エントロピー

最初に確率の話を簡単にし、その後にエントロピーを定義します。
エントロピーを定義したところから、log は $\log_2$ を表し、自然対数は $\log_e$ か ln と表記しています。

確率変数のことは無視して確率の用語を与えていきます (事象の結果が実数になっていると思えば同じ)。ある事象 X において起きる結果 x_i $(i = 1, 2, \ldots, n)$ があるとします。x_i がその事象における結果の全てであるなら、x_i になる確率を $P(x_i)$ と書くと、確率の定義から

$$\sum_{i=1}^{n} P(x_i) = 1$$

例えば、事象がサイコロを振ることなら x_i はサイコロの目なので、$x_1 = 1,\ x_2 = 2, \ldots, x_6 = 6$ として

$$P(x_1) + P(x_2) + \cdots + P(x_6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

と書けます。

2 つの事象 X, Y があり、それぞれの起きる結果を x_i, y_j $(j = 1, 2, \ldots, m)$ とします。x_i と y_j が同時に起きる確率を結合確率 (joint probability) や同時確率と言い、$P(x_i, y_j)$ と書かれます。x_i, y_j の並びに意味はないので $P(y_j, x_i)$ と書いても同じです。X, Y がそれぞれ無関係な事象 (お互いの結果に影響を与えない事象) であるなら、x_i と y_j が同時に起きる確率は

$$P(x_i, y_j) = P(x_i)P(y_j)$$

となります。このようなとき、X と Y は独立と言われます。

例えば、2 個のサイコロを振ってもお互いに影響しあわないので独立ですが、丸印のボールとバツ印のボールが入っている箱からボールを 2 人が交互に取り出す場合ではどの印のボールになるのかの確率はお互いに影響しあいます。どの印のボールを取り出したかによって、後の人が取り出すボールの確率は影響を受けるからです (先に丸印を取り出せば、後の人が丸印を取り出す確率が減る)。このようなときは独立ではないです。

$P(x_i, y_j)$ は片方の全部の和を取ると、片方だけの確率になります。独立な場合はそのまま

$$\sum_{j=1}^{m} P(x_i, y_j) = P(x_i) \sum_{j=1}^{m} P(y_j) = P(x_i)$$

X, Y が独立でなく関係しているときでは

$$\sum_{j=1}^{m} P(x_i, y_j) = P(x_i, y_1) + P(x_i, y_2) + \cdots + P(x_i, y_m)$$

これは Y の全ての結果に対して x_i となる確率を与えています（y_1 のときに x_i が起きる確率、y_2 のときに x_i が起きる確率として足しているため)。つまり、これは 2 つの事象 X, Y において x_i が起きる確率です。なので、X, Y において x_i, y_j が起きる確率は

$$P(x_i) = \sum_{j=1}^{m} P(x_i, y_j) \ , \ P(y_j) = \sum_{i=1}^{n} P(x_i, y_j) \tag{1}$$

と書けます。

事象 X, Y があり、X で x_i と与えられているときに Y が y_j となる確率を $P(y_j|x_i)$ と書き、条件付き確率 (conditional probability) と呼びます。例えば、サイコロにおいて偶数が出ることを前提にした 4 が出る確率 1/3 (3 個の偶数の中の 1 個がでる確率) は条件付き確率です。条件付き確率は、与えられた x_i に対して y_j が起きる確率 (x_i が起きる前提での y_j が起きる確率) なので

$$\sum_{j=1}^{m} P(y_j|x_i) = 1$$

また、X, Y が独立であるなら、どの x_i に対しても y_j が起きる確率は同じなので

$$P(y_j|x_i) = P(y_j)$$

となります。

例えば、X, Y に対してそれぞれ x_1, x_2 と y_1, y_2 としたとき、x_1 が起きたときに y_1, y_2 が起きる確率が $3/4, 1/4$ であるなら (y_1, y_2 しか起きないから足せば 1)、条件付き確率は

$$P(y_1|x_1) = \frac{3}{4} \ , \ P(y_2|x_1) = \frac{1}{4}$$

一方で、$P(x_i, y_j)$ は x_i と y_j が起きる確率なので、$P(x_1) = 2/3$、$P(x_2) = 1/3$ であるなら

$$P(x_1, y_1) = \frac{2}{3} \times \frac{3}{4} = \frac{1}{2} \ , \ P(x_1, y_2) = \frac{2}{3} \times \frac{1}{4} = \frac{1}{6} \quad (\sum_{j=1}^{2} P(x_1, y_j) = \frac{1}{2} + \frac{1}{6} = \frac{2}{3} = P(x_1))$$

となります。

これらを見比べると

$$P(y_1|x_1) = P^{-1}(x_1)P(x_1, y_1) \ , \ P(y_2|x_1) = P^{-1}(x_1)P(x_1, y_2)$$

となっています。これを一般化して、x_i に依存する係数 $C(x_i)$ によって

$$P(y_j|x_i) = C(x_i)P(x_i, y_j)$$

と書けるとします。y_j の和を取ると、左辺は条件付き確率の定義から 1、右辺は $P(x_i)$ になるので

$$\sum_{j=1}^{m} P(y_j|x_i) = C(x_i) \sum_{j=1}^{m} P(x_i, y_j)$$

$$1 = C(x_i)P(x_i)$$

$$C(x_i) = \frac{1}{P(x_i)}$$

よって、結合確率、条件付き確率、確率による関係として

$$P(y_j|x_i) = \frac{P(x_i, y_j)}{P(x_i)}$$

$$P(x_i, y_j) = P(y_j|x_i)P(x_i) \tag{2}$$

これを確率の積の法則と言ったりします。同じことを $P(x_i|y_j)$ でも行うと

$$P(x_i|y_j) = C'(y_j)P(x_i, y_j)$$

$$\sum_{i=1}^{n} P(x_i|y_j) = C'(y_j) \sum_{i=1}^{n} P(x_i, y_j)$$

$$C'(y_j) = \frac{1}{P(y_j)}$$

から

$$P(x_i|y_j) = \frac{P(x_i, y_j)}{P(y_j)}$$

$$P(x_i, y_j) = P(x_i|y_j)P(y_j) \tag{3}$$

(2),(3) から

$$P(x_i|y_j)P(y_j) = P(y_j|x_i)P(x_i)$$

$$P(x_i|y_j) = \frac{P(y_j|x_i)P(x_i)}{P(y_j)}$$

として、条件付き確率の関係が求まります。これをベイズ (Bayes) の定理と言います。

ベイズの定理を使った例を示しておきます。大きなボールと小さなボールを的に当てるとします。大きなボールを使えば必ず的に当たり、小さなボールでは 3/4 で当たるとし、大きなボールは 3/4、小さなボールは 1/4 の確率で常に選ばれるとします。的にあたったときに小さなボールが使われた確率を求めます。

直接的に求めるなら、ボールの個数を N とすれば、大きなボールが選ばれて的に当たる確率 3/4 と小さなボールが選ばれて的に当たる確率 3/16 から的に当たる確率は 15/16 なので、$15N/16$ 個当たります。的に小さなボールが当たる個数は $3N/16$ なので、的に当たったときに小さなボールが使われていた確率は

$$\frac{3N}{16}\frac{16}{15N} = \frac{1}{5}$$

となります。

確率の計算規則とベイズの定理から求めます。x_s を当たった場合、x_f を外した場合、大きなボールを y_1、小さなボールを y_2 として、求めたい条件付き確率を $P(y_2|x_\mathrm{s})$ とします。まず、それぞれのボールでの的に当たる確率は条件付き確率として

$$P(x_\mathrm{s}|y_1) = 1\ ,\ P(x_\mathrm{s}|y_2) = \frac{3}{4}$$

どちらのボールが選ばれるかの確率は $P(y_1) = 3/4,\ P(y_2) = 1/4$ なので、全体の的に当たる確率は

$$P(x_\mathrm{s}) = P(x_\mathrm{s}, y_1) + P(x_\mathrm{s}, y_2) = P(x_\mathrm{s}|y_1)P(y_1) + P(x_\mathrm{s}|y_2)P(y_2) = \frac{3}{4} + \frac{3}{4}\frac{1}{4} = \frac{15}{16}$$

ベイズの定理を使うと、的に当たるときに小さなボールが使用された確率は

$$P(y_2|x_\mathrm{s}) = \frac{P(x_\mathrm{s}|y_2)P(y_2)}{P(x_\mathrm{s})} = \frac{3}{4}\frac{1}{4}\frac{16}{15} = \frac{1}{5}$$

と求まります。

確率の話はこれで終わりにして、エントロピーの話に移ります。

まず設定を作ります。事象 X, Y があり、X を送信側、Y を受信側と呼ぶことにし、送信側で X での x_i が確率 $P(x_i)$ で起き、受信側で Y での y_j が確率 $P(y_j)$ で起きるとします。このとき、送信側で x_i が起き、受信側で y_j が起きる確率は条件付き確率から $P(y_j|x_i)$ で与えられ、これらの確率の関係は (2) で与えられます。

この状況において、送信側と受信側がそれぞれに持っている量が存在すると考えるなら、送信側の量を I_s、受信側での量を I_r(送信側の x_i が与えられているとして)、全体の量を I として

$$I = I_\mathrm{s} + I_\mathrm{r}$$

これらを特徴づける量を今は確率しか持っていないので、確率から作ります。作るのは簡単で (2) は積ですが、対数を取れば積は和になることから

$$\begin{aligned} P(x_i, y_j) &= P(y_j|x_i)P(x_i) \\ \log_a P(x_i, y_j) &= \log_a [P(y_j|x_i)P(x_i)] \\ &= \log_a P(y_j|x_i) + \log_a P(x_i) \\ I(x_i, y_j) &= I(y_j|x_i) + I(x_i) \end{aligned} \tag{4}$$

I の変数は確率の表記に合わせています。$I(x_i)$ は送信側で発生した量、$I(y_j|x_i)$ は送信側で x_i が起きたときに受信側で発生する量と言えます。そうすると、$I(x_i, y_j)$ はこのやり取りで生じた全体的な量と言えます。また、(3) からも

$$I(x_i, y_j) = I(x_i|y_j) + I(y_j)$$

となり、受信側から見たような関係になります。特に、送信側の確率と受信側の確率が独立であるなら $P(y_j|x_i) = P(y_j)$ なので

$$I(y_j|x_i) = \log_a P(y_j|x_i) = \log_a P(y_j) = I(y_j)$$

から

$$I(x_i, y_j) = I(y_j) + I(x_i)$$

となり、送信側と受信側のそれぞれの確率による量の和が結合確率による量となります。
確率は 1 以下なので対数は必ず負になることから

$$I(x_i) = -\log_a P(x_i)$$

と定義したものを情報量 (information) と言います。情報量は、確率が高いほど少なく、低いほど多くなっており、確定している事象 (確率が 1 の場合) では 0 です。情報量の単位は対数の底 a を決めることで与えられています。底を 2 としたときの情報量の単位はビット (bit, binary digit)、e としたときではナット (nat, natural unit of information)、10 としたときではディット (dit, decimal digit) と呼ばれます。これらは底の変換で繋がっており、例えばビットとナットの関係は底の変換から

$$\log_e x = \log_e 2 \cdot \log_2 x \quad (\log_a x = \log_a b \cdot \log_b x)$$

と与えられます。
情報量は確率を変数としているので、平均を

$$H_a(X) = h_a(P(x_1), P(x_2), \ldots, P(x_n)) = \sum_{i=1}^{n} P(x_i) I(P(x_i)) = -\sum_{i=1}^{n} P(x_i) \log_a P(x_i)$$

と与えられます。事象 X での x_i $(i = 1, 2, \ldots, n)$ の全ての和を取っているので H_a の変数に X が使われます (細かく言うと、$P(X = x_i)$ の確率を持つ確率変数 X)。$H_a(X)$ を X の情報エントロピー (information entropy) やシャノンエントロピー (Shanonn entropy)、もしくはエントロピーと呼びます。熱力学のエントロピーと混同するようなことはないので、エントロピーと呼んでいきます。
$x \log x$ は $x \to 0$ で 0 になることから、確率が 0 のときと 1 のときエントロピーは 0 とされます。言い換えれば、事象 X では 1 つのことのみが確定して起きるなら (例えば $P(x_3) = 1$ で残りが 0)、$H_a(X) = 0$ になります。$x = 0$ の極限のはロピタルの定理から求められます。ロピタルの定理は

$$\lim_{x \to c} \frac{f(x)}{g(x)} = \lim_{x \to c} \frac{f'(x)}{g'(x)} \quad (f'(x) = \frac{df}{dx},\ g'(x) = \frac{dg}{dx})$$

$$\lim_{x \to c} f(x) = \pm\infty\ ,\ \lim_{x \to c} g(x) = \pm\infty$$

$\lim_{x \to 0} \log_a x = -\infty,\ \lim_{x \to 0} 1/x = \infty$ からロピタルの定理が使えるので

$$\lim_{x \to 0} x \log_a x = \lim_{x \to 0} \frac{\log_a x}{1/x} = \lim_{x \to 0} (\frac{d}{dx} \log_a x)(\frac{d}{dx} \frac{1}{x})^{-1} = -\lim_{x \to 0} \frac{x^2}{x \ln a} = -\frac{1}{\ln a} \lim_{x \to 0} x = 0$$

となります。
エントロピーと言っているときはほとんどの場合で底を $a = 2$ にしているので、ここでもこれ以降は log と書いているときは $\log_2$ を表すとします。また、自然対数 $\log_e$ は ln と表記します。底の変更によるエントロピーの対応は

$$H_b(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) = -(\log_b a) \sum_{i=1}^{n} P(x_i) \log_a P(x_i) = (\log_b a) H_a(X)$$

となるだけです。

エントロピーのよくある実用的な例は、事象 X において x_1, x_2 の 2 通りしか起きなく、x_1 の確率を p とした場合です。このときのエントロピーは

$$H = -p \log p - (1-p) \log[1-p]$$

$p = 0, 1$ のとき $H = 0$ となり、極値は

$$0 = \frac{dH}{dp} = -\log_2 p - \frac{1}{\log_e 2} + \log_2[1-p] + \frac{1}{\log_e 2} = -\log_2 p + \log_2[1-p]$$

から、$p = 1/2$ の地点での $H(1/2) = 1$ となる凹関数です ($p = 1/2$ を中心に対称)。よって、x_1, x_2 が同じ確率 $1/2$ のときに、エントロピーは最大値を持ちます。

一般的な場合でも起きる確率が全て等しいときエントロピーが最大になることは「最適な符号」で示したものとほぼ同じです。事象 X での x_i $(i = 1, 2, \ldots, n)$ の確率を p_i と書きます。対数の関係として

$$\sum_{i=1}^{n} x_i \log_a \frac{1}{x_i} \leq \sum_{i=1}^{n} x_i \log_a \frac{1}{y_i} \quad (\sum_{i=1}^{n} x_i = 1 \ , \ \sum_{j=1}^{n} x_i = 1)$$

これの等号は $x_i = y_i$ のときに成立します。今の場合に合わせて書けば、任意の確率 q_i $(i = 1, 2, \ldots, n)$ に対して

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i$$

任意なので $q_i = 1/n$ と選ぶと

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log \frac{1}{n} = (-\log \frac{1}{n}) \sum_{i=1}^{n} p_i = -\log \frac{1}{n}$$

そして、$p_i = q_i$ のときに等号になるので、$p_i = 1/n$ のとき最大になります。よって、エントロピーは

$$0 \leq H(X) \leq \log n$$

となり、X の確率が全て等しいとき最大値となります。また、ラグランジュの未定乗数法による求め方は下の補足で示しています。

全ての結果が同じ確率のときエントロピーが最大になることは、最も曖昧なときにエントロピーは最大になると言えます (明日の天気が晴れ 90% 、雨 10% より、晴れ 50% 、雨 50% のほうが曖昧)。このことと、確定しているときは 0 になることから、エントロピーは曖昧さ (uncertain) を表す量と言われます。

事象 X, Y での結合確率を使って

$$H(X, Y) = -\sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) \log P(x_i, y_j)$$

としたものを結合エントロピー (joint entropy) と呼びます。X では n 個、Y では m 個の結果があるとしています。結合確率の定義から

$$H(X,Y) \geq 0$$

$$H(X,Y) = H(Y,X)$$

対数内の和を取ると、(1) から

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log\sum_{k=1}^{m} P(x_i,y_k) = -\sum_{i=1}^{n} P(x_i)\log P(x_i) = H(X)$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log\sum_{k=1}^{n} P(x_k,y_j) = -\sum_{j=1}^{m} P(y_j)\log P(y_j) = H(Y)$$

となります。

結合確率と条件付き確率から

$$H(Y|X) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_i)\log P(y_j|x_j) \tag{5}$$

としたものを条件付きエントロピー (conditional entropy) と呼びます。結合確率が使われているのは、条件付き確率が x_i を前提とした確率になっているためです。条件付き確率による情報量は

$$I(y_j|x_i) = -\log P(y_j|x_i)$$

これは x_i が与えられているとしているので、平均は Y での和を取ることになり

$$H(Y|x_i) = \sum_{j=1}^{m} P(y_j|x_i)I(y_j|x_i) = -\sum_{j=1}^{m} P(y_j|x_i)\log P(y_j|x_i)$$

さらに X での平均を取るために $P(x_i)$ をかけて和を取れば

$$\begin{aligned}\sum_{i=1}^{n} P(x_i)H(Y|x_i) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i)P(y_j|x_i)\log P(y_j|x_i)\\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i,y_j)\log P(y_j|x_i)\end{aligned}$$

となり、条件付きエントロピーとなります。もしくは、(4) から

$$
\begin{aligned}
I(y_j|x_i) &= I(x_i, y_j) - I(x_i) \\
P(x_i, y_j)I(y_j|x_i) &= P(x_i, y_j)I(x_i, y_j) - P(x_i, y_j)I(x_i) \\
\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)I(y_j|x_i) &= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)I(x_i, y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)I(x_i) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)I(x_i, y_j) - \sum_{i=1}^{n} P(x_i)I(x_i) \\
&= H(X, Y) - H(X)
\end{aligned}
$$

となるので、左辺を条件付きエントロピーとしているとも言えます。これは確率の積の法則をエントロピーにしたものになっています。また、結合エントロピーから

$$
\begin{aligned}
H(X, Y) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log P(x_i, y_j) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log P(y_j|x_i)P(x_i) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log P(y_j|x_i) - \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log P(x_i) \\
&= -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)\log P(y_j|x_i) - \sum_{i=1}^{n} P(x_i)\log P(x_i) \\
&= H(Y|X) + H(X)
\end{aligned}
$$

としても同じように求まります。

$H(X|Y)$ でも定義は同じで

$$
\begin{aligned}
H(X|Y) &= \sum_{j=1}^{m} P(y_j)H(X|y_j) = -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i|y_j)P(y_j)\log P(x_i|y_j) \\
H(X|Y) &= \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)I(x_i, y_j) - \sum_{j=1}^{m} P(y_j)I(y_j) = H(X, Y) - H(Y)
\end{aligned}
$$

見て分かるように、$H(Y|X) \neq H(X|Y)$ です。

エントロピーと条件付きエントロピーから作られる量として、相互情報量 (mutual information) があります。これは

$$
I(X;Y) = H(X) - H(X|Y)
$$

と定義されます。$I(X:Y)$ と表記されることもあります。条件付きエントロピーの関係から

$$
I(X;Y) = H(X) - H(X|Y) = H(X) - (H(X,Y) - H(Y)) = H(Y) - (H(X,Y) - H(X)) = H(Y) - H(Y|X)
$$

となるので、相互情報量は X, Y の入れ替えで同じです。

相互情報量は確率で書くと

$$\begin{aligned} I(X;Y) = & -\sum_{i=1}^{n} P(x_i) \log P(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_i) \log P(x_i|y_j) \\ = & -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) \log P(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_i) \log \frac{P(x_i, y_j)}{P(y_j)} \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_i) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \end{aligned} \tag{6}$$

これからも相互情報量は $I(X;Y) = I(Y;X)$ と分かります。

相互情報量の意味に触れておきます。(6) の 1 行目を

$$\begin{aligned} & -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j) \log P(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_i) \log P(x_i|y_j) \\ = & -\sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)(\log P(x_i) - \log P(x_i|y_j)) \\ = & \sum_{i=1}^{n}\sum_{j=1}^{m} P(x_i, y_j)(I(x_i) - I(x_i|y_j)) \end{aligned}$$

と変形すると、X の情報量から Y が起きる前提での X の情報量を引いたものがいます。なので、Y が起きたと知ることで X の情報量が減ったときの量と言えます (Y を知ることで X の曖昧さが減ったときの量)。このことから、相互情報量は X, Y の関係性を定量化したものと言えます。実際に、X, Y が独立なら $P(x_i, y_j) = P(x_i)P(y_j)$ なので、相互情報量は 0 です (独立なら Y が起きたことを知っても X については何も分からない)。他にも、Y では y のみが確実に起きる場合では $P(x_i|y) = P(x_i)$ なので、このときも相互情報量は 0 です。y が確実に起きるために、y が起きたことを知ったところで X については何も分からないからです。

もしくは

$$I(x_i) - I(x_i|y_j) = I(x_i) + I(y_j) - I(x_i, y_j)$$

と書くと、X, Y の情報量の和と X, Y が同時に起きるときの情報量の差になっているので、X, Y の関係性を表す部分だけが残っていると言えます。

もう 1 つエントロピーを定義します。事象 X に対する 2 つの確率 $P(x_i), Q(x_i)$ があるとして

$$H(P||Q) = -\sum_{i} P(x_i) \log \frac{Q(x_i)}{P(x_i)} = \sum_{i} P(x_i)(\log P(x_i) - \log Q(x_i))$$

としたものを相対エントロピー (relative entropy) やカルバック・ライブラー情報量 (Kullback-Leibler divergence) と呼びます。英語では divergence となっていますが、無限大への発散やベクトルの発散とは無関係です。2 つの確率 P, Q による差になっているので divergence でなく distance が使われることもありますが、数学での距離の意味にはなっていません。

確率の変数が N 個の場合での表記にも触れておきます。結合エントロピーで N 個の結合確率にすればいいだけなので

$$H(X_1, X_2, \ldots, X_N) = -\sum_{i_1, i_2, \ldots, i_N} P(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_N}^{(N)}) \log P(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_N}^{(N)})$$

$x_{i_k}^{(k)}$ は X_k での結果とし、Σ での i_k はそれぞれの結果の数の範囲で取るという意味です。例えば、X_3 の結果が 4 個ならなら $x_1^{(3)}, x_2^{(3)}, x_3^{(3)}, x_4^{(3)}$ が結果です。これだと表記が煩わしいので、確率の変数は

$$P(x_1, x_2, \ldots, x_N)\ , \sum_{i_1, i_2, \ldots, i_N} = \sum_{x_1, x_2, \ldots, x_N}$$

と書かれることが多いです。もしくは、これを使うと $H(X)$ での X の結果 $x_1, x_2, \ldots, x_N$ と混同しそうなときは

$$P(x^{(1)}, x^{(2)}, \ldots, x^{(N)})\ , \sum_{x^{(1)}, x^{(2)}, \ldots, x^{(N)}}$$

と表記します。また、ベクトルのように太字にして

$$H(\boldsymbol{X}) = H(X_1, X_2, \ldots, H_N)$$
$$P(\boldsymbol{x}) = P(x_{i_1}^{(1)}, x_{i_2}^{(2)}, \ldots, x_{i_N}^{(N)}) = P(x^{(1)}, x^{(2)}, \ldots, x^{(N)}) = P(x_1, x_2, \ldots, x_N)$$

と書きます。$\boldsymbol{X}$ は確率変数のベクトルということで確率ベクトル (randam vector) や確率変数ベクトルと呼ばれます。

最後に、関数に条件を与えるとエントロピーが出てくることを示します。表記を先に与えておきます。確率を変数とする実数値関数を F とし、事象 X, Y での結合確率を変数にするときは $F(X, Y)$、条件付き確率が変数のときは $F(Y|X)$ のように表記し

$$F(X) = F(P(x_1), P(x_2), \ldots P(x_n))$$
$$F(X, Y) = F(P(x_1, y_1), P(x_1, y_2), \ldots P(x_n, y_m))$$
$$F(y_j|x_i) = F(P(y_j|x_i))$$
$$F(Y|x_i) = F(P(y_1|x_i), P(y_2|x_i), \ldots, P(y_m|x_i))$$
$$F(Y|X) = \sum_{i=1}^{n} P(x_i) F(Y|x_i)$$

X, Y が独立なときは

$$F(Y|x_i) = F(P(y_1), P(y_2), \ldots, P(y_m)) = F(Y)$$
$$F(Y|X) = F(Y) \sum_{i=1}^{n} P(x_i) = F(Y)$$

となります。

関数 F は次の性質を持つとします。

(i) 確率が全て等しいとき最大値を持つ。

(ii) 事象 X, Y に対して $F(X,Y) = F(Y|X) + F(X)$。

(iii) $F(p_1, p_2, \ldots, p_n, 0) = F(p_1, p_2, \ldots, p_n)$。

(iv) F は連続。

(i) はエントロピーの性質、(ii) はエントロピーを導出するときに使った関係です。(iii) は確率が 0 のものを加えても変化しないとしているだけです。(iv) は数学的な要請です。このとき、F がエントロピーになることを示します。

(i) を表すために

$$A(n) = F(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$$

と定義します。$A(n)$ が増加関数であることは簡単に分かります。$n = 2$ とすれば、(iii) から

$$A(2) = F(\frac{1}{2}, \frac{1}{2}) = F(\frac{1}{2}, \frac{1}{2}, 0)$$

$n = 3$ では、F は確率が全て等しいとき (変数が全て同じとき) に最大なので

$$A(3) = F(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \geq F(\frac{1}{2}, \frac{1}{2}, 0) = A(2)$$

これは任意の n で成立するので、$A(n) \leq A(n+1)$ と分かります。

事象 X_1, X_2 が独立であるとし、$F(X_1) = A(r),\ F(X_2) = A(r)$ となっているとします。(ii) から

$$F(X_1, X_2) = F(X_2|X_1) + F(X_1) = F(X_2) + F(X_1)$$

X_1, X_2 の結果の個数が r で、確率が全て等しいとしているので

$$F(X_1, X_2) = 2A(r)$$

一方で、X_1, X_2 の確率は

$$(\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r})\ ,\ (\frac{1}{r}, \frac{1}{r}, \ldots, \frac{1}{r})$$

となっているために、結合確率における組み合わせの数は r^2 個で、結合確率は $1/r^2$ です。このため

$$F(X_1, X_2) = A(r^2)$$

と書けます。

同じように X_1, X_2, X_3 が独立であるなら

$$F(X_1, X_2, X_3) = F(X_2, X_3|X_1) + F(X_1) = F(X_2, X_3) + F(X_1) = F(X_2) + F(X_3) + F(X_1)$$

これも各結果の個数が r で、確率が全て等しいなら

$$F(X_1, X_2, X_3) = 3A(r)$$

このときの結合確率は $1/r^3$ で、組み合わせは r^3 個なので

$$F(X_1, X_2, X_3) = A(r^3)$$

後は同じことの繰り返しなので、$X_1, X_2, \ldots, X_\alpha$ では

$$F(X_1, X_2, \ldots, X_\alpha) = \alpha A(r) F(X_1, X_2, \ldots, X_\alpha) = A(r^\alpha)$$

ここで、事象における結果の数 s と事象の数 β を

$$r^\alpha \leq s^\beta \leq r^{\alpha+1} \tag{7}$$

となるように選んだとします。A は増加関数と分かっているので

$$\begin{aligned} A(r^\alpha) \leq & A(s^\beta) \leq A(r^{\alpha+1}) \\ \alpha A(r) \leq & \beta A(s) \leq (\alpha+1) A(r) \\ \alpha \leq & \beta \frac{A(s)}{A(r)} \leq \alpha + 1 \\ \frac{\alpha}{\beta} \leq & \frac{A(s)}{A(r)} \leq \frac{\alpha}{\beta} + \frac{1}{\beta} \end{aligned}$$

なので

$$\frac{A(s)}{A(r)} - \frac{\alpha}{\beta} \leq \frac{1}{\beta} \tag{8}$$

これとは別に (7) の対数 (ここでの log は底を指定していない) を取ってみると

$$\begin{aligned} \alpha \log r \leq & n \log s \leq (\alpha+1) \log r \\ \alpha \leq & \beta \frac{\log s}{\log r} \leq \alpha + 1 \\ \frac{\alpha}{\beta} \leq & \frac{\log s}{\log r} \leq \frac{\alpha}{\beta} + \frac{1}{\beta} \end{aligned}$$

となるので

$$\frac{\log s}{\log r} - \frac{\alpha}{\beta} \leq \frac{1}{\beta} \tag{9}$$

(8),(9) に三角不等式

$$|x+y| \leq |x| + |y|$$

を使うと

$$\begin{aligned} |\frac{A(s)}{A(r)} - \frac{\log s}{\log r}| &= |\frac{A(s)}{A(r)} - \frac{\alpha}{\beta} + \frac{\alpha}{\beta} - \frac{\log s}{\log r}| \\ &\leq |\frac{A(s)}{A(r)} - \frac{\alpha}{\beta}| + |\frac{\alpha}{\beta} - \frac{\log s}{\log r}| \\ &\leq \frac{2}{\beta} \end{aligned}$$

左辺は α, β を含んでなく、β は (7) を満たしさえすれば任意なので、十分大きいと取れて

$$\frac{A(s)}{A(r)} = \frac{\log s}{\log r} \;\Rightarrow\; A(s) = K \log s \quad (K > 0) \tag{10}$$

K は定数で、A は増加関数なので正です。

次に、確率が等しくない場合を見ていきます。事象 X での確率が、λ_i, λ を整数として

$$P(x_i) = \frac{\lambda_i}{\lambda}\ ,\ \sum_{i=1}^{n} P(x_i) = \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} = 1 \quad (\lambda \neq 0)$$

と与えられているとします。これは確率が有理数になっている場合です。事象 Y による結果を

$$y_1^{(1)}, y_2^{(1)}, \ldots, y_{\lambda_1}^{(1)}, y_1^{(2)}, y_2^{(2)}, \ldots, y_{\lambda_2}^{(2)}, \ldots, y_1^{(n)}, y_2^{(n)}, \ldots, y_{\lambda_n}^{(n)}$$

と並べたとします。(10) を利用するために、このとき条件付き確率を

$$\begin{aligned} P(y_k^{(i)}|x_i) &= \frac{1}{\lambda_i} \quad (1 \leq k \leq \lambda_i) \\ P(y_k^{(c)}|x_i) &= 0 \quad (1 \leq k \leq \lambda_c\ ,\ c \neq i) \end{aligned}$$

と与えたとします。これは $y_1^{(i)}$ から $y_{\lambda_i}^{(i)}$ までは確率 $1/\lambda_1$、それ以外は 0 としているだけです。$i = 1$ とすれば

$$\begin{aligned} P(y_k^{(1)}|x_1) &= \frac{1}{\lambda_1} \quad (1 \leq k \leq \lambda_1) \\ P(y_k^{(c)}|x_1) &= 0 \quad (1 \leq k \leq \lambda_c\ ,\ c \neq 1) \end{aligned}$$

このときは

$$
\begin{aligned}
F(Y|x_1) &= F(P(y_1^{(1)}|x_1), P(y_2^{(1)}|x_1), \ldots, P(y_{\lambda_n}^{(n)}|x_1)) \\
&= F(\frac{1}{\lambda_1}, \frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_1}, 0, \ldots, 0) \\
&= F(\frac{1}{\lambda_1}, \frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_1}) \\
&= A(\lambda_1) \\
&= K \log \lambda_1
\end{aligned}
$$

他の x_i でも同様なので $F(Y|x_i) = K \log \lambda_i$ となり

$$
F(Y|X) = \sum_{i=1}^{n} P(x_i) F(Y|x_i) = K \sum_{i=1}^{n} P(x_i) \log \lambda_i = K \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \log \lambda_i
$$

と求まります。

一方で、結合確率は

$$
\begin{aligned}
P(x_i, y_k^{(i)}) &= P(y_k^{(i)}|x_i) P(x_i) = \frac{1}{\lambda_i} \frac{\lambda_i}{\lambda} = \frac{1}{\lambda} \\
P(x_i, y_k^{(c)}) &= P(y_k^{(c)}|x_i) P(x_i) = 0 \quad (c \neq i)
\end{aligned}
$$

そうすると、結合確率による F は

$$
F(X,Y) = F(P(x_1, y_1^{(1)}), P(x_1, y_2^{(1)}), \ldots, P(x_1, y_{m_1}^{(1)}), \ldots, P(x_n, y_{m_n}^{(n)})) = F(\frac{1}{\lambda}, \frac{1}{\lambda} \ldots, \frac{1}{\lambda}) = K \log \lambda
$$

これらを (ii) に入れれば

$$
\begin{aligned}
F(X) = F(X,Y) - F(Y|X) &= K \log \lambda - K \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \log \lambda_i \\
&= -K \sum_{i=1}^{n} \frac{\lambda_i}{\lambda} \log \frac{\lambda_i}{\lambda} \\
&= -K \sum_{i=1}^{n} P(x_i) \log P(x_i) \qquad (11)
\end{aligned}
$$

となり、エントロピーになります。

有理数としてますが、そのまま実数にできます。無理数は有理数による数列の無限大の極限なので、有理数による数列

$$
\{P^{(1)}(x_i), P^{(2)}(x_i), \ldots, P^{(N)}(x_i)\}
$$

を作ったとき、その極限として無理数 $P(x_i)$ の式

$$\lim_{N\to\infty}\sum_{i=1}^{n}P^{(N)}(x_i)\log P^{(N)}(x_i)=\sum_{i=1}^{n}P(x_i)\log P(x_i)$$

が存在します。そして、(iv) で F は連続としているので極限は 1 つに決まることから、無理数に対しても (11) です。よって、実数 $P(x_i)$ に対して

$$F(X)=-K\sum_{r=1}^{n}P(x_r)\log P(x_r)$$

となります。

・補足

ラグランジュの未定乗数法を使って、エントロピーの最大値が $\log n$ になることを示します。ラグランジュ乗数を λ、拘束条件を g として、関数 F を

$$F=H(p_1,p_2,\ldots,p_n)+\lambda g(p_1,p_2,\ldots,p_n)$$

確率を p_i とし、変数をはっきりさせるためにエントロピーの変数を p_i で書いています。拘束条件は確率を全て足せば 1 になるということから

$$\sum_{i=1}^{n}p_i=1\ \Rightarrow\ g(p_1,p_2,\ldots,p_n)=1-\sum_{i=1}^{n}p_i$$

このとき、H の極値を与える $(p_1,p_2,\ldots,p_n)$ は

$$\frac{\partial F}{\partial p_i}\Big|_{p_i=x_i}=0$$

として求められます。極値を与える p_i を x_i としています。

対数の底の変換でエントロピーは定数倍しか変わらないので、微分で余計な係数が出てこない自然対数で行います。偏微分は、$i=k$ の項以外は 0 になるので

$$\begin{aligned}\frac{\partial F}{\partial p_k}=\frac{\partial H}{\partial p_k}+\lambda\frac{\partial g}{\partial p_k}&=\frac{\partial}{\partial p_k}(-\sum_{i=1}^{n}p_i\ln p_i)+\lambda\frac{\partial}{\partial p_k}(1-\sum_{i=1}^{n}p_i)\\&=-\frac{\partial}{\partial p_k}(p_k\ln p_k)-\lambda\\&=-\ln p_k-1-\lambda\end{aligned}\tag{12}$$

また、もう 1 回偏微分を行うと

$$\frac{\partial^2 F}{\partial p_j\partial p_k}=0\quad(j\neq k)$$

$$\frac{\partial^2 F}{\partial p_k\partial p_k}=-\frac{1}{p_k}<0$$

このため、極値は極大値です ($p_k = 0$ のときは $p_k \ln p_k \to 0$)。
(12) が 0 になればいいので

$$\ln p_k = -\lambda - 1$$

$$p_k = e^{-\lambda-1}$$

k の依存性がないので

$$\sum_{k=1}^{n} p_k = e^{-\lambda-1} n$$

これが 1 なので、λ は

$$e^{-\lambda-1} n = 1$$

$$\ln e^{-\lambda-1} = \ln \frac{1}{n}$$

$$-\lambda - 1 = -\ln n$$

$$\lambda = \ln n - 1$$

と求まり、極値を与える p_k は

$$-\lambda - 1 = -\ln n + 1 - 1 = -\ln n$$

から

$$p_k = e^{-\ln n} = \frac{1}{n}$$

よって、全ての確率が等しいときにエントロピーは最大となり

$$H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) = -\sum_{i=1}^{n} p_i \ln p_i = \ln n$$

底を b に変更しても

$$(\log_b e) H_e = (\log_b e) \log_e n$$

$$H_b = \log_b n$$

となるだけなので、同じです。