Proiect la Probabilități și Statistică

An II, Informatică, grupele 241 și 243

Precizări importante

1. Proiectele se realizează în echipă de 2-4 persoane. Fiecare echipa va desemna un lider care va fi precizat în documentație.

2. Liderul echipei va trimite pe adresa **simona.cojocea@fmi.unibuc.ro** până la data de **1 februarie 2026 ora 22:00** *o singură arhivă* care va conține fișierele sursă ale proiectului împreună cu documentația.

3. Documentația este **obligatorie** și lipsa ei atrage *necorectarea proiectului*.

4. Documentația trebuie să conțină :

- numele membrilor echipei
- descrierea problemei
- aspecte teoretice folosite în rezolvarea problemei care depășesc nivelul cursului
- reprezentări grafice și orice altă formă multimedia oportună proiectului
- precizări privind pachete software folosite și surse de inspirație
- codul și comentarea acestuia, precum și a soluției prezentate
- identificarea unor eventuale dificultăți în realizarea cerințelor
- probleme care au rămas deschise în urma implementării actuale
- concluzii

OBS: Documentația se dorește o prezentare *completă* a proiectului astfel încât evaluarea acestuia să poată fi făcută facil și nu o documentație uzuală pentru un produs software.

5. Dacă se realizează cerințe suplimentare față de cele date, cerințe care să fie relevante, se poate obține un bonus de **5p**, fără însă ca nota finală asociată laboratorului să poată depăși **50 p**.
6. Orice informație care nu este expres precizată în enunț este lăsată la latitudinea voastră(ex. alegerea parametrilor repartițiilor), dar orice ipoteză suplimentară utilizați trebuie să fie specificată în documentație.
7. Proiectul se dorește a fi o aplicație Shiny(Shiny - Welcome to Shiny) care să folosească modelarea probabilistică pentru rezolvarea unei probleme din zona IT. Partea de interfață în Shiny valorează **10 puncte**, iar rezolvarea cerințelor împreună cu documentația valorează **20 de puncte**. În cazul în care nu vă descurcați cu construirea interfeței, se acceptă și un proiect fără Shiny, pierzând însă punctajul aferent interfeței.

Analiza probabilistică a performanței unui serviciu online cu trafic aleator și impact economic

Se consideră o platformă online (aplicație web / API / microserviciu) care deservește utilizatori finali. Platforma procesează cereri (***requests***) venite de la un număr variabil de clienți, fiecare cerere fiind supusă *incertitudinii*: timpi de răspuns variabili, eșecuri temporare, reîncercări (***retry***), timeout-uri și **politici de backoff**.

- **Trafic zilnic**

Numărul de clienți care accesează platforma într-o zi este aleator, notat $K_d =$ numărul de clienți activi în ziua d, și este influențat de factori externi (sezonalitate, campanii, popularitate). Acest trafic determină încărcarea sistemului și, implicit, performanța

- **Procesarea cererilor**

Fiecare client generează cereri. O cerere poate: reuși sau eșua, fi reluată de cel mult $N_{\max}$ ori,

avea ***timeout*** dacă depășește un prag de timp, aplica ***backoff*** între ***retry***-uri.

Pentru o cerere, definim:

$\boldsymbol{S_i}$ - timpul de răspuns la încercarea i;

$\boldsymbol{U_i} \in \{0,1\}$ - succes/eșec la încercarea i;

$\boldsymbol{B_i}$ - backoff între încercări;

$\boldsymbol{N}$ - numărul total de încercări;

$\boldsymbol{T}$ - timpul total până la succes sau abandon;

$\boldsymbol{I}$ - indicator de succes final.

- **Experiența utilizatorului și churn**

Un utilizator poate părăsi aplicația(churn): aleator, fără o cauză direct observabilă sau condiționat de performanță, de exemplu dacă, într-o fereastră de timp sau într-un număr de cereri consecutive, prea multe cereri nu sunt rezolvate.

- **Impact economic**

Fiecare cerere reușită produce un câștig, fiecare utilizator pierdut produce o pierdere (cost de achiziție + venituri viitoare ratate), iar nerespectarea SLA poate produce penalități.

Scopul proiectului este de a înțelege, prin modelare probabilistică și prin simulare în R, relația dintre trafic, performanță tehnică și impact economic.

Cerințe

1. Modelarea traficului zilnic (variabile aleatoare discrete)

a) Modelați K_d folosind, pe rând, cel puțin două distribuții discrete (ex.: Poisson, Binomială).
b) Generați prin simulare eșantioane mari care să reprezinte traficul zilnic pentru o perioadă de câțiva ani și reprezentați histogramele asociate acestora. Interpretați comparativ histogramele obținute pe luni și pe ani.
c) Estimați empiric media și varianța traficului pentru fiecare an și comparați cu valorile teoretice.
d) Interpretați diferențele între modele (trafic redus vs plafonat).

2. Modelarea timpilor de răspuns (variabile aleatoare continue)

a) Modelați S, pe rând, cu o distribuție asimetrică (Exponențială/Gamma) și respectiv cu o distribuție Normală (eventual trunchiată la valori pozitive).

b) Construiți histogramele pentru S și suprapuneți peste acestea densitățile teoretice.

c) Calculați media, varianța, mediana, valoarea modală și interpretați rezultatele obținute.

d) Discutați diferența dintre medie și mediană în contextul latențelor.

3. Cereri, retry-uri și evenimente

Definiți evenimentele:

- $A = \{I = 1\}$ (succes);
- $B = \{T \leq t_0\}$ (SLA);
- $C = \{N \leq n_0\}$;
- $D =$ {cel puțin un eșec}.

a) Estimați empiric: $P(A), P(B), P(C), P(A \cap B), P(A \cup D)$
b) Verificați numeric formulele pentru reuniune/intersecție
c) Explicați de ce probabilitatea empirică aproximează bine probabilitatea teoretică.

4. Variabile aleatoare bidimensionale discrete

Considerați variabila bidimensională (N, F), unde F este numărul de eșecuri. Determinați:
a) Distribuția comună empirică;

b) Distribuțiile marginale;

c) Un test empiric de independență;

d) O modalitate de vizualizare (tabel/heatmap) și interpretare.

5. Variabile aleatoare bidimensionale(discrete și continue)

Considerați variabila bidimensională (N, T).

a) Reprezentați grafic variabila bidimensională (N, T).

b)Calculați mediile, variațele, covarianța și coeficientul de corelație

c) Interpretați corelația (retry-uri vs latență totală).

6. Probabilități condiționate și condiționări

a) Estimați $P(A \mid N \leq n_0)$, $P(B \mid A)$.
b) Calculați:$E(T \mid I = 1)$,$E(T \mid I = 0)$.
c) Interpretați rezultatele din perspectiva experienței utilizatorului.

7. Independență vs dependență

a) Simulați două scenarii: timpi S_i independenți vs dependenți (latența crește după eșecuri).
b) Comparați distribuția și varianța lui T în cele două scenarii.
c) Formulați concluzii privind riscul și stabilitatea sistemului.

8. Inegalități probabilistice (garanții worst-case)

Pentru $T \geq 0$:

a) Verificați numeric inegalitățile Markov și Cebîșev (empiric versus teoretic).
b) Pentru variabila **număr de eșecuri/încercări** verificați o inegalitate de tip Chernoff.
c) Interpretați utilitatea acestor limite când distribuțiile exacte sunt necunoscute.
d) Pentru o funcție convexă φ (ex.: x^2, e^x) verificați numeric $\varphi(E(T)) \leq E(\varphi(T))$(inegalitatea lui Jensen)
e) Interpretați rezultatul de la d) în contextul riscului (penalizarea valorilor extreme).

9. Aproximare normală și agregare

a) Pentru sume/agregări zilnice (ex.: total latență pe zi sau profit zilnic), studiați oportunitatea aproximării cu o distribuție normală prin simulare.

b) Comparați histograma agregatului cu o normală ajustată și precizați când aproximarea este adecvată.

10. Churn (pierderea utilizatorilor)

Pierderea utilizatorilor se realizează prin două mecanisme: ***aleator***(cu o probabilitate constantă q) și respectiv ***condițional***, dacă într-o fereastră de m cereri, cel puțin k eșuează.

a) Modelați probabilistic cele două scenarii.

b) Estimați probabilitatea de pierdere a utilizatorului.

c) Comparați scenariile și interpretați.

11. Impact economic

a) Definiți o v.a. pentru profitul zilnic(câștig per succes, pierdere per churn, penalități SLA).
b) Estimați media, varianța, și (opțional) intervale de încredere pentru profit.

c) Analizați compromisurile tehnico-economice.

12. Vizualizare statistică

a) Histograme pentru T și profit.

b)Boxplot-uri pentru T condiționat de succes/eșec și pentru scenarii diferite.
c) Interpretați mediană, IQR, outlieri.

13. Analiză de sinteză

În raport cu problema modelată, comentați:

a) Rolul probabilității empirice
b) Ce informații aduc condiționările
c) Utilitatea inegalităților probabilistice
d) Legătura dintre performanța tehnică și impactul economic
e) Ce parametri influențează cel mai mult rezultatele finale și ce ați modifica pentru îmbunătățirea sistemului.