Muestra mínima

Samuel Rocha, Mauricio Gonzalez, David Lopez

Octubre 2015

Información y Entropía

Cuando se observa un evento y sus posibles resultados, es posible determinar la totalidad de información que nos puede dar dicho evento o podemos determinar que tanta información tiene un libro, una película o una pintura. Aunque parezca que es calcular , es posible determinarlo de manera objetiva y determinar la cantidad de información que contine el objeto de análisis para, después, determinar el valor intrínseco de la información que contiene.

La forma de calcular la información tiene dos posturas, la teroría estadística de la información y la teoría algorítmica de la información. Las dos buscan lo mismo, pero se determina de manera diferente.

Teoría estádistica de la información

Para Shannon una fuente de información es un ente abstracto que produce símbolos de un alfabto finito. Los simbolos que se producen por la funte son una sucesión estacionaria de variables aleatorias, la probabilidad de que sea producida una cierta secuencia de símbolos no cambia con respecto al tiempo. El modelo que propuso Shannon para la fuente es una cadena de Markov ergódica, es decir, hay una distribución de probablidades límite para los símbolos del alfabeto. En una cadena de Markov en un sistema pueden existir ciertos estados (en este caso, la aparición de un símbolo) y existe una cierta probablidad de que pase de un estado a otro(existe, con cierta probabilidad de que , después de un símbolo, aparezca otro). Si la cadena de Markov es ergódica, las probabilidades de pasar de un estado a otro son siempre las mismas.

Un ejemplo es si enviamos siempre mensajes tomados de textos en español, la probabilidad de que aparezca la letra “x” permanece constante, ya uqe la fuente es ergódica, pero si enviamos mensajes en ingles y en español de manera alternada, la probabilidad de de aparición de “x” cambia con el tiempo puesto que la fuente no es ergódica.

Shannon define la cantidad de información en bits, en cada símbolo s_i del alfabeto de una fuente de información S se define como:

$$I(s_i) = -log_2(p_i)$$

donde p_i es la probabilidad de que el símbolo s_i sea producido por S y log_2 es el logaritmo base 2. De esta forma la información de una secuencia de símbolos es igual a la suma de lainformación de cada símbolo, siempre y cuando los símbolos sean estadísticamente independientes.

Visto de esta forma, la información contenida en un símbolo es mayor mientras menos frencuente sea el símbolo. El concepto de información está asociado a incertidumbre. Entre menor sea la probabilidad de encontrar el mensaje, mayor es su información.

Un ejemplo puede ser una canción popular. Normalmente estas canciones consisten en un conjunto de temas que se repiten y normalmente es predecible. A diferencia de una obra de Mozart o Beethoven, no se repiten de la misma manera y son más dificiles de predecir. Las piezas populares, ofrecen menos información que la música barroca o la música clásica.

Después de haber definido la información de un símbolo y un mensaje, toca definir la entropía de la informaciión como el valor esperado de la cantidad de información de los símbolos producidos por una fuente. Laa entropía de la fuente S es:

$$H(S) = \sum_{i=1}^{n} P_i I(s_i) = \sum_{i=1}^{n} P_i log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} P_i log_2 p_i$$

La cantida de información de un símbolo es mayor si el seimbolo es menos probable y la entropía de la fuente es mayor mientras más uniforme sea la distribución de probabilidades de los símbolos de un alfabeto. Para cualquier fuente S con un alfabeto de n símbolos se cumole:

$$H(S) \leq log_2 n$$

En el caso de una fuente S con símbolos con la misma probabilidad, se cumple la ecuación de arriba. L entropía es por lo tanto una medida de incertidumbre o desorden de la fuente. Cuanto más impredecible es el comportamiento de S, mayor será su entropia. Por el contrario, mientras menos sea la entropía, más organizada es S. Dicho de otra forma, si $H(S)$ es pequeña significa que podemos encontrar regularidades en el objeto de estudio.

Teoría algorítmica de la información

Hay algunos puntos en común en la forma en que la Teoría estadística de la información(TEI), pero una de las diferencias es que la TEI supone que se conocen de antemano las probabilidades p_i de los símbolos s_i de los mensajes. Normalmente no sucede así. Las probabilidades se aproximan por las proporciones de cada símbolo en un mensaje dado.

Por ejemplo. la probablidad de aparición de la letra “a” en un texto escrito en español puede aproximarse contando la cantidad de veces que dicha letra aparece en el mensaje y dividiendo esto entre el número total de caracteres del mensaje, pero ésto puede variar dependiendo de donde tomemos el total de caracteres. La probabilidad de ocurriencia de una letra no necesariamente es igual que en otro capítulo.

Otra diferencia, es que en ésta teoría es posible suponer que las letras del alfabeto son bits o parejas de bits. Esto permite elegir la longitud de secuencias de bits para usarlas como símbolos, siempre y cuando los bits de los símbolos elegidos sean estadísticamente independientes. Normalmente esto no sucede y en la aplicación de TEI se tienen que hacer aproximaciones.

Proceso para reducción en una base de datos

Si los datos se analizan como una secuancia de símbolos que componen mensajes, cada vector de atributos representa un mensaje y cada valor del atributo corresponde a un símbolo. De esta forma, es posible estimar la entropía contenida en el mensaje. Con esto se puede verificar que la información de una muetra de la población contiene la misma información.

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

$$H(X) = -\sum_{i=1}^{m}(\sum_{j=1}^{n} \frac{\delta(S_i, v_i)}{n})log(\sum_{i=1}^{m} \frac{\delta(S_i, v_i)}{n})$$

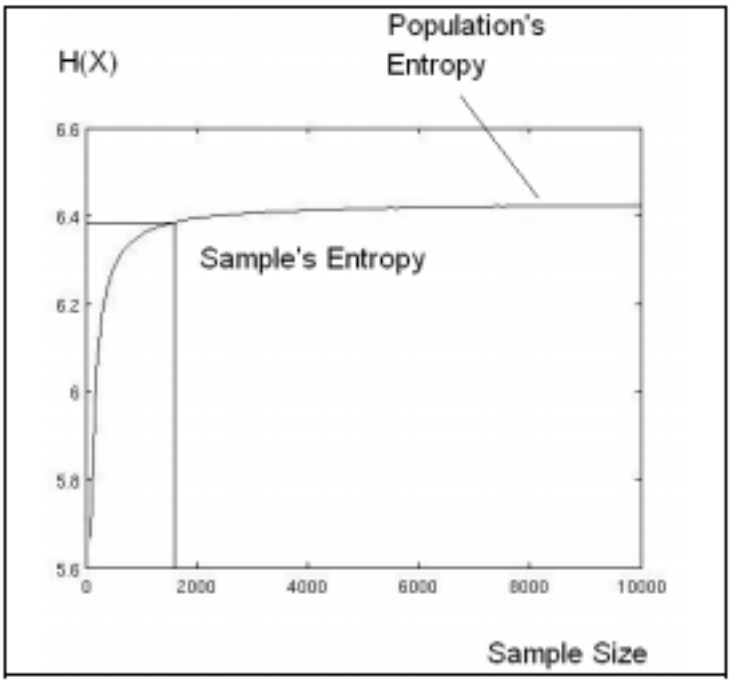
Donde X es el mensaje, p_i es la probabilidad de ocurrencia del símbolo i, m es el número de elementos en los datos, Si es el símbolo i-ésimo, v_j es el valor j-ésimo de los datos y $\delta(S_i, v_i)$ es 0 si v es igual a s, o es 1 si v=s.

El objetivo es aproximar el valor de entropía poblacional por la entropia de una muestra seleccionada de tal manera que tenga la misma entropía de la población para evitar utilizar toda la base de datos. El metodo consiste en tratar cada atributo $t $ para obtener una muestra M_t como sigue:

Inicialmente M_t está vacio. Luego se procesde a extraer elementos de manera aleatoria de la población para el atributo t de manera iterativa y sumamos esos elementos a M_t. En cada iteración de i, la entropía se calcula y se compara con el la calculada anteriormente como sigue:

$$\delta(H_i) = H(i) - H(i-1)$$

Mientras $\delta(Hi)$ se acerca al parámetro ϵ de la entropia de M_t, M_t se acerca asintoticamente a la entropía de la población.En éste punto, se puede saber el tamaño de M_t para determinar el tamaño de la muestra que reprente aproximadamente la entropía de la población total.



Este proceso se hace para cada atributo. Cuando el valor M_t final se haya calculado, se toma la muestra de M como la más grande de las calculadas para mantener la representatividad de todos los atributos.

Para validar la muestra, se debe asegurar que las entropias tienen la misma o mayor entropia que la calculada anteriormente.

Distribuciones Multimodales

Una distribución multimodal, es una distribución de distribución con más de un pico o moda que pueden ser llamados máximos locales o globales. Si tiene dos modas es bimodal; si tiene tres modas, es trimodal según la clasificación establecida por Galtung.

Aunque hay muchas sugerencias, no hay un común acuerdo en los estadísticos para cuantificar los parámetros de una distribución general bimodal. Un estadístico es el establecido por Ashman:

$$D = \frac{2^{1/2}|\mu_1 - \mu_2|}{(\sigma_1^2 + \sigma_2^2)^{1/2}}$$

donde μ_1, μ_2 son las medias y σ_1, σ_2 son las desviaciones estandar. Para una mezcla de dos normales se requiere una D>2 para decir con exactitud que se cuentan con dos distribuciones separadas.

Se pueden determinar coeficientes para analizar la ampitud, el parametro de bimodalidad e indices de bimodalidad para enriquecer el estudio de las distribuciones.

Una mezcla de dos distribuciones unimodales con diferentes medias no necesariamente es bimodal. Normalmente se usa el ejemplo de usar la altura de hombres y mujeres para mostrar la distribución bimodal, pero la diferencia en la media y su desviación estandar es pequeña como para decir que son bimodales. Esto se debe a que a que una mezcla de distribuciones normales se considera bimodal si y solo si la diferencia de las medias es de las distribuciones es mayor a que la suma de sus desviaciones estandar.

Algoritmo de Ascenso

El propósito de este algoritmo es expresar el comportamiento de una variable dependiente y como función de un conjunto de variables independientes $v_1, ..., v_n$. Escogemos un aproximante de la forma:

$$y = c_1 X_1 + ... + c_m X_m.$$

Donde las X_i son funciones de las variables dependientes; es decir, $X_i = f_i(v_1, ..., v_n)$. Como en muchos otros algoritmos, buscamos escoger los coeficientes c de modo que se minimice el error de aproximación en algún sentido. El Algoritmo de Ascenso busca minimizar la norma minimax; es decir, si para cada observación y_i tomamos el error de aproximación $\varepsilon_i = |f_i - y_i|$ con esto, definimos $\varepsilon = max\{\varepsilon_1, ..., \varepsilon_n\}$. Es este error el que se busca minimizar en el Algoritmo de Ascenso.
El Algoritmo tiene 2 etapas:
Primero, se toma un subconjunto de tamaño M de los datos originales; a este conjunto le llamaremos el conjunto interior. Para este conjunto, se obtienen los coeficientes del mejor aproximante en el sentido minimax. Calculamos , para el conjunto interior, el máximo de los errores y lo llamaremos ε_θ.

Luego, con este aproximante se calculan los valores estimados para el resto de las observaciones; es decir, se calculan los valores ajustados para las observaciones en el complemento del conjunto interior (llamado a su vez conjunto exterior).
Ahora, como lo que buscamos es minimizar el máximo de los errores sobre todo el conjunto de observaciones, calculamos los errores para los valores predichos en el conjunto exterior y con estos calculamos el máximo de dichos errores, lo llamamos ε_φ. Si resulta que $\varepsilon_\theta \geq \varepsilon_\varphi$, entonces habremos terminado, ya que encontramos el mínimo dentro del conjunto interior. De no ser este el caso, entonces necesitamos intercambiar un objeto

del conjunto interior por uno del conjunto exterior; esto debe realizarse de manera cuidadosa y será lo que discutiremos a continuación.

Fundamentos Matemáticos

Para un aproximante $y = c_1X_1 + ... + c_mX_m$, tenemos los errores $\varepsilon_i = |f_i - y_i|$ y $\varepsilon_\theta = \max(\varepsilon_1, ..., \varepsilon_M)$. Entonces, es claro que

$$\varepsilon_i = |f_i - (c_1X_{i1} + ... + c_mX_{im})|$$

Podemos tomar el valor positivo de ε_i y hacer

$$\varepsilon_i + c_1X_{i1} + ... + c_mX_{im} = f_i.$$

Sean $s_1, ..., s_M$ constantes tales que $\varepsilon_i = s_i\varepsilon_\theta$. Tenemos entonces que,

$$\begin{aligned} s_1\varepsilon_\theta + c_1X_{11} + c_2X_{12} + ... + c_mX_{1m} &= f_1 \\ s_2\varepsilon_\theta + c_2X_{21} + c_2X_{22} + ... + c_mX_{2m} &= f_2 \\ \vdots &= \vdots \\ s_M\varepsilon_\theta + c_1X_{M1} + c_2X_{M2} + ... + c_mX_{MM} &= f_M. \end{aligned}$$

Luego, podemos aplicar la Regla de Cramer para obtener ε_θ:

$$\varepsilon_\theta = \frac{\begin{bmatrix} f_1 & X_{11} & \dots & X_{1M} \\ \cdots & \cdots & \cdots & \cdots \\ f_M & X_{M1} & \cdots & X_{Mm} \end{bmatrix}}{s_1 \begin{bmatrix} X_{21} & \cdots & X_{2M} \\ \cdots & \cdots & \cdots \\ X_{M1} & \cdots & X_{Mm} \end{bmatrix} - ... + s_M \begin{bmatrix} X_{21} & \cdots & X_{2M} \\ \cdots & \cdots & \cdots \\ X_{M1} & \cdots & X_{Mm} \end{bmatrix}}$$

Utilizamos la notación X_{i*} para denotar que el i-ésimo renglón y la primer columna del denominador anterior fueron eliminadas; de este modo, tenemos que

$$\varepsilon_\theta = \frac{\Delta}{s_1|X_{1*}| - s_2|X_{2*}| + s_3|X_{3*}| - ...}$$

O, de manera equivalente,

$$\varepsilon_\theta = \frac{\Delta}{s_1\Delta_{1*} - s_2\Delta_{2*} + s_3\Delta_{3*} - ...}$$

Donde Δ es,

$$\begin{bmatrix} f_1 & X_{11} & \dots & X_{1M} \\ \cdots & \cdots & \cdots & \cdots \\ f_M & X_{M1} & \cdots & X_{Mm} \end{bmatrix}$$

De manera más compacta tenemos que,

$$\varepsilon_\theta = \frac{\Delta}{\sum_{i=1}^{M} s_i(-1)^{i-1}\Delta_{i*}}.$$

Entonces, si lo que queremos es minimizar ε_θ, tenemos que maximizar el denominador de la expresión anterior; para lograr esto, todos los productos contenidos en esta expresión deben ser maximizados y todos los sumandos deben tenerel mismo signo.
Para que los $s_i\Delta_i$ sean maximizados, debemos tomar el máximo valor de las s_i y además lograr que $s_i = sgn((-1)^{i-1}\Delta_{i*})$.
La primera de estas condiciones se logra si y sólo si $s_i = 1$ para toda i, y para lograr la segunda condición debemos encontrar los signos de los Δ_{i*} y para esto recurrimos al Teorema de los Cofactores, el cual dice lo

siguiente: Si los elementos de una columna de un determinante se multiplican por el cofactor de una columna diferente y se suman, el resultado es cero.
Finalmente, para encontrar los coeficientes minimax para el conjunto interior, debemos resolver el siguiente sistema:

$$\begin{bmatrix} s_1 & X_{11} & X_{12} & \cdots & X_{1m} \\ s_2 & X_{21} & X_{22} & \cdots & X_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ s_M & X_{M1} & X_{M2} & \cdots & X_{MM} \end{bmatrix} \begin{bmatrix} \varepsilon_\theta \\ c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_m \end{bmatrix}.$$

Hasta ahora, sólo hemos encontrado los coeficientes minimax para el conjunto interior, y ya sabemos que una vez encontrados estos, hay que calcular los valores estimados para el conjunto exterior y ver cuál de los errores es mayor. En el caso en que el error máximo para el conjunto exterior sea mayor, hay que intercambiar un elemento del conjunto exterior por uno del conjunto interior y volver a calcular. La pregunta es cómo escoger estos elementos a intercambiar.
Denotemos como A a la matriz que incorpora la columna de las s_i, entonces por el Teorema del Cofactor tenemos que,

$$\sum_{i=1}^{M} K_i A_i = 0.$$

Luego, Si denotamos a la j -ésimo renglón de A como A^j, entonces

$$A^j = \sum_{i=1, i\neq j}^{M} -\frac{K_i}{K_j} A^i.$$

Lo que propone el algoritmo es escoger el índice I_E del elemento del conjunto externo en el cual se alcanza el error máximo y entonces tratar de expresar el vector A^{I_E} como combinación lineal de los vectores del conjunto interno; es decir

$$A^{I_E} = \sum_{i=1}^{M} \lambda_i A^i,$$

lo que es igual a,

$$A^{I_E} - \sum_{i=1}^{M} \lambda_i A^i = 0.$$

Entonces,

$$A^{I_E} - \lambda_j A^j - \sum_{i=1, i\neq j}^{M} \lambda_i A^i = 0.$$

Utilizamos el Teorema del Cofactor para ver que,

$$A^{I_E} + \left(\sum_{i=1, i\neq j}^{M} \frac{\lambda_j K_i}{K_j} - \lambda_i \right) A^i = 0.$$

Entonces, buscamos seleccionar j tal que

$$\frac{\lambda_j K_i}{K_j} - \lambda_i \geq 0.$$

Luego,

$$\lambda = A^{I_E} B.$$

Por otro lado, calculamos el vector en el cual se maximiza y sea I_I su índice

$$\sigma_{I_E} \frac{\lambda_j}{A_j^{-1}}$$

Finalmente, intercambiamos los vectores I_E y I_I y se repiten todos los pasos anteriores hasta que se cumpla la condición de paro. Este resultado es importante porque en las siguientes instancias de B se pueden obtener en un orden $O(M^2)$ ciclos de reloj. Por lo que el algoritmo de ascenso se ajusta bien a resolver ecuaciones de la forma:

$$y = c_1 X_1 + ... + c_m X_m.$$

Algoritmo de Levenberg-Marquardt

El algoritmo de Levenberg-Marquardt sirve para ajustar modelos no lineales de regresión con pérdida cuadrática. Es decir, resuelve problemas de optimización del tipo:

$$\underset{\theta}{\text{minimizar}} \quad g(\theta) \stackrel{def}{=} \sum_{i=1}^{N} (y_i - f(x_i, \theta))^2,$$

donde $x_i \in \mathbb{R}^p$ es el vector de variables de regresión de la observación i, $y_i \in \mathbb{R}$ es la variable de respuesta asociada a x_i, $\theta \in \mathbb{R}^t$ es el parámetro desconocido a encontrar y f es una función con la que modelamos y_i a partir de x_i. Cuando f es lineal, el problema anterior se convierte en Mínimos Cuadrados Ordinarios (lineales), es decir, en regresión lineal. Antes de proseguir, notemos que lo anterior se puede reescribir en forma matricial haciendo:

$$X = (x_1, \ldots, x_N)^T \in \mathbb{R}^{Nxp}$$
$$Y = (y_1, \ldots, y_N)^T \in \mathbb{R}^{Nx1}$$
$$f(X, \theta) = (f(x_1, \theta), \ldots, f(x_N, \theta))^T \in \mathbb{R}^{Nx1}$$
$$g(\theta) = (Y - f(X, \theta))^T (Y - f(X, \theta))$$

Dado que en general no se puede despejar θ de la ecuación $\nabla g(\theta) = 0$ (condiciones necesarias de primer orden), se procede iterativamente. Comenzando con una aproximación inicial θ, queremos obtener una mejor, $\theta_+ = \theta + \delta$. Si lográramos resolver el problema

$$\underset{\delta}{\text{minimizar}} \quad g(\theta + \delta) = (Y - f(X, \theta + \delta))^T (Y - f(X, \theta + \delta)),$$

de hecho tendríamos la solución al problema original haciendo $\theta^* = \theta_+ = \theta + \delta^*$. Sin embargo, resolver el problema modificado es igual de difícil que el original. Una alternativa es aproximar la solución del problema anterior sustituyendo $f(x_i, \theta + \delta)$ por su aproximación de Taylor de primer orden, $f(x_i, \theta) + J\delta$, donde

$$J \stackrel{def}{=} Jf(X, \theta) = [\nabla f(x_1, \theta), \ldots, \nabla f(x_N, \theta)] \in \mathbb{R}^{Nxt},$$

es decir, queremos encontrar θ_+ resolviendo

$$\underset{\delta}{\text{minimizar}} \quad h(\theta + \delta) \stackrel{def}{=} (Y - f(X, \theta) + J\delta)^T (Y - f(X, \theta) + J\delta)$$

La diferencia entre este problema y el original es que es lineal en los parámetros a optimizar (i.e. en δ), de modo que podemos resolverlo notando que

$$Jh(\delta) = 2(J^T (Y - f(X, \theta) - J^T J\delta),$$

y por lo tanto, para encontrar δ basta igualar lo anterior a cero, o equivalentemente, resolver el siguiente problema lineal:

$$(J^T J)\delta = J^T (Y - f(X, \theta))$$

Utilizar repetidamente lo anterior para encontrar mejores y mejores soluciones es conocido como el algoritmo de Gauss-Newton.

Un problema que tiene el algoritmo de Gauss-Newton es que cuando la matriz $J^T J$ está mal condicionada o es singular, puede haber problemas numéricos a la hora de resolver para δ y la solución podría no ser única o inestable de calcular numéricamente. La contribución de Levenberg consiste en sumarle un múltiplo de la matriz indentidad para garantizar que la matriz siempre sea estrictamente positiva definida:

$$(J^T J + \lambda I)\delta = J^T(Y - f(X, \theta))$$

La elección del parámetro λ se hace de manera heurística pero esencialmente lo que se busca hacer es que cuando $J^T J$ esté bien condicionada, λ sea pequeño, mientras que cuando sea singular o esté mal condicionada, λ sea suficientemente grande para asegurar que el sistema lineal tenga una única solución.

El algoritmo de Levenberg-Marquardt se debe a la contribución final hecha por Marquardt, que propuso utilizar una matriz con las mismas entradas de $J^T J$ en la diagonal y cero en las demás:

$$(J^T J + \lambda \operatorname{diag}(J^T J))\delta = J^T(Y - f(X, \theta))$$

La idea detrás de esto es tomar en cuenta la escala de los parámetros, de modo que la curvatura es afectada más fuertemente en las variables en las que más se necesita. La elección de λ nuevamente es heurística.

->

Referencias

<-

Kuri-Morales, A., Galaviz-Casas, J., “Algoritmos Genéticos”.
Kuri-Morales, A., “Unbiased Generation of Polynomial Functions”.
Kuri-Morales, A., “Polynomial Multivariate Approximation with Genetic Algorithms”. Haykin S & Network, N. “A comprehensive foundetion. Neural Networks,2” Mark F Schilling, Is Human Height bimodal?