Unidad 4: Algoritmos para flujos de datos

Estimación de momentos

Blanca Vázquez

Abril 2020

- **Flujo de datos:** es un conjunto de m elementos provenientes de algún universo de tamaño n, ejemplo: 3,5,8,5,9,5,7,5,9,6,1,4,2,5...
- **Algoritmos**: tienen como objetivo estimar propiedades de los flujos de datos, ejemplo: promedio, mediana, número de elementos distintos, conteo, filtrado,...

¿Problema?

- Tamaño limitado de memoria
- Datos secuenciales
- Procesar rápidamente cada dato recibido

¿Problema?

- Tamaño limitado de memoria
- Datos secuenciales
- Procesar rápidamente cada dato recibido

¡Esto ha tratado de resolverse desde los años 70!,
pero en los últimos 10 años ha recobrado popularidad

Popularidad

- Redes más rápidas
- Almacenamientos más accesibles
- Surgimiento de plataformas

- Encontrar el **número de elementos distintos** en un flujo de datos con elementos repetidos

Dado un strem s de elementos $x_1, x_2, ..., x_n$, encontrar el número de elementos distintos n, donde n = $|\{x_1, x_2, ..., x_n\}|$

Ejemplo:
Dado $s = \{a, b, a, c, d, b, d\}$ entonces n es igual a
n = $|\{a, b, c, d\}| = 4$

- Es una **generalización del problema del conteo de elementos distintos**
- Objetivo: calcular **’momentos’**, es decir, estimar la distribución de frecuencias de diferentes elementos en un flujo de datos.

Ejemplo

Conjunto universal de flujos de datos
1,5,9,6,7,1,5,6,4,2,3,6,8,9,6,4,6
Flujo de datos a procesar
6,8,9,6,4,6
Conjunto de elementos seleccionados del universo

m_i = es el número ocurrencias del *ith* elemento

$$m_6 = 3$$

Cálculo de momentos

- m_i es el número de ocurrencias de i en un stream.
- El momento k^{th} de un flujo de datos es:

$$\sum_{i \in A} (m_i)^k$$

$$\sum_{i \in A} (m_i)^k$$

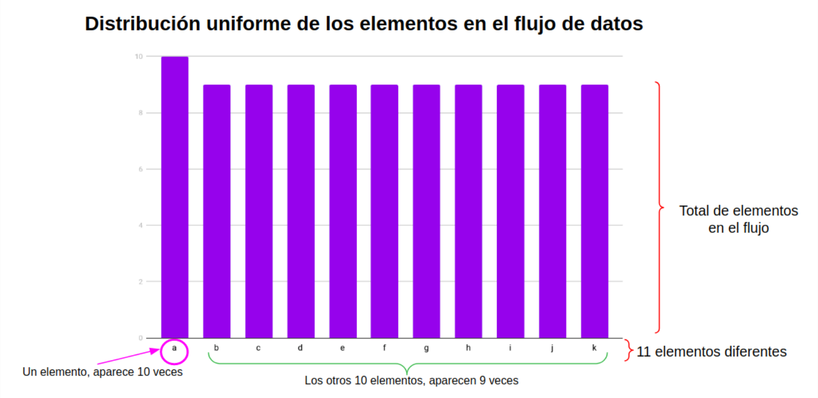
- **Momento 0**: número de elementos distintos (usar algoritmos de conteo)
- **Momento 1**: es la suma de m_i (es el tamaño del flujo de datos)
  $s = \{a, b, a, c\}$, por lo tanto: $m_a = 2$,$m_b = 1$,$m_c = 1$, aplicando la sumatoria = $(2 + 1 + 1)^1 = 4$

$$\sum_{i \in A} (m_i)^k$$

- **Momento 2**: es la suma de los cuadrados de m_i, mide la irregularidad de la distribución de los elementos en el flujo de datos (también conocido como el 'número sorpresa')

- Supongamos que tenemos un flujo de datos de tamaño n = 100, en el cual aparecen 11 elementos diferentes

Ejemplo: cálculo del 2do momento



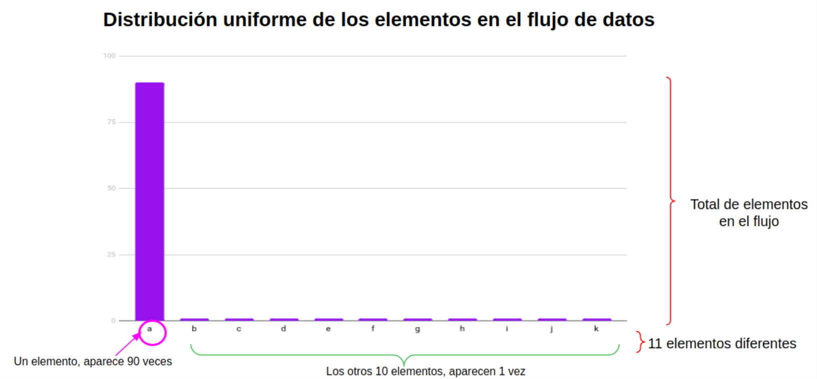
Ejemplo: cálculo del 2do momento

$$\sum_i m_i^2$$

$$= (1 * 10^2) + (10 * 9^2)$$
$$= (1 * 100) + (10 * 81)$$
$$= 100 + 810$$
$$= 910$$

El cálculo del 2do momento es 910 (o también conocido como número sorpresa)

Ejemplo2: cálculo del 2do momento



Ejemplo2: cálculo del 2do momento

$$\sum_i m_i^2$$

$$= (1 * 90^2) + (10 * 1^2)$$
$$= (1 * 8100) + (10)$$
$$= 8100 + 10$$
$$= 8100$$

Aspectos importantes

- No es necesario aplicar el cálculo de momentos, cuándo en memoria podemos mantener los flujos de datos.
- En caso contrario, se necesita estimar el k momento para mantener un número limitado de valores en memoria y calcular un estimado de estos valores.

- Algoritmo definido por Noga Alon, Yossi Matias, y Mario Szegedy, 1996.
- Se utiliza para el cálculo de momentos en flujos de datos
- Se enfoca en aproximar la suma de las entradas al cuadrado de un vector definido por un FD.
- Trabaja para todos los momentos

- Supongamos: que no tenemos espacio suficiente para contar todas las ocurrencias (m_i) para todos los elementos en un flujo de datos... aún con datos limitados es posible calcular los momentos

Para calcular los 2dos momentos, usando el algoritmo de AMS es necesario definir:

- Para cada elemento X en el flujo de datos, se almacena:
  - El elemento como tal, al cual se refiere como $X.element$
  - Un valor entero, $X.value$ el cual es el valor de la variable. Para determinar este valor, se escoge una posición entre 1 y n del flujo de datos (de manera aleatoria).

Importante: un $X.element$ se inicializa con $X.value = 1$, por lo tanto, cada vez que encontremos una ocurrencia sumamos 1 al valor.

Ejemplo: cálculo del 2do momento

$s = \{a,b,c,b,d,a,c,d,a,b,d,c,a,a,b\}$

n = 15 (tamaño de s) >>>> *momento 1*

Dato	Frecuencia
a	5
b	4
c	3
d	3

$$\sum_i m_i^2$$

Cálculo del 2do momento:

$= (1 * 5^2) + (1 * 4^2) + (2 * 3^2)$
$= 25 + 16 + 18$
$= 59$

$$s = \{a,b,c,b,d,a,c,d,a,b,d,c,a,a,b\}$$

- Cálculo del 2do momento usando el algoritmo de AMS
    1. Supongamos que seleccionamos 3 variables aleatorias: X_1, X_2 y X_3
    2. Asumimos las siguientes posiciones para las 3 variables aleatorias: 3, 8 y 13.

Ejemplo: usando el algoritmo AMS

$$s = \{a,b,c,b,d,a,c,d,a,b,d,c,a,a,b\}$$

- Cálculo del 2do momento usando el algoritmo de AMS
  1. Asumimos las siguientes posiciones para las 3 variables: 3, 8 y 13.
  2. En la posición 3, encontramos el elemento 'c', al cual llamamos: $X_1.element$ = c
  3. En la posición 8, encontramos el elemento 'd', al cual llamamos: $X_2.element$ = d
  4. En la posición 13, encontramos el elemento 'a', al cual llamamos: $X_3.element$ = a

Ejemplo: usando el algoritmo AMS

$s = \{a,b,c,b,d,a,c,d,a,b,d,c,a,a,b\}$

$X_1.element = c$	$X_1.value = 1$	$X_1.value = 2$	$X_1.value = 3$
$X_2.element = d$	$X_2.value = 1$	$X_2.value = 2$	
$X_3.element = a$	$X_3.value = 1$	$X_3.value = 2$	

El valor final es: $X_1.value = 3$, $X_2.value = 2$ y $X_3.value = 2$

Ejemplo: usando el algoritmo AMS

s = {a,b,c,b,d,a,c,d,a,b,d,c,a,a,b}

Fórmula para calcular el 2do momento: $n * (2 * X.value - 1)$

- Para X_1, 15 * (2 * 3 - 1) = 75
- Para X_2, 15 * (2 * 2 - 1) = 45
- Para X_3, 15 * (2 * 2 - 1) = 45
- El 2do momento es: (75 + 45 + 45) / 3 = 55 *(este valor es cercano a lo obtenido previamente)*

Ejemplo: usando el algoritmo AMS

- De forma general, el 2do momento del valor esperado de la variable (2 * X.value + 1) es el promedio sobre todas las posiciones *i* entre 1 y n valores.

$$E(2 * X.value + 1) = \frac{1}{n} \sum_{i=1}^{n} n * (2 * c(i) - 1)$$

Simplificando (cancelamos los factores 1/n y n)

$$E(2 * X.value + 1) = \sum_{i=1}^{n} (2c(i) - 1)$$

Estimando el 3er momento

- La fórmula general para estimar el 3 momento es: $n * (3v^2 - 3v + 1)$, donde v = X.value
- Mientras que para estimar los k momentos, $n * (v^k - (v-1)^k)$

Tratando con flujos infinitos

- Asumimos que el tamaño del flujo n es una constante
- En la práctica, n crece con el tiempo
- Un problema, es ¿cómo seleccionar las posiciones para la variables de manera óptima?
- Si seleccionamos los primeros valores recibidos, podemos sesgar los resultados en favor de las 1eras posiciones
- Si seleccionamos los últimos valores, el cálculo del momento sería complejo

Referencias

- Data Streams Tutorial. Andrew McGregor. University of Massachusetts, Amherst, 2011 en https://people.cs.umass.edu/ mcgregor/slides/11-michigan.pdf