Análisis de datos masivos

Estimación de momentos

Blanca Vázquez

8 de octubre de 2024

Estimación de momentos en flujos de datos

- Generalización del problema del conteo de elementos distintos.
- Objetivo: estimar los momentos en un flujo de datos, lo cual se obtiene mediante la distribución de frecuencias de los diferentes elementos.

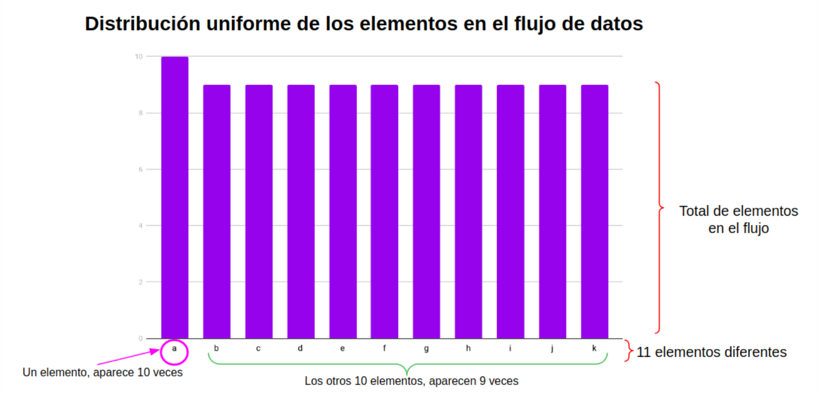
Momentos

- Sea m_i es el número de ocurrencias del elemento i en el flujo, el momento k-ésimo está definido por

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

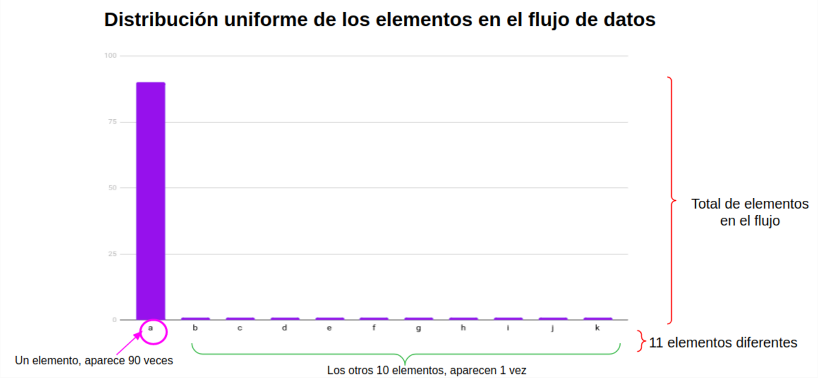
  donde $\mathbb{U}$ es el conjunto universal.
  - El momento 0 es el número de elementos distintos
  - El momento 1 es la suma de m_i, es decir, el tamaño del flujo de datos
  - El momento 2 es la suma de los cuadrados de m_i, también conocido como número sorpresa

Ejemplo (1)



Distribución uniforme de los elementos en el flujo de datos
10
8
6
4
2
0
a
b
c
d
e
f
g
h
i
j
k
Total de elementos en el flujo
11 elementos diferentes
Un elemento, aparece 10 veces
Los otros 10 elementos, aparecen 9 veces

Ejemplo (2)



Distribución uniforme de los elementos en el flujo de datos
100
75
50
25
0
a
b
c
d
e
f
g
h
i
j
k
Total de elementos en el flujo
11 elementos diferentes
Un elemento, aparece 90 veces
Los otros 10 elementos, aparecen 1 vez

Algoritmo de Alon-Matias-Szegedy (AMS)

- Algoritmo para el cálculo de momentos en flujos de datos, definido por Noga Alon, Yossi Matias y Mario Szegedy.
- Se enfoca en aproximar la suma de las entradas al cuadrado de un vector definido por un flujo de datos.
- Permite calcular cualquier momento aún si no es posible almacenar todas las cuentas m_i de todos los elementos.

- Dado un flujo de tamaño n constante, se toman K variables $X_1, X_2, \ldots, X_K$ seleccionando K posiciones en el flujo de forma aleatoria y uniforme.

- Para calcular los momentos, las variables almacenan:
  - Un elemento, al cual se refiere como $X_k.elemento$.
  - Un valor entero $X_k.valor$ el cual es el valor de la variable.
  - $X_k.valor$ se inicializa con 1 y cada vez que encontremos una ocurrencia de $X_k.elemento$, le sumamos 1 a $X_k.valor$
- El segundo momento de cualquier variable X_k se estima con $n \cdot (2 \cdot X_k.valor - 1)$.

Ejemplo: cálculo del 2do momento

- Considera el flujo $a, b, c, b, d, a, c, d, a, b, d, c, a, a, b$

Elemento (e)	Número de ocurrencias (m_e)
a	5
b	4
c	3
d	3

- El momento 0 es: 4
- El primero momento es $5 + 4 + 3 + 3 = 15$
- El segundo momento es $5^2 + 4^2 + 3^2 + 3^2 = 59$

Ejemplo: cálculo del segundo momento con AMS

- Considera el flujo $a, b, c, b, d, a, c, d, a, b, d, c, a, a, b$

1. Supongamos que seleccionamos 3 variables aleatorias: X_1, X_2 y X_3
2. Supongamos que las posiciones para las 3 variables son 3, 8 y 13.

Ejemplo: cálculo del segundo momento con AMS

1. Supongamos que las posiciones para las 3 variables son 3, 8 y 13.
2. En la posición 3, encontramos el elemento c, al cual llamamos: $X_1.elemento = c$
3. En la posición 8, encontramos el elemento d, al cual llamamos: $X_2.elemento = d$
4. En la posición 13, encontramos el elemento a, al cual llamamos: $X_3.elemento = a$

Ejemplo: cálculo del segundo momento con AMS

$X_1.elemento = c$	$X_1.valor = 1$	$X_1.valor = 2$	$X_1.valor = 3$
$X_2.elemento = d$	$X_2.valor = 1$	$X_2.valor = 2$	
$X_3.elemento = a$	$X_3.valor = 1$	$X_3.valor = 2$	

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

Ejemplo: cálculo del segundo momento con AMS

- Para estimar el segundo momento usamos $n \cdot (2 \cdot X_k.valor - 1)$
  - Para X_1: $15 \cdot (2 \cdot 3 - 1) = 75$
  - Para X_2: $15 \cdot (2 \cdot 2 - 1) = 45$
  - Para X_3: $15 \cdot (2 * 2 - 1) = 45$
- Promediando las estimaciones de cada variable tenemos $(75 + 45 + 45)/3 = 55$

- El valor esperado de cualquier variable es el segundo momento del flujo del que fue generada
- Sea $e(i)$ el elemento que aparece en la posición i en el flujo y $c(i)$ el número de veces que aparece este elemento de la posición i a la n, el valor esperado del estimador del segundo momento es

$$\mathrm{E}\left[n \cdot (2 \cdot X_k.valor - 1)\right] = \frac{1}{n}\sum_{i=1}^{n} n \cdot (2 \cdot c(i) - 1) = \sum_{i=1}^{n} (2 \cdot c(i) - 1)$$

Otra forma de calcula el valor esperado es: agrupando los términos de todas las posiciones que tienen el mismo elemento.

- Para el ejemplo anterior, la letra a aparece m_a veces. Si tomamos los términos de la última posición hacia la primera tendríamos

$$
\begin{gathered}
2 \cdot 1 - 1 = 1 \\
2 \cdot 2 - 1 = 3 \\
2 \cdot 3 - 1 = 5 \\
\vdots \\
2 \cdot m_a - 1
\end{gathered}
$$

- Por lo tanto, podemos reescribir el valor esperado para cada elemento $e \in \mathbb{U}$ de la siguiente manera

$$\mathrm{E}\left[n \cdot (2 \cdot X_k.valor - 1)\right] = \sum_{e \in \mathbb{U}} \left[1 + 3 + 5 + \cdots + (2 \cdot m_e - 1)\right]$$

- Dado que $[1 + 3 + 5 + \cdots + (2 \cdot m_e - 1)] = (m_e)^2$

$$\mathrm{E}\left[n \cdot (2 \cdot X_k.valor - 1)\right] = \sum_{e} (m_e)^2$$

- Para el estimador del segundo momento tenemos

$$n \cdot (2 \cdot X_k.valor - 1) = n \cdot (X_k.valor^2 - (X_k.valor - 1)^2)$$

- En general, el estimador del i-ésimo momento es

$$n \cdot (X_k.valor^i - (X_k.valor - 1)^i)$$

Estimación para flujos infinitos (1)

- Hasta ahora hemos considerado que n es constante, sin embargo, en la práctica no lo es
- ¿Cómo seleccionamos las posiciones para la variables?
  - Si seleccionamos los primeros valores recibidos, podemos sesgar los resultados en favor de las primeras posiciones.
  - Si seleccionamos los últimos valores, el cálculo del momento sería complejo.

Estimación para flujos infinitos (2)

- Estrategia de selección de posiciones
    1. Se toman las primeras s posiciones del flujo como variables.
    2. Se elige la posición $n > s$ con probabilidad $\frac{s}{n}$
        - Si es elegida, se selecciona de forma aleatoria y uniforme una de las s variables y se reemplaza por la de la posición n
        - En caso contrario se mantienen las posiciones de las s variables