Seminar 11

Partea 3: Mixuri de probabilitate

11.1 Mixturi de probabilitate

Fie $p_1, p_2, \ldots, p_n \in (0,1)$ astfel încât $p_1 + p_2 + \cdots + p_n = 1$. Dacă $F_1, F_2, \ldots, F_n$ sunt funcţiile de repartiţie ale variabilelor aleatoare $X_1, X_2, \ldots, X_n$, atunci funcţia

$$F = p_1F_1 + p_2F_2 + \cdots + p_nF_n$$

este o funcţie de repartiţie, numită repartiţie compusă.

Analog, dacă $f_1, f_2, \ldots, f_n$ sunt densităţile de probabilitate ale variabilelor aleatoare $X_1, X_2, \ldots, X_n$, atunci

$$f = p_1f_1 + p_2f_2 + \cdots + p_nf_n$$

este o densitate de probabilitate, numită densitate compusă.

Definiţia 11.1.1 O variabilă aleatoare X ce are densitatea de probabilitate compusă f sau funcţia de repartiţie compusă F, se numeşte *mixtură de distribuţii de probabilitate* sau, mai simplu, *mixtură de probabilitate*.

Dacă variabila aleatoare X are densitatea $f = p_1f_1+p_2f_2+\ldots+p_nf_n$, aceasta înseamnă că X are densitatea f_1 cu probab. p_1, ..., X are densitatea f_n cu probab. p_n. Reprezentând fiecare densitate f_k prin indicele său k, asociem unei densităţi compuse o variabilă aleatoare discretă:

$$H = \begin{pmatrix} 1 & 2 & \ldots & k & \ldots & n \\ p_1 & p_2 & \ldots & p_k & \ldots & p_n \end{pmatrix}. \tag{11.1}$$

Selectând la întâmplare o valoare a lui H este echivalent cu a selecta la întâmplare, cu aceeaşi probabilitate, o densitate de probabilitate din cele n.

Exemplul 1. O variabilă aleatoare X a cărei densitate de probabilitate

$$f = p_1f_1 + p_2f_2 + \cdots + p_nf_n, \text{unde } f_i(x) = \begin{cases} \dfrac{1}{\theta_i}e^{-x/\theta_i}, & \text{dacă } x \geq 0, \\ 0, & \text{dacă } x < 0, \end{cases} \tag{11.2}$$

este compusa a n densităţi ale distribuţiei exponenţiale de parametrii $\theta_1, \theta_2, \ldots, \theta_n$ se numeşte variabilă aleatoare hiperexponenţială. Variabilele aleat. hiperexponenţiale modelează durata serviciului procesorului. Acestea se folosesc în simularea reţelelor de cozi.

Dăm în continuare un exemplu de aplicare a mixturii în analiza algoritmilor:

Exemplul 2. Considerăm instrucţiunea `if-else`:

```
if(B) then I1 else I2;
```

Fie X_1, X_2 variabilele aleatoare exponenţial distribuite care dau timpul de execuţie al grupului de instrucţiuni I_1, respectiv I_2. Fie $p \in (0,1)$ probabilitatea ca expresia booleană `B` să fie adevărată. Densitatea de probabilitate a variabilei ce dă timpul total de execuţie al blocului `if-else` este $f = pf_1 + (1-p)f_2$, unde f_1, f_2 sunt densităţi exponenţiale.

Dacă $p = 0.75$ şi $M(X_1) = 20$ milisecunde, $M(X_2) = 40$ milisecunde, să determinăm media timpului de execuţie pentru `if-else`.

Se ştie că dacă $X \sim$Exp(θ), atunci $M(X) = \theta$. Prin urmare, $\theta_1 = 20$, respectiv $\theta_2 = 40$ şi, deci, densitatea de probabilitate a timpului de execuţie a lui `if-else` este

$$f(x) = \begin{cases} 0.75\dfrac{1}{20}e^{-x/20} + 0.25\dfrac{1}{40}e^{-x/40}, & \text{dacă } x \geq 0, \\ 0, & \text{dacă } x < 0. \end{cases}$$

Astfel, media timpului de execuţie este

$$\begin{aligned} M(X) &= \int_{-\infty}^{\infty} xf(x)dx \\ &= \int_{0}^{\infty} x\left(0.75\frac{1}{20}e^{-x/20} + 0.25\frac{1}{40}e^{-x/40}\right)dx \\ &= 0.75M(X_1) + 0.25M(X_2). \end{aligned}$$

Mixturile de densităţi normale, numite şi *mixturi Gaussiene*, se folosesc, de exemplu, în analiza şi procesarea imaginilor, recunoaşterea formelor, în clusterizare etc.

Pe lângă mixturile de distribuţii de probabilitate continue se folosesc şi mixturi de distribuţii discrete.

Exemplul 3. Facebook monitorizează atitudinea unui user faţă de postările pe wall-uri şi îi asociază un număr de reacţii ce sunt modelate de o mixtură Poisson. Pentru a înţelege mai uşor, discutăm cazul cel mai simplu când userul are reacţia R_1 cu probabilitatea p_1 şi reacţia R_2 cu probabilitatea p_2, $p_1 + p_2 = 1$. Reacţia R_1 constă din linkuri la articole din *Times New Roman* cu rata λ_1/oră şi reacţia R_2, ce constă din like-uri la pozele amicilor, cu rata λ_2/oră. Astfel, numărul de reacţii/manifestări ale userului pe oră este o variabilă aleatoare X ce are ca distribuţie de probabilitate mixtura Poisson:

$$P_X(k) := P(X = k) = p_1 e^{-\lambda_1}\frac{\lambda_1^k}{k!} + p_2 e^{-\lambda_2}\frac{\lambda_2^k}{k!}.$$

ATENŢIE!!! Mixtura de distribuţii de probabilitate **nu** înseamnă că variabila X este de forma $X = p_1X_1 + p_2X_2$, cu $X_i \sim Poiss(\lambda_i)$, $i = 1, 2$, ci că X are distribuţia Poisson de rată λ_1 cu probab. p_1, respectiv X are distribuţia Poisson de rată λ_2 cu probab. p_2.

O mixtura modelează foarte bine comportamentul uman, care nu este constant, ci în funcţie de diverse circumstanţe "umanul" are o reacţie sau alta.

Exercitiu: Dacă $p_1 = 0.6$ şi $p_2 = 0.4$, iar $\lambda_1 = 8$, $\lambda_2 = 10$, să se calculeze probabilitatea ca $(X < 3)$ (probabilitatea evenimentului ca într-o oră userul să înregistreze mai puţin de 3 manifestări din cele monitorizate) şi numărul mediu de manifestări pe oră.

Mixturile de distribuţii m_i-Erlang, $i = \overline{1,n}$, de parametrii $\theta_1, \theta_2, \ldots, \theta_n$,

$$f(x) = p_1 \frac{x^{m_1-1}e^{-x/\theta_1}}{\theta^{m_1}(m_1-1)!} + p_2 \frac{x^{m_2-1}e^{-x/\theta_2}}{\theta^{m_2}(m_2-1)!} + \cdots + p_n \frac{x^{m_n-1}e^{-x/\theta_n}}{\theta^{m_n}(m_n-1)!}$$

se folosesc ca modele pentru aplicaţii în reţele wireless şi sisteme de calcul mobil. O variabilă aleatoare X ce are o astfel de densitate de probabilitate se zice că are distribuţie hyper-Erlangen. Exemplu de astfel de model: reţelele wireless de generaţia a treia oferă servicii integrate de telefonie, date, multimedia etc. Dacă o reţea wireless cu structură celulară oferă n tipuri de servicii şi rata medie a sosirii apelurilor de tip i într-o celulă, în unitatea de timp, este λ_i, iar $f_i(x)$ este densitatea de probabilitate a duratei de servire a cererii de tip i într-o celulă, atunci

$$f(x) = \underbrace{\frac{\lambda_1}{\sum_{j=1}^{n} \lambda_j}}_{p_1} f_1(x) + \underbrace{\frac{\lambda_2}{\sum_{j=1}^{n} \lambda_j}}_{p_2} f_2(x) + \cdots + \underbrace{\frac{\lambda_n}{\sum_{j=1}^{n} \lambda_j}}_{p_n} f_n(x)$$

este densitatea de probabilitate a duratei de servire a celulei respective.

Exploatând interpretarea dată mai sus distribuţiei compuse, putem da următoarea modalitate de generare a N valori de observaţie asupra variabilei aleatoare X ce are distribuţie de probabilitate compusă:

```
for i = 1 : N {
k ← simulatorul variabilei aleatoare discrete H;
x_i ← simulatorul distribuţiei f_k;
}
```

În şirul generat (x_i), $i = \overline{1,N}$, o proporţie de aproximativ $100\,p_1\%$ valori vor fi din legea f_1, $100\,p_2\%$ valori vor fi din legea f_2 etc.