Πιθανοτική ανάλυση και τυχαιοκρατικοί αλγόριθμοι

Πιθανοτική ανάλυση και τυχαιοκρατικοί αλγόριθμοι

Οδηγος Προσληψης(n)

βέλτιστος $= 0$ // ο υποψήφιος υπ' αριθμ. 0 είναι ένας εικονικός υποψήφιος χωρίς κανένα προσόν

για $i = 1$ έως n

    εξετάζουμε τον υποψήφιο i

    αν ο υποψήφιος i είναι καταλληλότερος από τον *βέλτιστος*

        βέλτιστος $= i$

        προσλαμβάνουμε τον υποψήφιο i

- c_i : η εξέταση ενός υποψηφίου έχει χαμηλό κόστος, έστω c_i,
- c_h : το κόστος c_h της πρόσληψης είναι αρκετά υψηλότερο
- Το συνολικό κόστος που προκύπτει από τον αλγόριθμο είναι $O(c_i n + c_h m)$.
- Χειρότερη περίπτωση : $O(nc_h)$ – όταν όλοι οι υποψήφιοι έρχονται με αυστηρώς αύξουσα σειρά ποιότητας.
- Σε αυτό το είδος του προβλήματος, είναι λογικό να αναμένουμε ότι οι υποψήφιοι δεν θα έρχονται πάντα με αύξουσα σειρά ποιότητας.
- Είναι πιο λογικό να θεωρήσουμε τη μέση περίπτωση.

Πιθανοτική ανάλυση

- Πιθανοτική ανάλυση
  - Η χρήση των πιθανοτήτων στην ανάλυση των προβλημάτων
  - Πρέπει να γνωρίζουμε ή να κάνουμε υποθέσεις σχετικά με την κατανομή της εισόδου.
  - Έπειτα, αναλύουμε τον αλγόριθμο, υπολογίζοντας το χρόνο εκτέλεσης στη μέση περίπτωση, όπου παίρνουμε ένα μέσο όρο ως προς όλες τις πιθανές εισόδους.
  - Προσοχή:
    - Για κάποια προβλήματα, δεν μπορούμε να βρούμε μια ρεαλιστική κατανομή για τις εισόδους του προβλήματος.
    - Σε αυτή την περίπτωση, δεν μπορούμε να χρησιμοποιήσουμε πιθανοτική ανάλυση.

Πιθανοτική Ανάλυση

- Για το πρόβλημα της πρόσληψης, θεωρούμε ότι οι υποψήφιοι έρχονται με τυχαία διάταξη.
- Καταχώρησε μία τάξη (rank) σε κάθε υποψήφιο
- rank(i) δηλώνει τη θέση του υποψήφιου i στην τελική κατάταξη και είναι ένας μοναδικός ακέραιος στο εύρος 1 και n.
- Η λίστα $<$ rank(1), rank(2), . . . , rank(n) $>$ είναι μία μετάθεση των αριθμών $< 1, 2, . . . , n >$.
- Οι υποψήφιοι έρχονται με τυχαία σειρά

  = Η λίστα των τάξεων είναι εξίσου πιθανόν να είναι μία από της n ! μεταθέσεις.

- Ισοδύναμα, οι τάξεις σχηματίζουν μία **ομοιόμορφη τυχαία μετάθεση**: κάθε μία από τις n ! πιθανές μεταθέσεις εμφανίζονται με την ίδια πιθανότητα.

Πιθανοτική Ανάλυση

- Προκειμένου να χρησιμοποιήσουμε την πιθανοτική ανάλυση, θα πρέπει να γνωρίζουμε κάτι σχετικά με την κατανομή της εισόδου.
- Όμως, σε πολλές περιπτώσεις, πολύ λίγα είναι γνωστά για την κατανομή εισόδου.
- Ακόμα και αν εμείς γνωρίζουμε κάτι σχετικά με την κατανομή, μπορεί να μην είμαστε σε θέση να την εκφράσουμε με ένα μοντέλο.
- Παρόλα αυτά, μπορούμε να χρησιμοποιήσουμε την τυχαιότητα και τις πιθανότητες με τη μετατροπή της συμπεριφοράς ενός τμήματος του αλγορίθμου σε τυχαίο.

Τυχαιοκρατικός Αλγόριθμος

- Για το πρόβλημα της πρόσληψης:
  - Αντί να βασιζόμαστε στο να έρχονται οι υποψήφιοι με τυχαία σειρά, έχουμε τον έλεγχο της διαδικασίας και επιβάλλουμε μία τυχαία διάταξη.
  - Το πρακτορείο εύρεσης εργασίας μας στέλνει προκαταβολικά τη λίστα όλων των *n* υποψηφίων
  - Κάθε μέρα, διατάσουμε με τυχαίο τρόπο τη λίστα αυτή πριν τις συνεντεύξεις.

```
RANDOMIZED-HIRE-ASSISTANT(n)
randomly permute the list of candidates
best = 0            // candidate 0 is a least-qualified dummy candidate
for i = 1 to n
    interview candidate i
    if candidate i is better than candidate best
        best = i
        hire candidate i
```

Τυχαιοκρατικός Αλγόριθμος

- Τι κάνει ένα αλγόριθμο τυχαιοκρατικό;
  - Ένας αλγόριθμος είναι τυχαιοκρατικός αν η συμπεριφορά του προσδιορίζεται όχι μόνο από την είσοδο του αλλά από τις τιμές που παράγονται από μία γεννήτρια τυχαίων αριθμών.
  - $RANDOM(a, b)$ επιστρέφει ένα ακέραιο αριθμό r, όπου $a \leq r \leq b$ και κάθε μία από τις $b - a + 1$ πιθανές τιμές της r είναι εξίσου πιθανές.
  - Μία κλήση στη $RANDOM(3, 7)$ επιστρέφει ένα από τα 3, 4, 5, 6, or 7, κάθε ένα με πιθανότητα 1/5.
  - Στην πράξη, RANDOM υλοποιείται με **μία γεννήτρια ψευδοτυχαίων αριθμών**, η οποία είναι μία ντετερμινιστική μέθοδος για την παραγωγή αριθμών που φαίνονται στατιστικά τυχαίοι.

Δείκτριες τυχαίες μεταβλητές

- Οι δείκτριες τυχαίες μεταβλητές είναι ένας εύχρηστος τρόπος για μετατροπή μεταξύ πιθανοτήτων και αναμενόμενης τιμής.
- Δοθέντος ένος δειγματοχώρου S και ενός γενονότος A, η δείκτρια τυχαία μεταβλητή που σχετίζεται με το συμβάν A ορίζεται ως:

$$\mathrm{I}\{A\} = \begin{cases} 1 & \text{εάν προκύψει } A\ , \\ 0 & \text{εάν δεν προκύψει } A\ . \end{cases}$$

Π.χ. S = {H, T}, P(H) = P(T) = 1/2,

- Η δείκτρια τυχαία μεταβλητή X_H αντιστοιχεί στο γεγονός να φέρουμε κορώνα στη ρίψη του κέρματος

$$\begin{aligned} X_H &= \mathrm{I}\{H\} \\ &= \begin{cases} 1 & \text{if } H \text{ occurs}\ , \\ 0 & \text{if } T \text{ occurs}\ . \end{cases} \end{aligned}$$

Δείκτριες τυχαίες μεταβλητές

Το αναμενόμενο πλήθος των κορωνών σε ένα γύρισμα κέρματος = αναμενόμενη τιμή της X_H

$$
\begin{aligned}
E[X_H] &= E[I\{H\}] \\
&= 1 \cdot \mathrm{P}(H) + 0 \cdot \mathrm{P}(T) \\
&= 1 \cdot (1/2) + 0 \cdot (1/2) \\
&= 1/2
\end{aligned}
$$

- Έτσι, το αναμενόμενο πλήθος των κορωνών για ένα γύρισμα είναι 1/2.
- Επομένως αν $X_i = I\{\text{το νόμισμα } i \text{ είναι κορώνα}\}$ τότε

  $X = X_1 + X_2 + X_3 + X_4 + \ldots + X_n$ (πλήθος κορωνών για n ρίψεις)
- Επομένως, $E(X) = E(X_1) + E(X_2) + E(X_3) + E(X_4) + E(X_n) = n/2$

Δείκτριες τυχαίες μεταβλητές

Λήμμα. Δοθέντος ενός δειγματοχώρου Σ και ενός γεγονότος A στο S, έστω $X_A = I\ \{A\}$. Τότε $E\ [X_A] =$ P(A).

Απόδειξη

$$E\ [X_A] = E\ [I\ \{A\}]$$
$$= 1 \cdot \mathrm{P}(A) + 0 \cdot \mathrm{P}(\overline{\mathrm{A}})$$
$$= \mathrm{P}(A),$$

όπου $\overline{\mathrm{A}}$ δηλώνει το $S - A$, το συμπλήρωμα του A.

Η ανάλυση του προβλήματος πρόσληψης με τη χρήση των δείκτριων τυχαίων μεταβλητών

- Έστω X μία τυχαία μεταβλητή της οποίας η τιμή ισούται με το πλήθος των φορών που προσλαμβάνουμε ένα νέο άτομο.

$$\mathrm{E}[X] = \sum_{x=1}^{n} x \Pr\{X = x\}$$

$$X_i = \mathrm{I}\{\text{candidate } i \text{ is hired}\} = \begin{cases} 1 & \text{if candidate } i \text{ is hired}, \\ 0 & \text{if candidate } i \text{ is not hired}, \end{cases}$$

$$\mathrm{E}[X_i] = \Pr\{\text{ο υποψήφιος } i \text{ προσλαμβάνεται}\}$$

$$\mathrm{E}[X_i] = 1/i \qquad X = X_1 + X_2 + X_3 + X_4 + \ldots + X_n$$

$$\begin{aligned} \mathrm{E}[X] &= \mathrm{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathrm{E}[X_i] \\ &= \sum_{i=1}^{n} 1/i \\ &= \ln n + O(1) \end{aligned}$$

Τυχαία μετάθεση της εισόδου

```
PERMUTE-BY-SORTING(A)
n = A.length
let P[1..n] be a new array
for i = 1 to n
    P[i] = RANDOM(1, n^3)
sort A, using P as sort keys
```

```
RANDOMIZE-IN-PLACE(A)
n = A.length
for i = 1 to n
    swap A[i] with A[RANDOM(i, n)]
```

Ανάλυση της Permute-By-Sorting

- Για i = 1, 2, . . . , n, έστω E_i το ενδεχόμενο το στοιχείο A[i] να λάβει τον i-οστό μικρότερο αριθμό προτεραιότητας.
- Ο στόχος: να υπολογίσουμε την πιθανότητα να συμβεί το ενδεχόμενο E_i, για όλα τα i, η οποία είναι

$$\Pr\{E_1 \cap E_2 \cap E_3 \cap \cdots \cap E_{n-1} \cap E_n\} = \Pr\{E_1\} \cdot \Pr\{E_2 \mid E_1\} \cdot \Pr\{E_3 \mid E_2 \cap E_1\} \cdot \Pr\{E_4 \mid E_3 \cap E_2 \cap E_1\} \cdots \Pr\{E_i \mid E_{i-1} \cap E_{i-2} \cap \cdots \cap E_1\} \cdots \Pr\{E_n \mid E_{n-1} \cap \cdots \cap E_1\} .$$

$$\begin{aligned} \Pr\{E_1 \cap E_2 \cap E_3 \cap \cdots \cap E_{n-1} \cap E_n\} &= \left(\frac{1}{n}\right)\left(\frac{1}{n-1}\right)\cdots\left(\frac{1}{2}\right)\left(\frac{1}{1}\right) \\ &= \frac{1}{n!} , \end{aligned}$$

Permute-By_Sorting

- Η πιθανότητα όλες οι προτεραιότητες να είναι διαφορετικές:
- $P_{diff} = 1*(1- 1/n^3)(1-2/n^3)(1-3/n^3)\ldots (1- (n-1)/n^3) \geq (1- n/n^3)\ldots. (1- n/n^3) = (1-1/n^2)^n \geq 1-1/n$

  Η τελευταία ανισότητα ισχύει γιατί $(1-a)^n \geq 1-n*a$ $(a<1)$ (Η συνάρτηση $f(x)=(1-a)^x + ax-1$ είναι αύξουσα για $x \geq 1$ και $f(1)=0$)
- Αν προκύψουν ίδιες οι προτεραιότητες, ξανατρέχουμε τον αλγόριθμο.
- Πόσες επαναλήψεις χρειάζονται για να προκύψουν μοναδικές προτεραιότητες;
- Η πιθανότητα να χρειαστούν k επαναλήψεις:

  $$P_k =(1- P_{diff})^{k-1} P_{diff}$$ (γεωμετρική κατανομή)
- Επομένως, το μέσο πλήθος προσπαθειών είναι $1/ P_{diff} \leq 1/(1-1/n)=n/(n-1)$

Τυχαία μετάθεση της εισόδου

- Ανάλυση της Randomize-In-Place
- Θα αποδείξουμε την αναλλοίωτη συνθήκη:
- Ακριβώς πριν από την i-οστή επανάληψη, για κάθε δυνατή $(i-1)$-μετάθεση των n στοιχείων, η υποσυστοιχία $A[1 \,..\, i-1]$ έχει πιθανότητα $(n-i+1)!/n!$ να περιέχει αυτήν την $(i-1)$-μετάθεση.
- Έστω E_1 το ενδεχόμενο οι πρώτες $i-1$ επαναλήψεις να έχουν τη συγκεκριμένη $(i-1)$-μετάθεση $(x_1, \ldots, x_{i-1})$ στην υποσυστοιχία $A[1 \,..\, i-1]$
- E_2 το ενδεχόμενο το στοιχείο που θα τοποθετηθεί στη θέση A[i] με την i-οστή επανάληψη να είναι το x_i.
- $\Pr\{E_2 \cap E_1\} = \Pr\{E_2 \mid E_1\} \Pr\{E_1\} = 1/(n-i+1) * (n-i+1)!/n! =$

  $(n-i)!/n!$.

Το άμεσης απόκρισης πρόβλημα της πρόσληψης

```
ON-LINE-MAXIMUM(k, n)
bestscore = −∞
for i = 1 to k
    if score(i) > bestscore
        bestscore = score(i)
for i = k + 1 to n
    if score(i) > bestscore
        return i
return n
```

S: το ενδεχόμενο να επιλέξουμε τον καταλληλότερο υποψήφιο

S_i: το ενδεχόμενο να επιλέξουμε τον καταλληλότερο όταν αυτός είναι ο i-οστός.

$$\Pr\{S\} = \sum_{i=k+1}^{n} \Pr\{S_i\} .$$

B_i : το ενδεχόμενο ο καλύτερος υποψήφιος να βρίσκεται στη θέση i

O_i : το ενδεχόμενο να μην επιλεγεί κανείς υποψήφιος από τις θέσεις k + 1 έως i − 1

$\Pr\{S_i\} = \Pr\{B_i \cap O_i\} = \Pr\{B_i\} \Pr\{O_i\}$. $\Pr\{B_i\} = 1/n$ $\Pr\{O_i\} = k/(i-1)$

Το άμεσης απόκρισης πρόβλημα της πρόσληψης

$$\begin{aligned}
\Pr\{S\} &= \sum_{i=k+1}^{n} \Pr\{S_i\} \\
&= \sum_{i=k+1}^{n} \frac{k}{n(i-1)} \\
&= \frac{k}{n} \sum_{i=k+1}^{n} \frac{1}{i-1} \\
&= \frac{k}{n} \sum_{i=k}^{n-1} \frac{1}{i} .
\end{aligned}$$

$$\int_{k}^{n} \frac{1}{x}\,dx \le \sum_{i=k}^{n-1} \frac{1}{i} \le \int_{k-1}^{n-1} \frac{1}{x}\,dx .$$

$$\frac{k}{n}(\ln n - \ln k) \le \Pr\{S\} \le \frac{k}{n}(\ln(n-1) - \ln(k-1)) ,$$

k = n/e για μέγιστη πιθανότητα πρόσληψης του καλύτερου

Το πρόβλημα του συλλέκτη κουπονιών

- Άλμπουμ που πρέπει να συμπληρώσετε b αυτοκόλλητα.
- Κάθε φακελάκι που αγοράζετε περιέχει ένα αυτοκόλλητο και έστω ότι περιέχει ένα από τα b αυτοκόλλητα με την ίδια πιθανότητα.
- Πόσα φακελάκια πρέπει να αγοράσουμε για να συμπληρώσουμε το άλμπουμ.
- Το πρόβλημα μπορεί να ειδωθεί ως ένα πρόβλημα σφαιρών και δοχείων (balls and bins)
- Συγκεκριμένα, αν έχουμε b δοχεία, πόσες σφαίρες πρέπει να ρίξουμε για να μην έχουμε άδεια δοχεία;
- Οι ρίψεις γίνονται με τυχαίο τρόπο, κάθε σφαίρα μπορεί να καταλήξει σε οποιαδήποτε δοχείο με την ίδια πιθανότητα.

Το πρόβλημα του συλλέκτη κουπονιών (Συν.)

- Η διαδικασία χωρίζεται σε b φάσεις.
- Η 1η φάση τελειώνει με την πρώτη ρίψη (άρα 1 δοχείο μη άδειο)
- Η 2η φάση τελειώνει όταν ένα άδειο δοχείο δεχθεί την πρώτη σφαίρα (2 δοχεία μη άδεια)
- Η i-οστη φάση τελειώνει όταν ένα άδειο δοχείο δεχθεί την πρώτη σφαίρα (i δοχεία μη άδεια)
- Έστω n_i οι ρίψεις που χρειάζονται για να ολοκληρωθεί η i-οστη φάσεις. Ισχύει

$$\mathrm{E}[n_i] = \frac{b}{b-i+1}$$

Γεωμετρική κατανομή

Το πρόβλημα του συλλέκτη κουπονιών (Συν.)

- Το συνολικό πλήθος ρίψεων:

  $n=n_1+n_2+\ldots..+ n_b$

- Ισχύει

$$
\begin{aligned}
\mathrm{E}[n] &= \mathrm{E}\left[\sum_{i=1}^{b} n_i\right] \\
&= \sum_{i=1}^{b} \mathrm{E}[n_i] \\
&= \sum_{i=1}^{b} \frac{b}{b-i+1} \\
&= b\sum_{i=1}^{b} \frac{1}{i} \\
&= b(\ln b + O(1))
\end{aligned}
$$