ECONOMETRÍA I
Actuaría. Facultad de Ciencias.
Carlos Vladimir Rodríguez-Caballero

Entropía y teoría de la información

1. Máxima verosimilitud

Había sido utilizado antes por Gauss, Laplace, Thiele y F. Y. Edgeworth, aunque fue finalmente recomendado, analizado y popularizado por R. A. Fisher entre 1912 y 1922.

Sea x variable aleatoria definida $(\Omega, F, \Re)$. La función de distribución de x admite una función de densidad f con respecto a R. $(F(x) : R(\mathrm{x} \leq x))$.
Contexto: La función de densidad f pertenece a una familia paramétrica $f(x;\theta)$ con $f(x;\theta_0) = f(x)$

La siguiente proposición es el resultado clave para fundamentar máxima verosimilitud.

Proposición 1 *Para cada parámetro admisible θ tenemos la desigualdad siguiente*

$$E_R[logf(x;\theta)] \leq E_R[logf(x;\theta_0)] = E_R[logf(x)]$$

Demostración

$$E_R[\log f(x;\theta) - \log f(x;\theta_0)]$$
$$= \int_{-\infty}^{\infty} f(x;\theta) f(x;\theta_0) dx - \int_{-\infty}^{\infty} \log f(x;\theta_0) f(x;\theta_0) dx$$
$$= \int_{-\infty}^{\infty} f(x;\theta_0) \log \left(\frac{f(x;\theta)}{f(x;\theta_0)} \right) dx$$

teniendo en cuenta que $\log x \leq x - 1$

$$\leq \int_{-\infty}^{\infty} f(x;\theta_0) \left\{ \frac{f(x;\theta)}{f(x;\theta_0)} - 1 \right\} dx$$
$$= \int_{-\infty}^{\infty} (f(x;\theta) - f(x;\theta_0)) dx = 0$$

Ahora bien, de $E_R[\log f(x;\theta)] \leq E_R[\log f(x;\theta_0)]$, tenemos

$$\max_\theta E_R[\log f(x;\theta)] = E_R[\log f(x;\theta_0)]$$

□

2. Entropía

Entropía es un concepto usado originalmente en termodinámica, mecánica estadística y luego en teoría de la información. Se concibe como una medida del desorden o una medida de la incertidumbre, aunado a ello, la información tiene que ver con cualquier proceso que permite acotar, reducir o eliminar la incertidumbre; resulta que el concepto de información y el de entropía están ampliamente relacionados entre sí, aunque se tardó años en el desarrollo de la mecánica estadística y la teoría de la información para hacer esto aparente.

El punto de vista que se explica aquí data como una formulación que hace la teoría de la información. Por ello la entropía se llama frecuentemente entropía de Shannon, en honor a Claude E. Shannon.

La entropía asociada a la variable aleatoria X es un número que depende directamente de la distribución de probabilidad de X e indica como es de predictible el resultado del proceso sujeto a incertidumbre o experimento. Desde un punto de vista matemático cuanto más plana sea la distribución de probabilidad más difícil será acertar cual de las posibilidades se dará en cada instancia. Una distribución es plana (tiene alta entropía) cuando todos los valores de X tienen probabilidades similares, mientras que es poco plana cuando algunos valores de X son mucho más probables que otros (se dice que la función es más puntiguda en los valores más probables). En una distribución de probabilidad plana (con alta entropía) es difícil poder predecir cuál es el próximo valor de X que va a presentarse, ya que todos los valores de X son igualmente probables.

Shannon ofrece una definición de entropía que satisface las siguientes afirmaciones:

- La medida de información debe ser proporcional (continua). Es decir, el cambio pequeño en una de las probabilidades de aparición de uno de los elementos de la señal debe cambiar poco la entropía.
- Si todos los elementos de la señal son equiprobables a la hora de aparecer, entonces la entropía será máxima.

La información que aporta un determinado valor (símbolo), x_i, de una variable aleatoria discreta X, se define como:

$$I(x_i) = \log_2 \frac{1}{p(x_i)} = -\log_2 p(x_i)$$

cuya unidad es el bit cuando se utiliza el logaritmo en base 2 (por ejemplo, cuando se emplea el logaritmo neperiano se habla de nats). A pesar del signo negativo en la última expresión, la información tiene siempre signo positivo (lo cual queda más claro en la primera expresión).

La entropía determina el límite máximo al que se puede comprimir un mensaje usando un enfoque símbolo a símbolo sin ninguna pérdida de información (demostrado analíticamente por Shannon), el límite de compresión (en bits) es igual a la entropía multiplicada por el largo del mensaje. También es una medida de la información promedio contenida en cada símbolo del mensaje. Su cálculo se realiza a partir de su distribución de probabilidad p(x) mediante la siguiente fórmula:

$$H(X) = \mathbb{E}(I(X)) = \sum_{i=1}^{n} p(x_i) \log_a \left(\frac{1}{p(x_i)} \right) = -\sum_{i=1}^{n} p(x_i) \log_a p(x_i)$$

Propiedades de la entropía

1. $0 <= H <= loga(m)$. Es decir, la entropía H esta acotada superiormente (cuando es máxima) y no supone perdida de información.

2. Dado un procesos con posibles resultados $\{A1, .., An\}$ con probabilidades relativas $p1, \ldots, pn$, la función $H(p_1, \ldots, p_n)$, es máxima en el caso de que $p_1 = \cdots = p_n = 1/n$,

3. Dado un procesos con posibles resultados $\{A1, .., An\}$ con probabilidades relativas $p1, \ldots, pn$, la función $H(p_1, \ldots, p_n)$, es nula en el caso de que $pi = 0$ para cualquier i.

3. Divergencia Kullback-Leibler

En teoría de la probabilidad la divergencia de Kullback-Leibler es un indicador de la similitud entre dos funciones de distribución. Dentro de la teoría de la información también se la conoce como divergencia de la información, ganancia de la información o entropía relativa. Se trata de una divergencia y no una métrica (distancia) por no ser simétrica.

La divergencia de Kullback-Leibler entre dos funciones de distribución P y Q suele representarse así:

$$D_{\mathrm{KL}}(P\|Q)$$

Si P es una medida de probabilidad que es absolutamente continua con respecto a otra, Q, (condición necesaria para que $D_{\mathrm{KL}}(P\|Q)$ sea finito) y si $\frac{dP}{dQ}$ es la derivada de Radon-Nikodym de P con respecto a Q, se define la divergencia de Kullback-Leibler desde P hasta Q de la forma

$$D_{KL}(P\|Q) = \int_X \log \frac{dP}{dQ}\, dP = \int_X \frac{dP}{dQ} \log \frac{dP}{dQ}\, dQ$$

A la última integral se le conoce con el nombre de **entropía relativa**. De la misma manera, si Q es absolutamente continua con respecto a P, entonces

$$D_{\mathrm{KL}}(P\|Q) = -\int_X \log \frac{dQ}{dP}\, dP$$

La divergencia de Kullback-Leibler no depende de la medida μ. Cuando esta medida es la de medida de Lebesgue sobre el eje real, resulta

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

Lo que facilita su uso en la estadística y la econometría.

4. Resultado importante

Supongamos dos funciones de distribución con densidad

$$F_p(x) = \int p(z)dz \qquad \text{y} \qquad F_q(x) = \int q(z)dz$$

$$D_{\mathrm{KL}}(P\|Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)}\, dx$$

Por otro lado sabemos que

$$\max_\theta E_R[\log f(x;\theta)] = \max_\theta \int_{\infty}^{\infty} f(x,\theta_0) \log \frac{p(x)}{q(x)} dx$$

De dónde se obtiene que es lo mismo que

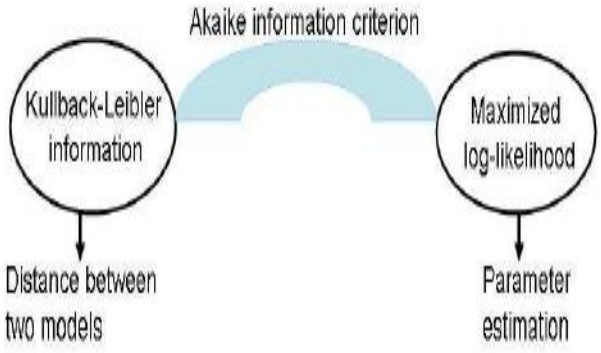
$$\min_\theta D_{\mathrm{KL}}(f(x;\theta_0)\|f(x;\theta)) = 0$$

Es decir, maximizar la función de verosimilitud es (aproximadamente) equivalente a encontrar el parámetro θ que minimiza la divergencia de Kullback-Leibler entre la distribución real y la familia de distribuciones parametrizadas por dicho parámetro.

5. Criterio de Información de Akaike (AIC)

5.1. Información Kullback-Leibler

Definimos la información de Kullback-Leibler como la pérdida de información que se presenta cuando un modelo es usado para aproximar la realidad completa de un evento o



lo que se esté estudiando. De hecho no es diferente a la presentada en la sección dedicada a la divergencia K-L.

Para el caso continuo:

$$I(f, g(\cdot|\theta)) = \int_{\Omega} f(x)\log\left(\frac{f(x)}{g(x|\theta)}\right) dx$$

$$= \int_{\Omega} f(x)\log(f(x))dx - \underbrace{\int_{\Omega} f(x)\log\left(g(x|\theta)\right) dx}_{\text{relativo a la información K-L}}$$

donde f es la realidad completa en términos de la distribución de probabilidad, g es el modelo aproximado en términos de la distribución de probabilidad y θ es el vector de parámetros en el modelo aproximado g.

- $I(f, g) \leq 0$, comúnmente
- $I(f, g) = 0$ si y solo si $f = g$ casi donde sea.

5.2. Motivación del Criterio de Información Akaike(1973)

- f es desconocida
- Los parámetros θ en g deben ser estimados desde los datos empiricos y
- Los datos y son generados desde $f(x)$, es decir, son realizaciones para la variable aleatoria x
- $\hat{\theta}(x)$: estimador de θ. Es una variable aleatoria.

- $I(f, g(\cdot|\hat{\theta}(y)))$ es una variable aleatoria.
- Necesitamos usar la esperanza de la información K-L para medir la divergencia entre g y f.

Obtengamos pues el AIC, buscamos

$$\text{minimizar}_{g\in G} E_y\left(I(f, g(\cdot|\hat{\theta}(y)))\right)$$

Sabemos que

$$E_y\left(I(f, g(\cdot|\hat{\theta}(y)))\right) = \int_\Omega f(x)\log(f(x))dx - \underbrace{\int_\Omega f(y)\left[\int_\Omega f(x)\log\left(g(x|\hat{\theta}(y))\right)dx\right]dy}_{E_y E_x[\log(g(x|\hat{\theta}(y)))}$$

donde, G es la colección de modelos admisibles en términos de funciones de densidad, $\hat{\theta}$ es el estimador de máxima verosimilitud basado en el modelo g y los datos y y y es la muestra aleatoria de la función de densidad $f(x)$

Por ende, el criterio de selección se obtendrá a partir de

$$\text{Maximizar}_{g\in G} E_y E_x[\log(g(x|\hat{\theta}(y)))]$$

Akaike demostró un estimador insesgado $E_y E_x[\log(g(x|\hat{\theta}(y)))]$ para grandes muestras y un **buen** modelo es

$$\log(L(\hat{\theta}(y)) - k$$

donde, L es la función de verosimilitud,$\hat{\theta}$ es el estimador de ML de θ y k número de parámetros estimados. Por buen modelo se debería entender aquel modelo que es cercano a f en el sentido de tener un valor K-L pequeño.

- **Caso de máxima verosimilitud**

$$\text{AIC=-2 log } \underbrace{L(\hat{\theta}|y)}_{\text{sesgo}} + \underbrace{2k}_{\text{varianza}}$$

El mejor modelo será el que tenga el mínimo AIC

- **Caso de mínimos cuadrados**

$$\text{AIC=n log } \left(\frac{RSS}{n}\right) + 2k$$

Suponiendo que los errores se distribuyen normal iid. Donde RSS son los residuales estimados del modelo ajustado.

5.2.1. Otros criterios

- **Criterio de información de Schwarz (o Bayesiano)**

$$\text{SIC=log } \left(\frac{\hat{U}'\hat{U}}{n}\right) + \frac{k}{n}\, log\ n$$

- **Criterio de Hannan-Quinn**

$$\text{HQC=n log } \left(\frac{\hat{U}'\hat{U}}{n}\right) + 2k\ log\ log\ n$$

5.2.2. Ventajas y desventajas

- Valido tanto para modelos anidados como no anidados
- Compara modelos con diferentes distribución de errores
- Evita múltiples pruebas
- No puede ser usado para comparar diferentes conjuntos de datos

6. Referencias

[1] Akaike H.(1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory, P.N. Petrov y F. Csaki. Akad. Kiado, Budapest. pp 267-281

[2] Schützenberger, M.P. (1954). Contribution aux statistiques de la théorie de l'information. Institut de statistique de l'Université de Paris.

[3] Hannan, E. J., and B. G. Quinn (1979) The Determination of the Order of an Autoregression, Journal of the Royal Statistical Society, B, 41, 190-195.

[4] Schwarz, Gideon E. (1978). Estimating the dimension of a model. Annals of Statistics 6 (2): 461-464