2 \in D(X_2)} P_{X_1X_2X_3X_4}(x_1, x_2, x^*_3 x^*_4)} \end{aligned} \quad (6)$$

問 4:$P_{X_1|X_3X_4}(x_1 | x^*_3 x^*_4)$ の場合はどうなるだろうか。

一般の事後確率

$$P_{H|V}(x_{j_1}, \ldots, x_{j_t} | x^*_{i_1}, x^*_{i_2}, \ldots, x^*_{i_s})$$

の計算の問題についても、同様に適切な周辺化計算により事後確率の評価が可能である。

なお、本節では離散型確率変数の場合に注目して説明を行ったが、連続型確率変数の場合も、離散型確率変数と同様に事後確率密度関数を評価できる。ただし、和の代わりに積分計算が必要となり、次節で述べる計算量的な困難性は離散型の場合と同様である。

3　計算量の壁

すでに見たように多変数の場合の事後分布は「原理的には」周辺化計算により求めることができる。問題は、周辺化に必要される和計算の計算量である。周辺化計算には、不必要な確率変数を分布から消去するために消去すべき確率変数についての (すべての実現値に渡る) 総和を計算しなければならない。複数の確率変数を消去する必要がある場合には、登場する多重和の計算に膨大な計算量が必要になる。

例えば、n 変数の確率変数 $X_1, X_2, \ldots, X_n$ が 2 値を取るものとする。その同時分布から、X_1 に関する周辺確率分布 $P_{X_1}(x_1)$ を求めるためには、

$$P_{X_1}(x_1) = \sum_{x_2 \in \{0,1\}} \cdots \sum_{x_n \in \{0,1\}} P_{X_1 \cdots X_n}(x_1, \ldots, x_n) \tag{7}$$

という多重和の評価が必要となる。この多重和を式の通りに計算するとすれば、2^{n-1} 通りの $(x_1, \ldots, x_n)$ について和を取ることとなり、その計算量は $O(2^n)$ となる (同時分布の評価は定数時間で可能であると仮定した場合)。すなわち、変数の数について指数的に計算時間が増大することになるため、確率変数の個数が少ない (n が小さい) 場合にのみ、この種の「単純な多重和アルゴリズム」は実用的に意味を持つことになる。

周辺化操作が消去変数の数について指数時間かかるとすれば、多数の確率変数が登場する大規模な問題について、事後確率を評価するという問題はお手上げである。音声・画像認識、通信など様々な問題に登場する系に含まれる確率変数は数千から数万にのぼる場合も多いこの**周辺化計算における「計算量の壁」が長らく事後確率評価に基づく情報処理アルゴリズムの発展を阻害してきた**と言ってもよい。

ただし、上の議論は、周辺化の計算を「定義の通りナイーブに行った場合」の議論であり、もし、同時分布がある特殊な性質を持つ場合には、その性質を生かした効率の良い多重和の計算手法が存在する可能性はある。実際、近年、同時分布の特殊性を生かした事後確率計算を非常に効率良く行うアルゴリズム群が開発されてきている。本講義の後半では、そのようなアルゴリズムについて解説を行う。

もう一つの問題点は「同時分布自身の表現」である。ふたたび、n 変数の確率変数 $X_1, X_2, \ldots, X_n$ が 2 値を取る状況を考えてみよう。この場合、同時分布 $P_{X_1 \cdots X_n}(x_1, \ldots, x_n)$ を計算機内で直接表現するためには、その表 (表 4 のような表) を記憶しておけばよい。問題は、表のエントリー数 (行数) は n について指数的に増加する点にある。例えば、$n = 50$ で表のエントリー数は 2^{50} となり、高々50 変数の系でもこのような単純な手法では同時分布を正確に表現できる望みはない。高速な周辺化アルゴリズムとともに、「**同時分布のコンパクトな表現**」が現実的なアルゴリズムの実現のためには不可欠である。