Complementi di Algoritmi e Strutture Dati

Reservoir Sampling

Docente: *Nicolò Cesa-Bianchi* versione 22 marzo 2025

Si consideri il problema di mantenere una struttura dati che, ad ogni istante di tempo, contenga k elementi estratti a caso con probabilità uniforme da uno stream di elementi in ingresso. In particolare, vogliamo sviluppare un algoritmo che soddisfi il seguente invariante: per ogni $t \geq k$, ognuno dei primi t elementi dello stream è contenuto nella struttura dati con probabilità pari a $\frac{k}{t}$. Per esempio, vogliamo stimare le percentuali delle varie tipologie di oggetti (libri, elettronica, abbigliamento, eccetera) 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 tendenzialemente uguale a quella nello stream.

Studiamo il problema nel modello streaming: ad ogni istante di tempo $t = 1, 2, \ldots$ l'algoritmo può accedere soltanto al t-esimo elemento x_t dello stream. Chiediamo inoltre che l'algoritmo lavori in spazio $\Theta(k)$.

Il seguente semplice algoritmo soddisfa tutte le proprietà richieste.

Algoritmo 1 (reservoir sampling)

Input: Intero k

$R = \emptyset$ ▷ inizializza la riserva
for $t = 1, 2, \ldots$ **do**
    Leggi il prossimo elemento x_t nello stream
    if $t \leq k$ **then**
        Aggiungi x_t a R
    else
        Con probabilità $\frac{k}{t}$, sostituisci un elemento a caso in R con x_t
    end if
end for

Nel caso in cui lo stream avesse lunghezza nota N, potremmo aggiungere alla riserva ogni elemento dello stream in modo indipendente con probabilità $\frac{k}{N}$. Questo garantirebbe la proprietà che ogni elemento dello stream è contenuto nella riserva con la stessa probabilità, ma il numero di elementi effettivamente inseriti nella riserva potrebbe essere maggiore o minore di k.

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: $\mathbb{P}\big(x_i \in R_t\big) = \frac{k}{t}$ 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 $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A \mid B)$.

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

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

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

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

e dimostriamo

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

Se $i = t+1$ allora (1) vale per costruzione (riga 7 dell'algoritmo). Se invece $i \leq t$, dato che $x_i \in R_{t+1}$ implica $x_i \in R_t$, abbiamo che $\mathbb{P}\big(x_i \in R_{t+1}\big) = \mathbb{P}\big(x_i \in R_{t+1}, x_i \in R_t\big)$. Quindi possiamo scrivere

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

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

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

dove

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

per costruzione dell'algoritmo e

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

dato che x_i ha probabilità uniforme di essere selezionato dalla riserva per far posto a x_{t+1}. Quindi,

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

che conclude la dimostrazione. □