Εξόρυξη Δεδομένων
Πρώτη Σειρά Ασκήσεων

Ταφλαμπάς Νικόλαος 4500

Άσκηση 1

Ερώτημα Α)

1)

Ο αλγόριθμος, αρχικοποιεί μιας λίστα Κ θέσεων για την συλλογή τυχαίων δειγμάτων (δεξαμενή ή reservoir). Έπειτα, αρχίζει να διαβάζει την σειρά δεδομένων μας, τελικού και αγνώστου μεγέθους Ν.

Τα πρώτα Κ δεδομένα της σειράς καταγράφονται στην δεξαμενή. Τα υπόλοιπα δεδομένα, έχουν πιθανότητα K/i, να αντικαταστήσουν κάποιο στοιχείο της δεξαμενής, με i= τον αριθμό των στοιχείων που έχουμε προσπέραση.

2)

Θεωρώ:

$$N = \text{"}\alpha\rho\iota\theta\mu o\varsigma\ \sigma\tau o\iota\chi\varepsilon\iota\omega\nu\ \gamma\iota\alpha\ \delta\varepsilon\iota\gamma\mu\alpha\tau o\lambda\eta\psi\iota\alpha\text{"}$$

$$R = \text{"}\Delta\varepsilon\xi\alpha\mu\varepsilon\nu\eta/Reservoir\text{"}$$

$$K = \text{"}\mu\varepsilon\gamma\varepsilon\theta o\varsigma\ \delta\varepsilon\xi\alpha\mu\varepsilon\nu\acute{\eta}\varsigma\text{"}$$

$$I = \text{"}M\varepsilon\tau\rho\eta\tau\eta\varsigma\ \pi\varepsilon\rho\alpha\sigma\mu\acute{\varepsilon}\nu\omega\nu\ \sigma\tau o\iota\chi\varepsilon\iota\omega\nu\text{"}$$

$$X = \text{"}\Delta\varepsilon\iota\gamma\mu\alpha\text{"}$$

Αρχικά, θα υπολογίσουμε μερικές πιθανότητες:

$$P(\text{“να επιλεγεί ένα στοιχείο”}) = K/I = P(X \in R)$$

$P(\text{“να αντικαταστήσει ένα συγκεκριμένο στοιχείο της δεξαμενής}) = 1/K$=P(X->R[x])

Αρχικά, ας δούμε για I > K. Αρκεί να δείξουμε ότι η πιθανότητα για ένα στοιχείο να είναι στο τέλος του αλγορίθμου μέσα στην δεξαμενή, πρέπει να είναι K/N

$$P(\text{“}\nu\alpha\ \alpha\nu\acute{\eta}\kappa\varepsilon\iota\ \sigma\tau o\ \tau\acute{\varepsilon}\lambda o\varsigma\ \sigma\tau\eta\nu\ \delta\varepsilon\xi\alpha\mu\varepsilon\nu\acute{\eta}\text{”}) =$$

$$P(\text{“}\nu\alpha\ \varepsilon\pi\iota\lambda\varepsilon\chi\theta\varepsilon\acute{\iota}\text{”}) * P(\text{“}\nu\alpha\ \mu\eta\nu\ \alpha\lambda\lambda\acute{\alpha}\xi\varepsilon\iota\ \sigma\varepsilon\ \kappa\acute{\alpha}\pi o\iota o\ \mu\varepsilon\lambda\lambda o\nu\tau\iota\kappa\acute{o}\ \gamma\acute{\upsilon}\rho o\text{”}) =$$

$$P(X \in T) * [P(\text{"να μην επιλεχθεί το στοιχείο i"})\ \text{Ή}\ P(\text{"να επιλεχθει αλλα να μην αλλαξει το στοιχείο μας"})$$

$$P(X \in R') = P(X \in R) * \prod_{i=x+1}^{N} [(1 - P(X \in R) + (P(X \in R) * \mathrm{P(X\!-\!> R[x])})]$$

$$P(X \in R') = \frac{k}{i} * \left[\left(1 - \frac{k}{i+1}\right) + \left(\frac{k}{i+1} * \frac{k-1}{k}\right)\right] * \ldots.$$

$$= \frac{k}{i} * \left[\frac{i}{i+1} * \frac{i+1}{i+2} * \ldots * \frac{N-1}{N}\right]$$

$$= \frac{k}{N}$$

Αρά όντως ισχύει. Τώρα για τα πρώτα K στοιχεία, θεωρούμε:

$$\frac{K}{i} = 1 <=> K = i$$

Έπειτα, κάνοντας την από πάνω διαδικασία για K/i = 1, καταλήγουμε με:

$$P(X \in R') = \frac{i}{N} <=> \frac{K}{N}$$

Πράγμα που ισχύει για κάθε ένα από τα πρώτα K στοιχεία. Έτσι καταλήγουμε ότι ΟΛΑ τα στοιχεία έχουν πιθανότητα K/N να επιλεχθούν.

Ερώτημα Β)

Σαφώς, θα βασιστούμε στον κλασικό Reservoir Sampling για Κ=1.

Ο αλγόριθμος μας, αρχικά, ορίζει έναν μετρητή περασμένου βάρους, που θα κρατάει το άθροισμα όλων των βαρών των στοιχείων που διαβάζουμε.

Με το πέρασμα ενός στοιχείου, αυξάνουμε τον μετρητή κατά w_i και ελέγχουμε την πιθανότητα w_i/W οπού w_i το βάρος του στοιχείου και W η τιμή του μετρητή μας. Άμα πέτυχει, τότε αλλάζουμε το στοιχείο του Reservoir στο στοιχείο που ελέγχουμε.

Το ερώτημα, φυσικά, είναι αν δουλεύει. Αν όντως λειτουργεί ορθά, τότε η πιθανότητα ΚΑΘΕ στοιχείου, πρέπει να είναι w_i/W', όπου W' είναι το σύνολο όλων των βαρών μας

Αρχικά, ας υποθέσουμε τυχαίο στοιχείο i, με ακόλουθες πιθανότητες:

$$P(\text{“}να\ επιλεχθει\ το\ στοιχειο\ i\text{”}) = \frac{w_i}{w_i + \sum_{k<i} w_k} = \frac{w_i}{\sum_{k \le i} w_k}$$

Τώρα έστω στοιχείο j που έρχεται μετά το i. Ξέροντας βάση το από πάνω την πιθανότητα να επιλεχθεί, θα βρούμε την πιθανότητα να ΜΗΝ επιλεχθεί:

$$P(\text{να ΜΗΝ επιλεχθεί το στοιχείο j}) = 1 - \frac{w_j}{\sum_{k \le j} w_k}$$

$$= \frac{\sum_{k\le j} w_k - w_j}{\sum_{k\le j} w_k} = \frac{\sum_{k\le i} w_k}{\sum_{k\le j} w_k}$$

Αυτό μπορεί να γραφτεί για κάθε στοιχείο μετά το j. Άρα η πιθανότητα του i να μείνει στην δεξαμενή είναι:

$$P(i \in R') = P(\text{“}\nu\alpha\ \varepsilon\pi\iota\lambda\varepsilon\chi\theta\varepsilon\iota\ \tau o\ \sigma\tau o\iota\chi\varepsilon\iota o\ i\text{”}) * \prod_{k=i+1}^{N} P(\text{να ΜΗΝ επιλεχθεί το στοιχείο k})$$

$$= \frac{w_i}{\sum_{k\le i} w_k} * \frac{\sum_{k\le i} w_k}{\sum_{k\le j} w_k} * \dots * \frac{\sum_{k\le N-1} w_k}{W'} = \frac{w_i}{W'}$$

Αυτό μπορεί να αποδειχθεί για κάθε στοιχείο i, συνεπώς ο αλγόριθμος μας δουλεύει σωστα!