

BAYESOVA TEORIJA

NEŽA KRŽAN

1. BAYESOVA STATISTIKA

Bayesova statistika je statistična veja, ki nam s pomočjo matematičnih pristopov omogoča uporabo verjetnosti pri reševanju statističnih problemov. V svoje modele vključuje pogojno verjetnost, katero izračunamo z uporabo Bayesovega pravila.

Zlasti Bayesovo sklepanje razlaga verjetnost kot merilo verjetnosti ali zaupanja, ki ga lahko ima posameznik glede nastanka določenega dogodka. O nekem dogodku lahko že imamo predhodno prepričanje oziroma apriorno prepričanje, ki pa se lahko spremeni, ko se pojavijo novi dokazi. Bayesova statistika nam daje matematične modele za vključevanje naših apriornih prepričanj in dokazov za ustvarjenje novih prepričanj oziroma za pridobitev aposteriornega prepričanja, ki se lahko uporabi za kasnejše odločitve.

Bayesova analiza je standardna metoda za posodabljanje verjetnosti po opazovanju več dokazov, zato je zelo primerna za sintezo dokazov. Vsakdo, ki mora presoditi o hipotezi, kot je "krivda" (vključno s preiskovalci pred sojenjem, sodniki, porotami), neformalno začne z nekim predhodnim prepričanjem o hipotezi in ga posodablja, ko se dokazi ponovno pojavijo. Včasih lahko obstajajo celo objektivni podatki, na katerih temelji predhodna verjetnost. Pri uporabi Bayesovega sklepanja morajo statistiki utemeljiti predhodne predpostavke, kadar je to mogoče, na primer z uporabo zunanjih podatkov; v nasprotnem primeru morajo uporabiti razpon vrednosti predpostavk in analizo občutljivosti, da preverijo zanesljivost rezultata glede na te vrednosti.

1.1. Bayesovo pravilo. Bayesovo sklepanje temelji na Bayesovim pravilom, ki izraža verjetnost nekega dogodka z verjetnostjo dveh dogodkov in obratno pogojne verjetnosti. Pogojna verjetnost predstavlja verjetnost dogodka, glede na drug dogodek.

Definicija 1.1. Pogojna verjetnost dogodka H , glede na dogodek E , je

$$(1) \quad P(H|E) = \frac{P(H \cap E)}{P(E)},$$

ob predpostavki, da je $P(E) > 0$.

Formula (1) pove, da je verjetnost dogodka H ob pogoju, da se je zgodil dogodek E , enaka razmerju verjetnosti, da se zgodita oba dogodka in verjetnosti, da se je zgodil dogodek E . Potem pogojno verjetnost uporabimo še v števcu formule (1) in dobimo Bayesovo pravilo:

$$(2) \quad P(H|E) = \frac{P(E|H) \times P(H)}{P(E)},$$

verjetnost dogodka E lahko še razpišemo in dobimo:

$$(3) \quad P(H|E) = \frac{P(E|H) \times P(H)}{P(E|H)P(H) + P(E|\neg H)P(\neg H)}.$$

Obstaja še ena formulacija Bayesovega pravila, ki olajša izračune in je pogosto uporabljena pri Bayesovi analizi DNK dokazov:

$$(4) \quad \frac{P(H|E)}{P(\neg H|E)} = \frac{P(E|H)}{P(E|\neg H)} \times \frac{P(H)}{P(\neg H)}.$$

1.2. Bayesovo posodabljanje. Bayesovo pravilo se razlikuje od Bayesovega posodabljanja. Prvo je matematični izrek, drugo pa logična trditev, kako se sčasoma posodablja apriorne verjetnosti dokazov glede na novo zbrane dokaze oziroma prepričanja.

Bayesovo posodabljanje pravi:

Trditev 1.2. Če se dogodek E zgodi ob času $t_1 > t_0$, potem je $P_1(H) = P_0(H|E)$.

Ob času t_0 dogodku H dodelimo verjetnost $P_0(H)$; to se imenuje predhodna verjetnost oziroma apriorna verjetnost. Ko se zgodi dogodek E ob času t_1 , ki vpliva na naša prepričanja o dogodku H, Bayesovo posodabljanje pravi, da je potrebno apriorno verjetnost dogodka H v času t_1 enačiti z pogojno verjetnostjo dogodka H glede na dogodek E v času t_0 .

Recimo, da je dogodek H neka hipoteza oziroma prepričanje o zločinu in dogodek E dokazi, zbrani za ta zločin. Pri Bayesovem posodabljanju je videti, kot da je dokaz E nesporno resničen. Z drugimi besedami, predpostavka je, da moramo imeti po zbiranju dokazov E stopnjo zaupanja v E enako 1, torej če so dokazi zbrani v času t_1 , je $P_1(E) = 1$.

2. BEYESOVA ANALIZA

V kazenskih zadevah želimo vedeti, ali je obtoženec kriv ali ne. Če imamo torej na voljo dokaz E, nas zanima pogojna verjetnost

$$P(kriv|E),$$

pri čemer nam je lahko v pomoč Bayesovo pravilo. To v teoriji drži, čeprav je v praksi izračun verjetnostne krivde lahko preveč zapleten. Ampak z Bayesovim pravilom lahko ocenimo verjetnosti vmesnih trditev oziroma dokazov, ki so ključnega pomena za ugotavljanje obtoženčeve krivde. Najbolj pogosta uporaba Bayesovega pravila je pri ugotavljanju, ali je obtoženec vir sledi DNK-ja s kraja zločina.

2.1. Poenostavljena Bayesova analiza. Naj bo:

S ... trditev, da je obtoženec vir sledi DNK s kraja zločina;

M ... trditev, da se obtoženčev DNK ujema z DNK-jem s kraja zločina;

f ... funkcija pogostosti ujemanja DNK z DNK-jem s kraja zločina.

Želimo vedeti, kakšna je verjetnost S glede na M, tj. $P(S|M)$.

Bayesovo pravilo lahko uporabimo na naslednji način:

$$\frac{P(S|M)}{P(\neg S|M)} = \frac{P(M|S)}{P(M|\neg S)} \times \frac{P(S)}{P(\neg S)}.$$

Verjetnosti $P(S)$ in $P(\neg S)$ je težko oceniti, ker ne vemo kakšna je množica osumljencev. Smiselno bi bilo, da zanju upoštevamo interval predhodnih verjetnosti in ocenimo njihov vpliv na verjetnost trditve S in njene negacije. Nato moramo določiti vrednost $P(M|S)$, ki je običajno enaka ena - če bi obtoženec dejansko pustil sledove, bi laboratorijske analize pokazale ujemanje (to imenujemo lažno negativni rezultat); to je sicer poenostavitev, saj se lahko zgodi, da analize ne pokažejo ujemanja, čeprav je obtoženec pustil sledi. Potrebujemo še verjetnost $P(M|\neg S)$ (verjetnost, da se bo našlo ujemanje, če obtoženec ni vir sledi na kraju zločina). To je običajno enakovredno pogostosti ujemanja DNK-ja z DNK-jem s kraja zločina (tj. f); tudi to je poenostavitev, saj se lahko zgodi, da obtoženec nima enakega DNK profila, vendar so laboratorijske analize pokazale, da ga ima (to imenujemo lažno pozitivni rezultat). Sledi:

$$\frac{P(S|M)}{P(\neg S|M)} = \frac{1}{f} \times \frac{P(S)}{P(\neg S)}.$$

Ker poenostavljena Bayesova analiza ne upošteva možnosti lažno pozitivne in negativne laboratorijske analize si pogledimo še izpopolnjeno Bayesovo analizo.

2.2. Izpopolnjena Bayesova analiza. Za upoštevanje možnosti laboratorijskih napak, bomo namesto M uvedli spremenljivko M_p .

M_p ... poročano ujemanje laboratorijske analize;

M_t ... trditev, da obstaja dejansko ujemanje v DNK-ju;

$\neg M_t$... trditev, da obstaja neujemanje v DNK-ju.

Sledi:

$$P(M_p|\neg S) = P(M_p|M_t)P(M_t|\neg S) + P(M_p|\neg M_t)P(\neg M_t|\neg S).$$

Sedaj je $P(M_t|\neg S)$ enako f in zato $P(\neg M_t|\neg S)$ enako $1 - f$. $P(M_p|\neg M_t)$ opisuje verjetnost lažno pozitivnih rezultatov laboratorija (oznaka FP) in $P(M_p|M_t)$ verjetnost resničnih pozitivnih rezultatov laboratorija (oznaka FN). Sledi:

$$P(M|\neg S) = [(1 - FN) \times f] + [FP \times (1 - f)].$$

Formula pokaže, da za pravilno oceno verjetnosti $P(M_p|\neg S)$ potrebujemo statistično oceno pogostosti profila DNK in stopnje napak laboratorijskih analiz, ki pa so redko na voljo.

Druga poenostavitev je, da predpostavimo, da je $P(M_p|S) = 1$, pri čemer ni upoštevana možnost lažnega negativnega rezultata. Kot zgoraj, imamo:

$$P(M_p|S) = P(M_p|M_t)P(M_t|S) + P(M_p|\neg M_t)P(\neg M_t|S).$$

Če je $P(M_p|S) = 1$, je $P(\neg M_p|S) = 0$ in $P(M_p|S) = 1 - FN$. Potem sledi, da je:

$$\frac{P(M_p|S)}{P(M_p|\neg S)} = \frac{1 - FN}{[(1 - FN) \times f] + [FN \times (1 - f)]}.$$

Za predstavo, kako stopnje napak vplivajo na razmerje verjetnosti, predpostavim, da je pogostost profila DNK 1 proti milijardi. Predpostavim tudi, da sta stopnji lažno pozitivnih in negativnih laboratorijskih rezultatov (tj. FP in FN) enako 0,01. Če je razmerje verjetnosti enako $\frac{1}{f}$, je milijarda. S formulo pa dobimo:

$$\frac{P(M_p|S)}{P(M_p|\neg S)} = \frac{1 - 0,01}{[(1 - 0,01) \times 0,000000001] + [0,01 \times (1 - 0,000000001)]} \approx 99.$$

Relativno majhne stopnje napak lahko bistveno zmanjšajo dokazno vrednost DNK dokazov, saj močno zmanjšajo razmerje verjetnosti; v našem primeru smo iz milijarde prišli na približno 100. Vpliv stopnje laboratorijskih napak kaže, da ne glede na to, kako nizka se izkaže pogostost profila, bo ta relativno nepomembna, če pogostosti ne spremlja ocena stopnje laboratorijskih napak. Bayesovo pravilo nam omogoča, da ta vidik upoštevamo.

Ker profili DNK predstavlja del naše genetske zasnove in imajo ljudje, ki so v sorodu, večjo verjetnost, da bodo imeli enak profil DNK, kot ljudje, ki niso v sorodu morajo forenzični strokovnjaki svoje izjave vedno opredeliti z navedbo, da njihove ocene pogostosti veljajo za populacijo nepovezanih posameznikov. To spremenljivost pogostosti profila lahko v Bayesovem okviru upoštevamo na dva načina: s spremembo predhodne verjetnosti in s spremembo pogostosti profila. Izvedli bi lahko tudi različne izračune: enega za populacijo nepovezanih posameznikov in drugega za populacijo sorodnih posameznikov.

2.3. Primer - Taksi podjetja. Za lažje razumevanje Bayesovega pravila si pogledimo primer.

Obstajata dve taksi podjetji, Zeleni Taksi in Modri taksi, katerih vozila so po barvana zeleno oziroma modro. Podjetje Zeleni taksi pokriva 85 odstotkov trga, podjetje Modri taksi pa preostanek. Predpostavimo še, da v okolici ni drugih taksij podjetij. V meglenem dnevu taksi trči v mimoidočega pešca in ga poškoduje, vendar odpelje s kraja nesreče. Priča nesreče poroča, da je bilo vozilo modre barve. Priča ima prav le v 80 odstotkih primerov, kar pomeni, da je njegova zanesljivost enaka 0,8. Kolikšna je verjetnost, da je bil taksi, ki je povzročil nesrečo, modre barve glede na poročilo priče?

Vpeljimo oznake:

Z ... hipoteza, da je bil taksi zelen,

M ... hipoteza, da je bil taksi moder,

W_m ... dokaz, t.j. poročanje priče, da je bil taksi moder.

Problem je določiti pogojno verjetnost hipoteze M , ob pogoju, da je dokaz W_m resničen, torej $P(M|W_m)$.

Za uporabo Bayesovega pravila potrebujemo tri elemente: verjetnost dokazov glede na hipotezo, verjetnost dokazov in verjetnost hipoteze. Podjetje Zeleni taksi pokriva 85 odstotkov trga, zato je verjetnost $P(Z) = 0,85$. Po definiciji verjetnosti,

je potem $P(M) = 1 - P(Z) = 0,15$. Vemo tudi, da ima priča v 80 odstotkih primerov prav, torej $P(W_m|M) = 0,8$ in $P(W_m|Z) = 0,2$. Izračunajmo še verjetnost dokaza $P(W_m)$:

$$P(W_m) = P(W_m|B)P(M) + P(W_m|Z)P(Z) = 0,8 \times 0,15 + 0,2 \times 0,85 = 0,29.$$

Sedaj lahko uporabimo Bayesovo pravilo:

$$P(M|W_m) = \frac{P(W_m|M)P(M)}{P(W_m|M)P(M) + P(W_m|Z)P(Z)} = \frac{0,8 \times 0,15}{0,29} = \frac{12}{29} \approx 0,41.$$

Verjetnost, da je bil taksi glede na pričanje dejansko modre barve, je precej majhna, tudi če ima priča v 80 odstotkih primerov prav. Razlog za to je, da je verjetnost M , ne glede na dokaze, majhna ($P(M) = 0,15$). Spodnja tabela kaže, da lahko s spreminjanjem verjetnosti M dobimo različne pogojne verjetnosti $P(M|W_m)$, pri čemer je zanesljivost priče nespremenjena:

$P(M)$	$P(Z)$	$P(W_m M)$	$P(M W_m)$
0,15	0,85	0,8	0,41
0,25	0,75	0,8	0,57
0,35	0,65	0,8	0,68
0,45	0,55	0,8	0,76
0,50	0,50	0,8	0,80
0,55	0,45	0,8	0,83
0,65	0,35	0,8	0,88
0,75	0,25	0,8	0,92
0,85	0,15	0,8	0,95