Curs 10

Cristian Niculescu

1 Estimări de verosimilitate maximă

1.1 Scopurile învăţării

1. Să poată să definească funcţia de verosimilitate pentru date dintr-un model parametric.
2. Să poată să calculeze estimarea de verosimilitate maximă a parametrului necunoscut (parametrilor necunoscuţi).

1.2 Introducere

Presupunem că ştim că avem date constând din valori $x_1, ..., x_n$ dintr-o repartiţie exponenţială. Întrebarea rămâne: care repartiţie exponenţială?!
Ne-am referit uneori la repartiţia exponenţială sau la repartiţia binomială sau la repartiţia normală. De fapt, repartiţia exponenţială $\exp(\lambda)$ nu este o singură repartiţie, ci mai degrabă o familie de repartiţii cu un parametru. Fiecare valoare λ defineşte o repartiţie diferită din familie, cu pdf
$f_\lambda(x) = \lambda e^{-\lambda x}$ pe $[0, \infty)$. Similar, o repartiţie binomială $\text{bin}(n, p)$ este determinată de cei 2 parametri n şi p, iar o repartiţie normală $N(\mu, \sigma^2)$ este determinată de cei 2 parametri μ şi σ^2 (sau echivalent, μ şi σ). Familiile parametrizate de repartiţii sunt adesea numite repartiţii parametrice sau modele parametrice.
Adesea suntem puşi în faţa situaţiei de a avea date aleatoare despre care ştim (sau credem) că provin dintr-un model parametric, ai cărui parametri nu-i ştim. De exemplu, într-o alegere între 2 candidaţi, datele sondajului sunt dintr-o repartiţie Bernoulli(p) cu parametru necunoscut p. În acest caz ne-ar plăcea să folosim datele pentru a estima valoarea parametrului p, deoarece aceasta prezice rezultatul alegerii. Similar, presupunând că lungimea gestaţiei are o repartiţie normală, ne-ar plăcea să folosim datele lungimilor gestaţionale dintr-o selecţie aleatoare de sarcini pentru a trage concluzii despre valorile parametrilor μ şi σ^2.
Scopul nostru până acum a fost să calculăm probabilitatea datelor provenind

dintr-un model parametric cu parametri cunoscuţi. Deducţia statistică întoarce asta pe dos: estimăm probabilitatea parametrilor dat fiind un model parametric şi date observate din el. Valorile parametrului sunt văzute în mod natural ca ipoteze, deci de fapt estimăm probabilitatea diverselor ipoteze ştiind datele.

1.3 Estimări de verosimilitate maximă

Sunt mai multe metode de a estima parametrii necunoscuţi din date. Întâi considerăm estimarea de verosimilitate maximă (MLE, prescurtare de la maximum likelihood estimate), care răspunde la întrebarea:
Pentru ce valoare a parametrului au datele observate cea mai mare probabilitate?
MLE este un exemplu de estimare punctuală deoarece dă o singură valoare pentru parametrul necunoscut. 2 avantaje ale MLE sunt că este adesea uşor de calculat şi este de acord cu intuiţia noastră în exemple simple.
Exemplul 1. Aruncăm o monedă de 100 de ori. Dat fiind că au fost 55 de aversuri, aflaţi estimarea de verosimilitate maximă pentru probabilitatea p a aversului la o singură aruncare.
Înainte de a rezolva problema, stabilim nişte notaţii şi termeni.
Putem gândi calcularea numărului de aversuri în 100 de aruncări ca un experiment. Pentru o valoare dată a lui p, probabilitatea de a obţine 55 de aversuri în acest experiment este probabilitatea binomială

$$P(55 \text{ aversuri}) = C_{100}^{55} p^{55} (1-p)^{45}.$$

Probabilitatea de a obţine 55 de aversuri depinde de valoarea lui p, deci hai să-l includem pe p folosind notaţia probabilităţii condiţionate:

$$P(55 \text{ aversuri } |p) = C_{100}^{55} p^{55} (1-p)^{45}.$$

Citim $P(55 \text{ aversuri } |p)$
"probabilitatea a 55 de aversuri dat fiind p,"
sau mai precis
"probabilitatea a 55 de aversuri dat fiind că probabilitatea unui avers la o singură aruncare este p."
Iată câţiva termeni standard în statistică:
Experiment: Aruncăm moneda de 100 de ori şi calculăm numărul de aversuri.
Date: Datele sunt rezultatul experimentului. În acest caz sunt "55 de aversuri".

Parametru (Parametri) de interes: Suntem interesaţi de valoarea parametrului necunoscut p.
Verosimilitatea sau funcţia de verosimilitate: aceasta este $P(\text{date}|p)$. Observăm că este o funcţie de date şi parametrul p. În acest caz verosimilitatea este

$$P(55 \text{ aversuri } |p) = C_{100}^{55} p^{55}(1-p)^{45}.$$

Observaţii: 1. Verosimilitatea $P(\text{date}|p)$ se schimbă dacă dacă parametrul de interes p se schimbă.
2. Atenţie la definiţie! O sursă tipică de confuzie este a confunda verosimilitatea $P(\text{date}|p)$ cu $P(p|\text{ date})$. Ştim că $P(\text{date}|p)$ şi $P(p|\text{ date})$ sunt de obicei diferite.
Definiţie: Ştiind datele, estimarea de verosimilitate maximă (MLE) pentru parametrul p este valoarea care maximizează verosimilitatea $P(\text{date}|p)$. Adică, MLE este valoarea lui p pentru care datele sunt cele mai probabile.
Răspuns: Am văzut că verosimilitatea este

$$P(55 \text{ aversuri } |p) = C_{100}^{55} p^{55}(1-p)^{45}.$$

Folosim notaţia $\hat{p}$ pentru MLE. Folosim analiza matematică pentru a o afla, luând derivata funcţiei de verosimilitate şi egalând-o cu 0.

$$\frac{d}{dp} P(\text{date } |p) = C_{100}^{55}(55p^{54}(1-p)^{45} - 45p^{55}(1-p)^{44}) = 0.$$

Rezolvând ecuaţia obţinem

$$\begin{aligned} &55p^{54}(1-p)^{45} = 45p^{55}(1-p)^{44} \\ &55(1-p) = 45p \text{ (deoarece } p \in (0,1)) \\ &55 = 100p \\ &\text{MLE este } \hat{p} = 0.55. \end{aligned}$$

Observaţii: 1. MLE pentru p s-a dovedit a fi exact fracţia de aversuri pe care am văzut-o în datele noastre.
2. MLE este calculată din date. Adică, este o statistică.
3. Oficial ar trebui să verificăm că punctul critic este într-adevăr un maxim. Putem face aceasta cu testul derivatei a 2-a.

1.3.1 Log verosimilitate

Adesea este mai uşor să lucrăm cu logaritmul natural al funcţiei de verosimilitate. Pentru scurtime acesta este numit simplu log verosimilitate. Deoarece $\ln(x)$ este o funcţie strict crescătoare, maximele verosimilităţii şi log verosimilităţii

coincid.
Exemplul 2. Refacem exemplul precedent folosind log verosimilitatea.
Răspuns: Aveam verosimilitatea $P(55 \text{ aversuri } |p) = C_{100}^{55} p^{55}(1-p)^{45}$. De aceea log verosimilitatea este

$$\ln(P(55 \text{ aversuri } |p)) = \ln(C_{100}^{55}) + 55\ln(p) + 45\ln(1-p).$$

Maximizarea verosimilităţii este acelaşi lucru cu maximizarea log verosimilităţii. Verificăm că analiza matematică ne dă acelaşi răspuns ca mai sus:

$$\begin{aligned}
\frac{d}{dp}(\text{log verosimilitate}) &= \frac{d}{dp}[\ln(C_{100}^{55}) + 55\ln(p) + 45\ln(1-p)] \\
&= \frac{55}{p} - \frac{45}{1-p} = 0 \\
&\Longrightarrow 55(1-p) = 45p \\
&\Longrightarrow \hat{p} = 0.55.
\end{aligned}$$

1.3.2 Verosimilitate maximă pentru repartiţii continue

Pentru repartiţii continue, folosim funcţia densitate de probabilitate pentru a defini verosimilitatea.
Exemplul 3. Becuri
Presupunem că durata de funcţionare a unui anumit tip de becuri este modelată de o repartiţie exponenţială cu parametru necunoscut λ. Testăm 5 becuri şi aflăm că au duratele de funcţionare de 2, 3, 1, 3 respectiv 4 ani. Care este MLE pentru λ?
Răspuns: Avem nevoie să fim atenţi cu notaţia noastră. Cu 5 valori diferite cel mai bine este să folosim indici. Fie X_i durata de funcţionare al celui de-al i-lea bec şi fie x_i valoarea pe care o ia X_i. Atunci fiecare X_i are pdf $f_{X_i}(x_i) = \lambda e^{-\lambda x_i}$. Presupunem că duratele de funcţionare ale becurilor sunt independente, deci pdf comună este produsul densităţilor individuale:

$$f(x_1, x_2, x_3, x_4, x_5|\lambda) = (\lambda e^{-\lambda x_1})(\lambda e^{-\lambda x_2})(\lambda e^{-\lambda x_3})(\lambda e^{-\lambda x_4})(\lambda e^{-\lambda x_5}) = \lambda^5 e^{-\lambda(x_1+x_2+x_3+x_4+x_5)}.$$

Observăm că scriem aceasta ca o densitate condiţionată, deoarece depinde de λ. Văzând datele ca fixate şi λ ca variabilă, această densitate este funcţia de verosimilitate. Datele noastre au avut valorile

$$x_1 = 2,\ x_2 = 3,\ x_3 = 1,\ x_4 = 3,\ x_5 = 4.$$

Deci funcţiile de verosimilitate şi log verosimilitate cu datele noastre sunt

$$f(2,3,1,3,4|\lambda) = \lambda^5 e^{-13\lambda},\ \ \ln(f(2,3,1,3,4|\lambda)) = 5\ln(\lambda) - 13\lambda.$$

În sfârşit, folosim analiza matematică pentru a afla MLE:

$$\frac{d}{d\lambda}(\text{log verosimilitate}) = \frac{5}{\lambda} - 13 = 0 \implies \hat{\lambda} = \frac{5}{13}.$$

Observaţii: 1. În acest exemplu am folosit o literă mare pentru o variabilă aleatoare şi litera mică corespunzătoare pentru valoarea pe care o ia.
2. MLE pentru λ s-a dovedit a fi inversa mediei de selecţie $\overline{x}$, deci $X \sim \exp(\hat{\lambda})$ satisface $E(X) = \overline{x}$.

Exemplul 4. Repartiţii normale

Presupunem că datele $x_1, x_2, ..., x_n$ sunt dintr-o repartiţie $N(\mu, \sigma^2)$, unde μ şi σ sunt necunoscute. Aflaţi estimarea de verosimilitate maximă pentru perechea (μ, σ^2).

Răspuns: Formulăm aceasta în termeni de variabile aleatoare şi densităţi. Fie $X_1, ..., X_n$ variabile aleatoare i.i.d. $N(\mu, \sigma^2)$ şi fie x_i valoarea pe care o ia X_i. Densitatea pentru fiecare X_i este

$$f_{X_i}(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Deoarece X_i sunt independente pdf comună a lor este produsul pdf-urilor individuale:

$$f(x_1, ..., x_n|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum\limits_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Pentru datele fixate $x_1, ..., x_n$, verosimilitatea şi log verosimilitatea sunt

$$f(x_1, ..., x_n|\mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum\limits_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}},$$

$$\ln(f(x_1, ..., x_n|\mu, \sigma)) = -n\ln(\sqrt{2\pi}) - n\ln(\sigma) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Deoarece $\ln(f(x_1, ..., x_n|\mu, \sigma))$ este o funcţie de 2 variabile μ, σ, folosim derivatele parţiale pentru a afla MLE. Valoarea uşor de aflat este $\hat{\mu}$:

$$\frac{\partial \ln f(x_1, ..., x_n|\mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0 \implies \sum_{i=1}^{n} x_i = n\mu \implies \hat{\mu} = \frac{\sum\limits_{i=1}^{n} x_i}{n} = \overline{x}.$$

Pentru a afla $\hat{\sigma}$ derivăm şi rezolvăm în raport cu σ:

$$\frac{\partial \ln f(x_1, ..., x_n|\mu, \sigma)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0 \implies \hat{\sigma^2} = \frac{\sum\limits_{i=1}^{n}(x_i-\mu)^2}{n}.$$

Ştim deja $\hat{\mu} = \overline{x}$, deci folosim aceasta ca valoare pentru μ în formula pentru $\hat{\sigma}$. Obţinem estimările de verosimilitate maximă

$$\begin{aligned} \hat{\mu} &= \overline{x} && = \text{media datelor} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 && = \text{dispersia datelor.} \end{aligned}$$

Exemplul 5. Repartiţii uniforme
Presupunem că datele noastre $x_1, ..., x_n$ sunt independente dintr-o repartiţie uniformă $U(a, b)$. Aflaţi MLE pentru a şi b.
Răspuns: Acest exemplu este diferit de cele precedente prin aceea că nu folosim analiza matematică pentru a afla MLE. Densitatea pentru $U(a, b)$ este $\frac{1}{b-a}$ pe $[a, b]$. De aceea funcţia noastră de veromisimilitate este

$$f(x_1, ..., x_n|a, b) = \begin{cases} \left(\frac{1}{b-a}\right)^n, & \text{dacă toţi } x_i \text{ sunt în intervalul } [a, b] \\ 0, & \text{altfel.} \end{cases}$$

Aceasta este maximizată făcând $b - a$ cât mai mic posibil. Singura restricţie este aceea că intervalul $[a, b]$ trebuie să conţină toate datele. Astfel, MLE pentru perechea (a, b) este

$$\hat{a} = \min(x_1, ..., x_n), \;\; \hat{b} = \max(x_1, ..., x_n).$$

Exemplul 6. Metoda captură/recaptură
Metoda captură/recaptură este un mod de a estima mărimea unei populaţii în sălbăticie. Metoda presupune că fiecare animal din populaţie este egal posibil de a fi capturat de o cursă.
Presupunem că 10 animale sunt capturate, etichetate şi eliberate. Câteva luni mai târziu, 20 de animale sunt capturate, examinate şi eliberate. 4 dintre aceste 20 sunt găsite etichetate. Estimaţi mărimea populaţiei sălbatice folosind MLE pentru probabilitatea că un animal sălbatic este etichetat.
Răspuns: Parametrul nostru necunoscut n este numărul de animale din sălbăticie. Datele noastre sunt că 4 dintre cele 20 de animale capturate ultima dată au fost etichetate (şi că sunt 10 animale etichetate). Funcţia de verosimilitate este

$$P(\text{date } | n \text{ animale}) = \frac{C_{n-10}^{16} \cdot C_{10}^{4}}{C_n^{20}}.$$

(Numărătorul este numărul de moduri de a alege 16 animale din cele $n - 10$ neetichetate ori numărul de moduri de a alege 4 din cele 10 animale etichetate. Numitorul este numărul de moduri de a alege 20 de animale din întreaga populaţie de n.) Putem folosi R (cu comanda `dhyper(16,n-10,10,20)` pentru diverşi n) pentru a calcula că funcţia de verosimilitate este maximizată când $n = 50$. Aceasta ar trebui să aibă un sens. Acesta spune că cea mai bună estimare a noastră este când fracţia din toate animalele care sunt etichetate este 10/50, care este egală cu fracţia din animalele capturate ultima oară care sunt etichetate (4/20).

Exemplul 7. Hardy-Weinberg. Presupunem că o genă particulară apare ca una din 2 alele (A şi a), unde alela A are frecvenţa θ în populaţie. Adică, o copie aleatoare a genei este A cu probabilitatea θ şi a cu probabilitatea $1 - \theta$. Deoarece un genotip diploid este format din 2 gene, probabilitatea fiecărui genotip este dată de:

genotip	AA	Aa	aa
probabilitate	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

Presupunem că testăm un eşantion aleator de oameni şi aflăm că: k_1 sunt AA, k_2 sunt Aa şi k_3 sunt aa. Aflaţi MLE pentru θ.

Răspuns: Funţia de verosimilitate este dată de

$$P(k_1, k_2, k_3|\theta) = C_{k_1+k_2+k_3}^{k_1} \cdot C_{k_2+k_3}^{k_2} \cdot C_{k_3}^{k_3} \cdot \theta^{2k_1}(2\theta(1-\theta))^{k_2}(1-\theta)^{2k_3}.$$

Log verosimilitatea este dată de

$$\text{constantă } + 2k_1 \ln(\theta) + k_2 \ln(\theta) + k_2 \ln(1-\theta) + 2k_3 \ln(1-\theta)$$

Egalăm derivata cu 0:

$$\frac{2k_1 + k_2}{\theta} - \frac{k_2 + 2k_3}{1-\theta} = 0$$

Rezolvând ecuaţia în θ, aflăm că MLE este

$$\hat{\theta} = \frac{2k_1 + k_2}{2k_1 + 2k_2 + 2k_3},$$

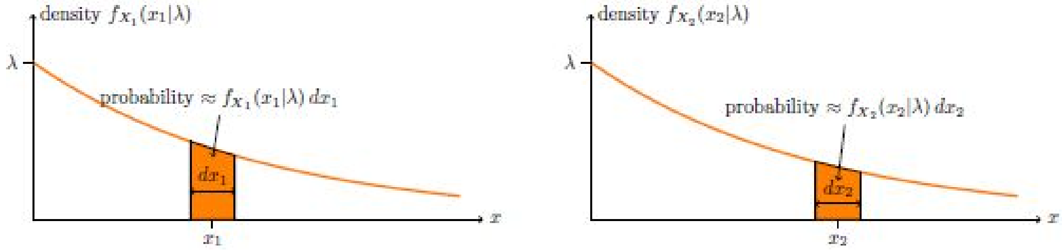
care este pur şi simplu fracţia de alele A dintre toate genele din populaţia selectată.

1.4 De ce folosim densitatea pentru a afla MLE pentru repartiţii continue

Ideea pentru estimarea de verosimilitate maximă este de a afla valoarea parametrului (parametrilor) pentru care datele au cea mai mare probabilitate.

Exemplul 8. Presupunem că avem 2 becuri a căror durată de funcţionare

are o repartiţie exp(λ). Presupunem de asemenea că măsurăm independent duratele lor de funţionare şi obţinem datele $x_1 = 2$ ani şi $x_2 = 3$ ani. Aflaţi valoarea lui λ care maximizează probabilitatea acestor date.
Răspuns: Principalul paradox cu care avem de-a face este că pentru o repartiţie continuă probabilitatea unei singure valori, să spunem $x_1 = 2$, este 0. Rezolvăm acest paradox reamintind că o singură măsurătoare înseamnă de fapt un domeniu de valori, de exemplu aici putem verifica becul o dată pe zi. Deci data $x_1 = 2$ ani înseamnă de fapt că x_1 este undeva într-un domeniu de o zi în jurul a 2 ani.
Dacă domeniul este mic, îl numim dx_1. Probabilitatea că X_1 este în domeniu este aproximativ $f_{X_1}(x_1|\lambda)dx_1$. Acest fapt este ilustrat de figura de mai jos. Valoarea x_2 a datelor este tratată exact la fel.



Relaţia uzuală între densitate şi probabilitate pentru domenii mici.

Deoarece datele sunt colectate independent, probabilitatea comună este produsul probabilităţilor individuale, adică

$$P(X_1 \text{ în domeniu}, X_2 \text{ în domeniu}|\lambda) \approx f_{X_1}(x_1|\lambda)dx_1 \cdot f_{X_2}(x_2|\lambda)dx_2.$$

Folosind valorile $x_1 = 2$ şi $x_2 = 3$ şi formula pentru o pdf exponenţială avem

$$P(X_1 \text{ în domeniu}, X_2 \text{ în domeniu}|\lambda) \approx \lambda e^{-2\lambda}dx_1 \cdot \lambda e^{-3\lambda}dx_2 = \lambda^2 e^{-5\lambda}dx_1dx_2.$$

Acum, deoarece avem o probabilitate autentică, putem căuta valoarea pentru λ care o maximizează. Privind formula de mai sus vedem că factorul dx_1dx_2 nu joacă niciun rol în aflarea maximului. Deci pentru MLE îl eliminăm şi numim pur şi simplu densitatea verosimilitate:

$$\text{verosimilitatea} = f(x_1, x_2|\lambda) = \lambda^2 e^{-5\lambda}.$$

Valoarea lui λ care maximizează aceasta este aflată la fel ca în exemplul de mai sus. Ea este $\hat{\lambda} = 2/5$.

1.5 Proprietăţi ale MLE

MLE se comportă bine la transformări. Adică, dacă $\hat{p}$ este MLE pentru p şi g este o funcţie injectivă, atunci $g(\hat{p})$ este MLE pentru $g(p)$. De exemplu,

dacă $\hat{\sigma}$ este MLE pentru deviaţia standard σ, atunci $(\hat{\sigma})^2$ este MLE pentru dispersia σ^2.
În plus, MLE este asimptotic nedeplasat şi are asimptotic dispersie minimă. Pentru a explica aceste noţiuni, observăm că MLE este ea însăşi o variabilă aleatoare deoarece datele sunt aleatoare şi MLE este calculată din date. Fie $x_1, x_2, ...$ un şir de date dintr-o repartiţie cu parametrul p. Fie $\hat{p}_n$ MLE pentru p bazată pe datele $x_1, ..., x_n$.
Asimptotic nedeplasat înseamnă că atunci când numărul datelor creşte, media MLE converge la p. În simboluri: $E(\hat{p}_n) \to p$ când $n \to \infty$. Desigur, am vrea ca MLE să fie aproape de p cu probabilitate mare, nu doar în medie, deci cu cât mai mică este dispersia MLE, cu atât mai bine. Asimptotic dispersie minimă înseamnă că, atunci când numărul de date creşte, MLE are dispersia minimă dintre toţi estimatorii nedeplasaţi ai lui p. În simboluri: pentru orice estimator nedeplasat $\tilde{p}_n$ şi $\varepsilon > 0$ avem $Var(\tilde{p}_n) + \varepsilon > Var(\hat{p}_n)$ când $n \to \infty$.

2 Actualizare Bayesiană cu a priori discrete

2.1 Scopurile învăţării

1. Să poată aplica teorema lui Bayes pentru a calcula probabilităţi.
2. Să poată defini şi identifica rolurile probabilităţii a priori, verosimilităţii (termenul Bayes), probabilităţii a posteriori, datelor şi ipotezei în aplicarea teoremei lui Bayes.
3. Să poată folosi un tabel de actualizare Bayesiană pentru a calcula probabilităţile a posteriori.

2.2 Recapitularea teoremei lui Bayes

Reamintim că teorema lui Bayes ne permite să "inversăm" probabilităţile condiţionate. Dacă $\mathcal{H}$ şi $\mathcal{D}$ sunt evenimente, atunci:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}.$$

2.3 Terminologie şi teorema lui Bayes în formă tabelară

Exemplul 1. Sunt 3 tipuri de monede care au probabilităţi diferite de avers când sunt aruncate.
Tipul A: monedele sunt corecte cu probabilitatea aversului 0.5.
Tipul B: monedele sunt cu probabilitatea aversului 0.6.
Tipul C: monedele sunt cu probabilitatea aversului 0.9.

Presupunem că am un sertar cu 5 monede: 2 de tipul A, 2 de tipul B şi una de tipul C. Iau aleator o monedă din sertar. Fără a v-o arăta, o arunc o dată şi obţin avers. Care este probabilitatea ca ea să fie de tipul A? Tipul B? Tipul C?
Răspuns. Fie A, B şi C evenimentul că moneda aleasă a fost de tipul A, tipul B, respectiv tipul C. Fie $\mathcal{D}$ evenimentul că aruncarea este avers. Problema cere

$$P(A|\mathcal{D}),\ P(B|\mathcal{D}),\ P(C|\mathcal{D}).$$

Terminologie:
Experiment: Luăm aleator o monedă din sertar, o aruncăm şi înregistrăm rezultatul.
Date: rezultatul experimentului nostru. În acest caz evenimentul
$\mathcal{D}$ = "avers". Gândim $\mathcal{D}$ ca datele care dau probe pentru sau împotriva fiecărei ipoteze.
Ipoteze: testăm 3 ipoteze: moneda este de tipul A, B sau C.
Probabilitate a priori: probabilitatea fiecărei ipoteze înaintea aruncării monedei (colectării datelor). Deoarece sertarul are 2 monede de tipul A, 2 de tipul B şi una de tipul C avem

$$P(A) = 0.4,\ P(B) = 0.4,\ P(C) = 0.2.$$

Verosimilitate: (Aceasta este aceeaşi verosimilitate pe care am folosit-o pentru MLE.) Funcţia de verosimilitate este $P(\mathcal{D}|\mathcal{H})$, i.e., probabilitatea datelor presupunând că ipoteza este adevărată. Cel mai adesea considerăm datele ca fixate şi lăsăm ipotezele să varieze. De exemplu, $P(\mathcal{D}|A)$ = probabilitatea aversului dacă moneda este de tipul A. În cazul nostru verosimilităţile sunt

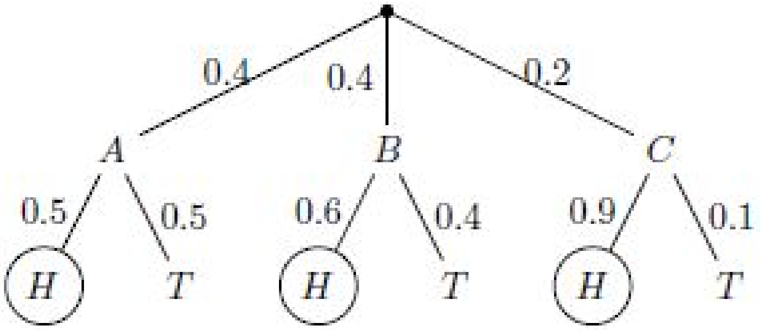
$$P(\mathcal{D}|A) = 0.5,\ P(\mathcal{D}|B) = 0.6,\ P(\mathcal{D}|C) = 0.9.$$

Numele verosimilitate este atât de bine stabilit în literatură că trebuie să vi-l predăm. Dar în limbajul colocvial verosimilitate şi probabilitate sunt sinonime. Aceasta duce adesea la confundarea funcţiei de verosimilitate cu probabilitatea unei ipoteze. De aceea preferăm să folosim numele termenul Bayes.
Probabilitatea a posteriori: probabilitatea (a posteriori) a fiecărei ipoteze cunoscând datele din aruncarea monedei.

$$P(A|\mathcal{D}),\ P(B|\mathcal{D}),\ P(C|\mathcal{D}).$$

Aceste probabilităţi a posteriori sunt cele cerute de problemă.
Folosim teorema lui Bayes pentru a calcula fiecare din probabilităţile a posteriori. Reamintim că datele $\mathcal{D}$ sunt că aruncarea a fost avers.

Organizăm probabilităţile într-un arbore:



Arborele de probabilităţi pentru alegerea şi aruncarea unei monede.

Teorema lui Bayes spune, de exemplu, $P(A|\mathcal{D}) = \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})}$. Numitorul $P(\mathcal{D})$ este calculat cu legea probabilităţii totale:

$$P(\mathcal{D}) = P(\mathcal{D}|A)P(A)+P(\mathcal{D}|B)P(B)+P(\mathcal{D}|C)P(C) = 0.5{\cdot}0.4+0.6{\cdot}0.4+0.9{\cdot}0.2 = 0.62.$$

Acum fiecare dintre cele 3 probabilităţi a posteriori poate fi calculată:

$$P(A|\mathcal{D}) = \frac{P(\mathcal{D}|A)P(A)}{P(\mathcal{D})} = \frac{0.5\cdot 0.4}{0.62} = \frac{0.2}{0.62} \approx 0.3226,$$
$$P(B|\mathcal{D}) = \frac{P(\mathcal{D}|B)P(B)}{P(\mathcal{D})} = \frac{0.6\cdot 0.4}{0.62} = \frac{0.24}{0.62} \approx 0.3871,$$
$$P(C|\mathcal{D}) = \frac{P(\mathcal{D}|C)P(C)}{P(\mathcal{D})} = \frac{0.9\cdot 0.2}{0.62} = \frac{0.18}{0.62} \approx 0.2903.$$

Observăm că probabilitatea totală $P(\mathcal{D})$ este aceeaşi la fiecare numitor şi este suma celor 3 numărători. Putem organiza toate acestea într-un tabel de actualizare Bayesiană ("hypothesis" = "ipoteză", "prior" = "a priori", "likelihood" = "verosimilitate", "numerator" = "numărător", "posterior" = "a posteriori"):

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
A	0.4	0.5	0.2	0.3226
B	0.4	0.6	0.24	0.3871
C	0.2	0.9	0.18	0.2903
total	1		0.62	1

Numărătorul Bayes este produsul dintre (probabilitatea) a priori şi verosimilitate. Vedem în fiecare din calculele de mai sus cu formula lui Bayes că probabilitatea a posteriori este obţinută împărţind numărătorul Bayes la $P(\mathcal{D}) = 0.62$. Vedem de asemenea că legea probabilităţii totale spune că $P(\mathcal{D})$ este suma elementelor de pe coloana numărătorului Bayes.
Actualizare Bayesiană: Procesul trecerii de la probabilitatea a priori $P(\mathcal{H})$ la a posteriori $P(\mathcal{H}|\mathcal{D})$ se numeşte actualizare Bayesiană. Actualizarea Bayesiană utilizează datele pentru a modifica înţelegerea noastră pentru probabilitatea fiecărei ipoteze posibile.

2.3.1 Lucruri importante de observat

1. Sunt 2 tipuri de probabilităţi: Tipul 1 este probabilitatea standard a datelor, de exemplu, probabilitatea aversurilor este $p = 0.9$. Tipul 2 este probabilitatea ipotezelor, de exemplu, probabilitatea că moneda aleasă este de tipul A, B sau C. Al 2-lea tip are valori a priori (înainte de date) şi a posteriori (după date).
2. Probabilităţile a posteriori (după date) pentru fiecare ipoteză sunt în ultima coloană. Vedem că moneda B este acum cea mai probabilă, cu toate că probabilitatea ei a scăzut de la 0.4 la 0.3871. Între timp, probabilitatea tipului C a crescut de la 0.2 la 0.2903.
3. Coloana numărătorului Bayes determină coloana probabilităţii a posteriori. Pentru a o calcula pe ultima, pur şi simplu am rescalat numărătorii Bayes astfel încât sumele lor să fie 1.
4. Dacă tot ce vrem este să aflăm cea mai probabilă ipoteză, numărătorul Bayes funcţionează la fel de bine ca a posteriori.
5. Suma de pe coloana verosimilităţii nu este 1. Funcţia de verosimilitate *nu* este o funcţie de probabilitate.
6. Probabilitatea a posteriori reprezintă un ”compromis” între verosimilitate şi a priori. Când calculăm a posteriori, o a priori mare poate fi scăzută de o verosimilitate mică şi o a priori mică poate fi mărită de o verosimilitate mare.
7. Estimarea de verosimilitate maximă (MLE) pentru exemplul 1 este ipoteza C, cu o verosimilitate $P(\mathcal{D}|C) = 0.9$. MLE este utilă, dar puteţi vedea din acest exemplu că nu este toată povestea, deoarece tipul B are cea mai mare probabilitate a posteriori.

Cu terminologia la îndemână, putem exprima teorema lui Bayes în diverse moduri:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})}$$

$$P(\text{ipoteză}|\text{date}) = \frac{P(\text{date}|\text{ipoteză})P(\text{ipoteză})}{P(\text{date})}.$$

Cu datele fixate, numitorul $P(\mathcal{D})$ serveşte doar la normalizarea probabilităţii a posteriori la 1. Astfel, putem exprima teorema lui Bayes ca o afirmaţie despre proporţionalitatea a 2 funcţii de $\mathcal{H}$ (i.e., a ultimelor 2 coloane ale tabelului).

$$P(\text{ipoteză}|\text{date}) \propto P(\text{date}|\text{ipoteză})P(\text{ipoteză}).$$

Aceasta conduce la cea mai elegantă formă a teoremei lui Bayes în contextul actualizării Bayesiene:

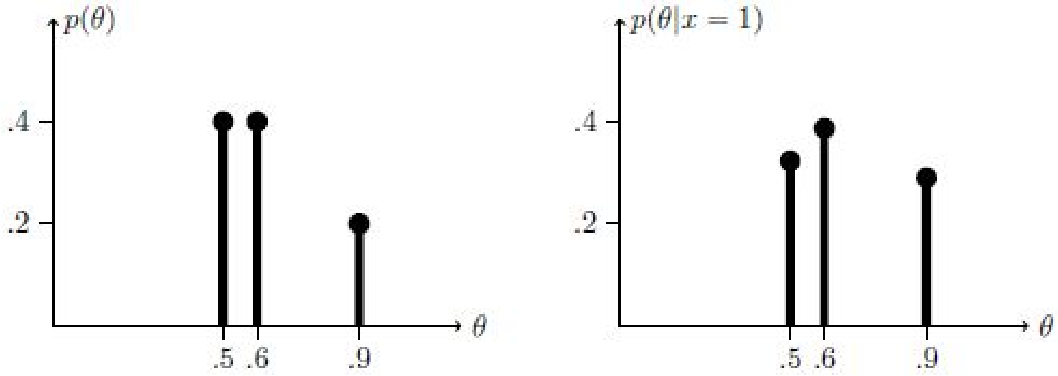
$$\text{a posteriori} \propto \text{verosimilitate} \times \text{a priori}.$$

2.3.2 Funcţii masă de probabilitate a priori şi a posteriori

Notaţii standard:
θ este valoarea ipotezei.
$p(\theta)$ este funcţia masă de probabilitate a priori a ipotezei.
$p(\theta|\mathcal{D})$ este funcţia masă de probabilitate a posteriori a ipotezei cunoscând datele.
$p(\mathcal{D}|\theta)$ este funcţia de verosimilitate. (Aceasta nu este o pmf!)
În exemplul 1 putem reprezenta cele 3 ipoteze A, B şi C prin $\theta = 0.5,\ 0.6,\ 0.9$. Pentru date $x = 1$ înseamnă avers şi $x = 0$ revers. Atunci probabilităţile a priori şi a posteriori din tabel definesc funcţiile masă de probabilitate a priori şi a posteriori. (”poster” = ”a posteriori”)

Hypothesis	θ	prior pmf $p(\theta)$	poster pmf $p(\theta\|x=1)$
A	0.5	$P(A) = p(0.5) = 0.4$	$P(A\|\mathcal{D}) = p(0.5\|x=1) = 0.3226$
B	0.6	$P(B) = p(0.6) = 0.4$	$P(B\|\mathcal{D}) = p(0.6\|x=1) = 0.3871$
C	0.9	$P(C) = p(0.9) = 0.2$	$P(C\|\mathcal{D}) = p(0.9\|x=1) = 0.2903$

Iată reprezentările pmf-urilor a priori şi a posteriori din exemplu:



Pmf a priori $p(\theta)$ şi pmf a posteriori $p(\theta|x=1)$ pentru exemplul 1.

Dacă datele ar fi diferite, atunci coloana verosimilităţii din tabnelul de actualizare Bayesiană ar fi diferită. Putem planifica pentru date diferite construind întregul tabel de verosimilitate în avans. În exemplul cu moneda sunt 2 posibilităţi pentru date: aruncarea este avers sau aruncarea este revers. Astfel, tabelul de verosimilitate complet are 2 coloane de verosimilitate:

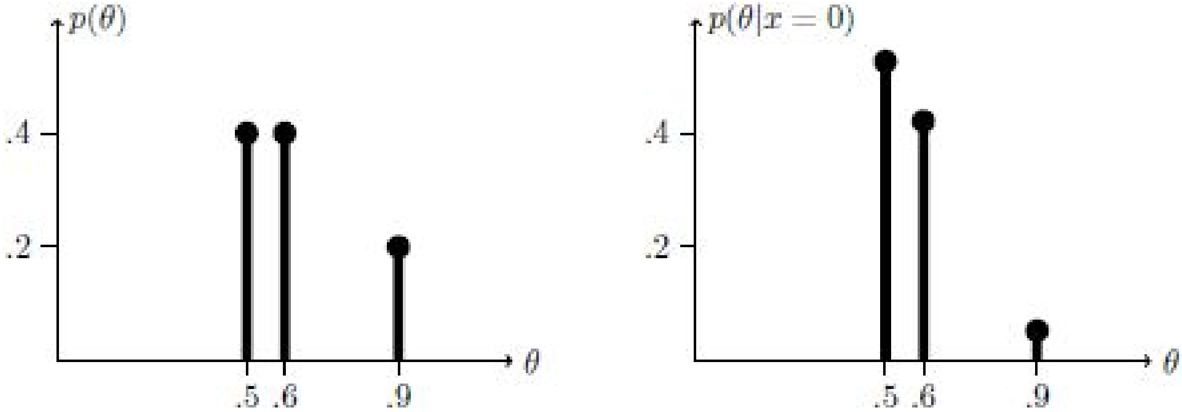
hypothesis	likelihood $p(x\|\theta)$	
θ	$p(x=0\|\theta)$	$p(x=1\|\theta)$
0.5	0.5	0.5
0.6	0.4	0.6
0.9	0.1	0.9

Exemplul 2. Folosind notaţia $p(\theta)$, etc., refaceţi exemplul 1 presupunând că aruncarea a fost revers.

Răspuns: Deoarece datele s-au schimbat, coloana verosimilităţii din tabelul de actualizare Bayesiană este acum pentru $x = 0$. Adică, trebuie să luăm coloana $p(x=0|\theta)$ din tabelul de verosimilitate.

hypothesis	prior	likelihood	Bayes numerator	posterior
θ	$p(\theta)$	$p(x=0\,\|\,\theta)$	$p(x=0\,\|\,\theta)p(\theta)$	$p(\theta\,\|\,x=0)$
0.5	0.4	0.5	0.2	0.5263
0.6	0.4	0.4	0.16	0.4211
0.9	0.2	0.1	0.02	0.0526
total	1		0.38	1

Acum probabilitatea tipului A a crescut de la 0.4 la 0.5263, în timp ce probabilitatea tipului C a scăzut de la 0.2 la 0.0526. Iată reprezentările corespunzătoare:



2.3.3 Hrană pentru minte

Presupunem că în exemplul 1 nu ştim câte monede de fiecare tip sunt în sertar. Aruncăm una aleasă aleator şi obţinem avers. Cum veţi decide care ipoteză (tip de monedă), dacă este vreuna, a fost cea mai susţinută de date?

2.4 Reactualizare

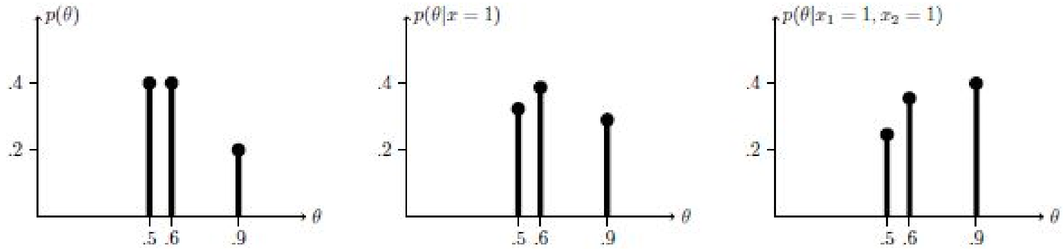
În deducţia Bayesiană, după actualizarea de la a priori la a posteriori, putem să luăm mai multe date şi să actualizăm iar. Pentru a 2-a actualizare, a posteriori din primele date devine a priori pentru ultimele date.

Exemplul 3. Presupunem că am ales o monedă ca în exemplul 1. O aruncăm o dată şi obţinem avers. Apoi aruncăm aceeaşi monedă şi obţinem iar avers. Care este probabilitatea ca moneda să fi fost de tipul A? Tipul B? Tipul C?

Răspuns: Tabelul devine mare:

hypothesis	prior	likelihood 1	Bayes numerator 1	likelihood 2	Bayes numerator 2	posterior 2
θ	$p(\theta)$	$p(x_1 = 1\|\theta)$	$p(x_1 = 1\|\theta)p(\theta)$	$p(x_2 = 1\|\theta)$	$p(x_2 = 1\|\theta)p(x_1 = 1\|\theta)p(\theta)$	$p(\theta\|x_1 = 1, x_2 = 1)$
0.5	0.4	0.5	0.2	0.5	0.1	0.2463
0.6	0.4	0.6	0.24	0.6	0.144	0.3547
0.9	0.2	0.9	0.18	0.9	0.162	0.3990
total	1				0.406	1

Al 2-lea numărător Bayes este calculat înmulţind primul numărător Bayes cu a 2-a verosimilitate; deoarece suntem interesaţi doar de a posteriori finală, nu trebuie să normalizăm până la ultimul pas. După cum este arătat în ultima coloană şi reprezentare, după 2 aversuri ipoteza tipul C a luat în sfârşit conducerea.



A priori $p(\theta)$, prima a posteriori $p(\theta|x_1 = 1)$, a 2-a a posteriori $p(\theta|x_1 = 1, x_2 = 1)$

2.5 Eroarea ratei de bază

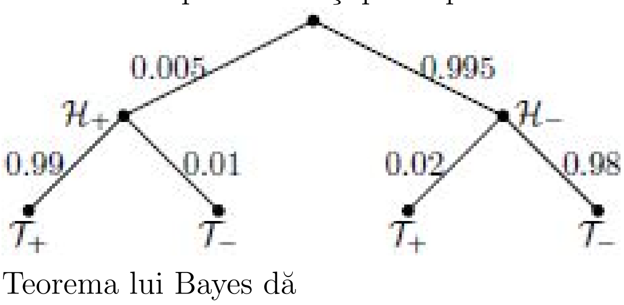
Exemplul 4. Un test pentru o boală este atât sensibil cât şi specific. Prin aceasta înţelegem că el este de obicei pozitiv când este testată o persoană care are boala şi de obicei negativ când este testat cineva care nu are boala. Presupunem că rata adevărat pozitivă este 99% şi rata fals pozitivă este 2%. Presupunem că răspândirea bolii în populaţia totală este 0.5%. Dacă o persoană aleatoare este testată pozitiv, care este probabilitatea ca ea să aibă boala?
Răspuns: Ca o recapitulare, întâi facem calculul folosind un arbore. Apoi refacem calculul folosind un tabel.
Folosim notaţiile stabilite anterior pentru ipoteze şi date: fie $\mathcal{H}_+$ ipoteza (evenimentul) că persoana are boala şi $\mathcal{H}_-$ ipoteza că nu o are. Analog, fie $\mathcal{T}_+$ şi $\mathcal{T}_-$ datele unui test pozitiv, respectiv negativ. Ni se cere să calculăm $P(\mathcal{H}_+|\mathcal{T}_+)$. Ni s-a dat

$$P(\mathcal{T}_+|\mathcal{H}_+) = 0.99,\ P(\mathcal{T}_+|\mathcal{H}_-) = 0.02,\ P(\mathcal{H}_+) = 0.005.$$

Din acestea putem calcula ratele fals negativă şi adevărat negativă:

$$P(\mathcal{T}_-|\mathcal{H}_+) = 0.01,\ P(\mathcal{T}_-|\mathcal{H}_-) = 0.98.$$

Toate aceste probabilităţi pot fi prezentate într-un arbore.



Teorema lui Bayes dă

$$P(\mathcal{H}_+|\mathcal{T}_+) = \frac{P(\mathcal{T}_+|\mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+)} = \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.02 \cdot 0.995} \approx 0.1992 \approx 20\%.$$

Acum refacem calculele folosind un tabel de actualizare Bayesiană:

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{T}_+\|\mathcal{H})$	$P(\mathcal{T}_+\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{T}_+)$
$\mathcal{H}_+$	0.005	0.99	0.00495	0.19920
$\mathcal{H}_-$	0.995	0.02	0.01990	0.80080
total	1	NO SUM	0.02485	1

Tabelul arată că probabilitatea a posteriori $P(\mathcal{H}_+|\mathcal{T}_+)$ ca o persoană cu test pozitiv să aibă boala este de aproximativ 20%. Aceasta este mult sub cea a sensibilităţii testului (99%), dar mult mai mare decât răspândirea bolii în populaţie (0.5%).