2010 年 6 月 11 日

速修 確率論と情報理論

医者の藪野竹庵氏は自分のところへくる患者について

$$A:\begin{cases} A_1: & \text{熱がある} \\ A_2: & \text{熱がない} \end{cases} \qquad B:\begin{cases} B_1: & \text{風邪を引いている} \\ B_2: & \text{風邪ではない} \end{cases}$$

という 2 種類の事象の割合を調べ，以下の表を作った．

	B_1 （風邪）	B_2 （風邪なし）	$p(A_i)$
A_1 （熱あり）	0.55	0.05	0.60
A_2 （熱なし）	0.10	0.30	0.40
$p(B_j)$	0.65	0.35	

この割合は，事象 (A_i, B_j) の確率 $p(A_i, B_j)$ であると思ってさしつかえない．この表から，熱があるかないかを知ってしまえば，風邪であるかないかはほぼ検討がつくことがわかる．まず必要な確率の知識を整理しておこう．確率分布として次のものを考える．

$$\begin{cases} \text{同時確率分布:} & p(A_i, B_j) \\ \text{周辺確率分布:} & p(A_i), p(B_j) \\ \text{条件付確率分布:} & p(B_j|A_i), p(A_i|B_j) \end{cases}$$

$p(A_i, B_j)$ のように，A_i と B_j とが組になって同時に起こる確率を示すものを，同時確率分布と呼ぶ．そのうちの一方だけに着目した確率分布 $p(A_i)$ や $p(B_i)$ は，上の表を行または列について加え合わせると出てくる．この $p(A_i)$ や $p(B_i)$ は，表の周辺に並んでいるので，周辺分布と呼ぶ．条件付確率分布 $p(B_j|A_i)$ とは A に関する事象が A_i であったときの B_j の確率である．$p(A_i|B_j)$ は逆に B が B_j であることがわかっているときの A_i の確率である．これらの間には次の関係が成立する．

$$\sum_{i,j} p(A_i, B_j) = 1, \ \sum_i p(A_i) = \sum_j p(B_j) = 1 \tag{1}$$

$$\sum_j p(B_j|A_i) = \sum_i p(A_i|B_j) = 1 \tag{2}$$

$$p(A_i) = \sum_j p(A_i, B_j), \ p(B_j) = \sum_i p(A_i, B_j) \tag{3}$$

$$p(A_i, B_j) = p(A_i)p(B_j|A_i) = p(B_j)p(A_i|B_j) \tag{4}$$

$$p(B_j|A_i) = \frac{p(B_j)p(A_i|B_j)}{p(A_i)} = \frac{p(A_i, B_j)}{p(A_i)} \tag{5}$$

$$p(A_i|B_j) = \frac{p(A_i)p(B_j|A_i)}{p(B_j)} = \frac{p(A_i, B_j)}{p(B_j)} \tag{6}$$

式 (4) は，A_i と B_j との同時確率は，A_i の起こる確率に，A_i が起こるという条件下での B_j の起こる確率を掛けたものであることを示す．式 (5,6) は Bayes の公式と呼ばれるもので，式 (4) から導き出せる．これから使う確率の式は以上の式だけでいい（はずである）．具体的に，熱があった場合に風邪である確率 $p(B_1|A_1)$，熱がない場合に風邪である確率 $p(B_1|A_2)$，を計算してみる．

$$p(B_1|A_1) = \frac{p(A_1, B_1)}{p(A_1)} = \frac{0.55}{0.60} = 0.92, \ p(B_1|A_2) = \frac{p(A_2, B_1)}{p(A_2)} = \frac{0.10}{0.40} = 0.25 \tag{7}$$

であるから，熱があるかないかを知ってしまえば，風邪であるかないかはほぼ検討がつく．

奇特な藪野先生は，自分のところへくる患者について，以下のような表も作成していた．

	B_1 （風邪）	B_2 （風邪なし）	$p(C_i)$
C_1 （晴れ）	0.39	0.21	0.60
C_2 （曇りか雨）	0.26	0.14	0.40
$p(B_j)$	0.65	0.35	

この表をパッと見ただけでは，晴れの日には曇りか雨の日に比べ風邪の患者が来やすいかどうかは，よくわからない．そこで,「晴れの日」という条件のもと，風邪の患者が来る確率 $p(B_1|C_1)$ と,「曇りか雨の日」という条件のもと，風邪の患者が来る確率 $p(B_1|C_2)$ を計算してみる．

$$p(B_1|C_1) = \frac{p(C_1, B_1)}{p(C_1)} = \frac{0.39}{0.60} = 0.65,\ p(B_1|C_2) = \frac{p(C_2, B_1)}{p(C_2)} = \frac{0.26}{0.40} = 0.65 \tag{8}$$

なんと同じ確率となった．つまり，晴れであっても曇りか雨てあっても，風邪である患者がくる確率は変わらない，ということである．これは，実は C と B が確率的に 独立な場合（風邪と天気が無関係の場合）であり，このとき

$$p(C_i, B_j) = p(C_i)p(B_j) \tag{9}$$

$$p(B_j|C_i) = p(B_j),\ p(C_i|B_j) = p(C_i), \tag{10}$$

という式が成立している．実際，例えば,

$$p(C_1, B_1) = 0.39 = p(C_1)p(B_1) = 0.60 \times 0.65 \tag{11}$$

$$p(C_1, B_2) = 0.21 = p(C_1)p(B_2) = 0.60 \times 0.35 \tag{12}$$

となっている．

さて，熱があるかないかを知ってしまえば，風邪であるかないかはほぼ検討がつく．これは「熱のあるなし」という事象が，風邪であるかないかの情報をもっていることである．逆に，風邪であるかないかという事象が，熱があるかないかの情報をもっているとも言える．確率を使うことにより，情報の量を定められそうである．

定義：情報量

確率 p の事象が実際に起ったことを知らせる情報に含まれている情報量を $-\log_2 p$ ビットと定義する．

$A_1, \cdots, A_n$ の n 個の事象があって，それぞれ $p_1, p_2, \cdots, p_n$ の確率で生じる場合を考えよう．確率は足して 1 になるので

$$\sum_{i=1}^{n} p_i = 1$$

である．ここで，どの事象が起こったかを教えてもらうことにする．得られる情報の量は，どの A が生じたかで異なってくる．例えば，A_1 が起これば $-\log p_1$，A_2 ならば $-\log p_2, \cdots$ という情報が得られる．ここで，A_i の起こる確率は p_i だから，得られる情報の量の期待値は $-\log p_i$ を 確率 p_i で平均したもの

$$I = -\sum_{i=1}^{n} p_i \log p_i \tag{13}$$

である．

われわれが情報をほしいのは，不確定な状況を確定したいからである．この場合，どういう情報がもらえるかは事前にわかるはずがなく，したがって，もらえる情報量そのものはわからない．わかるのはもらえる情報量の期待値だけである．この値は，不確実な状況を確定するのに要する平均情報量だといってもよい．この状況の不確定度を表す量がエントロピーである．

定義： エントロピー

n 個の事象がそれぞれ確率 $p_1, p_2, \cdots, p_n$ で発生するとき，どれが発生したかの不確定度を

$$H(p_1, p_2, \cdots, p_n) = -\sum_{i=1}^{n} p_i \log p_i$$

と定義し，これをエントロピーと呼ぶ．

複合事象のエントロピーも同様に考えることができる．具体的に，藪野先生の例で計算してみよう．熱と風邪の複合事象のエントロピーは

$$H(A, B) = -0.55 \log 0.55 - 0.05 \log 0.05 - 0.1 \log 0.1 - 0.3 \log 0.3 = 1.54 \text{ ビット}$$

である．一方，熱のあるなしのエントロピーは

$$H(A) = -0.6 \log 0.6 - 0.4 \log 0.4 = 0.97 \text{ ビット}$$

風邪についてのエントロピーは

$$H(B) = -0.65 \log 0.65 - 0.35 \log 0.35 = 0.93 \text{ ビット}$$

である．

ここで $H(A)$ や $H(B), H(A, B), H(A) + H(B)$ などの大小関係を考えてみよう．$H(A)$ や $H(B)$ は，熱や風邪のどちらか一方だけに着目した 1 種類の事象系の不確定度であるから，A と B と両方に関する確定度 $H(A, B)$ よりは小さいことが予想される．また，事象系 A と事象系 B の間に密接な関係がある場合には，A が何であるかがわかれば B が何であるかはだいたいの見当がついてしまう．熱の有無と風邪かいなかは密接に関係している．だから，$H(A, B)$ は A と B とを別々に考えたときの不確定度の和 $H(A) + H(B)$ より一般に小さいだろう．しかし，事象系 A と事象系 B とがまったく関係のない場合には $H(A, B)$ は $H(A) + H(B)$ に等しいものと思われる．この例でも，

$$H(A), H(B) \leq H(A, B) \leq H(A) + H(B)$$

が成り立っている．風邪と熱とを組み合わせた不確定度は，両者を別々に考えたそれぞれの不確定度よりは大きく，その和よりは小さい．

このようなエントロピー間の大小関係，さらに風邪と熱のような密接に関係している事象系の間の情報関係を知るには，条件付きエントロピーなる概念を導入する必要がある．

熱があるかないか知ってしまえば，風邪であるかないかはほぼ見当がつく．このように，一方の事柄が何であるかわかっている，という条件のもとでの他方の事柄の不確定度を条件付きエントロピーという．

A が何であるかを知ったときの，B が何であるかについての不確定度を調べよう．いま，A が A_i であることがわかったとする．このとき，B として $B_1, B_2, \cdots, B_m$ が起こる確率は，それぞれ条件付確率 $p(B_1|A_i), p(B_2|A_i), \cdots, p(B_m|A_i)$ となる．したがって，このときの不確定度を表すエントロピーは

$$H(B|A_i) = -\sum_{j=1}^{m} p(B_j|A_i) \log p(B_j|A_i)$$

と書ける．ところで，A はいつも A_i が起こるわけではない．A_i の起こる確率は $p(A_i)$ である．だから，A が何であるかを知ったあとの B についての不確定度を表すエントロピーは，$H(B|A_i)$ をすべての A_i について平均したもの

$$H(B|A) = \sum_{i=1}^{n} p(A_i) H(B|A_i) = -\sum_{i,j} p(A_i) p(B_j|A_i) \log p(B_j|A_i)$$

と書ける．このエントロピーの平均値 $H(B|A)$ のことを条件付きエントロピーという．

さきほどの，藪野先生の例で，熱の有無がわかったいう条件付の病気のエントロピー，つまり検温後にもなお残る不確定度を求めてみよう．Bayes の公式を用いると，条件付き確率は

$$p(B_1|A_1) = \frac{0.55}{0.6} = 0.92,\ p(B_2|A_1) = \frac{0.05}{0.6} = 0.08,$$

$$p(B_1|A_2) = \frac{0.1}{0.4} = 0.25,\ p(B_2|A_2) = \frac{0.3}{0.4} = 0.75$$

である．したがって，

$$H(B|A_1) = 0.41$$

$$H(B|A_2) = 0.81$$

これより

$$H(B|A) = 0.57$$

が求まる．検温前の B についてのエントロピー $H(B)$ は 0.93 であったから，検温をすませれば，それだけで不確定度はかなり減ることになる．検温によって得られる風邪についての情報量の平均値は，エントロピーの減少分

$$I = H(B) - H(B|A) = 0.36 \text{ ビット}$$

である．$H(B|A)$ は $H(B|A_1)$ と $H(B|A_2)$ の平均値であり，検温の結果が何であるかは問わないで，とにかく検温したときの，検温だけでは確定できない残りの不確定度の期待値を表すことに注意したい．検温の結果によって，エントロピーは $H(B|A)$ よりも大きくも小さくもなるが，このエントロピーを平均すれば $H(B|A)$ となる．

文献： 情報理論，甘利俊一, ダイヤモンド社，1970.
本解説は，この教科書の中身の一部分を使っております．