Proiect

Grupele 311, 321

Notă: Rezolvarea problemelor de mai jos va fi realizată în **R** (scripturile trebuie să fie comentate) şi va fi însoţită de un document text (.pdf sau .docx) care să conţină comentarii şi concluzii, acolo unde sunt cerute.

Punctaj: 1. 0.5p , 2. 0.75p, 3. 0.25p 4. 0.25p 5. 0.25p **BONUS:** 0.5 p

1 Problemă

1. Generaţi 10000 de variabile aleatoare folosind **metoda transformării inverse** pentru repartiţiile definite mai jos:

   a) Repartiţia logistică are densitatea de probabilitate $f(x) = \frac{e^{-\frac{x-\mu}{\beta}}}{\beta\left(1+e^{-\frac{x-\mu}{\beta}}\right)^2}$ și funcția de repartiție $F(x) = \frac{1}{1+e^{-\frac{x-\mu}{\beta}}}$.

   b) Repartiţia Cauchy are densitatea de probabilitate $f(x) = \frac{1}{\pi\sigma}\frac{1}{1+\left(\frac{x-\mu}{\sigma}\right)^2}$ și funcția de repartiție $F(x) = \frac{1}{2} + \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right)$.

Comparaţi rezultatele obţinute cu valorile date de funcţiile `rlogis` şi respectiv `rcauchy` (funcţiile de repartiţie predefinite în R pentru repartiţiile logistică şi respectiv Cauchy). Ilustraţi grafic aceste rezultate.

2. Folosiţi **metoda respingerii** pentru a genera observaţii din densitatea de probabilitate definită prin $f(x) \propto e^{-\frac{x^2}{2}}\left[\sin(6x)^2 + 3\cos(x)^2\sin(4x)^2 + 1\right]$[1] parcurgând paşii următori:

   a) Reprezentaţi grafic $f(x)$ şi arătaţi că aceasta este mărginită de $Mg(x)$ unde $g(x)$ este densitatea de probabilitate a repartiţiei normale standard. Determinaţi o valoare potrivită pentru constanta M, chiar dacă nu este optimă[2].

   b) Generaţi 2500 de observaţii din densitatea de mai sus folosind metoda respingerii.

   c) Deduceţi, pornind de la rata de acceptare a acestui algoritm, o aproximare a *constantei de normalizare* a lui $f(x)$, apoi comparaţi histograma valorilor generate cu reprezentarea grafică a lui $f(x)$ normalizată.

3. **Metoda Monte Carlo pentru aproximarea unor integrale**

Punctul de plecare al metodei Monte Carlo pentru aproximarea unei integrale este nevoia de a evalua expresia $\mathbb{E}_f[h(X)] = \int_\chi h(x)f(x)\,dx$, unde χ reprezintă mulţimea de valori a variabile aleatoare X (care este, de obicei, suportul densităţii f).

Principiul metodei Monte Carlo este de a aproxima expresia de mai sus cu media de selecţie $\bar{h}_n = \frac{1}{n}\sum_{j=1}^n h(X_j)$ pornind de la un eşantion $X_1, X_2, \ldots, X_n$ din densitatea f, întrucât aceasta converge a.s. către $\mathbb{E}_f[h(X)]$, conform legii numerelor mari. Mai mult, atunci când $h(X)^2$ are medie finită viteza de convergenţă a lui $\bar{h}_n$ poate fi determinată întrucât convergenţa este de ordin $\mathcal{O}(\sqrt{n})$ iar varianţa aproximării este $Var(\bar{h}_n) = \frac{1}{n}\int_\chi \left(h(x) - \mathbb{E}_f[h(X)]\right)^2 f(x)\,dx$, cantitate care poate fi de asemenea aproximată prin $v_n = \frac{1}{n^2}\sum_{j=1}^n \left(h(X_j) - \bar{h}_n\right)^2$.

Mai precis, datorită teoremei limită centrală, pentru un n suficient de mare expresia

[1] Notația $\propto$ înseamnă că $f(x)$ este proporțională cu expresia din dreapta
[2] **Indiciu:** Folosiţi funcţia `optimise` din R

$$\frac{\bar{h}_n - \mathbb{E}_f[h(X)]}{\sqrt{v_n}}$$

poate fi aproximată cu o normală standard, ceea ce conduce la posibilitatea construirii unui test de convergenţă şi a unor margini pentru aproximarea lui $\mathbb{E}_f[h(X)]$.

Cerință:

Pentru funcţia $h(x) = (\cos(50x) + \sin(20x))^2$ construiţi o aproximare a integralei acesteia pe intervalul [0,1] după cum urmează: valoarea integralei poate fi văzută ca fiind media funcţiei $h(X)$ unde X este repartizată uniform pe $[0, 1]$. Urmărind algoritmul dat de metoda Monte Carlo construiţi programul R care determină aproximarea acestei integrale. Comparaţi rezultatul obţinut cu cel analitic și cu cel numeric obținut folosind funcția `integrate`. Ataşaţi reprezentările grafice pe care le consideraţi utile pentru a putea observa eficienţa metodei.

4. Pornind de la un set de date din cele oferite de R (*fiecare student îşi alege singur setul de date*) realizaţi o analiză de tipul "statistică descriptivă" a acestora (medie, cuartile, histogramă, boxplot, boxplot comparativ, identificare de outlieri, etc.). Documentaţi corespunzător acest proces şi explicaţi ce concluzii puteţi trage în urma acestei analize asupra datelor. Includeţi în fişier şi o descriere a datelor şi sursa lor.
5. Reprezentaţi grafic densitatea de probabilitate şi funcţia de repartiţie (în câte două grafice alăturate) pentru un set de parametri la alegere (cel puţin 4 cu reprezentările realizate în acelaşi grafic) pentru repartiţiile **Student**, **Fisher** şi χ^2. Identificaţi proprietăţile lor şi daţi un exemplu, construind un program în R, de utilizarea lor în testarea unor ipoteze statistice.

BONUS:

6. Construiţi un script R util în unul din experimentele de Machine Learning disponibile aici: https://studio.azureml.net/ (vezi exemplu la ultimul laborator !!!)