APRENDIZAJE AUTOMÁTICO: TERCER CURSO DEL GRADO
DE ING. INFORMÁTICA EN COMPUTACION

Aprendizaje estadístico: Teoría de la Información

César Hervás-Martínez
Grupo de Investigación AYRNA

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2019-2020

INTRODUCCION
TEORIA DE LA INFORMACION ESTADISTICA

1 Cantidad de información

2 Entropía de una variable

3 Divergencia de Kullback–Leibler

4 Cantidad de información mútua

CANTIDAD DE INFORMACIÓN: Ejemplo

Sea una urna con 9 bolas negras y 1 bola blanca. Se efectúan extracciones sin reemplazamiento y sean los sucesos A={sacar una bola blanca} y B={sacar una bola negra} con probabilidades P(A)= 1/10 y P(B) =9/10

Se saca una bola blanca. El suceso A proporciona una alta información, ya que la incertidumbre sobre la siguiente extracción desaparece, puesto que P(B/A)=9/9=1 y P(A/A)=0

Se saca una bola negra. El suceso B proporciona una información pequeña, ya que la incertidumbre acerca de la siguiente extracción se mantiene puesto que P(A/B)=1/9 y P(B/B)= 8/9

Cantidad de información como medida de reducción de la incertidumbre: Ejemplo

Al lanzar un dado si nos dicen que ha salido un numero menor que 2, suceso A, tenemos mas información (reduce mas la incertidumbre) que si nos dicen que ha salido un numero múltiplo de 2, suceso B, puesto que si las probabilidades de los sucesos que salga la puntación i-ésima son equiprobables, esto es, P(Ei)=1/6, entonces

P(E1/A)=1, P(E2/A)=0, P(E3/A)=0, P(E4/A)=0, P(E5/A)=0, P(E6/A)=0;
mientras que

P(E1/B)=0, P(E2/B)=1/3, P(E3/B)=0, P(E4/B)=1/3, P(E5/B)=0, P(E6/B)=1/3

CANTIDAD DE INFORMACIÓN

Definición.- Sea X una variable aleatoria con posibles valores $x_1, \ldots x_n$ y probabilidades asociadas $p(X=x_1), \ldots, p(X=x_n)$, definimos la Cantidad de Información asociada a cada valor de la v.a como

$$I(x_i) = -\log_2 p(x_i)$$

$$Si\ p(x_i) = 0,\quad entonces\ \mathrm{I}(\mathrm{x_i}) \cong \infty$$

$$Si\ p(x_i) = 1/2,\quad entonces\ \mathrm{I}(\mathrm{x_i}) = 1$$

$$Si\ p(x_i) = 1,\quad entonces\ \mathrm{I}(\mathrm{x_i}) = 0$$

Cuanto mas probable es un suceso, menor cantidad de información aporta.

ENTROPÍA DE UNA VARIABLE

Sea X una variable aleatoria con posibles valores $x_1, \dots x_n$ y probabilidades asociadas $p(x_1), \dots, p(x_n)$, definimos

$$I(x_i) = -\log_2 p(x_i)$$

Sea I(X) la variable aleatoria Cantidad de Información asociada a X, con posibles valores $I(x_1), \dots I(x_n)$ y probabilidades asociadas $p(x_1), \dots, p(x_n)$.

Definición.- Se define la Entropía de Shannon (1948), H(X), de una variable aleatoria discreta X como la esperanza matemática de la cantidad de información I(X)

$$H(X) = E(I(X = x_i)) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

Si $p(x_i)=0$, la indeterminación $p(x_i)\log_2 p(x_i)$ se resuelve asignándole el valor 0 a la entropía.

Entropía de una variable aleatoria de Bernouilli

Sea X una variable aleatoria de Bernouilli de parámetro p, B(p). La función de probabilidad es

$$P(X = x) = p^{x}(1 - p)^{1-x} \quad \text{para} \;\; x \in \{0,1\},$$

$$P(X = 1) = p; \;\; \mathrm{P(X= 0)=1\text{-}p}$$

$$H(X) = E(l(X)) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

$$H(X) = -p\log_2 p - (1 - p)\log_2(1 - p)$$

Si p= 0; H(x)= -0log$_2$0-1log$_2$1= 0

Si p = 0,50; H(X) = −0,5 log$_2$ 0,5 − 0,5 log$_2$ 0,5 = 1

Si p = 0,60; H(X) = −0,6 log$_2$ 0,6 − 0,4 log$_2$ 0,4 = 0,97

Si p = 0,90; H(X) = −0,9 log$_2$ 0,9 − 0,1 log$_2$ 0,1 = 0,468

Si p= 1; H(x) = -1log$_2$1-0log$_2$0 = 0

La máxima incertidumbre se obtiene para una probabilidad de éxito de 0,5

Propiedades de la Entropia

Se verifica: 1) $$0 \le H(X) \le \log_2 n$$

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i);\ 0 \le \mathrm{p}(\mathrm{x}_i) \le 1,\ \text{mientras que}$$

$$-\infty < \log_2 p(x_i) \le 0,\quad \text{luego } 0 \le \mathrm{H(X)}$$

2) $$H(X) = 0 \Leftrightarrow \exists\ x_i \ \ con \ \ p(x_i) = 1$$

Esto significa que la variable aleatoria es singular y toma un solo valor con probabilidad 1.

3) Por otra parte Si X es una variable aleatoria uniforme discreta, es decir

$P(X = x_i) = 1/n$ para todo $i = 1, \dots, n$, entonces $H(X) = \log_2 n$

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) = -\sum_{i=1}^{n}\frac{1}{n}\log_2\frac{1}{n} = \sum_{i=1}^{n}\frac{1}{n}\log_2 n = \log_2 n$$

Luego, si una variable aleatoria tiene una distribución uniforme, discreta o continua tiene máxima incertidumbre

ENTROPÍA CONDICIONADA

Sea X una v. a. con valores x1, . . . , xn , y con probabilidades p(x1), . . . , p(xn)

Sea Y una v.a. con valores y1, . . . , ym , y con probabilidades p(y1), . . . , p(ym)

Sea (X, Y) una v.a. bidimensional con (x1, y1), . . . , (x1, ym), . . . , (xn, y1), . . . , (xn, ym) y con probabilidades p(x1, y1), . . . , p(x1, ym), . . . , p(xn, y1), . . . , p(xn, ym)

Sea X|Y = yj una v.a. condicionada con probabilidades p(x1|yj), . . . , p(xn|yj)

Definición.- Entonces la Entropía de la v. a. bidimensional conjunta (X, Y) es:

$$H(X,Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i, y_j)$$

ENTROPÍA CONDICIONADA

Definición.- La entropía de la v.a. X condicionada al valor Y = y_j

$$H(X \mid Y = y_j) = -\sum_{i=1}^{n} p(x_i \mid y_j) \log_2 p(x_i \mid y_j)$$

Definición.- La entropía de la v. a. X condicionada a la v. a. Y

$$H(X \mid Y) = \sum_{j=1}^{m} p(y_j) H(X \mid Y = y_j) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(x_i \mid y_j)$$

ENTROPÍA DE UNA VARIABLE

Proposición.- Se verifica que:

H(X,Y) = H(X) + H(Y|X)

Demostración.-

$$H(X)+H(Y\mid X)=-\sum_{i=1}^{n}p(x_i)\log_2 p(x_i)-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(y_j\mid x_i)$$

Ahora

$$-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(y_j\mid x_i)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2\frac{p(x_i,y_j)}{p(x_i)}$$

$$=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(x_i,y_j)+\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(x_i)$$

pero

$$\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(x_i)=\sum_{i=1}^{n}p(x_i)\log_2 p(x_i)$$

luego

$$H(X)+H(Y\mid X)=-\sum_{i=1}^{n}p(x_i)\log_2 p(x_i)-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(x_i,y_j)$$

$$+\sum_{i=1}^{n}p(x_i)\log_2 p(x_i)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log_2 p(x_i,y_j)=H(X,Y)$$

Proposición.- Si X e Y son variables aleatorias independientes, esto es, si $P(x_i, y_j) = P_1(x_i)\, P_2(y_j)$ entonces:

$$H(X|Y) = H(X)$$

$$H(Y|X) = H(Y)$$

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

DIVERGENCIA DE KULLBACK–LEIBLER

Mide la distancia entre dos distribuciones de probabilidad – Una de las cuales actúa como referencia, por ejemplo p es la probabilidad "a priori" y q la probabilidad "a posteriori"– definidas sobre la misma variable aleatoria X, se denomina también "entropía relativa" o "entropía cruzada".

Se puede interpretar como el incremento de información necesaria para cambiar la distribución "a priori" p en una distribución "a posteriori" q. Para una variable discreta X, se define en la forma.

$$KL(p \| q) = D_{K-L}(p,q) = \sum_{i=1}^{n} q(x_i) \log_2 \frac{q(x_i)}{p(x_i)}$$

Propiedades de la divergencia de KULLBACK–LEIBLER

1) **Suponemos que para cualquier valor x, si p(x)= 0 entonces q(x)= 0 y**
2) **También por convenio 0*$\log_2$(0)= 0.**

Además dado que

$$D_{K-L}(p,q)=\sum_{i=1}^{n} q(x_i)\log_2\frac{q(x_i)}{p(x_i)}$$

3)

$$D_{K-L}(p,q)\geq 0$$

4)

$$D_{K-L}(p,q)=0 \;\Leftrightarrow\; p(x_i)=q(x_i),\ \forall \mathrm{i=1,\ldots,n}$$

CANTIDAD DE INFORMACIÓN MÚTUA

En teoría de la probabilidad, y en teoría de la información, la Información Mutua o trans-información de dos v. a. X, Y es una cantidad que mide la dependencia mutua de las dos variables, es decir, mide la reducción de la incertidumbre (entropía) de una variable aleatoria, X, debido al conocimiento del valor de otra variable aleatoria Y.

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

CANTIDAD DE INFORMACIÓN MUTUA: Ejemplo

Ejercicio.-

Consideremos dos monedas: La A en la cual la probabilidad de cara es 1/2 , y la B en la cual la probabilidad de cara es igual a 1. Se elige una moneda al azar, se lanza dos veces y se anota el numero de caras obtenidas.

Definimos dos variables aleatorias.

X denota la moneda escogida, con valores A con probabilidad ½ y B con probabilidad ½

Y denota el número de caras obtenidas, con valores 0 (sacar cruz en A y cruz en B) 1 (sacar cruz en A y cara en B) y 2 (sacar cara en A y cara en B) y con probabilidades, 0, 1/2 y 1/2

CANTIDAD DE INFORMACIÓN MUTUA: Ejemplo

Las entropías asociadas a cada variable son

$$H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1;$$

$$\mathrm{H(Y)} = -0\log_2(0) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$$

La entropía de la variable X condicionada al valor de Y es

$$H(X \mid Y) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i \mid y_j)$$

La cantidad de información mutua es entonces

I(X, Y) = H(X) − H(X|Y) = 1 − 0,4509 = 0,5491

X denota la moneda escogida, en la A la probabilidad de cara es 1/2 , en la B la probabilidad de cara es 1,
Y denota el número de caras en 2 lanzamientos de la moneda

$P(Y = 2|X = A) = 1/4$; $P(Y = 1|X = A) = 1/2$; $P(Y = 0|X = A) = 1/4$;

$P(Y = 2|X = B) = 1$; $P(Y = 1|X = B) = 0$; $P(Y = 0|X = B) = 0$

$P(X = A, Y = 0) = P(X=A)P(Y=0/X=A)=1/2*1/4=1/8$;

$P(X = B, Y = 0) = 0$;

$P(X = A, Y = 1) = P(X=A)P(Y=1/X=A)= 1/2*1/2=1/4$;

$P(X = B, Y = 1) = 0$;

$P(X = A, Y = 2) = 1/2*1/4=1/8$; $P(X = B, Y = 2) = 1/2*1=1/2$

$P(Y = 0) = P(X=A)P(Y=0/X=A) + P(B)P(Y=0/B)=1/2*1/4+1/2*0=1/8$;

$P(Y = 1) = P(X=A)P(Y=1/X=A) + P(B)P(Y=1/B)=1/2*1/2+0= 1/4$;

$P(Y = 2) = P(X=A)P(Y=2/X=A) + P(B)P(Y=2/B)= 1/2*1/4+1/2*1= 5/8$

$P(X = A|Y = 0) = 1$; $P(X = B|Y = 0) = 0$; $P(X = A|Y = 1) = 1$;

$P(X = B|Y = 1) = 0$; $P(X = A|Y = 2) = 1/8:5/8=1/5$; $P(X = B|Y = 2) = 1/2:5/8=4/5$

$H(X|Y = 2)= -P(X|Y=2)*\log_2 P(X|Y=2)= -(1/5)*\log_2(1/5)-(4/5)*\log_2(4/5)= 0,7215$

$H(X|Y = 2) = 0,7215$; $H(X|Y = 1) = 0$; $H(X|Y = 0) = 0$

$H(X|Y) = P(Y = 0)*H(X|Y = 0) + P(Y = 1)*H(X|Y = 1) + P(Y = 2) *H(X|Y = 2)=0+0+$
$+5/8 * 0,7215 = 0,4509$, luego $I(X, Y) = H(X) - H(X|Y) = 1 - 0,4509 = 0,5491$

CANTIDAD DE INFORMACIÓN MÚTUA

Proposición.-

$$I(X,Y) = H(X) - H(X \mid Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

Demostración.-

$$I(X,Y) = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) + \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i \mid y_j)$$

Pero la segunda sumatoria es

$$\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(y_j \mid x_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 \frac{p(x_i, y_j)}{p(y_j)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(x_i, y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 p(y_j)$$

Luego, podemos definir la Cantidad de Información Mutua como

$$I(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\left[\log_2 p(x_i, y_j) - (\log_2 p(x_i) + \log_2 p(y_j))\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i, y_j)\log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)}$$

CANTIDAD DE INFORMACIÓN MÚTUA

En el caso continuo, esto es, si las variables aleatoria son continuas, reemplazamos la suma con una integral definida doble :

$$I(X;Y) = \iint_{Y,X} p(x,y)\log\left(\frac{p(x,y)}{P(x)\,p(y)}\right)dxdy$$

CANTIDAD DE INFORMACIÓN MÚTUA: Propiedades

Se verifica:

1) I(X,Y) = I(Y,X)

2) I(X,Y) = D_{K-L}(p(x, y), p(x)p(y))

3) $$I(X,Y\mid Z)=\sum_{k=1}^{r} p(z_k) I(X,Y\mid Z=z_k)=$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{r} p(x_i,y_j,z_k)\log_2 \frac{p(x_i,y_j\mid z_k)}{p(x_i\mid z_k)\,p(y_j\mid z_k)}$$

4) I(X,Y|Z) = H(X|Z) + H(Y|Z) − H(X,Y|Z)

I(X,Y|Z) = 0 Sii X e Y son condicionalmente independientes dado Z

Sean X e Y son condicionalmente independientes dado Z

Sii p(x|y, z) = p(x|z) para todo x, y, z

APRENDIZAJE AUTOMATICO: TERCER CURSO DEL GRADO DE ING. INFORMÁTICA EN COMPUTACION

Aprendizaje estadístico: Teoría de la Información

GRACIAS POR SU ATENCIÓN

César Hervás-Martínez
Grupo de Investigación AYRNA

Departamento de Informática y Análisis Numérico
Universidad de Córdoba
Campus de Rabanales. Edificio Einstein.
Email: chervas@uco.es

2019-2020