Trabajo Integrador 1

Informe

Por:
Barriga Nahuel,
Bedini Tomas,
Hernandez Julieta.

Fecha de entrega: 17/10/2023

Introducción

En el presente trabajo integrador se analizaron distintas fuentes de información brindadas por la cátedra. Para cada una de ellas, se calculó las probabilidades condicionales del contenido del archivo, se determinó si se trataba de una fuente de memoria nula o no nula y se calculó su entropía. Dependiendo de qué tipo de fuente se trataba, se calculó las probabilidades y la entropía de la extensión de orden N o se obtubo su vector estacionario.

La finalidad de este trabajo es, a partir del análisis de las distintas fuentes, entender mejor las características de estas y las propiedades de los códigos.

Análisis de la fuente de informacion

Obtención de la matriz de probabilidades condicionales

Para determinar si se trataba de una fuente de memoria nula o no nula, primero se analizó la relación entre sus símbolos. Para ello, se elaboró una matriz con las probabilidades condicionales mediante el siguiente algoritmo:

A partir de la lectura y guardado de los símbolos del archivo en la lista “datos”, se toma esta secuencia de datos binarios y se calculan las probabilidades condicionales de transición entre los estados binarios 0 y 1. A medida que se avanza en la lista “datos” dentro del ciclo for, la matriz “MT” (previamente inicializada en cero) va incrementando el elemento de la columna del símbolo actual y la fila del símbolo previo, quedando la posición (previo,actual). Además, en cada ciclo se incrementa el contador “cuentaBin” de incidencias del símbolo correspondiente.

Luego, de procesar toda la secuencia de datos, se calcula la suma de elementos de cada columna. Estas sumas representan el número total de transiciones desde un estado al otro. Finalmente, se calculan las probabilidades de transición dividiendo cada columna de la matriz “MT” por la suma previamente calculada. El resultado se almacena en la matriz “PTM”.

Determinación del tipo de fuente y su entropía

Teniendo la matriz de transición, la cual informa la probabilidad de que ocurra el símbolo i siendo j el actual, evaluamos si los símbolos son estadísticamente independientes o si dependen de los símbolos anteriormente ocurridos.

Dado que los símbolos son estadísticamente independientes si la probabilidad del símbolo de cada fila es igual o muy cercana en cada columna, elaboramos una función booleana que recorre cada fila comparando las probabilidades de dicha fila en todas las columnas con cierto margen de tolerancia (optamos por usar 0.02). Si ambas filas de la matriz cumplen con esta condición, la función devuelve True, lo que implica una fuente de memoria nula. Caso contrario, implica una fuente de memoria no nula.

La entropía es una medida fundamental en teoría de la información que cuantifica la incertidumbre o la cantidad de información en una fuente de datos. A partir de la fórmula,

$$H(S) = \sum_s P(S_i) log \frac{1}{P(S_i)}$$

se calculó la entropía de la fuente. Para ello, obtuvimos las probabilidades de cada palabra, dividiendo su frecuencia por la cantidad total de palabras, guardando los resultados en la lista "prob". Luego, recorrimos la lista acumulando la suma de la cantidad de información de cada palabra multiplicado por su probabilidad.

Para el cálculo de la cantidad de información, se utilizó algoritmo en base 2 porque sabíamos a priori que la fuente provenía de un archivo binario.

Fuente de memoria nula

Probabilidades - Entropía de extensión de orden N

La entropía de extensión se refiere a la incertidumbre promedio por símbolo de una secuencia de longitud N en una fuente de información.

$$H(S^n) = \sum_{S^n} P(\sigma i) log \frac{1}{P(\sigma i)}$$

$$H(S^n) = n * H(S)$$

El algoritmo implementado calcula la entropía de extensión N para una fuente de memoria nula generando todas las secuencias posibles de N bits y calculando las probabilidades condicionales de cada secuencia. Luego, utiliza estas probabilidades para calcular la entropía de Shannon para la secuencia de N bits dada.

Fuente de memoria no nula

Vector estacionario

Este vector representa la distribución de probabilidades de estados en un sistema que no cambia con el tiempo. Es decir, es un vector que permanece constante a medida que el sistema evoluciona a lo largo de las transiciones de estados. Al no depender de las condiciones iniciales, se puede obtener mediante la matriz de probabilidades condicionales calculada previamente.

Se implementó mediante un proceso iterativo que actualiza el vector hasta que converja a un estado estacionario, donde la diferencia entre el vector en dos iteraciones consecutivas sea menor que una tolerancia dada (0.00001) o hasta que se alcance un número máximo de iteraciones. Este proceso itera a través de cada estado en una cadena de Markov, calculando una nueva probabilidad para cada estado. Para cada estado actual, se ponderan las probabilidades de transición hacia otros estados y se suman. El resultado se redondea a 5 decimales y se almacena como la nueva probabilidad del estado actual en el vector. Este proceso se repite para todos los estados, actualizando así el vector de probabilidades en cada iteración.

Conclusiones

Este trabajo práctico ha representado una valiosa oportunidad para poner en práctica los conceptos teóricos relativos a la codificación y análisis de información en un contexto real. Al llevar a cabo la implementación de las diversas fases involucradas en el procesamiento de un código, hemos adquirido una comprensión más profunda de los principios subyacentes de la teoría de la información y cómo se aplican en la codificación de datos. Este ejercicio nos ha permitido consolidar nuestros conocimientos teóricos y apreciar cómo se traducen en la práctica, lo que enriquece nuestra comprensión del campo de la informática y la teoría de la información