

Bayesovské metody pro identifikaci rizika

1 Pravděpodobnost náhodného jevu

1.1 Axiomatická definice pravděpodobnosti

Každému jevu A , který je podmnožinou základního prostoru Ω , přiřazujeme reálné číslo $P(A)$, které nazýváme **pravděpodobností jevu** A , přičemž platí:

(A1): Pravděpodobnost jevu A je nezáporné číslo, tzn. $P(A) \geq 0$.

(A2): Pravděpodobnost jistého jevu Ω je rovna 1, tzn. $P(\Omega) = 1$.

(A3): Jsou-li jevy $A_1, \dots, A_n$ vzájemně neslučitelné (tzn. $A_i \cap A_k = \emptyset$ pro $i \neq k$, $i, k = 1, 2, \dots, n$), potom pravděpodobnost sjednocení jevů $A_1, \dots, A_n$ je rovna součtu pravděpodobností jednotlivých jevů, tzn.

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

V teorii pravděpodobnosti se axiom (A3) obvykle uvádí pro posloupnost jevů $A_1, \dots, A_n, \dots$ (pro spočetnou množinu) vzájemně neslučitelných jevů, tzn.

$$P(A_1 \cup \dots \cup A_n \cup A_{n+1} \cup \dots) = P(A_1) + \dots + P(A_n) + P(A_{n+1}) + \dots$$

1.2 Podmíněná pravděpodobnost

Nechť $P(A)$ je pravděpodobnost jevu A při daném systému podmínek. Připojíme-li k tomuto systému další podmínku, tj. nastoupení jevu B , hovoříme o **podmíněné pravděpodobnosti** jevu A za předpokladu, že nastal jev B . Tato podmíněná pravděpodobnost $P(A|B)$ je dána vztahem

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0, \quad (1)$$

jev A nazýváme **podmíněným jevem**, o jevu B mluvíme jako o **hypotéze**.

Podmíněnou pravděpodobnost $P(A|B)$ lze interpretovat jako velikost ztráty, kterou utrpíme při nastoupení rizikového jevu A , za doprovodných okolností (doprovodných faktorů), které jsou shrnuty do jevu B .

V případě, že jev B nenastane, a tedy nastane jev $\overline{B}$, můžeme podmíněnou pravděpodobnost $P(A|\overline{B})$ považovat za velikost ztráty, kterou utrpíme, když nastoupí rizikový jev A při absenci doprovodných faktorů shrnutých v B . Potom podíl podmíněných pravděpodobností

$$R = R(A|B) = \frac{P(A|B)}{P(A|\overline{B})}$$

se nazývá **relativní riziko** (ztráty A vzhledem k B).

1.3 Nezávislost náhodných jevů

Pomocí podmíněné pravděpodobnosti lze zavést **nezávislost náhodných jevů**. Když platí, že

$$P(A|B) = P(A),$$

říkáme, že jevy A a B jsou nezávislé. Ekvivalentně lze nezávislost náhodných jevů A a B vyjádřit vztahem $P(A \cap B) = P(A)P(B)$. Podobným způsobem lze definovat také nezávislost více než dvou náhodných jevů.

2 Úplná pravděpodobnost, Bayesovy vzorce

Dále budeme předpokládat, že možné podmínky experimentu lze popsat k vzájemně neslučitelnými jevy $B_1, B_2, \dots, B_k$. Tyto jevy se často nazývají hypotézy a platí pro ně $\cup_{i=1}^k B_i = \Omega$ a $P(B_i) > 0$, $i = 1, 2, \dots, k$. Za těchto předpokladů lze odvodit následující tři vzorce:

1. **Vzorec pro celkovou pravděpodobnost:** Pro libovolný jev A platí

$$P(A) = \sum_{i=1}^k P(B_i)P(A|B_i). \quad (2)$$

2. **První Bayesův vzorec:** Jestliže $P(A) > 0$, pak pro $j = 1, 2, \dots, k$ platí

$$P(B_j|A) = \frac{P(B_j)P(A|B_j)}{\sum_{i=1}^k P(B_i)P(A|B_i)}. \quad (3)$$

3. **Druhý Bayesův vzorec:** Jestliže $P(A) > 0$ a C je libovolný náhodný jev, pak platí

$$P(C|A) = \frac{\sum_{i=1}^k P(B_i)P(A|B_i)P(C|A \cap B_i)}{\sum_{i=1}^k P(B_i)P(A|B_i)}. \quad (4)$$

Příklad. Uvažujme, že ve společnosti je 40 % nekuřáků, 40 % slabých kuřáků (1–24 cigaret/den) a 20 % silných kuřáků (více než 24 cigaret/den). Podmíněné pravděpodobnosti, že osoba bude postižena infarktem myokardu za předpokladu, že je nekuřák, resp. slabý kuřák, resp. silný kuřák jsou po řadě 0,01, resp. 0,04, resp. 0,2. Na základě uvedených dat určete celkovou pravděpodobnost toho, že osoba bude postižena infarktem. Stanovte rovněž podmíněné pravděpodobnosti, že osoba stížená infarktem myokardu je nekuřákem, resp. slabým kuřákem, resp. silným kuřákem.

Označme A jev, že osoba bude postižena infarktem myokardu,

B_1 jev, že osoba je nekuřák,

B_2 jev, že osoba je slabý kuřák,

B_3 jev, že osoba je silný kuřák.

Ze zadání jsou známy tyto pravděpodobnosti:

$$P(B_1) = 0,4, P(B_2) = 0,4, P(B_3) = 0,2,$$

$$P(A|B_1) = 0,01, P(A|B_2) = 0,04, P(A|B_3) = 0,2.$$

Celkovou pravděpodobnost, že osoba bude postižena infarktem, spočítáme ze vzorce (2) pro celkovou pravděpodobnost:

$$P(A) = \sum_{i=1}^3 P(B_i)P(A|B_i) = 0,4 \cdot 0,01 + 0,4 \cdot 0,04 + 0,2 \cdot 0,2 = 0,06.$$

Užitím prvního Bayesova vzorce (3) určíme požadované podmíněné pravděpodobnosti:

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{\sum_{i=1}^3 P(B_i)P(A|B_i)} = \frac{0,4 \cdot 0,01}{0,06} = \frac{1}{15} \doteq 0,07,$$

$$P(B_2|A) = \frac{P(B_2)P(A|B_2)}{\sum_{i=1}^3 P(B_i)P(A|B_i)} = \frac{0,4 \cdot 0,04}{0,06} = \frac{4}{15} \doteq 0,26,$$

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

$$P(B_3|A) = \frac{P(B_3)P(A|B_3)}{\sum_{i=1}^3 P(B_i)P(A|B_i)} = \frac{0,2 \cdot 0,2}{0,06} = \frac{10}{15} \doteq 0,67.$$

Z výsledků jasně plyne, že riziko infarktu je pro silné kuřáky značně větší než pro nekuřáky. Z osob postižených infarktem je podle citované studie přibližně 67 % silných kuřáků, 26 % slabých kuřáků a pouze 7 % nekuřáků.

Příklad. V jisté studii se vyšetřuje, jak závisí přežití novorozence a délka těhotenství na věku matky a na skutečnosti, zda matka je kuřáčka. Uvažujme čtyři skupiny těhotných žen - matek:

B_1 ... věk do 30 let, kouří méně než 5 cigaret denně,

B_2 ... věk do 30 let, kouří 5 nebo více cigaret denně,

B_3 ... věk 30 a více let, kouří méně než 5 cigaret denně,

B_4 ... věk 30 a více let, kouří 5 nebo více cigaret denně.

Označení B_i použijeme rovněž pro samotný jev, že žena patří do příslušné skupiny B_i , $i = 1, 2, 3, 4$. Dále označme A jev, že délka těhotenství ženy je 260 nebo méně dní, a jev $\bar{A}$, že těhotenství ženy trvá déle než 260 dní. Konečně C značí jev, že se narodí živé dítě a že nezemře během novorozeneckého období. Pravděpodobnosti, že matka patří do příslušné skupiny, jsou $P(B_1) = 0,652$, $P(B_2) = 0,076$, $P(B_3) = 0,251$, $P(B_4) = 0,021$. Podmíněné pravděpodobnosti, že těhotenství bude krátké (do 260 dní včetně) za předpokladu, že matka patří do dané skupiny, jsou:

$$P(A|B_1) = 0,083, P(A|B_2) = 0,095,$$

$$P(A|B_3) = 0,111, P(A|B_4) = 0,107.$$

Podmíněné pravděpodobnosti novorozeneckého přežití za předpokladu, že těhotenství bylo kratší a současně matka patří do dané skupiny, jsou:

$$P(C|A \cap B_1) = 0,863, P(C|A \cap B_2) = 0,816,$$

$$P(C|A \cap B_3) = 0,782, P(C|A \cap B_4) = 0,733.$$

Určíme nyní celkovou pravděpodobnost toho, že doba trvání těhotenství je 260 dní nebo kratší. Dále stanovíme podmíněné pravděpodobnosti, že matka patří do příslušné skupiny za předpokladu, že těhotenství je kratší. Následně porovnáme úmrtnost novorozenců, jejichž délka těhotenství byla standardní, s úmrtností novorozenců se zkrácenou dobou těhotenství. Celkovou pravděpodobnost toho, že doba trvání těhotenství je 260 dní nebo kratší, spočítáme ze vzorce (2) pro celkovou pravděpodobnost:

$$P(A) = \sum_{i=1}^4 P(B_i)P(A|B_i) \doteq 0,0914.$$

Užitím prvního Bayesova vzorce (3) stanovíme podmíněné pravděpodobnosti, že matka patří do příslušné skupiny za předpokladu, že těhotenství je kratší:

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{\sum_{i=1}^4 P(B_i)P(A|B_i)} \doteq \frac{0,0652 \cdot 0,083}{0,0914} \doteq 0,591,$$

$$P(B_2|A) = \frac{P(B_2)P(A|B_2)}{\sum_{i=1}^4 P(B_i)P(A|B_i)} \doteq \frac{0,076 \cdot 0,095}{0,0914} \doteq 0,079,$$

$$P(B_3|A) = \frac{P(B_3)P(A|B_3)}{\sum_{i=1}^4 P(B_i)P(A|B_i)} \doteq \frac{0,251 \cdot 0,111}{0,0914} \doteq 0,305,$$

$$P(B_4|A) = \frac{P(B_4)P(A|B_4)}{\sum_{i=1}^4 P(B_i)P(A|B_i)} \doteq \frac{0,021 \cdot 0,107}{0,0914} \doteq 0,025.$$

Tyto podmíněné pravděpodobnosti porovnáme s pravděpodobnostmi $P(B_i)$, $i = 1, 2, 3, 4$ (pravděpodobnost, že matka patří do příslušné skupiny). Z tohoto srovnání je patrný především vliv věku na délku těhotenství, ale také negativní vliv kouření na nezkrácení doby těhotenství. Ve skupině všech rodiček tvoří mladé (do 30 let

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

věku) nekuřáčky (do 5 cigaret denně) 65,2 %, zatímco ve skupině rodiček, jejichž doba těhotenství byla kratší (do 260 dní), je to jen 59,1 %, což značí „relativní pokles“ o 9,4 %. U skupiny mladých kuřáček se jedná o „relativní nárůst“ o 3,9 %, u starších nekuřáček je to „relativní nárůst“ o 21,5 % a u starších kuřáček „relativní nárůst“ o 19 %.

Zajímavé může být zjištění, jak závisí pravděpodobnost, že se narodí živé dítě a nezemře během novorozeneckého věku, na tom, zda se jednalo či nejednalo o zkrácené těhotenství. K tomu můžeme využít druhého Bayesova vzorce (4). Tedy, pravděpodobnost, že se narodí živé dítě a nezemře během novorozeneckého věku za předpokladu, že se jedná o zkrácené těhotenství, je

$$P(C|A) = \frac{\sum_{i=1}^4 P(B_i)P(A|B_i)P(C|A \cap B_i)}{\sum_{i=1}^4 P(B_i)P(A|B_i)} \doteq 0,832.$$

Analogicky lze vypočítat pravděpodobnost, že se narodí živé dítě a nezemře během novorozeneckého věku za předpokladu, že se nejednalo o zkrácené těhotenství, tj.

$$P(C|\bar{A}) = \frac{\sum_{i=1}^4 P(B_i)P(\bar{A}|B_i)P(C|\bar{A} \cap B_i)}{\sum_{i=1}^4 P(B_i)P(\bar{A}|B_i)} \doteq 0,993.$$

Z výsledků je zřejmé, že úmrtnost novorozenců, jejichž délka těhotenství byla standardní, je pouze 0,7 %, zatímco ve skupině novorozenců se zkrácenou dobou těhotenství už je to 16,8 %.

Příklady k procvičení

1. Zkoumejme pravděpodobnost, že student je kuřák v souvislosti s tím, kolik jeho rodičů kouří. předpokládejme, že pravděpodobnost, že kouří oba rodiče je 0,33, že kouří jeden z rodičů 0,42 a že oba rodiče jsou nekuřáci 0,25. Dále předpokládejme, že podmíněné pravděpodobnosti, že student kouří za předpokladu, že pochází z rodiny, kde kouří oba rodiče, resp. jeden rodič, resp. žádný z rodičů, jsou po řadě 0,25, resp. 0,20, resp. 0,10. Určete celkovou pravděpodobnost toho, že student je kuřák, a podmíněné pravděpodobnosti toho, že student pochází z rodiny, kde kouří oba rodiče, resp. jeden rodič, resp. žádný z rodičů za předpokladu, že sám kouří.
2. Zkoumejme pravděpodobnost, že student je kuřák v souvislosti s tím, kolik jeho rodičů kouří. předpokládejme, že pravděpodobnost, že kouří oba rodiče je 0,33, že kouří jeden z rodičů 0,42 a že oba rodiče jsou nekuřáci 0,25. Dále předpokládejme, že podmíněné pravděpodobnosti, že student kouří za předpokladu, že pochází z rodiny, kde kouří oba rodiče, resp. jeden rodič, resp. žádný z rodičů, jsou po řadě 0,25, resp. 0,20, resp. 0,10. Určete celkovou pravděpodobnost toho, že student je kuřák, a podmíněné pravděpodobnosti toho, že student pochází z rodiny, kde kouří oba rodiče, resp. jeden rodič, resp. žádný z rodičů za předpokladu, že sám kouří.