

Curs 7

Cristian Niculescu

1 Estimări de verosimilitate maximă

1.1 Scopurile învățării

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

1.2 Introducere

Presupunem că știm că avem date constând din valori x_1, \dots, x_n dintr-o repartiție exponențială. Întrebarea rămâne: care repartiție exponențială?!

Ne-am referit uneori la repartiția exponențială sau la repartiția binomială sau la repartiția normală. De fapt, repartiția exponențială $\exp(\lambda)$ nu este o singură repartiție, ci mai degrabă o familie de repartiții cu un parametru. Fiecare valoare λ definește o repartiție diferită din familie, cu pdf $f_\lambda(x) = \lambda e^{-\lambda x}$ pe $[0, \infty)$. Similar, o repartiție binomială $\text{bin}(n, p)$ este determinată de cei 2 parametri n și p , iar o repartiție normală $N(\mu, \sigma^2)$ este determinată de cei 2 parametri μ și σ^2 (sau echivalent, μ și σ). Familiile parametrizate de repartiții sunt adesea numite [repartiții parametrice](#) sau [modele parametrice](#).

Adesea suntem puși în fața situației de a avea date aleatoare despre care știm (sau credem) că provin dintr-un model parametric, ai cărui parametri nu-i știm. De exemplu, într-o alegere între 2 candidați, datele sondajului sunt dintr-o repartiție 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 gestației are o repartiție normală, ne-ar plăcea să folosim datele lungimilor gestaționale dintr-o selecție aleatoare de sarcini pentru a trage concluzii despre valorile parametrilor μ și σ^2 .

Scopul nostru până acum a fost să calculăm [probabilitatea datelor](#) provenind

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

1.3 Estimări de verosimilitate maximă

Sunt mai multe metode de a estima parametrii necunoscuți 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 ușor de calculat și este de acord cu intuiția noastră în exemple simple.

Exemplul 1. Aruncăm o monedă de 100 de ori. Dat fiind că au fost 55 de aversuri, aflați estimarea de verosimilitate maximă pentru probabilitatea p a aversului la o singură aruncare.

Înainte de a rezolva problema, stabilim niște notații și termeni.

Putem gândi calcularea numărului de aversuri în 100 de aruncări ca un experiment. Pentru o valoare dată a lui p , probabilitatea de a obține 55 de aversuri în acest experiment este probabilitatea binomială

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

Probabilitatea de a obține 55 de aversuri depinde de valoarea lui p , deci hai să-l includem pe p folosind notația probabilității condiționate:

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

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 câțiva termeni standard în statistică:

Experiment: Aruncăm moneda de 100 de ori și calculăm numărul de aversuri.

Date: Datele sunt rezultatul experimentului. În acest caz sunt ”55 de aversuri”.

Parametru (Parametri) de interes: Suntem interesați de valoarea parametru-lui necunoscut p .

Verosimilitatea sau **funcția de verosimilitate:** aceasta este $P(\text{date}|p)$. Observăm că este o funcție de date și parametrul p . În acest caz verosimilitatea este

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

Observații: 1. Verosimilitatea $P(\text{date}|p)$ se schimbă dacă parametrul de interes p se schimbă.

2. Atenție la definiție! O sursă tipică de confuzie este a confunda verosimilitatea $P(\text{date}|p)$ cu $P(p | \text{date})$. Știm că $P(\text{date}|p)$ și $P(p | \text{date})$ sunt de obicei diferite.

Definiție: Știind 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 notația \hat{p} pentru MLE. Folosim analiza matematică pentru a o afla, luând derivata funcției de verosimilitate și 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 ecuația obținem

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

Observații: 1. MLE pentru p s-a dovedit a fi exact fracția 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 ușor să lucrăm cu logaritmul natural al funcției de verosimilitate. Pentru scurtime acesta este numit simplu **log verosimilitate**. Deoarece $\ln(x)$ este o funcție strict crescătoare, maximele verosimilității și log verosimilității

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 verosimilității este același lucru cu maximizarea log verosimilității. Verificăm că analiza matematică ne dă același răspuns ca mai sus:

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

1.3.2 Verosimilitate maximă pentru repartiții continue

Pentru repartiții continue, folosim funcția densitate de probabilitate pentru a defini verosimilitatea.

Exemplul 3. Becuri

Presupunem că durata de funcționare a unui anumit tip de becuri este modelată de o repartiție exponențială cu parametru necunoscut λ . Testăm 5 becuri și aflăm că au duratele de funcționare de 2, 3, 1, 3 respectiv 4 ani. Care este MLE pentru λ ?

Răspuns: Avem nevoie să fim atenți cu notația noastră. Cu 5 valori diferite cel mai bine este să folosim indici. Fie X_i durata de funcționare al celui de-al i -lea bec și 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 funcționare ale becurilor sunt independente, deci pdf comună este produsul densităților 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 condiționată, deoarece depinde de λ . Văzând datele ca fixate și λ ca variabilă, această densitate este funcția de verosimilitate. Datele noastre au avut valorile

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

Deci funcțiile de verosimilitate și log verosimilitate cu datele noastre sunt

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

În sfârșit, folosim analiza matematică pentru a afla MLE:

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

Observații: 1. În acest exemplu am folosit o literă mare pentru o variabilă aleatoare și litera mică corespunzătoare pentru valoarea pe care o ia.

2. MLE pentru λ s-a dovedit a fi inversa mediei de selecție \bar{x} , deci $X \sim \exp(\hat{\lambda})$ satisface $E(X) = \bar{x}$.

Exemplul 4. Repartiții normale

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

Răspuns: Formulăm aceasta în termeni de variabile aleatoare și densități. Fie X_1, \dots, X_n variabile aleatoare i.i.d. $N(\mu, \sigma^2)$ și 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:

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

Pentru datele fixate x_1, \dots, x_n , verosimilitatea și log verosimilitatea sunt

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

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

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

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

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

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

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

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

Exemplul 5. Repartiții uniforme

Presupunem că datele noastre x_1, \dots, x_n sunt independente dintr-o repartiție uniformă $U(a, b)$. Aflați MLE pentru a și 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 funcția noastră de verosimilitate este

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

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

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

Exemplul 6. Metoda captură/recaptură

Metoda captură/recaptură este un mod de a estima mărimea unei populații în sălbăticie. Metoda presupune că fiecare animal din populație este egal posibil de a fi capturat de o cursă.

Presupunem că 10 animale sunt capturate, etichetate și eliberate. Câteva luni mai târziu, 20 de animale sunt capturate, examinate și eliberate. 4 dintre aceste 20 sunt găsite etichetate. Estimați mărimea populației 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 (și că sunt 10 animale etichetate). Funcția 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 populație de n .) Putem folosi R (cu comanda `dhyper(16, n-10, 10, 20)`)

pentru diverși n) pentru a calcula că funcția 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 fracția din toate animalele care sunt etichetate este $10/50$, care este egală cu fracția 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 și a), unde alela A are frecvența θ în populație. Adică, o copie aleatoare a genei este A cu probabilitatea θ și 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 eșantion aleator de oameni și aflăm că: k_1 sunt AA, k_2 sunt Aa și k_3 sunt aa. Aflați MLE pentru θ .

Răspuns: Funcția 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 ecuația în θ , aflăm că MLE este

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

care este pur și simplu fracția de alele A dintre toate genele din populația selectată.

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

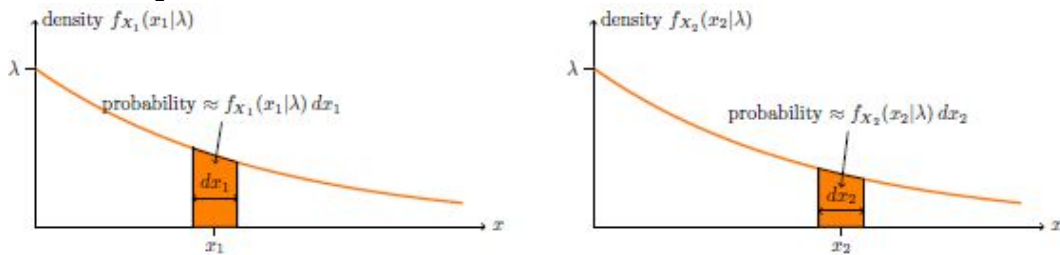
Ideea pentru estimarea de verosimilitate maximă este de a afla valoarea parametrului (parametrilor) pentru care datele au cea mai mare probabilitate.

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

are o repartiție $\exp(\lambda)$. Presupunem de asemenea că măsurăm independent duratele lor de funcționare și obținem datele $x_1 = 2$ ani și $x_2 = 3$ ani. Aflați valoarea lui λ care maximizează probabilitatea acestor date.

Răspuns: Principalul paradox cu care avem de-a face este că pentru o repartiție 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.



Relația uzuală între densitate și probabilitate pentru domenii mici.

Deoarece datele sunt colectate independent, probabilitatea comună este produsul probabilităților 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 $x_1 = 2$ și $x_2 = 3$ și formula pentru o pdf exponențială 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 maximumului. Deci pentru MLE îl eliminăm și numim pur și 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 Proprietăți ale MLE

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

dacă $\hat{\sigma}$ este MLE pentru deviația standard σ , atunci $(\hat{\sigma})^2$ este MLE pentru dispersia σ^2 .

În plus, MLE este **asimptotic nedeplasat** și are **asimptotic dispersie minimă**. Pentru a explica aceste noțiuni, observăm că MLE este ea însăși o variabilă aleatoare deoarece datele sunt aleatoare și MLE este calculată din date. Fie x_1, x_2, \dots un șir de date dintr-o repartiție cu parametrul p . Fie \hat{p}_n MLE pentru p bazată pe datele x_1, \dots, x_n .

Asimptotic nedeplasat înseamnă că atunci când numărul datelor crește, media MLE converge la p . În simboluri: $E(\hat{p}_n) \rightarrow p$ când $n \rightarrow \infty$. 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 crește, MLE are dispersia minimă dintre toți estimatorii nedeplasați ai lui p . În simboluri: pentru orice estimator nedeplasat \tilde{p}_n și $\varepsilon > 0$ avem $Var(\tilde{p}_n) + \varepsilon > Var(\hat{p}_n)$ când $n \rightarrow \infty$.

2 Actualizare Bayesiană cu a priori discrete

2.1 Scopurile învățării

1. Să poată aplica teorema lui Bayes pentru a calcula probabilități.
2. Să poată defini și identifica rolurile probabilității a priori, verosimilității (termenul Bayes), probabilității a posteriori, datelor și ipotezei în aplicarea teoremei lui Bayes.
3. Să poată folosi un tabel de actualizare Bayesiană pentru a calcula probabilitățile a posteriori.

2.2 Recapitularea teoremei lui Bayes

Reamintim că teorema lui Bayes ne permite să ”inversăm” probabilitățile condiționate. Dacă \mathcal{H} și \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 și teorema lui Bayes în formă tabelară

Exemplul 1. Sunt 3 tipuri de monede care au probabilități 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 și una de tipul C . Iau aleator o monedă din sertar. Fără a v-o arăta, o arunc o dată și obțin avers. Care este probabilitatea ca ea să fie de tipul A ? Tipul B ? Tipul C ?

Răspuns. Fie A, B și 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 și înregistrăm rezultatul.

Date: rezultatul experimentului nostru. În acest caz evenimentul $\mathcal{D} = \text{"avers"}$. 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 și una de tipul C avem

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

Verosimilitate: (Aceasta este aceeași verosimilitate pe care am folosit-o pentru MLE.) Funcția de verosimilitate este $P(\mathcal{D}|\mathcal{H})$, i.e., probabilitatea datelor presupunând că ipoteza este adevărată. Cel mai adesea considerăm datele ca fixate și lăsăm ipotezele să varieze. De exemplu, $P(\mathcal{D}|A) = \text{probabilitatea aversului dacă moneda este de tipul } A$. În cazul nostru verosimilitățile 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 și probabilitate sunt sinonime. Aceasta duce adesea la confundarea funcției de verosimilitate cu probabilitatea unei ipoteze. De aceea preferăm să folosim numele termenul Bayes.

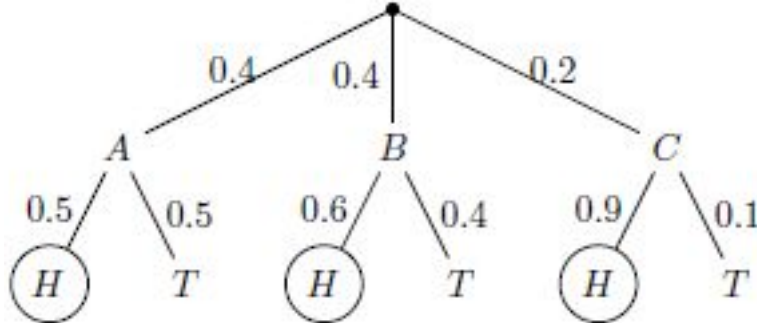
Probabilitatea a posteriori: probabilitatea (a posteriori) a fiecărei ipoteze cunoscând datele din aruncarea monedei.

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

Aceste probabilități a posteriori sunt cele cerute de problemă.

Folosim teorema lui Bayes pentru a calcula fiecare din probabilitățile a posteriori. Reamintim că datele \mathcal{D} sunt că aruncarea a fost avers.

Organizăm probabilitățile într-un arbore:



Arborele de probabilități pentru alegerea și 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 probabilității 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 probabilități 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 aceeași la fiecare numitor și 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 și verosimilitate. Vedem în fiecare din calculele de mai sus cu formula lui Bayes că probabilitatea a posteriori este obținută împărțind numărătorul Bayes la $P(\mathcal{D}) = 0.62$. Vedem de asemenea că legea probabilității 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 numește **actualizare Bayesiană**. Actualizarea Bayesiană utilizează datele pentru a modifica înțelegerea noastră pentru probabilitatea fiecărei ipoteze posibile.

2.3.1 Lucruri importante de observat

1. Sunt 2 tipuri de probabilități: 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) și a posteriori (după date).
 2. Probabilitățile 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 probabilității a posteriori. Pentru a o calcula pe ultima, pur și 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 funcționează la fel de bine ca a posteriori.
 5. Suma de pe coloana verosimilității nu este 1. Funcția de verosimilitate *nu* este o funcție de probabilitate.
 6. Probabilitatea a posteriori reprezintă un "compromis" între verosimilitate și a priori. Când calculăm a posteriori, o a priori mare poate fi scăzută de o verosimilitate mică și 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 puteți 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 $P(\mathcal{D})$ servește doar la normalizarea probabilității a posteriori la 1. Astfel, putem exprima teorema lui Bayes ca o afirmație despre proporționalitatea a 2 funcții 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 Funcții masă de probabilitate a priori și a posteriori

Notatii standard:

θ este [valoarea ipotezei](#).

$p(\theta)$ este [funcția masă de probabilitate a priori a ipotezei](#).

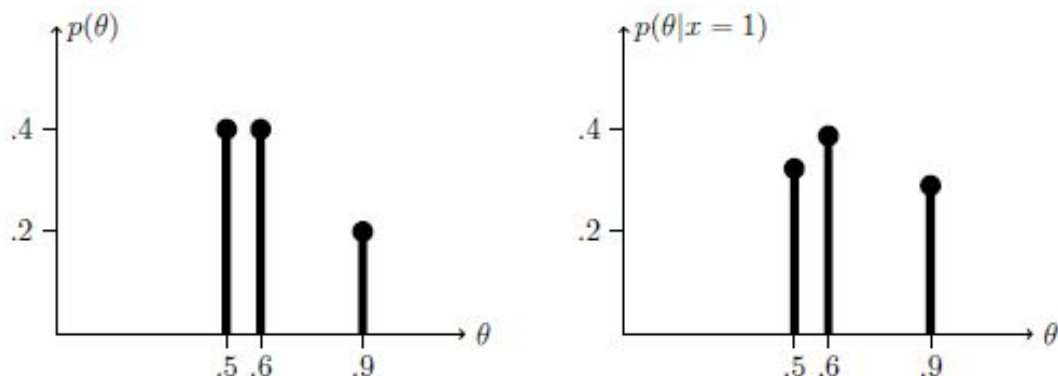
$p(\theta|\mathcal{D})$ este [funcția masă de probabilitate a posteriori a ipotezei cunoscând datele](#).

$p(\mathcal{D}|\theta)$ este [funcția de verosimilitate](#). (Aceasta nu este o pmf!)

În exemplul 1 putem reprezenta cele 3 ipoteze A , B și C prin $\theta = 0.5, 0.6, 0.9$. Pentru date $x = 1$ înseamnă avers și $x = 0$ revers. Atunci probabilitățile a priori și a posteriori din tabel definesc funcțiile masă de probabilitate a priori și 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 și a posteriori din exemplu:



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

Dacă datele ar fi diferite, atunci coloana verosimilității din tabelul de actualizare Bayesiană ar fi diferită. Putem planifica pentru date diferite construind întregul [tabel de verosimilitate](#) în avans. În exemplul cu moneda sunt 2 posibilități 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

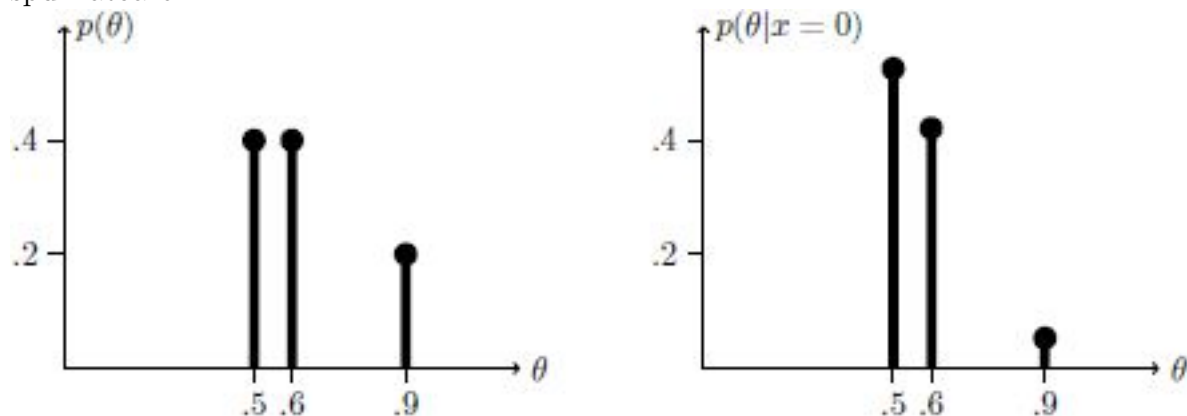
Exemplul 2. Folosind notația $p(\theta)$, etc., refaceți exemplul 1 presupunând că aruncarea a fost revers.

Răspuns: Deoarece datele s-au schimbat, coloana verosimilității 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 core-

spunzătoare:



2.4 Reactualizare

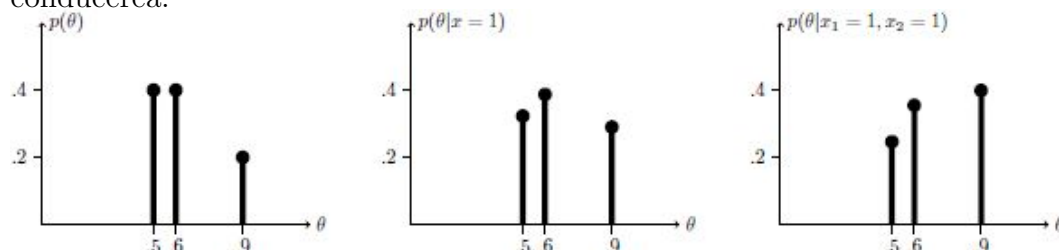
În deducția Bayesiană, după actualizarea de la a priori la a posteriori, putem să luăm mai multe date și să actualizăm iar. Pentru a 2-a actualizare, a posteriori din primele date devine a priori pentru ultimele date.

Exemplul 3. Presupunem că am ales o monedă ca în exemplul 1. O aruncăm o dată și obținem avers. Apoi aruncăm aceeași monedă și obținem 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 numărator Bayes este calculat înmulțind primul numărator Bayes cu a 2-a verosimilitate; deoarece suntem interesați doar de a posteriori finală, nu trebuie să normalizăm până la ultimul pas. După cum este arătat în ultima coloană și reprezentare, după 2 aversuri ipoteza tipul C a luat în sfârșit 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 și specific. Prin aceasta înțelegem că el este de obicei pozitiv când este testată o persoană care are boala și de obicei negativ când este testat cineva care nu are boala. Presupunem că rata adevărat pozitivă este 99% și rata fals pozitivă este 2%. Presupunem că răspândirea bolii în populația 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.

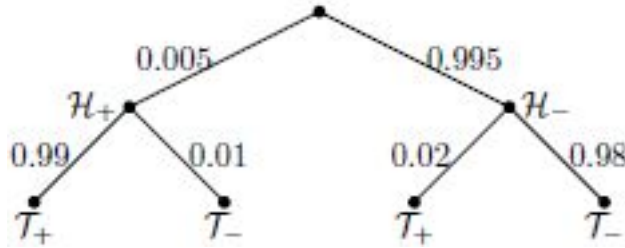
Folosim notațiile stabilite anterior pentru ipoteze și date: fie \mathcal{H}_+ ipoteza (evenimentul) că persoana are boala și \mathcal{H}_- ipoteza că nu o are. Analog, fie \mathcal{T}_+ și \mathcal{T}_- datele unui test pozitiv, respectiv negativ. Ni se cere să calculăm $P(\mathcal{H}_+|\mathcal{T}_+)$. Ni s-a dat

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

Din acestea putem calcula ratele fals negativă și adevărat negativă:

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

Toate aceste probabilități 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 sensibilității testului (99%), dar mult mai mare decât răspândirea bolii în populație (0.5%).

3 Actualizare Bayesiană: predicție probabilistică

3.1 Scopul învățării

Să poată să utilizeze legea probabilității totale pentru a calcula probabilitățile predictive a priori și a posteriori.

3.2 Introducere

Am văzut actualizarea probabilității ipotezelor bazată pe date. Putem de asemenea folosi datele pentru a actualiza probabilitatea fiecărui rezultat posibil al unui experiment viitor.

3.2.1 Predicție probabilistică; cuvinte de probabilitate estimativă (WEP)

Moduri de predicție:

Predicție: "Mâine va ploua."

Predicție folosind cuvinte de probabilitate estimativă (WEP): "Este probabil să plouă mâine."

Predicție probabilistică: "Mâine va ploua cu probabilitatea de 60% (și nu plouă cu probabilitatea de 40%)."

Fiecare tip de formulare este adecvat în diverse momente. Puteți vedea http://en.wikipedia.org/wiki/Words_of_Estimative_Probability pentru o

discuție despre utilizarea adecvată a cuvintelor de probabilitate estimativă. Articolul conține de asemenea o listă de *cuvinte nevăstuică* (*weasel words*) ca "ar putea (might)", "nu poate exclude (cannot rule out)", "este de conceput (it's conceivable)" care ar trebui evitate deoarece aproape sigur produc confuzie.

Sunt multe locuri unde vrem să facem o predicție probabilistică. Exemple:

Rezultate ale tratamentului medical

Prognoza meteo

Schimbarea climei

Pariuri sportive

Alegeri

...

Acestea sunt toate situații unde există incertitudine despre rezultat și am vrea o descriere cât mai precisă posibil a ce se poate întâmpla.

3.3 Probabilități predictive

Predicția probabilistică înseamnă atribuirea unei probabilități fiecărui rezultat posibil al unui experiment.

Reconsiderăm exemplul cu monedele: sunt 3 tipuri de monede care sunt de nedistins în afară de probabilitățile lor de avers la aruncare.

Monedele de tip A sunt corecte, cu probabilitatea aversului 0.5.

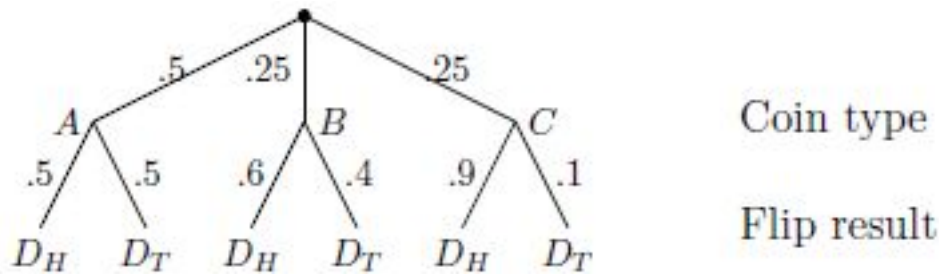
Monedele de tip B au probabilitatea aversului 0.6.

Monedele de tip C au probabilitatea aversului 0.9.

Avem un sertar cu 4 monede: 2 de tip A , una de tip B și una de tip C . Alegem o monedă din sertar. Fie A evenimentul "moneda aleasă este de tipul A ". Analog pentru B și C .

3.3.1 Probabilități predictive a priori

Înainte de a colecta datele, calculăm probabilitatea ca moneda aleasă de noi să dea avers (sau revers) dacă o aruncăm. Fie D_H evenimentul "avers" și D_T evenimentul "revers". Putem folosi [legea probabilității totale](#) pentru a determina probabilitățile acestor evenimente. Desenând un arbore sau calculând cu formula, obținem ("Coin type" = "tipul monedei", "Flip result" = "rezultatul aruncării"):



$$\begin{aligned}
 P(D_H) &= P(D_H|A)P(A) + P(D_H|B)P(B) + P(D_H|C)P(C) \\
 &= 0.5 \cdot 0.5 + 0.6 \cdot 0.25 + 0.9 \cdot 0.25 = 0.625 \\
 P(D_T) &= P(D_T|A)P(A) + P(D_T|B)P(B) + P(D_T|C)P(C) \\
 &= 0.5 \cdot 0.5 + 0.4 \cdot 0.25 + 0.1 \cdot 0.25 = 0.375.
 \end{aligned}$$

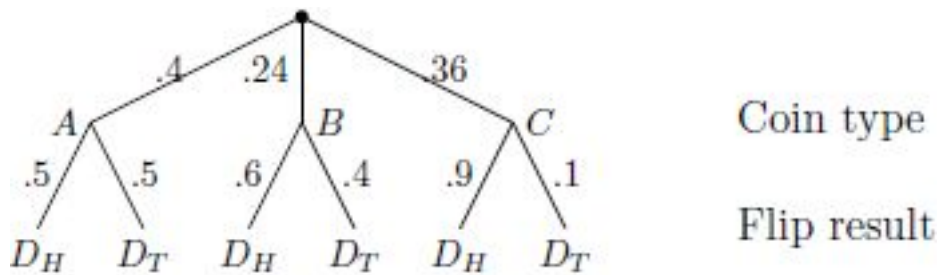
Definiție: Aceste probabilități dau o predicție (probabilistică) a ce se va întâmpla dacă moneda este aruncată. Deoarece ele sunt calculate înainte de a colecta orice dată, ele se numesc **probabilități predictive a priori**.

3.3.2 Probabilități predictive a posteriori

Presupunem că aruncăm moneda o dată și obținem avers. Acum avem datele D , pe care le putem folosi pentru a actualiza probabilitățile a priori ale ipotezelor noastre la probabilități a posteriori. Folosim un tabel Bayes pentru a facilita acest calcul:

hypothesis	prior	likelihood	Bayes	
			numerator	posterior
H	$P(H)$	$P(D H)$	$P(D H)P(H)$	$P(H D)$
A	0.5	0.5	0.25	0.4
B	0.25	0.6	0.15	0.24
C	0.25	0.9	0.225	0.36
total	1		0.625	1

Dacă am aruncat moneda o dată și am obținut avers, putem calcula probabilitatea că moneda aleasă de noi va da avers (sau revers) dacă o aruncăm a 2-a oară. Procedăm ca mai sus, dar folosind probabilitățile a posteriori $P(A|D)$, $P(B|D)$, $P(C|D)$ în locul probabilităților a priori $P(A)$, $P(B)$, $P(C)$.



$$P(D_H|D) = P(D_H|A)P(A|D) + P(D_H|B)P(B|D) + P(D_H|C)P(C|D) \\ = 0.5 \cdot 0.4 + 0.6 \cdot 0.24 + 0.9 \cdot 0.36 = 0.668$$

$$P(D_T|D) = P(D_T|A)P(A|D) + P(D_T|B)P(B|D) + P(D_T|C)P(C|D) \\ = 0.5 \cdot 0.4 + 0.4 \cdot 0.24 + 0.1 \cdot 0.36 = 0.332.$$

Definiție: Aceste probabilități dau o predicție (probabilistică) a ceea ce se va întâmpla dacă moneda este aruncată iar. Deoarece sunt calculate după colectarea datelor și actualizarea a priori la a posteriori, ele sunt numite [probabilități predictive a posteriori](#).

Observăm că aversul la prima aruncare crește probabilitatea aversului la a 2-a aruncare.

3.3.3 Recapitulare

Fiecare ipoteză dă o probabilitate diferită a aversului, deci probabilitatea totală a aversului este o medie ponderată. Pentru probabilitatea predictivă a priori a aversului, ponderile sunt date de probabilitățile a priori ale ipotezelor. Pentru probabilitatea predictivă a posteriori a aversului, ponderile sunt date de probabilitățile a posteriori ale ipotezelor.

Rețineți: Probabilitățile a priori și a posteriori sunt pentru ipoteze. Probabilitățile predictive a priori și a posteriori sunt pentru date. Ultimele [prezic](#) datele viitoare.