Complementi di Algoritmi e Strutture Dati

Il problema del Coupon Collector

Docente: *Nicolò Cesa-Bianchi* versione 19 giugno 2017

Il problema del coupon collector è definito come segue: sia $X_1, X_2, \dots$ una sequenza di variabili casuali indipendenti e uniformemente distribuite su n valori distinti $a_1, \dots, a_n$,

$$\mathbb{P}\big(X_t = a_i\big) = \frac{1}{n} \qquad i = 1, \dots, n \quad t \ge 1 \ .$$

Calcolare $\mathbb{E}[N]$, dove $N = \min\big\{k : \big(\forall i \le n\big)\big(\exists t \le k\big)\, X_t = a_i\big\}$. In altre parole, N è il minimo numero di realizzazioni $x_1, \dots, x_k$ sufficienti a osservare ciascun a_i almeno una volta.

Il nome coupon collector deriva dal problema di collezionare tutti gli n possibili buoni premio (coupon) contenuti in prodotti da acquistare (per esempio, scatole di cereali), dove ogni scatola di cereali contiene uno qualsiasi dei buoni premio con probabilità uniforme.

Un problema equivalente è il seguente: supponiamo che ad ogni lancio, una pallina cade con probabilità uniforme in una fra n possibili scatole. Quante palline devo lanciare in media affinché ce ne sia almeno una in ogni scatola?

Un'applicazione concreta del coupon collector è la seguente. Supponiamo di voler sapere gli identificativi degli n router attraversati da una sequenza di pacchetti. Mentre non c'è abbastanza spazio in un pacchetto per memorizzare tutti gli n identificativi, è facile memorizzare in un pacchetto l'identificativo di un router a caso fra quelli attraversati. Ci si chiede allora quanti pacchetti servono in media per ottenere gli identificativi di tutti gli n router.

Per analizzare il problema, suddividiamo $X_1, X_2, \dots$ in n blocchi di lunghezze $N_1, \dots, N_n$ dove N_i è il numero di estrazioni aggiuntive che mi servono per ottenere l'i-esimo valore distinto avendone già osservati $i-1$. Quindi

$$N = \sum_{i=1}^{n} N_i \ .$$

Le variabili casuali $N_1, \dots, N_n$ sono tutte Geometriche. In particolare, quando $i-1$ valori distinti sono già stati osservati, la probabilità di osservarne uno nuovo è

$$p_i = 1 - \frac{i-1}{n} = \frac{n-i+1}{n} \ .$$

Infatti, $p_1 = 1$ e questo implica $N_1 = 1$ deterministicamente, come è giusto che sia.

Ricordando che il valore atteso di una Geometrica di parametro p_i è $\mathbb{E}[N_i] = \frac{1}{p_i}$, per linearità del valore atteso abbiamo

$$\mathbb{E}[N] = \sum_{i=1}^{n} \mathbb{E}[N_i] = \sum_{i=1}^{n} \frac{1}{p_i} = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{i=1}^{n} \frac{1}{i} = n \ln n + \Theta(n)$$

dove l'ultima uguaglianza vale perché la somma armonica $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ è asintotica a $\ln n + \Theta(1)$.