Bayesova věta

Když pracujeme s podmíněnou pravděpodobností, tak se často hodí vzorec pro úplnou pravděpodobnost a Bayesův vzorec.

Definice: Řekneme, že jevy $A_1, A_2, \ldots \in \mathcal{A}$ tvoří *úplný systém jevů*, jestliže jsou po dvou disjunktní a $\cup A_i = \Omega$.

Věta: (o úplné pravděpodobnosti) Nechť $\{A_i\}$ je úplný systém jevů takový, že $P(A_i) > 0$ pro každé i. Potom platí

$$P(B) = \sum_i P(B \mid A_i)P(A_i).$$

Důkaz:

$$P(B) = P(B \cap (\cup_i A_i)) = P(\cup_i (B \cap A_i)) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i)P(A_i).$$

Zadání: Mámě dvě osudí A a B. Osudí A obsahuje 8 bílých a 7 černých míčků, zatímco v osudí B jsou 4 bílé a 6 černých. Nejprve náhodně vylosujeme osudí (A s pravděpodobností 2/3 a B s pravděpodobností 1/3), pak ze zvoleného osudí vybereme jeden míček. Je pravděpodobnější, že je bílý nebo černý?

Řešení: Pravděpodobnost, že bude černý, je $P(C) = P(C \mid A)P(A) + P(C \mid B)P(B) = \frac{7}{15} \cdot \frac{2}{3} + \frac{6}{10} \cdot \frac{1}{3} = \frac{23}{45}$. Bílý bude s pravděpodobností 22/45.

Věta: (Bayesova) Nechť $\{A_i\}$ je úplný systém jevů takový, že $P(A_i) > 0$ pro každé i. Jestliže je $P(B) > 0$, potom platí

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{\sum_i P(B \mid A_i)P(A_i)}.$$

Důkaz: Platí

$$P(A_j \mid B) = \frac{P(B \mid A_j)P(A_j)}{P(B)}$$

a stačí využít větu o podmíněné pravděpodobnosti k vyjádření $P(B)$ ve jmenovateli.

Zadání: Pokud v Monty Hallově problému předpokládáme (bez újmy na obecnosti), že soutěžící zvolil první dveře (v předchozím značení jev B_1), pak pro jevy A_i (auto je za dveřmi i) a C_j (moderátor otevře dveře j) dostáváme

$$\begin{aligned} P(A_1 \mid C_3) &= \frac{P(C_3 \mid A_1)P(A_1)}{P(C_3 \mid A_1)P(A_1) + P(C_3 \mid A_2)P(A_2) + P(C_3 \mid A_3)P(A_3)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3}} = \frac{1}{3}, \\ P(A_2 \mid C_3) &= 2/3, \\ P(A_3 \mid C_3) &= 0. \end{aligned}$$

Zadání: (papež pravděpodobně není mimozemšťan) Někdy se podmíněná pravděpodobnost $P(A \mid B)$ nesprávně zaměňuje s pravděpodobností $P(B \mid A)$. Následující příklad byl

probírán v několika příspěvcích publikovaných v roce 1996 v časopise Nature. Šance, že náhodně vybraný člověk bude papež, je zhruba jedna ku sedmi miliardám. To ale neznamená, že pravděpodobnost, že papež je člověk, je $(7 \cdot 10^9)^{-1}$, neboli $P(\text{papež} \mid \text{člověk}) \neq P(\text{člověk} \mid \text{papež})$.

Uvažujme pokus, kdy ze všech stvoření ve vesmíru vybereme jedno (rovnoměrně náhodně). Označme C jev, že vybrané stvoření je člověk; M, že je to mimozemšťan; a F, že je to papež (František I.). Víme, že $P(F \mid C) = (7 \cdot 10^9)^{-1}$, ale podle Bayesovy věty je

$$P(C \mid F) = \frac{P(F \mid C)P(C)}{P(F \mid C)P(C) + P(F \mid M)P(M)}.$$

Očekáváme, že pravděpodobnost $P(F \mid M)$ je zanedbatelná, a proto $P(C \mid F)$ je blízko jedné.

Jiný pohled je, že pravděpodobností varianta sylogismu nefunguje. Sylogismus se skládá ze tří částí: hlavní předpoklad, vedlejší předpoklad a závěr. Nejznámějším příkladem je: *Všichni lidé jsou smrtelní* (hlavní předpoklad), *Sokrates je člověk* (vedlejší předpoklad), z toho plyne *Sokrates je smrtelný* (závěr). Z hlavního předpokladu (náhodně vybraný člověk velmi pravděpodobně není papež) a vedlejšího předpokladu (František I. je papež) ale nemůžeme vyvodit, že František I. velmi pravděpodobně není člověk.

Když implikace $A \Rightarrow B$ platí v drtivé většině případu, tak to neznamená, že obměna $\neg B \Rightarrow \neg A$ platí ve většině případu. V našem případě je A člověk a B ne papež.

Zadání: (lékařská diagnostika) Dříve, než propukne nemoc D, lze její latentní existenci odhalit biologickým testem. U skrytě nemocné osoby je test pozitivní s pravděpodobností 0,999 (*senzitivita* testu). Oproti tomu u zdravé osoby je test negativní s pravděpodobností 0,99 (*specificita* testu). Zjištění tedy není jednoznačné, onemocnění nemusí odhaleno, nebo můžeme být vyvolán falešný poplach. Předpokládáme, že sledovanou nemoc má 1% populace (*prevalence* nemoci). Někdy se uvažuje *incidence* nemoci, což je počet nových případů onemocnění z celkového počtu obyvatel (neboli pouze osoby, u nichž není známo, zda nemocí trpí či nikoli). Jestliže u náhodně vybrané osoby dal test pozitivní výsledek, jaká je pravděpodobnost, že tato osoba má dané onemocnění?

Řešení: Ze zadání je $P(+ \mid D) = 0{,}999$ a $P(- \mid D^c) = 0{,}99$. Z toho plyne, že $P(- \mid D) = 0{,}001$ a $P(+ \mid D^c) = 0{,}01$. Dále víme, že $P(D) = 0{,}01$. Proto podle Bayesovy věty je

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} = \frac{0{,}999 \cdot 0{,}01}{0{,}999 \cdot 0{,}01 + 0{,}01 \cdot 0{,}99} \doteq 0{,}502.$$

Na první pohled se může zdát takový výsledek překvapující. Častá odpověď (nejen u laiků, ale i u lékařů) je, že hledaná pravděpodobnost je 0,999, což je spolehlivost testu. Na základě průzkumu mezi lékaři se ukázalo, že vysvětlení tohoto výsledku je názornější pomocí četností. Kdyby byl test aplikován na 100 000 lidí, pak by zhruba 1 000 mělo sledovanou nemoc a 99 000 ne. U 1% zdravých lidí dá test nesprávný výsledek, takže jich 990 označí jako nemocné. Z 1 000 nemocných bude mít jedna negativní test. Dohromady máme $990 + 999 = 1\,989$ pozitivních testů, z nichž jen 999 patří skutečně nemocným, to dává pravděpodobnost $999/1\,989 \doteq 0{,}502$.

Podívejme se, co se stane, když osoba, která měla pozitivní test, postoupí ještě jeden

test (nezávislý na tom prvním). Potom

$$\begin{aligned} P(D \mid 1+,2+) &= \frac{P(D,1+,2+)}{P(1+,2+)} = \frac{P(D,1+,2+)}{P(D,1+,2+)+P(D^c,1+,2+)} \\ &= \frac{P(2+ \mid D,1+)P(D,1+)}{P(2+ \mid D,1+)P(D,1+)+P(2+ \mid D^c,1+)P(D^c,1+)} \\ &= \frac{P(2+ \mid D,1+)P(D \mid 1+)}{P(2+ \mid D,1+)P(D \mid 1+)+P(2+ \mid D^c,1+)P(D^c \mid 1+)}, \end{aligned}$$

což je vlastně Bayesova věta pro pravděpodobnosti $P_1(A) = P(A \mid 1+)$, pak totiž

$$P_1(D \mid 2+) = \frac{P_1(2+ \mid D)P_1(D)}{P(2+ \mid D)P_1(D)+P(2+ \mid D^c)P_1(D^c)}.$$

Po dosazení je

$$P(D \mid 1+,2+) = P_1(D \mid 2+) = \frac{0{,}999 \cdot 0{,}502}{0{,}999 \cdot 0{,}502 + 0{,}01 \cdot 0{,}498} \doteq 0{,}9902.$$

Zadání: (vyšetření těhotných žen) U těhotných žen se v 1. trimestru (obvykle v 16. týdnu těhotenství) provádí AFP (alfa-fetoproteinový) test pro detekci vrozených vývojových vad. Má ukázat případné zvýšené riziko vrozených vad (Downův syndrom, rozštěp rtu a páteře, defekty břišní stěny apod.) Označme jako D jev, že je přítomna vrozená vada. Četnost vad závisí na věku matky a dalších faktorech, zhruba můžeme uvažovat $P(D) = 1/500$, tedy zhruba 2 postižené děti na 1000 těhotenství. Nechť jev A značí, že AFP test je pozitivní (indikuje vadu). Předpokládejme, že senzitivita testu je 80% (tj. $P(A \mid D) = 0{,}8$) a specificita je 90% (tj. $P(A^c \mid D^c) = 0{,}9$). Potom podle Bayesovy věty máme

$$P(D \mid A) = \frac{P(A \mid D)P(D)}{P(A \mid D)P(D)+P(A \mid D^c)P(D^c)} = \frac{0{,}8/500}{0{,}8/500 + 49{,}9/500} = \frac{8}{507} \doteq 0{,}0158.$$

Špatný výsledek AFP testu negativně ovlivňuje psychiku rodičů, přitom po pozitivním AFP testu je pravděpodobnost vrozené vady přibližně 1/63. Pozitivní AFP test znamená, že se žena musí objednat na amniocentézu. Jedná se o diagnostický úkon, který se provádí ambulantně mezi 16. a 20. týdnem těhotenství. Podstatou je vpíchnutí jehly do dělohy a odběr vzorku plodové vody. Analýza vzorku trvá 3 týdny. Senzitivita metody je 99%, tedy $P(C \mid D) = 0{,}99$. Specificitu neznáme, řekněme, že je 100%, tedy $P(C^c \mid D^c) = 1$. Při pozitivní amniocentéze lékař matce navrhne umělé přerušení těhotenství (k tomu dochází až v 19.–23. týdnu). Vedlejší účinky amniocentézy jsou, že s pravděpodobností 1/100 dojde k potratu. Pravděpodobnost, že postižené dítě bude detekováno a potraceno, je

$$P(A \mid D)P(C \mid D,A) = P(A \mid D)P(C \mid D) = 0{,}8 \cdot 0{,}99 = 0{,}792,$$

což znamená, že na 1 000 těhotenství bude $2 \cdot 0{,}792 = 1{,}584$ potracených postižených dětí. Pravděpodobnost, že zdravé dítě bude potraceno při amniocentéze je

$$P(A \mid D^c) \cdot \frac{1}{100} = \frac{0{,}1}{100} = 0{,}001,$$

což znamená, že na 1 000 těhotenství připadá $998 \cdot 0{,}001 = 0{,}998$ potracených zdravých dětí.