第7回カルバック情報量

荒木 純道

2012/11/15

カルバック情報量

- 確率密度関数同士の”近さ”を数量化するもの。

2 つの確率密度関数 $p(x), q(x), x \in \mathbf{R}^N$ の**カルバック情報量**(Kullback information):

カルバック情報量

$$D(p \| q) = \int_{\mathbf{R}^N} p(x) \log \frac{p(x)}{q(x)} dx$$

ただし、p, q は正で連続、積分が計算できる場合を考える。

カルバック情報量

- カルバック情報量についてつぎの性質が成り立つ。
  1. $D(p \| q) \geq 0$
  2. $D(p \| q) = 0$ ならば、$p(x) = q(x)$

  よって、p と q の距離に似た尺度。ただし、
  3. $D(p \| q) \neq D(q \| p)$ 、非対称

D の性質：証明

関数 $F(t) = \log t + \frac{1}{t} - 1, \quad t > 0$ は

$$F'(t) = \frac{1}{t} - \frac{1}{t^2} = \frac{t-1}{t^2}$$

より、t=1 で最小値0 をとる。

$$D(p \| q) = \int_{\mathbf{R}^N} p(x) F\left(\frac{p(x)}{q(x)}\right) dx$$

と書けるから、$D > 0$ であり、$D = 0$ となるのは $F\left(p(x)/q(x)\right) = 0$

つまり， $p(x) = q(x)$ のときに限る。

例

- 平均が0 で分散が異なる正規分布

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{x^2}{2\sigma_1^2}\right) \qquad q(x) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{x^2}{2\sigma_2^2}\right)$$

の間のカルバック情報量は

$$\begin{aligned} D(p \| q) &= \int_{-\infty}^{\infty} p(x) \log\left(\frac{\sigma_2}{\sigma_1} \exp\left(-\frac{x^2}{2\sigma_1^2} + \frac{x^2}{2\sigma_2^2}\right)\right) dx \\ &= \log\frac{\sigma_2}{\sigma_1} + \frac{1}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) \int_{-\infty}^{\infty} x^2 p(x) dx \\ &= \frac{1}{2}\left(\log\frac{\sigma_2^2}{\sigma_1^2} + \frac{\sigma_2^2}{\sigma_1^2} - 1\right) \\ &= \frac{1}{2} F\left(\frac{\sigma_2^2}{\sigma_1^2}\right) \end{aligned}$$

よって、$\sigma_1 = \sigma_2$ のときに限り最小値D=0 をとる。

有限集合あるいは可算集合

- 高々可算集合 Ω 上での定義：

任意の $\omega (\in \Omega)$ のついて $p(\omega), q(\omega)$ とする。

$$D(p \| q) = \sum_{w \in \Omega} p(\Omega) \log \frac{p(\omega)}{q(\omega)}$$

ただし、確率関数 $p(\omega), q(\omega)$ は正とする。

尤度関数

- 多数のデータ $x_i (i = 1, \cdots, n)$ から、それらの背後にある確率密度関数 $p(x)$ を推定したい場合。

$p(x)$ は未知→ $D(p \| q)$ は計算できない。その代わりに

$$L(q) = \prod_{i=1}^{n} q(x_i) \qquad \textit{Likelihood} \quad \textit{Function}$$

をなるべく大きくするような $q(x)$ を求める。但し、$q(x) \geq 0, \int q(x) dx = 1$

なぜならば

$$\begin{aligned} \frac{1}{n} \log L(q) &= \frac{1}{n} \sum_{i=1}^{n} \log q(x_i) \simeq \int p(x) \log q(x) dx \\ &= \int p(x) \log p(x) dx - \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \int p(x) \log p(x) dx - D(p \| q) \end{aligned}$$

よって、$L(q)$ をなるべく大きくする

$\Leftrightarrow \frac{1}{n} \log L(q)$ をなるべく大きくする

$\Rightarrow D(p \| q)$ をなるべく小さくする。

問1

関数 $F(t)=\log t+\frac{1}{t}-1$ を用いて、$D(p\|q)$ が

$$D(p\|q)=\int_{\mathbf{R}^N} p(x)F\left(\frac{p(x)}{q(x)}\right)dx$$

と書けることを示せ。

$$\int_{\mathbf{R}^N} p(x)F\left(\frac{p(x)}{q(x)}\right)dx=\int_{\mathbf{R}^N} p(x)\left(\log\frac{p(x)}{q(x)}+\frac{p(x)}{q(x)}-1\right)dx$$

$$=D(p\|q)+\int_{\mathbf{R}^N} q(x)dx-\int_{\mathbf{R}^N} p(x)dx$$

$$=D(p\|q)$$

問2

平均の異なる正規分布

$$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \qquad q(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-M)^2}{2\sigma^2}\right)$$

の間のカルバック情報量を求めよ。

また、それが0 となるのはとういう時か示せ。

$$D(p\|q)=\int_{-\infty}^{\infty} p(x)\log\left(\exp\left(-\frac{(x-m)^2}{2\sigma^2}+\frac{(x-M)^2}{2\sigma^2}\right)\right)dx$$

$$=\frac{1}{2\sigma^2}\int_{-\infty}^{\infty} p(x)((x-M)^2-(x-m)^2)dx$$

$$=\frac{1}{2\sigma^2}\left(2(m-M)\int_{-\infty}^{\infty} xp(x)dx-(m^2-M^2)\right)$$

$$=\frac{(m-M)^2}{2\sigma^2}$$

よって $m=M$ のとき最小値0 をとる。

応用の例

カルバック情報量 $D(p\|q)$

→シャノン・エントロピー

→相互情報量(通信容量)

→仮説検定(Stein補題)

→大偏位理論(Sanov定理)