Curs 2/0: Introducere în Teoria Probabilităților: Problema zilei de nastere și Eșantionare

Conf.dr. Maria Jivulescu

Departamentul de Matematică
UPT

Cuprins

- Recapitulare din cursul anterior
- Problema zilei de naștere(Birthday Paradox)
- Scheme clasice de probabilitate
  - Schema bilei revenite
  - Schema bilei nerevenite

Recapitulare

Dată o experiența aleatoare, tripletul $(\Omega, \mathcal{K}, P)$ **spatiu de probabilitate**, unde Ω-spațiul tuturor realizărilor acestei exp, $\mathcal{K} \subset \mathcal{P}(\Omega)$ o familie (admisibila) de evenimente, iar $P : \mathcal{K} \to [0, 1]$ funcția probabilitate

Ev. A	Multime A	$P(A) \in [0, 1]$
Ev. sigur	Ω	1
Ev. imposibil	$\emptyset$	0
Ev. contrar lui A	$\complement_\Omega A$	$P(\complement_\Omega A) = 1 - P(A)$
Ev. reuniune	$A \cup B$	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Ev. intersectie	$A \cap B$	$P(A \cap B) =?$
Ev. mutual exclusive	$A \cap B = \emptyset$	$P(A \cap B) = 0$
Ev. diferenta	$A \setminus B$	$P(A \setminus B) = P(A) - P(A \cap B)$

Formalizarea problemei

Fie n care persoane participă la un reuniune. Care este probabilitatea ca cel puțin două dintre ele să aibă aceeași zi de naștere?

- probabilitatea de a găsi într-un grup de 23 de persoane cel puțin două cu aceeași zi de naștere este mai mare decât 1/2:

$$p_{23} = 0.507297$$

- Rezultat este contrar intuiției umane$\rightarrow$ *the birthday paradox*.

n	1	5	10	20	23
p(n)	0	2.7%	11.7%	41.1%	50.7%

Rezolvare:

- Presupunem că anul are 365 de zile
- A mulțimea participanților la reuniune, $A = \{1, 2, \dots, n\}$
- B mulțimea codurilor pentru zilele anului, $B = \{1, 2, 3, \dots, 365\}$
- Ω mulțimea tuturor posibilităților pentru zilele de naștere ale celor n persoane
  $\Omega = \{z_1 z_2 \dots z_n, z_i \in B\}$
  Ω - mulțimea aplicațiilor de la A la B, adică $\Omega = B^A$
- $|\Omega| = 365^n$
- cele 365 zile de naștere din n–liste sunt echiprobabile.

Problema zilei de nastere

- E_n evenimentul: cel puțin două persoane din cele n care participă la reuniune au aceeași zi de naștere.
- $P(E_n) = |E_n|/|\Omega|$.
- $P(E_n) = 1 - P(\complement_\Omega E_n)$
- $\complement_\Omega E_n$: printre cele n persoane nu există două persoane cu aceeași zi de naștere
- $\complement_\Omega E_n = \{f \in B^A \mid f \text{ este injecție}\}$ (o injecție asociază la oricare două persoane diferite, zile de naștere diferite)
- $P(\complement_\Omega E_n) = \dfrac{\text{numărul injecțiilor de la A la B}}{\text{numărul aplicațiilor de la A la B}} = \frac{A_{365}^n}{365^n} = \prod\limits_{k=1}^{n-1}(1 - \frac{k}{365})$
- $P(E_n) = 1 - P(\complement_\Omega E_n) = 1 - \prod_{k=1}^{n-1}\left(1 - \frac{k}{365}\right)$

Problema zilei de nastere

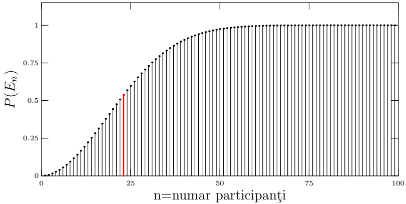

Figure: Ilustrarea $(n, P(E_n))$. Pentru $n \geq 23$, $P(E_n) > 0.5$, $n = 12, \ldots, 100$.

Generalizarea problemei zilei de naștere

Generalizare

Fie date n obiecte numerotate $1, 2, \ldots, n$ și N containere, $N \geq n$.
Pe rând, obiectele sunt atribuite la întâmplare containerelor (sunt aruncate la întâmplare în containere, fiecare obiect poate ateriza în oricare container cu aceeași probabilitate.)
Dacă într-un container au căzut cel puțin două obiecte spunem că s-a produs o *coliziune*.
Să se determine probabilitatea să se producă cel puțin o coliziune după aruncarea celor n obiecte în cele N containere.

- Probabilitatea se calculează similar cu probabilitatea să participe la reuniune cel puțin două persoane cu aceeași zi de naștere.
- În cazul coliziunii 365 se înlocuiește cu N.

Problema coliziunilor

Propozitie

Pentru $N \geq n \geq 2$ probabilitatea a cel puțin unei coliziuni $C(N, n)$ este

$$P(C(N,n)) = 1 - \prod_{k=1}^{n-1}(1 - \frac{k}{N}), \quad P(C(N,n)) \geq 1 - e^{-n(n-1)/2N}$$

.

Deci, numărul de obiecte ce trebuie aruncate in N containere pentru o obtine o coliziune cu probabiltate P este:

$$n(P) \leq 1 + \sqrt{2N \ln \frac{1}{1-P}}$$

Concluzie: Intr-o căutare a unu obiect, este mult mai ușor să găsim obiecte identice, decât un obiect particular.

Algoritmi de coliziune= constituie două liste de elemente, caută un element ce apare în ambele, adică identifică o coliziune.

Dorim să îmbunătățim performanța algoritmului de verificarea a echivalenței a două polinoame. Cum?

- alegem $r \in \{1, 2, \ldots, 1000d\}$; avem $P(\text{raspuns gresit}) \leq 1/10^3$, dar limită de memorie ale calculatorului.
- reluăm algoritmul de mai multe ori, considerând valori aleatoare diferite pentru a testa identitatea $F(x) = G(x)$.
  - dacă se repetă algoritmul de un număr de ori și la orice rulare se generează r pentru care $F(r) \neq G(r)$, atunci $F(x) \neq G(x)$;
  - dacă se repetă algoritmul de un număr de ori și determinăm un r pentru care $F(r) \neq G(r)$, atunci $F(x) \neq G(x)$;
  - dacă $F(r) = G(r)$, la toate rulările algoritmului, atunci $F(x) = G(x)$.
- Cum facem generarea numerelor $r \in \{1, 2, \ldots 100d\}$ in rularile algoritmului?

Esantionare cu înlocuire

Metoda 1: se bazează pe **Schema bilei revenite**

- nu ne amintim ce numere au fost testate în rulările anterioare ale algoritmului;
- presupunem că reluăm algoritmul de k ori;
- la fiecare rulare, se genereaza un numă aleator, uniform din $\{1, 2, \ldots, 100d\}$, indiferent de alegerile anterioare
- la fiecare reluare a algoritmului avea prb de eroare $d/100d$
- definim E_i: la iterația i s-a găsit o rădăcină a polinomului $F - G$; avem $P(E_1) = P(E_2) \leq 1/100$;
- probabilitatea ca la ambele iterații să se găsească o rădăcină a lui $F - G$ este $P(E_1 \cap E_2) \leq (1/100)^2$
- probabilitatea ca la k iterații să se găsească rădăcini ale lui $F - G$ este $P(E_1 \cap \ldots E_k) \leq (1/100)^k$;

Acst rezultat $\rightarrow$ noțiunea de independenta a evenimentelor!

$$E_1, \ldots E_k \text{ independente} : P(E_1 \cap E_2) = P(E_1)P(E_2)$$

Esantionare fără înlocuire

Metoda 2: se bazează pe **Schema bilei nerevenite**

- dacă la una din iterații s-a generat numărul r, nu vom mai permite ca acest număr să fie ales la iterațiile ulterioare ale algoritmului
- în acest caz, $P(E_1 \cap E_2)$ depinde de ce s-a obținut la iterația 1 $\Rightarrow$ necesitatea introducerii noțiunii de **probabilitate condiționată**

$$P(E_2|E_1) := \frac{P(E_1 \cap E_2)}{P(E_1)}$$

- $P(E_1 \cap E_2 \cap \ldots \cap E_k) =$
  $P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)\ldots P(E_k|E_1 \cap E_2 \ldots \cap E_{k-1})$

Concluzii cu care să rămânem

- Problema zilei de naștere (birthday paradox): formularea problemei + soluție
- eșantionare cu (fără) înlocuire
- necesitatea extinderii teorei pentru a cuprinde situații deosebite: independență și condiționare

Vă mulțumesc pentru atenție!
Întrebări?