

Reservoir Sampling

Si consideri il problema di campionare in modo efficiente k elementi a caso da uno stream di lunghezza ignota $N \gg k$. Per esempio, vogliamo farci un'idea delle query inviate a Google o dei prodotti acquistati su Amazon in un dato periodo di tempo. Se indichiamo con x_1, \dots, x_N gli elementi dello stream, vogliamo quindi che ogni x_i venga campionato con probabilità $\frac{k}{N}$.

Un algoritmo efficiente per questo problema è di tipo *streaming*. Ovvero, scandisce l'array una sola volta dall'inizio alla fine usando una memoria ausiliaria sublineare —in questo caso specifico, una memoria ausiliaria pari a $\Theta(k)$. Il seguente semplice algoritmo soddisfa queste proprietà.

Algorithm 1 (reservoir sampling)

Input: Intero k , stream di elementi x_1, x_2, \dots

```

1:  $R = \emptyset$  ▷ inizializza la riserva
2: for  $t = 1, 2, \dots$  do
3:   if  $t \leq k$  then
4:     Aggiungi  $x_t$  a  $R$ 
5:   else
6:     Con probabilità  $\frac{k}{t}$ , sostituisci un elemento a caso in  $R$  con  $x_t$ 
7:   end if
8: end for

```

Output: Riserva R

L'algoritmo di reservoir sampling garantisce che, dopo aver osservato i primi $t \geq k$ elementi dello stream, ogni elemento x_i per $i \leq t$ sia stato campionato con probabilità $\frac{k}{t}$.

Teorema 1 Sia R_t il contenuto della riserva dopo che sono stati osservati i primi t elementi dello stream. Per ogni $i \geq 1$ e per ogni $t \geq \max\{k, i\}$ vale $\mathbb{P}(x_i \in R_t) = \frac{k}{t}$.

Per la dimostrazione useremo più volte il fatto che, per ogni coppia di eventi A, B tale che $\mathbb{P}(B) > 0$, vale $\mathbb{P}(A) = \mathbb{P}(B)\mathbb{P}(A | B)$.

DIMOSTRAZIONE. La dimostrazione è per induzione su i .

Caso base: $t = \max\{k, i\}$. Allora

$$\mathbb{P}(x_i \in R_t) = \begin{cases} 1 & \text{se } t = \max\{k, i\} = k, \\ k/t & \text{se } t = \max\{k, i\} = i. \end{cases}$$

Nel primo caso $1 = \frac{k}{t}$ dato che $t = k$. Il secondo caso vale per la linea 6 dell'algoritmo. Quindi il caso base è verificato.

Ipotesi induttiva: $t > \max\{k, i\}$ e $\mathbb{P}(x_i \in R_t) = \frac{k}{t}$ per ogni $i < t$. Allora, per ogni $i < t$,

$$\mathbb{P}(x_i \in R_{t+1}) = \mathbb{P}(x_i \in R_t)\mathbb{P}(x_t \in R_{t+1} | x_i \in R_t) = \frac{k}{t}\mathbb{P}(x_t \in R_{t+1} | x_i \in R_t)$$

dove abbiamo applicato l'ipotesi induttiva nell'ultimo passo. Ora si osservi che

$$\begin{aligned}\mathbb{P}(x_i \in R_{t+1} \mid x_i \in R_t) &= 1 - \mathbb{P}(x_i \notin R_{t+1} \mid x_i \in R_t) \\ &= 1 - \mathbb{P}(x_{t+1} \in R_{t+1})\mathbb{P}(x_i \notin R_{t+1} \mid x_{t+1} \in R_{t+1}, x_i \in R_t) \\ &= 1 - \frac{k}{t+1} \frac{1}{k} = \frac{t}{t+1}\end{aligned}$$

dove $\mathbb{P}(x_i \notin R_{t+1} \mid x_{t+1} \in R_{t+1}, x_i \in R_t) = \frac{1}{k}$ dato che x_i ha probabilità uniforme di essere selezionato dalla riserva per far posto a x_{t+1} . Quindi,

$$\mathbb{P}(x_i \in R_{t+1}) = \frac{k}{t} \frac{t}{t+1} = \frac{k}{t+1}$$

che conclude la dimostrazione. □