Proyecto CompuStat: Criptografía

Marcos Olguín Martínez

6 de diciembre de 2015

Contents

1. Introducción

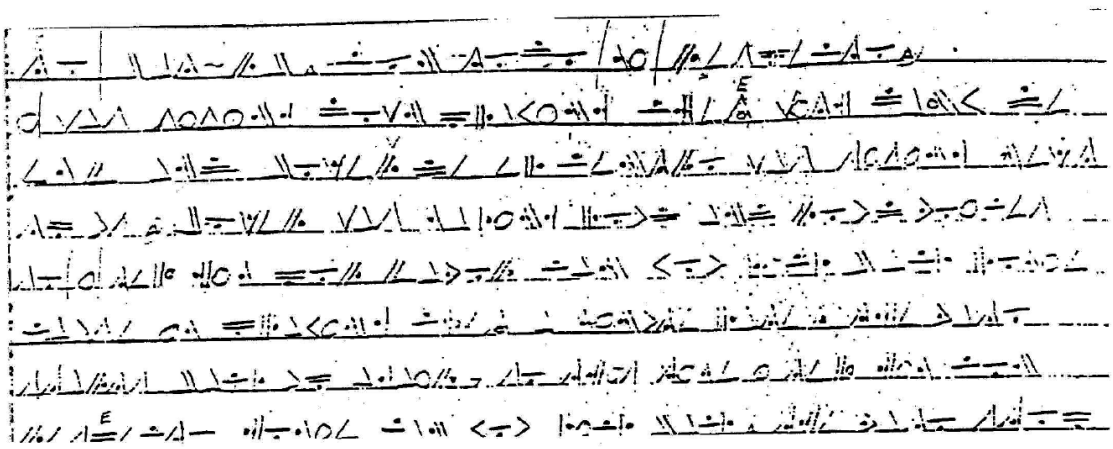
Definición de Criptografía en Wikipedia: *Tradicionalmente se ha definido como el ámbito de la criptología el que se ocupa de las técnicas de cifrado o codificado destinadas a alterar las representaciones lingüísticas de ciertos mensajes con el fin de hacerlos ininteligibles a receptores no autorizados. Estas técnicas se utilizan tanto en el Arte como en la Ciencia. Por tanto, el único objetivo de la criptografía era conseguir la confidencialidad de los mensajes. Para ello se diseñaban sistemas de cifrado y códigos. En esos tiempos la única criptografía existente era la llamada* ***criptografía clásica****.*

La aparición de la Informática y el uso masivo de las comunicaciones digitales, han producido un número creciente de problemas de seguridad. Las transacciones que se realizan a través de la red pueden ser interceptadas, y por tanto, la seguridad de esta información debe garantizarse. Este desafío ha generalizado los objetivos de la criptografía para ser la parte de la criptología que se encarga del estudio de los algoritmos, protocolos (se les llama protocolos criptográficos), y sistemas que se utilizan para proteger la información y dotar de seguridad a las comunicaciones y a las entidades que se comunican.

En el presente trabajo abordaremos la criptografía mediante técnicas de simulación con MCMC como parte del proyecto final de la materia Estadística Computacional.

Muchos problemas científicos básicos son resueltos mediante la simulación. Promedios calculados a partir de la caminata dan respuestas útiles a problemas formalmente intratables, hay muchos ejemplos de ello, particularmente nos enfocaremos en uno que sucedió en una prisión estatal de EEUU.

El departamento de estadística de Stanford tiene un servicio de consultoría. Un día, un psicólogo de una prisión estatal de EEUU les llevó un mensaje codificado, aquí mostramos parte de ese mensaje:



Se sugiere que el código fue una simple sustitución de cifras, donde cada simbolo representa una letra, número, marca de puntuación o espacio. Así, hay una función desconocida *f*.

$$f : \{\text{code space}\} \longrightarrow \{\text{usual alphabet}\}$$

Una aproximación estándar para descifrar es usar la estadística de la escritura Inglesa para adivinar las probables elecciones de *f*, probarlos, y ver si los mensajes tienen sentido.

Para obtener las estadísticas, descargaremos un texto grande que venga en Ingles (por ejemplo, War and Peace) y registrar la transición de primer orden: la proporción de simbolos de texto consecutivos desde x a y. Esto da una matriz $M(x, y)$ de trancisión. Uno entonces podría asociar la plausibilidad de f via:

$$Pl(f) = \prod_i M(f(s_i), f(s_{i+1}))$$

donde s_i corre sobre simbolos consecutivos en el mensaje codificado. Funciones f los cuales tienen altos valores de Pl(f) son buenos candidatos para descifrar. Maximizar las *f's* fue alcanzado mediante el siguiente algoritmo de Monte Carlo:

- Iniciar con una propuesta preliminar, diremos *f*.
- Calcular Pl(f).
- Cambiar a f^*; haciendo una transposición aleatoria de los valores de f asignados a los dos símbolos.
- Calcular Pl(f^*); si este es más grande que Pl(f) aceptamos f^*
- Si no lo es, lanzamos una moneda Pl(f^*)/Pl(f); si sale cara, aceptamos f^*.
- Si sale cruz, nos mantenemos con *f*.

El algoritmo continúa, intentando mejorar la actual f mediante transposiciones aleatórias. Los lanzamientos de la moneda le permiten ir a las fs menos probables, y evita que se atasque en máximos locales.

2. Cadenas de Markov

Sea χ un conjunto finito. Una *cadena de Markov* es definida como una matriz $K(x, y)$ con $K(x, y) \geq 0$, $\sum_y K(x, y) = 1$ para cada x. Así, cada fila es una medida de probabilidad entonces K puede dirigir un tipo de caminata aleatoria: de x, elegir y con probabilidad *K(x,y)*; de y elegir z con probabilidad *K(y,z)*, y así sucesivamente. Nos referimos a los resultados $X_0 = x, X_1 = y, X_2 = z, \ldots$ como una ejecución de cadenas iniciando en x. Desde las definiciones $P(X_1 = y|X_0 = x) = K(x, y)$, $P(X_1 = y, X_2 = z|X_0 = x) =$

$K(x,y)K(y,z)$. De esto, $P(X_2 = z|X_0 = x) = \sum_y K(x,y)K(y,z)$, y así sucesivamente. La n-potencia de la matriz tiene x, y entrada $P(X_n = y|X_0 = x)$.

Todas las cadenas de Markov consideradas en este documento tienen distribución estacionaria $\pi(x) > 0$, $\sum_x \pi(x) = 1$ con π satisfaciendo

$$\sum_x \pi(x)K(x,y) = \pi(y)$$

Entonces π es un eigenvector izquierdo de K con eigenvalor 1. Entonces π es estacionaria por la evolución. El teorema fundamental de las cadenas de Markov (un simple corolario del teorema de Peron-Frobenius) dice, bajo una simple condición de conectividad, π es único y potencias altas de K convergen al rango de una matríz con todas las filas iguales a π.

3. Algoritmo de Metropolis-Hastings

Sea χ un espacio de estados finito y $\pi(x)$ una probabilidad sobre χ (quizá especificado solo hasta una constante de normalización desconocida). Sea *J(x,y)* una matríz de Markov en χ con $J(x,y) > 0 \longleftrightarrow J(y,x) > 0$. En el inicio, J no está relacionada con π. El algoritmo Metropolis cambia de J hacia una nueva matríz de Markov *K(x,y)* con distribución estacionaria π. Se administra con una simple condición:

$$K(x,y) = \begin{cases} J(x,y) & \text{si } x \neq y, A(x,y) \geq 1 \\ J(x,y)A(x,y) & \text{si } x \neq y, A(x,y) < 1 \\ J(x,y) + \sum_{z:A(x,y)>1} J(x,z)(1 - A(x,z)) & \text{si } x = y \end{cases}$$

En la ecuación anterior, la tasa de aceptación es $A(x,y) = \pi(y)J(y,x)/\pi(x)J(x,y)$. Esta formula tiene una interpretación sencilla: de x, elejimos y con una probabilidad *J(x,y)*; si $A(x,y) \geq 1$, nos movemos a y; si *A(x,y) < 1*, lanzamos una moneda con esta probabilidad de éxito y nos movemos a y si el éxito ocurre; en cualquier otro caso, nos mantenemos en x. Notemos que la constante de normalización para π se anula en todos los cálculos. La nueva cadena satisface

$$\pi(x)K(x,y) = \pi(y)K(y,x)$$

y entonces

$$\sum_x \pi(x)K(x,y) = \sum_x \pi(y)K(y,x) = \pi(y)\sum_x K(y,x) = \pi(y)$$

tal que π es un eigenvector izquierdo con eigenvalor 1. Si la cadena está conectada, el teorema fundamental de las cadenas de Markov está en vigor. Después de muchos pasos en la cadena, la oportunidad de estar en y es de aproximadamente $\pi(y)$, sin importar en el estado inicial χ.

En el ejemplo de la criptografía χ son todas las funciones 1-1 del espacio de simbolos (decimos de tamaño m) al espacio alfabético (decimos de tamaño $n \geq m$). Entonces $|\chi| = n(n-1)\dots(n-m+1)$. Esto es grande, si por ejemplo, $m = n = 50$. La cadena propuesta $J(f, f^*)$ esta especificada por un intercambio aleatorio de dos símbolos,

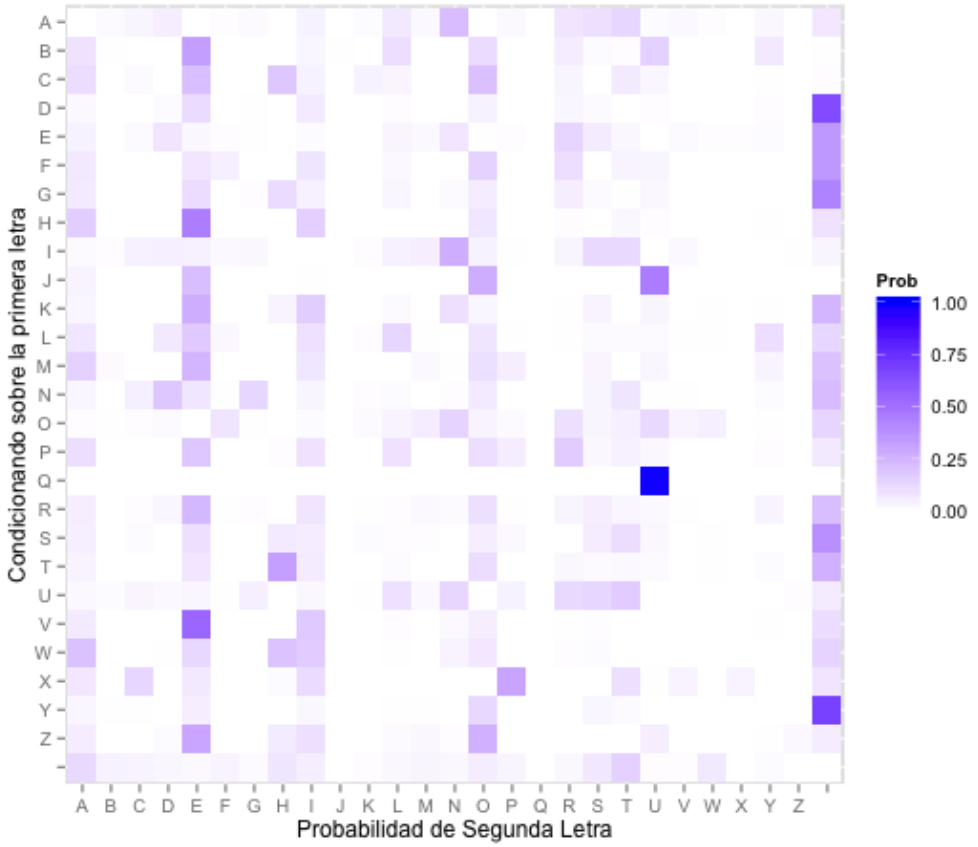
$$J(f,f^*) = \begin{cases} \frac{1}{n(n-1)(m-n+2)(m-n+1)} & \text{si } f, f^* \text{ difieren en al mayoría de los dos lugares} \\ 0 & \text{de otra manera} \end{cases}$$

Notemos que $J(f,f^*) = J(f^*,f)$ entonces $A(f,f^*) = \pi(f^*)/\pi(f)$

4. Análisis de resultados

Sabemos que la estrategia es utilizar un texto de referencia, como ya lo indicamos en la introducción, y para ello utilizaremos el texto *War and Peace* de Leo Tolstoy, el cual nos ayudará a crear las probabilidades de transición de una letra con la siguiente dentro del texto. Esto nos generará una matriz de 26x26, donde la i-ésima fila y la j-ésima columna es la probabilidad de la j-ésima letra que viene precedida de la i-ésima letra. Asumiendo estas probabilidades de transición de un paso son lo que importa, la verosimilitud de cualquier cartografía o mapeo es el producto de las probabilidades de transición observadas.

A continuación, nostramos la matriz de trancisión utilizando el texto `War and Peace`:



La última posición de la matriz (27va posición) se refiere a caracteres que no son letras, por ejemplo, comas, puntos, espacios, etc. Lo que se hizo en la matríz fue contar los casos por fila y después se normalizó dividiendo por el total de cada fila. Antes de normalizar se agregó un 1 a cada celda para evitar tener probabilidades de cero.

La solución que esperamos es la de mayor verosimilitud. Con el enfoque del **MCMC** nosotros utilizamos un mapeo aleatorio inicial de caracteres (letras). Después proponemos un nuevo par de letras de manera aleatória. Calculamos ambas verosimilitudes y se divide la nueva entre la anterior, si es el resultado es menor que 1, significa que el nuevo par tiene menor verosimilitud, entonces pasamos a un nuevo emparejamiento con una probabilidad igual a la tasa. Repetimos este proceso hasta que pensamos ya hemos encontrado una solución.

En el artículo revisado se indica que el algoritmo es suficientemente bueno con textos de hasta 2000 caracteres, sin embargo puede haber problemas para criptogramas con menor cifra. Además, se debe ejecutar varias veces con distintos puntos de inicio para poder encontrar una solución adecuada, es lo recomendable.

Ponemos varios ejemplos para valorar la eficiencia del código.

4.1 Ejemplo 1

Texto original.- [1] "THE OUTPUT OF THIS EXAMPLE IS BELOW. YOU CAN SEE THAT IT COMES CLOSE BUT IT DOESN'T QUITE FIND THE CORRECT MAPPING. I ATTRIBUTE THIS TO THE FACT THAT THE TEXT I WAS TRYING TO DECODE ONLY HAD 203 CHARACTERS.

Texto decodificado.- 2000 THE OUTPUT OW THAD EXIMPLE AD BELOF. YOU CIN DEE THIT AT COMED CLODE BUT AT SOEDN'T JUATE WANS THE CORRECT MIPPANG. A ITTRABUTE THAD TO THE WICT THIT THE TEXT A FID TRYANG TO SECOSE ONLY HIS 203 CHIRICTERD.

4.2 Ejemplo 2

Texto original.- [1] "ENTER HAMLET HAM TO BE OR NOT TO BE THAT IS THE QUESTION WHETHER TIS NOBLER IN THE MIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS FORTUNE OR TO TAKE ARMS AGAINST A SEA OF TROUBLES AND BY OPPOSING END"

Texto decodificado.- 1000 ENTER HAYLET HAY TO CE OR NOT TO CE THAT IS THE BUESTION WHETHER TIS NOCLER IN THE YIND TO SUFFER THE SLINGS AND ARROWS OF OUTRAGEOUS FORTUNE OR TO TAVE ARYS AGAINST A SEA OF TROUCLES AND CK OPPOSING END

4.3 Ejemplo 3

Texto original.- [1] "TWO HOUSEHOLDS, BOTH ALIKE IN DIGNITY, IN FAIR VERONA, WHERE WE LAY OUR SCENE, FROM ANCIENT GRUDGE BREAK TO NEW MUTINY, WHERE CIVIL BLOOD MAKES CIVIL HANDS UNCLEAN."

Texto decodificado.- 2000 MPO SOUGESOLNG, BOMS ALIDE IT NICTIMY, IT WAIR KEROTA, PSERE PE LAY OUR GHETE, WROF ATHIETM CRUNCE BREAD MO TEP FUMITY, PSERE HIKIL BLOON FADEG HIKIL SATNG UTHLEAT.

4.4 Ejemplo 4

Texto original.- [1] "WE OFFER YOU HERE SOME TEXTS ON VARIOUS INTERESTING SUBJECTS. YOU CAN PRACTISE READING COMPREHENSION AND AT THE SAME TIME YOU WILL CERTAINLY LEARN SOMETHING NEW."

Texto decodificado.- 2000 WE OFFER YOU HERE SOME TEXTS ON KARIOUS INTERESTIND SUZVECTS. YOU CAN BRACTISE REAGIND COMBREHENSION ANG AT THE SAME TIME YOU WILL CERTAINLY LEARN SOMETHIND NEW.