May, 2019

情報領域演習第二 第５回 K演習 (確率論)

演習問題１：確率表とベイズの定理

青・赤・黄の３色の袋がそれぞれ２、２、１個ずつある。３色の袋の中にはそれぞれ白球と黒球が表に書かれた個数入っている。いま、無作為に選ばれた袋から無作為に球を取り出したところ、白球であったという結果を知らされた。この白球が黄色の袋から取り出されたものである確率を、小問に従い、ベイズの定理を用いて求めよ。

	青袋	赤袋	黄袋
白球	2 個	1 個	4 個
黒球	3 個	4 個	1 個

1. 袋の周辺確率表を記せ。
2. 問題文内の条件付き確率表と、前問の周辺確率表から、ベイズの定理を用いて、「無作為に選ばれた袋から無作為に球を取り出したところ、白球であったという結果を知らされた。この白球が黄色の袋から取り出されたもの」である確率を求めよ。

ヒント：教科書 p.10-11 のベイズの定理を理解し、まずは $A_1, A_2, \ldots, A_n$ と B をこの問題ではどのように取れば良いかを考えよ。それから定理 1.3 のベイズの定理を適用すると良い。なお、この問題は教科書の p.12 の問題 1.6 である。

解説：この問題は、文章に書かれた確率に関する記述を読み取って条件付き確率表を作り、ベイズの定理を適用する問題です。求めたい確率は、問いの最後に記されている

$$\mathrm{P}_{\text{袋}\mid\text{球}}(\text{黄袋}\mid\text{白球}) \tag{1}$$

です。しかし問題文から読み取れる確率は、袋が指定されたときの中の球の確率 $\mathrm{P}_{\text{球}\mid\text{袋}}$ です。問われた確率を求めるためにベイズの定理を

$$\mathrm{P}_{\text{袋}\mid\text{球}}(\text{黄袋}\mid\text{白球}) = \frac{\mathrm{P}_{\text{球}\mid\text{袋}}(\text{白球}\mid\text{黄袋})\,\mathrm{P}_{\text{袋}}(\text{黄袋})}{\left(\begin{array}{r}\mathrm{P}_{\text{球}\mid\text{袋}}(\text{白球}\mid\text{青袋})\,\mathrm{P}_{\text{袋}}(\text{青袋}) \\ +\,\mathrm{P}_{\text{球}\mid\text{袋}}(\text{白球}\mid\text{赤袋})\,\mathrm{P}_{\text{袋}}(\text{赤袋}) \\ +\,\mathrm{P}_{\text{球}\mid\text{袋}}(\text{白球}\mid\text{黄袋})\,\mathrm{P}_{\text{袋}}(\text{黄袋})\end{array}\right)} \tag{2}$$

と適用したいです。

問題を方針が決まったら、必要な確率を順に求めていきます。例えば青袋から球を一つ取り出すとき、それが白球である確率は表から $\mathrm{P}_{\text{球}\mid\text{袋}}(\text{白球}\mid\text{青袋}) = 2/(2+3) = 0.4$ と求まります。他の確率も求めて一覧表にすると、

		袋の中の球の色	
		白球	黒球
袋の色	青袋	0.4	0.6
	赤袋	0.2	0.8
	黄袋	0.8	0.2

の通り。また袋をランダムに 1 つ選ぶときの周辺確率 $\mathrm{P}_{\text{袋}}(\text{色})$ は文章から

	青袋	赤袋	黄袋
確率	0.4	0.4	0.2

と求まります。これが小問 (1) の回答です。

ここまで準備が整えば、ベイズの定理から

$$\begin{aligned} \mathrm{P}_{\text{袋}\,|\,\text{球}}\left(\text{黄袋} \mid \text{白球}\right) &= \frac{0.8 \times 0.2}{0.4 \times 0.40.2 \times 0.4 + 0.8 \times 0.2+} = \frac{16}{16+8+16} \\ &= \frac{2}{5} = 0.4 \end{aligned} \tag{3}$$

を得ます。この計算ができるかどうか、を問うています。

この問題は個々の球に袋の色と球の色、という 2 つの属性がついています。ベイズの定理を使わなければ、次のようにも解けます。

1. この問題での白球の数は $2 \times 2 + 1 \times 2 + 4 \times 1 = 10$ 個。
2. 白球のうち黄色の袋に入っている数は $4 \times 1 = 4$ 個。
3. 従って、白球を得たときの黄色い袋からの条件付き確率は $4/10 = 0.4$。

これは、全ての場合を数え上げられるときのみに使える解法で、より複雑な問題には通用しないので、お勧めしないし、そもそも題意に沿わない解法です。

演習問題 2 ：累積分布関数と期待値の計算

累積分布関数が

$$F\left(x\right) = \begin{cases} 0 & x < 0 \\ 1/4 & x = 0 \\ 1/4 + x/2 & 0 < x < 1 \\ 1 & x \geq 1 \end{cases} \tag{4}$$

と与えられている確率分布がある。この分布につき、以下の問いに順に答えよ。

1. この確率分布の標本空間を記せ。
2. 累積分布関数のグラフを描け。(グラフの横軸の範囲は-1 から 2 までにとり、不連続点は、端点を含む場合に ●、端点を含まない場合には ○ で示すこと)
3. この確率分布に従う確率変数を X とするとき、X の関数 $h\left(X\right)$ の期待値を求める式を導出せよ。
4. 次の X の関数の期待値を求めよ。(a) $h\left(X\right) = X$。 (b) $h\left(X\right) = 2X - 1$。 (c) $h\left(X\right) = X^2$。 (d) $h\left(X\right) = \left(X - \mu\right)^2$。 (e) $h\left(X\right) = \exp\left(tX\right)$ 。ただし、t は任意の実数とする。

ヒント：確率分布を与えるのは、確率や期待値を計算して、分布の特徴を調べたり、将来の予測のために分布の中心 (平均) やばらつきの大きさ (標準偏差) を調べるため。確率分布の表現は主に確率表（標本空間が順序の定められていない離散集合の場合)、確率関数（標本空間が順序の定められた離散集合の場合)、累積分布関数（標本空間が順序の定められた集合の場合)、そして確率密度関数（標本空間がユークリッド空間もしくはそのアフィン部分空間の場合）などにより定式化される。どの表現を与えたときにも確率変数の期待値の計算式の定式化をまず考える。これは確率分布の累積分布関数が与えられている問題である。

解説：この問題は確率表、確率関数、累積分布関数、確率密度関数のいずれが与えられても、確率や期待値

を計算できるようになって欲しい、というメッセージを込めて、もっとも複雑な場合を選んでいます。宿題はこれを少し易しくしたものと、確率関数と確率密度関数を 1 題ずつとする予定です。この関数 $F(x)$ のグラフを描いてみると、$x \leq 0$ では 0、$x \geq 1$ では 1 となり、その間は確かに単調非減少な関数です。教科書に書いてある累積分布関数の条件は満たしています。しかし不連続点も 2 つあり、期待値の求め方に窮することでしょう。

しかし期待値を求めるにあたっての、考え方は連続集合上の確率分布ならば

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f(x)\, dx \tag{5}$$

離散集合上の確率分布ならば

$$\mathrm{E}[X] = \sum_{x=-\infty}^{\infty} x p(x)\, dx \tag{6}$$

で計算するのは同じです。それではどう考えていけば良いのでしょうか。

まず小問 (1) で、標本空間が $\{0\} \cup \{0 < x < 1\} \cup \{1\}$ と三つの部分集合の和集合 (OR 結合) であることを認識してもらいます。ここで素直に $\{x; 0 \leq x \leq 1\}$ と書かせてしまうと、次の小問からすぐに躓くかもしれません。

次の小問 (2) はグラフの座標 $(-1, 0)$ に ·、$(0, 0)$ に ○、$(0, 1/4)$ に ●、$(1, 3/4)$ に ○、$(1, 1)$ に ●、最後に $(2, 1)$ に · を描き、垂直関係にない隣り合った点どうしを実線で結んでください。

次の小問 (3) がこの問題の要です。与えられた累積分布関数 $F(x)$ が連続関数であれば

$$f(x) = \frac{d}{dx} F(x) \tag{7}$$

と微分して、確率密度関数 $f(x)$ を得ます。この関数は確率変数 X の関数 $h(x)$ の期待値を計算する際に

$$\mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\, dx \tag{8}$$

のように利用します。

もし、$F(x)$ が至る所で不連続であれば、すべての不連続点 x において

$$p(x) = \lim_{\epsilon \to +0} F(x+\epsilon) - \lim_{\epsilon \to -0} F(x+\epsilon) \tag{9}$$

のように、右側極限と左側極限の差を求め、その値を確率関数 $p(x)$ とします。この関数は、確率変数 X の関数 $h(X)$ の期待値を計算する際に

$$\mathrm{E}[h(X)] = \sum_{-\infty}^{\infty} h(x) p(x) \tag{10}$$

のように利用します。

ところが今回の問題では、$x = 0$ および $x = 1$ で不連続、それ以外の点では連続です。このような確率分布の期待値はどのように計算しましょう。実は、問題に与えられた累積分布関数は

$$F(x) = \begin{cases} 0 & x < 0 \\ I(x \geq 0) \times 1/4 & x = 0 \\ I(x \geq 0) \times 1/4 + \int_0^x 1 du/2 & 0 < x < 1 \\ I(x \geq 0) \times 1/4 + \int_0^1 1 du/2 + I(x \geq 1) \times 1/4 & x \geq 1 \end{cases} \tag{11}$$

という表現を持ちます。$I(A)$ は中が真の時に 1、偽の時に 0 をとる関数で、識別関数と呼ばれます。この表現から、期待値の計算式は

$$\mathrm{E}[h(X)] = h(0)\,P(X=0) + \int_0^1 h(x)\,1dx/2 + h(1)\,P(X=1) \tag{12}$$

となることが分かりますか？

小問 (4) は、$h(x)$ に具体的な関数を入れて計算すれば良いです。ただし、なるべく計算が簡単になるように整理し、てください。

$$\begin{aligned}
\mathrm{E}[X] &\equiv \mu \\
&= 0 \times P(X=0) + \int_0^1 x dx/2 + 1 \times P(X=1) \\
&= 0 + \left[\frac{x^2}{4}\right]_0^1 + \frac{1}{4} \\
&= 0 + \frac{1}{4} - \frac{0}{4} + \frac{1}{4} \\
&= \frac{1}{2}
\end{aligned} \tag{13}$$

確率変数の定数倍や定数の加減の期待値は、教科書 p.23 の定理 2.2 (1) を用いてください。分散も同じ定理の (2) を用いてください。

$$\mathrm{E}[2X-1] = 2\mathrm{E}[X] - 1 = 1 - 1 = 0 \tag{14}$$

$$\begin{aligned}
\mathrm{E}\left[X^2\right] &= 0 \times P(X=0) + \int_0^1 x^2 dx/2 + 1^2 \times P(X=1) \\
&= 0 + \left[\frac{x^3}{6}\right]_0^1 + \frac{1}{4} \\
&= \frac{1}{6} - \frac{0}{6} + \frac{1}{4} = \frac{5}{12}
\end{aligned} \tag{15}$$

これは分散です。ここまで計算してあると、教科書 p.23 の (2.10) 式の方が簡単です。

$$\begin{aligned}
\mathrm{E}\left[(X-\mu)^2\right] &\equiv \sigma^2 \\
&= \mathrm{E}\left[X^2\right] - \mu^2 = \frac{5}{12} - \frac{1}{4} = \frac{1}{6}
\end{aligned} \tag{16}$$

モーメント母関数ですが、計算してみるだけです。

$$
\begin{aligned}
\mathrm{E}\left[\exp\left(tX\right)\right] &= \exp\left(0\right) \times P\left(X=0\right) + \int_0^1 \exp\left(tx\right) dx/2 + \exp\left(t\right) \times P\left(X=1\right) \\
&= 1 \times \frac{1}{4} + \left[\frac{\exp\left(tx\right)}{t}\right]_0^1 /2 + e^t \times \frac{1}{4} \\
&= \frac{1+e^t}{4} + \frac{e^t/t - e^0/t}{2} \\
&= \frac{1}{4} + \frac{e^t}{4} + \frac{e^t}{2t} - \frac{1}{2t}
\end{aligned} \tag{17}
$$

演習問題 3 ：確率不等式

1. 非負の確率変数 X の期待値を μ とすると、X に関するマルコフの不等式は

$$
\Pr\left[X \geq a\right] \leq \frac{\mu}{a} \tag{18}
$$

となる。ある国の選挙の投票率の期待値が 0.4 であることが事前に分かっているとき、投票率が 0.60 以上になる確率の上限をこの不等式を用いて与えよ。

2. 確率変数 X の期待値を μ、分散を σ^2 とすると、X に関するチェビシェフの不等式は

$$
\Pr\left[|X-\mu| \geq a\right] \leq \frac{\sigma^2}{a^2} \tag{19}
$$

となる。ある国の選挙の投票率の期待値が 0.4、分散が 0.01 であることが事前に分かっているとき、投票率が 0.60 以上になる確率の上限をこの不等式を用いて与えよ。

ヒント：マルコフの不等式は、教科書には明示的には登場しないが、非負の確率分布の期待値 μ のみを知っている状況で、確率の範囲を限定できる。チェビシェフの不等式は、マルコフの不等式を用いても証明できる不等式で、確率分布の期待値 μ と分散 σ^2 を知っているいる状況で、確率の範囲を限定できる。

解説：マルコフの不等式は、そのまま当てはめれば良いです。

$$
\Pr\left[X \geq a\right] \leq \frac{\mu}{a} \tag{20}
$$

に $\mu = 0.4$ と $a = 0.60$ を代入すると

$$
\Pr\left[X \geq 0.60\right] \leq \frac{0.4}{0.60a} = \frac{2}{3} \tag{21}
$$

を得ます。こちらは素直です。

問題はチェビシェフの不等式の方です。

$$
\Pr\left[|X-\mu| \geq a\right] \leq \frac{\sigma^2}{a^2} \tag{22}
$$

まずは分かっている量を代入していきます。

$$
\Pr\left[|X-0.4| \geq a\right] \leq \frac{0.01}{a^2} \tag{23}
$$

問題はこの不等式を題意に合わせる際の a の値です。一部の学生が a に 0.60 を入れて右辺を評価します。

$$\Pr\left[|X - 0.4| \geq 0.60\right] \leq \frac{0.01}{0.60^2} \tag{24}$$

X が 0.60 以上という範囲は、$X - 0.4 \geq 0.20$ になるはずです。この問題では $a = 0.20$ を代入して

$$\Pr\left[|X - 0.4| \geq 0.20\right] \leq \frac{0.01}{0.20^2} = \frac{1}{4} \tag{25}$$

とするのが正しいです。なお、絶対値の記号がありますが、

$$\begin{aligned}\Pr\left[|X - 0.4| \geq 0.20\right] &= \Pr\left[X - 0.4 \geq 0.20\right] + \Pr\left[-X + 0.4 \leq -0.20\right] \\ &\geq \Pr\left[X - 0.4 \geq 0.20\right]\end{aligned}$$

のため、チェビシェフの不等式と合わせて

$$\Pr\left[X - 0.4 \geq 0.20\right] \leq \Pr\left[|X - 0.4| \geq 0.20\right] \leq \frac{1}{4} \tag{26}$$

となります。