

Reservoir Sampling

Si consideri il problema di campionare in modo efficiente k elementi a caso da uno stream di lunghezza ignota $N \geq k$. Se indichiamo con x_1, \dots, x_N gli elementi dello stream, vogliamo quindi che ogni x_i venga campionato con probabilità $\frac{k}{N}$. Per esempio, vogliamo stimare le percentuali delle varie tipologie di oggetti (libri, elettronica, abbigliamento, etc.) venduti su Amazon in un dato lasso di tempo. Se ogni oggetto venduto è campionato con la stessa probabilità, allora la distribuzione delle tipologie nel campione sarà tendenzialmente uguale a quella nello stream.

Un algoritmo efficiente per questo problema è di tipo *streaming*. Ovvero, esamina ogni elemento dello stream una sola volta e usa 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 $\frac{k}{t}$. Inoltre, come è immediato verificare dalla descrizione dell'algoritmo, il campione prodotto dall'algoritmo ha sempre grandezza k . Si noti che quest'ultima proprietà non sarebbe vera se N fosse noto ed ogni elemento dello stream venisse campionato in modo indipendente con probabilità $\frac{k}{N}$. E infatti l'algoritmo di reservoir sampling utilizza un processo di campionamento dipendente.

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

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

DIMOSTRAZIONE. La dimostrazione è per induzione su $t \geq k$.

Caso base: $t = k$. Allora $\mathbb{P}(x_i \in R_t) = 1 = \frac{k}{t}$ dato che $t = k$.

Caso generale: Fissato $t \geq k$ assumiamo l'ipotesi induttiva

$$\mathbb{P}(x_i \in R_t) = \frac{k}{t} \quad \text{per ogni } i \leq t.$$

e dimostriamo

$$\mathbb{P}(x_i \in R_{t+1}) = \frac{k}{t+1} \quad \text{per ogni } i \leq t+1. \quad (1)$$

Se $i = t+1$ allora (1) vale per costruzione (riga 6 dell'algoritmo). Se invece $i \leq t$, dato che $x_i \in R_{t+1}$ implica $x_i \in R_t$, abbiamo che $\mathbb{P}(x_i \in R_{t+1}) = \mathbb{P}(x_i \in R_{t+1}, x_i \in R_t)$. Quindi possiamo scrivere

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

dove abbiamo applicato l'ipotesi induttiva nell'ultimo passo. Ora si osservi che, dato $x_i \in R_t$, abbiamo che $x_i \notin R_{t+1}$ implica $x_{t+1} \in R_{t+1}$. Quindi possiamo scrivere

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

dove

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

per costruzione dell'algoritmo e

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

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. □