2020.4.15

情報統計力学 事前学習２

４月 20 日 (金) 午後１時ー２時半

離散型確率モデル

定義

標本空間 Ω が有限個あるいは加算無限個の要素からなり、Ω の $\sigma-$ **集合族** $\mathcal{F}$ が Ω の全ての部分集合から確率モデル $(\Omega, \mathcal{F}, P)$ を**離散型確率モデル**と呼ぶ。
このとき、$\Omega = \{\omega_1, \omega_2, \omega_3, \cdots, \}$ とすると、任意の $A \in \mathcal{F}$ に対して、確率 P の定義より

$$P(A) \quad = \quad \sum_{i, \omega_i \in A} P(\{\omega_i\}) \tag{1}$$

となる。
逆に、$\sum_{i=1}^{\infty} p_i = 1$ を満たす正数の組 $p_1, p_2, \cdots, (p_i \geq 0, i = 1, 2, \cdots)$ を与えたとき、**標本空間** Ω が加算無限個の標本 $\{\omega_i\}$ からなり、$P(A) = \sum_{i, \omega_i \in A} p_i$ として P を定めると、P が確率となることも分かる。

条件付き確率
ある事象 $B(\in \mathcal{F}), (P(B) > 0)$ が生起したとして、次に事象 $A(\in \mathcal{F})$ が生起する確率を以下のように定義する。

$$P(A|B) \quad = \quad \frac{P(A \cap B)}{P(B)}. \tag{2}$$

実際に、$P(\cdot|B)$ は、可測空間 $(\Omega, \mathcal{F})$ 上の確率の定義を満たすことが分かる。$P(\cdot|B)$ を事象 B が与えられたときの事象 A の条件付き確率と呼ぶ。

問題 以下の命題を証明せよ。

(1) $P(B) > 0$ なる事象と任意の事象 A に対して、

$$P(A \cap B) \quad = \quad P(B)P(A|B). \tag{3}$$

(2) 事象 $B_1, B_2, \cdots, B_n$ は互いに背反で、$\Omega = \cup_{i=1}^{n} B_i, P(B_i) > 0$ を満たすとする。このような事象列は全事象 Ω の**分割**と呼ばれる。**分割** $\{B_i\}_i$ が与えられたとき、任意の事象 A に対して

$$P(A) \quad = \quad \sum_{i=1}^{n} P(B_i)P(A|B_i) \tag{4}$$

が成り立つ。

ベイズの公式

$P(A) > 0, P(B) > 0$ とする。$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B|A) = \frac{P(B \cap A)}{P(A)}$ より、$P(A|B)P(B) = P(B|A)P(A)$ となる。従って、

$$P(B|A) \quad = \quad \frac{P(B)P(A|B)}{P(A)} \tag{5}$$

が成り立つ。これをベイズの公式と呼ぶ。
これは更に一般化することができる。

以下を示せ。

事象 $B_1, B_2, \cdots, B_n$ が全事象 Ω の分割とする。このとき、$P(A) > 0$ なる任意の事象 A に対して

$$P(B_k|A) \quad = \quad \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{n} P(B_i)P(A|B_i)} \tag{6}$$

が任意の $k = 1, 2, \cdots, n$ に対して成り立つ。

ベイズの公式の応用

マンモグラフィーによる乳がんの検診

マンモグラフィーによる検診について、実際に乳がんである場合に陽性となる確率が 90% とする。これを、P(陽性 | 乳がん) $= 0.9$ と表す。ベイズの定理により

$$P(\text{乳がん} \mid \text{陽性}) = \frac{P(\text{陽性} \mid \text{乳がん})P(\text{乳がん})}{P(\text{陽性})} \tag{7}$$

が成り立つ。よって、P(乳がん) と P(陽性) が分かれば、陽性と判定された場合に実際に乳がんである確率が分かる。
これまでの統計により、乳がんである確率は 0.8%、つまり、P(乳がん) $= 0.008$ である。また、マンモグラフィー検査で、乳がんでないのに陽性となる確率は、7% とする。つまり、P(陽性 | 乳がんでない) $= 0.07$。
P(陽性) を求めよう。まず、乳がんであるか乳がんでないかは背反事象であるから、

$$P(\text{乳がんでない}) \quad = \quad 1 - P(\text{乳がん}) = 1 - 0.008 = 0.992 \tag{8}$$

となる。一方、(4) より、

$$P(\text{陽性}) \quad = \quad P(\text{陽性} \mid \text{乳がん})P(\text{乳がん}) + P(\text{陽性} \mid \text{乳がんでない})P(\text{乳がんでない}) \tag{9}$$

が成り立つ。従って、

$$P(\text{陽性}) \quad = \quad 0.9 \times 0.008 + 0.07 \times 0.992 = 0.0072 + 0.0694 = 0.07664. \tag{10}$$

従って、有効数字 2 桁として、P(陽性) $= 0.077$. よって、

$$P(\text{乳がん} \mid \text{陽性}) \quad = \quad \frac{P(\text{陽性} \mid \text{乳がん})P(\text{乳がん})}{P(\text{陽性})} = 0.9 \times 0.008/0.077 = 0.0935.. \tag{11}$$

よって、検診で乳がんと判定されたとき、実際に乳がんである確率は 9% 程度であることが分かる！