TEORIA MATEMATICA DE LA INFORMACION.

Introducción.

La teoría matemática de la información estudia las relaciones entre un emisor y un receptor que se transmiten información a través de un canal. Para ello se intenta determinar la cantidad de información que se transmite por el mismo y, a diferencia de la criptografía, pretende que la recepción del mensaje sea lo más parecido al mensaje enviado, no pretende la ocultación del significado del mismo. Son sin embargo disciplinas complementarias, ya que el mensaje, si bien cifrado, puede estar expuesto al transmitirse por un canal público a perturbaciones. La teoría matemática de la información estudia los métodos para conseguir que el mensaje enviado sea igual al recibido, es decir, se encarga de estudiar la integridad del mensaje cuando este circula por un canal con ruido. La criptografía por su parte estudia los métodos para determinar la seguridad y confidencialidad del mismo.

Codificación y cifrado.

Tanto la codificación como el cifrado se basan en utilizar una aplicación inyectiva del conjunto de símbolos *M* del alfabeto inicial a un conjunto de palabras *P* que se pueden generar con el alfabeto *C,* siendo el código la imagen de *f*. Matemáticamente:

$$M \xrightarrow{f} P(C)$$

Se define un código extendido como aquel formado por la concatenación de las imágenes de *f* de los elementos de *M*. Es decir, la aplicación de *f* a cada uno de los símbolos de una palabra del alfabeto *M*. Si denominamos f^* a la aplicación extendida tenemos que $P(M) \xrightarrow{f^*} P(C)$ con lo que $f^*(a_1,...,a_k) = f^*(a_1)..f^*(a_k)$. Evidentemente la aplicación f^* debe ser a su vez inyectiva para tener una decodificación única.

Los códigos pueden ser de dos tipos, de palabras de longitud fija, denominados códigos en bloque, y de palabras de longitud variable. En el primer caso los códigos tienen una decodificación única, siendo en el segundo caso cierto solo cuando el código es instantáneo. Se define un código instantáneo como aquel en el que ninguna palabra del código es prefijo de otra.

En todo el resto del texto suponemos la utilización de codificación binaria. Las medidas serán pues en bits, aunque los resultados puedan aplicarse sin más problemas a otras medidas de información, únicamente cambiando la base de los algoritmos en función del número de símbolos del lenguaje.

Grado de Incertidumbre.

Se define el grado de incertidumbre de un suceso *s* a $I(s) = \log_n \frac{1}{s} = -\log_n s$. Este concepto nos indica la cantidad de información que hemos recibido a priori sobre el suceso *s*. Si solo existen dos estados equiprobables tenemos que $n = 2$ y decimos que el grado de incertidumbre se mide en bits. De aquí en adelante, y siempre que no se indique

explícitamente, se considerará que trabajamos con dos estados con lo que tendremos que para cualquier valor de *x* escribir $\log_2 x$ es lo mismo que si escribir $\log x$.

Entropía.

Se define como la esperanza matemática de la cantidad de información de una variable aleatoria. Más formalmente, sea *X* una variable aleatoria que toma sus valores en un conjunto $S = \{s_1, s_2, .., s_n\}$, con una función de distribución de probabilidades p_i. Se define la entropía de la variable *X* como la esperanza matemática de la incertidumbre de dicha variable con respecto al conjunto S, y se la denomina como *H(X)*.

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

Por convención asumimos que el valor de $p_i . \log p_i = 0$ si $p_i = 0$.

La entropía puede verse como el número promedio de bits (siempre que trabajemos con logaritmos base 2) que necesitaremos idealmente para representar todos los posibles sucesos de un conjunto. Por templo, supongamos que queremos codificar si un número es par o no (asumiremos que el cero es par para nuestra codificación), Ese será nuestra variable *X*, En este caso sabemos que $p_{par} = \frac{1}{2}$ y que $p_{impar} = \frac{1}{2}$ con lo que tenemos que $H(X) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = -(-\frac{1}{2} - \frac{1}{2}) = 1$. Con lo que sabemos que con un solo bit podremos codificar esta variable.

Lema de Gibbs.

Sean $p_1, p_2, .., p_n$ un conjunto de probabilidades y $q_1, q_2, .., q_n$ una secuencia de valores tal que $\sum_{i=1}^{n} q_i = 1$, se cumple siempre que $\sum_{i=1}^{n} p_i \log \frac{1}{p_i} \le \sum_{i=1}^{n} p_i \log \frac{1}{q_i}$ cumpliéndose la igualdad solo en el caso de que $p_i = q_i \ \forall i \in [1, n]$.

Propiedades de la entropía.

La entropía cumple las siguientes propiedades:

1) $H(X) \ge 0 \forall X$ cumpliéndose la igualdad solo en el caso de que la probabilidad de uno de los sucesos sea 1.
2) El valor máximo de la entropía de un suceso se alcanza cuando los valores del mismo son equiprobables.
3) $0 \le H(X) \le \log|n|$, siendo *n* la cardinalidad del conjunto *X*.

Entropía condicional.

Sean $p(a_i) = p(X = a_i)$, $p(b_j) = p(X = b_j)$ y $p(a_i, b_j) = p(X = a_i \text{ y } Y = b_j)$. Definimos $H(X,Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(a_i, b_j)\log\frac{1}{p(a_i, b_j)}$. La entropía de una variable X condicionada por la variable Y como $H(X/Y) = \sum_{j=1}^{m}\sum_{i=1}^{n} p(a_i, b_j)\log\frac{1}{p(a_i/b_j)}$, de la misma manera tenemos que $H(Y/X) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(a_i, b_j)\log\frac{1}{p(b_j/a_i)}$.

Propiedades de la entropía condicional.

Se cumplen las siguientes propiedades:

1. $H(X,Y) = H(X) + H(Y/X) = H(Y) + H(X/Y)$.
2. $0 \le H(X/Y) \le H(X)$.
3. $H(X,Y) \le H(X) + H(Y)$, cumpliéndose la igualdad solo cuando X e Y son variables independientes.
4. $H(X/Y) \le H(X)$ y $H(Y/X) \le H(Y)$.

Aplicación de la teoría matemática de la información a la criptografía.

En 1949 Shannon publicó una aproximación teórica a la criptografía basándose en sus trabajos sobre teoría de la información. Definimos el *ratio de un lenguaje, r,* para mensajes de longitud N como $r = \frac{H(X)}{N}$. El valor de r mide la cantidad media de bits de información por cada carácter. El *ratio absoluto* de un lenguaje se define como el máximo número de bits de información necesarios para codificar todos los caracteres partiendo del hecho de que todos los caracteres son equiprobables. Definimos el *ratio absoluto* como $R = \log_2 L$, siendo L el número de caracteres del alfabeto. Si suponemos el español de 27 letras con las siguientes probabilidades (sobre 100) obtenidas de un texto de 21000 caracteres [SOL16]:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	Ñ	O	P	Q	R	S	T	U	V	W	X	Y	Z
11,48	1,07	5,19	5,79	13,41	0,91	0,80	0,48	7,12	0,28	0,01	5,43	3,07	7,17	0,08	9,40	2,99	0,85	6,41	7,48	4,88	3,5	0,68	0,17	0,2	0,78	0,38

Obtenemos:

$$R = \log_2 27 = 4{,}7548875$$

$$H(X) = E(I(X)) = \sum_{i=1}^{n} p_i.\log_2\frac{1}{p_i} = -\sum_{i=1}^{27} p_i.\log_2 p_i = 4{,}004961321$$

Ocurre en todos los lenguajes que unos caracteres son de aparición más frecuente que otros, y ciertas combinaciones de caracteres son más comunes que otras y en algunos casos no se presentan estas combinaciones en el lenguaje. Esta redundancia por carácter, que definiremos como *redundancia absoluta,* se define como $D = R - r$.

Sea *K* la clave y *C* el mensaje cifrado, se define la *distancia de unicidad* como el mínimo número de caracteres cifrados tales que $H(K/C) \approx 0$. Un sistema se considera incondicionalmente seguro si $H(K/C)$ no se aproxima a cero en ningún caso.

La formula que nos da la distancia de unicidad *N* se obtiene de la siguiente manera. Sea *R* la ratio absoluta del lenguaje, existen pues 2^{RN} posibles mensajes de longitud *N*, existen pues 2^{rN} mensajes con significado y $2^{RN} - 2^{rN}$ mensajes sin significado. Suponiendo que todos los mensajes con sentido son equiprobables tenemos que $p = 2^{-rN}$, si consideramos también que el cifrado es aleatorio tenemos una probabilidad de obtener un mensaje con significado de $q = \frac{2^{rN}}{2^{RN}} = 2^{(r-R)N} = 2^{-DN}$. Por otra parte dado un texto cifrado tenemos una clave particular *k* que nos dará la solución correcta y $2^{H(K)} - 1$ claves que nos darán un resultado sin sentido. Tenemos pues que el número de soluciones falsas es de $\left[2^{H(K)} - 1\right]q = \left[2^{H(K)} - 1\right]2^{-DN} \approx 2^{H(K)-DN}$. Calculando el logaritmo de la expresión podemos considerar que una solución será suficientemente correcta si $H(K) - DN = 0$. Tenemos pues que el número mínimo de caracteres cifrados para que el cifrado sea teóricamente forzable es de $N = \frac{H(K)}{D}$. Solo conocemos el valor de *D* para el inglés. En este caso el valor es de 3,2. No sabemos de ningún texto en el que se haya calculado el resultado para el español, aunque eso no implica que no haya sido calculado.

Ciertamente es un resultado teórico interesante. Shannon suponía que el atacante disponía de una cantidad ilimitada de recursos, cosa que en la vida real no sucede. Veamos un par de aplicaciones de lo que hemos visto hasta ahora.

Consideremos el DES[1]. Como todos sabemos es un algoritmo de cifrado de 64 bits que utiliza una clave de 56 bits. Si consideramos que el mensaje está en inglés tendríamos que:

$$N = \frac{H(K)}{D} = \frac{56}{3{,}2} = 17{,}5,$$

Una aplicación más sencilla, y probablemente más útil, aunque no mucho más, de la entropía es determinar si un mensaje cifrado lo ha sido con una trasposición o una sustitución. Supongamos el siguiente mensaje: ESTAMOS MIRANDO UNA PRUEBA DE LA ENTROPIA.

Si calculamos la entropía del mensaje utilizando las probabilidades de cada una de las letras en español, ya que sabemos que éste es el idioma utilizado, obtendríamos:

Letra	E	S	T	A	M	O	I	R	N	D	U	P	B	L
Frecuencia	4	2	2	6	2	3	2	3	3	2	2	2	1	1
Probabil.	0,134	0,075	0,049	0,115	0,031	0,094	0,071	0,064	0,072	0,058	0,035	0,030	0,011	0,054

La entropía en este caso sería de 3,608.

Cifrando el mensaje con una tabla de trasposición obtenemos:

E	S	T	A	M
O	S	M	I	R
A	N	D	O	U
N	A	P	R	U
E	B	A	D	E
L	A	E	N	T

[1] Ejemplo obtenido de [DEN99]

R	O	P	I	A

El mensaje cifrado sería:

EOANELR SSNABAO TMDPAEP AIORDNI MRUUETA

Hemos separado el mensaje en columnas para una mejor comprensión. Evidentemente en un cifrado real no habría esa separación. Si calculamos ahora la entropía del mensaje utilizando los valores que hemos utilizado anteriormente, lógicamente esa entropía coincidirá con la del lenguaje dado que no ha habido alteración de los símbolos, sino simplemente un intercambio de posiciones. Por ejemplo, la A sigue siendo la A y aparece tantas veces como en el mensaje original, aunque en diferentes posiciones.

Ahora cifraremos el mensaje con una sustitución sencilla, simplemente adelantaremos una posición en el orden normal de cada letra, con lo que el mensaje cifrado quedaría como:

E	S	T	A	M	O	S	M	I	R	A	N	D	O	U	N	A	P	R	U	E	B	A	D	E	L	A	E	N	T	R	O	P	I	A
F	T	U	B	N	P	T	N	J	S	B	Ñ	E	P	V	Ñ	B	Q	S	V	F	C	B	E	F	M	B	F	Ñ	U	S	P	Q	J	B

La nueva tabla de frecuencias sería:

Letra	F	T	U	B	N	P	J	S	Ñ	E	V	Q	C	E	M
Frecuencia	4	2	2	6	2	3	2	3	3	1	1	2	1	2	1
Probabilidad	0,009	0,049	0,035	0,011	0,072	0,030	0,003	0,075	0,001	0,134	0,007	0,009	0,052	0,031	0,009

Calculando la nueva entropía tenemos que ésta es menor, 2,121.

Si lo pensamos fríamente éste será el caso general, ya que al cambiar las letras de mayor frecuencia de aparición por otras con menor frecuencia, el valor de la entropía también decrecerá. Vemos también que en el caso de la trasposición este valor es menor (3,608) que la entropía del lenguaje (4,005), aunque más parecido que en la sustitución. Esto ocurre por que el mensaje es corto, Para aplicar esta técnica el mensaje cifrado debería ser de una longitud considerable, y, viendo una simple tabla de frecuencias y comparándola con la frecuencia de aparición de las letras del lenguaje podríamos verlo mucho más rápidamente.