Analýza dat v neurologii

XXXII. Bayesovská vs klasická statistika v klinických aplikacích

L. Dušek, T. Pavlík,
J. Jarkovský, J. Koptíková
Institut biostatistiky a analýz
MU, Brno

✉
doc. RNDr. Ladislav Dušek, Dr.
Institut biostatistiky a analýz
MU, Brno
e-mail: dusek@cba.muni.cz

Předcházející díl seriálu otevřel významné téma tzv. bayesovské statistiky a bayesovských odhadů. Připomeňme zde, že jde o metodický koncept odhadující pravděpodobnost výskytu určitého jevu na základě znalosti jeho vztahu (asociace) s jiným jevem nebo s více jinými jevy. V nejjednodušším případě tak odhadujeme pravděpodobnost jevu *A* při nastání jevu *B* podle tzv. **Bayesovy věty:**

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

Hodnota ***P(A)*** zde představuje tzv. **apriorní pravděpodobnost** nastání jevu *A*, kterou známe nebo odhadujeme předem a výpočtem ji upřesňujeme. Dosazením do výše uvedeného vztahu získáváme tzv. **aposteriorní pravděpodobnost** nastání jevu *A* při nastání jevu *B*, tedy *P(A|B)*. Při výpočtu využíváme znalosti vztahu obou jevů, konkrétně znalosti **podmíněné pravděpodobnosti** výskytu jevu *B* při nastání jevu *A*, tedy *P(B|A)*. Pravděpodobnost jevu *B* ve vztahu doplňujeme dle tzv. věty o úplné pravděpodobnosti, tedy jako součet pravděpodobností nastání jevu *B* při nastání i nenastání jevu *A: P(B) = P(B|A)P(A) +P(B|not A)P(not A)*.

V díle č. XXXI seriálu jsme uvedli četné příklady klinické aplikace Bayesovy věty, která u řady problémů vede k velmi přesnému a dobře interpretovatelnému odhadu chování určitých jevů v cílových populacích. V tomto díle se pokusíme výklad rozšířit o další aplikace, které mají velmi blízko i k neurovědnímu výzkumu.

Příklady aplikace Bayesovy věty

V této kapitole uvedeme Bayesovu větu z jiného pohledu, než jak jsme ji představili v předchozím díle. Místo jevu *A* uveďme hypotézu *H* a místo jevu *B* evidenci *E*. Potom výše uvedený Bayesův vztah můžeme přepsat do tvaru, kdy odhadujeme aposteriorní pravděpodobnost *P(H|E)*, tedy pravděpodobnost platnosti hypotézy *H*, pokud máme k dispozici evidenci *E*. Apriorní pravděpodobnost *P(H)* získáme z literatury, z dostupných dat, z posudků expertů nebo v případě nejistoty ji nastavíme nerozhodně jako rovnu 0,5. Znalost vstupních pravděpodobností pro výpočet nemusí být přesná (ale samozřejmě by měla být co nejpřesnější), pokud potřebné informace o *P(H)* nemáme, lze jako pilotní vstup využít např. expertní odhady. Postupně, s rostoucí znalostí problému a zkoumané populace, výsledek zpřesňujeme. Dále musíme pro výpočet znát pravděpodobnost výskytu evidence *E* a pravděpodobnost platnosti evidence *E* při platnosti hypotézy *H*, tedy *P(E)* a *P(E|H)*. Bayesova věta je v tomto smyslu vyjádřena jako:

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

Obdobně můžeme např. zvažovat pravděpodobnost, že hypotéza *H* je platná při určitých pozorovaných datech (*D*), tedy pravděpodobnost *P(H|D)* apod. Taková zadání již dané téma nijak nerozvíjejí, jde o stále stejný výpočet aplikovaný v různých situacích. Z hlediska laického uživatele je ovšem mnohem důležitější otázka, kdy lze tento výpočet použít a kdy má smysl i jako alternativa tzv. klasické statistiky. Klasickým neboli frekventistickým způsobem rozumíme provádění odhadů na základě mnohonásobně opakovaných náhodných experimentů (viz též díl XXXI seriálu).

Použití bayesovského odhadu je smysluplné, pokud známe jev *B*, či evidenci *E*, který je ve známém vztahu ke zkoumanému jevu *A*, či hypotéze *H*. Využitím této informace zpřesňujeme odhad chování (pravděpodobnosti výskytu) jevu *A*. Pokud by *A* a *B* byly jevy nezávislé, pak by platilo, že *P(A|B)* = *P(A)* a naopak *P(B|A)* = *P(B)*, a výše uvedený vztah by ztratil smysl. V díle XXXI jsme uvedli příklad výpočtu, kde byla informace o pravděpodobnosti kuřáctví mezi onkologickými pacienty využita k odhadu pravděpodobnosti výskytu rakoviny u kuřáků apod. Na tomto příkladu krátce zopakujme i největší přidanou hodnotu bayesovských odhadů, tedy aplikovatelnost v situacích, kdy nemáme dostatečná vstupní data pro provedení odhadů klasickou statistikou. Klasická statistika by totiž zde položený úkol řešila provedením studie zaměřené na kuřáky, u kterých by byla zkoumána přítomnost zhoubného nádoru. Taková studie by ovšem byla velmi náročná, časově i finančně, a nadto by zatěžovala nádorovou diagnostikou i zdravé kuřáky. Přitom dle výše uvedeného Bayesova vztahu údaj o pravděpodobnosti výskytu rakoviny (jev *A*) u kuřáka (jev *B*), tedy *P(A|B)*, získáme, pokud jsme schopni získat apriorní údaje o:

- *P(A)* a *P(B)*, což jsou data dostupná například z oficiálních populačních statistik,
- *P(B|A)*, tedy pravděpodobnost výskytu kuřáků mezi již diagnostikovanými onkologickými pacienty; získání tohoto údaje je jistě jednodušší (např. ze záznamů v nemocnicích) než přímý odhad opačné podmíněné pravděpodobnosti.

Zadání: Existuje diagnostický marker, který nabývá pozitivní hodnoty u 98 % pacientů s určitou chorobou. Nicméně pozitivní hodnotu markeru má rovněž 5 % zdravých lidí. Jaká je pravděpodobnost, že pacient s pozitivní hodnotou markeru má skutečně danou chorobu, pokud pochází z populace s následující prevalencí onemocně ní: 0,1 %; 1 %; 10 %; 20 %?

A Nemocný s danou chorobou.

$\overline{A}$ Zdravý člověk.

B Pozitivní marker.

$$P(B \mid A) = 0{,}98$$

$$P(B \mid \overline{A}) = 0{,}05$$

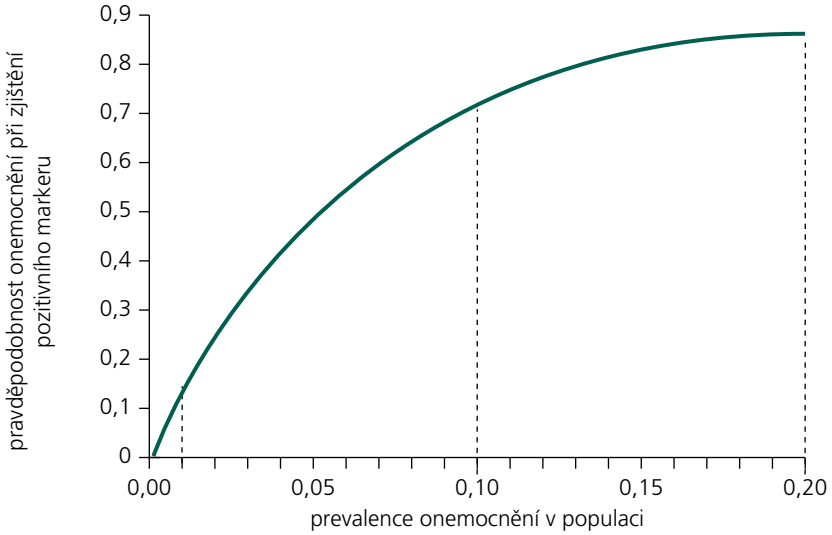
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \overline{A})P(\overline{A})}$$

$P(A) = 0{,}001$, $P(\overline{A}) = 0{,}999$ ⇒ $P(A \mid B) = 0{,}019$

$P(A) = 0{,}010$, $P(\overline{A}) = 0{,}990$ ⇒ $P(A \mid B) = 0{,}165$

$P(A) = 0{,}100$, $P(\overline{A}) = 0{,}900$ ⇒ $P(A \mid B) = 0{,}685$

$P(A) = 0{,}200$, $P(\overline{A}) = 0{,}800$ ⇒ $P(A \mid B) = 0{,}831$



Příklad 1. Využití Bayesovy věty pro odhad pravděpodobných výsledků diagnostického testu v populacích s různou prevalencí diagnostikované choroby.

Za určitých okolností je provedení klasického statistického měření výskytu sledovaného jevu doslova nemožné, a aplikace bayesovských odhadů tudíž není pouze alternativou klasických postupů. Téměř učebnicovou aplikací Bayesovy věty je odhad pravděpodobných výsledků diagnostického testu v populacích s různou prevalencí diagnostikované choroby. V praxi by bylo nemožné opakovat validační studie diagnostického testu ve všech populacích lišících se pouze prevalencí dané choroby. Příklad 1 ukazuje několik variant těchto výpočtů pro různě nastavené diagnostické hodnoty testů a prevalenci sledované choroby v cílové populaci.

Věta o úplné pravděpodobnosti a naivní bayesovský klasifikátor

Dosud vysvětlované příklady pracovaly s nejjednodušší možnou variantou, kdy zkoumáme pravděpodobnost výskytu jevu *A* (binární proměnná typu ano/ne) při výskytu jevu *B* (opět proměnná typu ano/ne). Anebo pravděpodobnost platnosti hypotézy *H* (platí/neplatí) při nastání určité evidence *E* (přítomna/nepřítomna). V praxi se ale často setkáme se situací, kdy pravděpodobnost výskytu jevu *A* sledujeme při výskytu více různých jevů B_1, …, B_k, což zkráceně zapisujeme jako B_i, i = 1, …, k. Předpokládejme pro jednoduchost, že jednotlivé jevy B_i jsou vzájemně nezávislé. Potom opět platí věta o úplné pravděpodobnosti (zde na rozdíl od vztahu uvedeného výše vyjádřena ve smyslu úplné pravděpodobnosti jevu *A*):

$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots\ldots\ldots + P(A|B_k)P(B_k)$

Doufáme, že jsme neodradili čtenáře, kteří nemají rádi komplikované matematické vztahy. Aplikace výše uvedené rovnice je stále ještě laicky zvládnutelná. Pokusíme se to dokumentovat na příkladu. Na léčbě určité nemoci se mohou podílet tři léčebné modality, přičemž každá může s určitou pravděpodobností způsobit určitou komplikaci (jev *A*). Tři aplikované modality (B_1, B_2, B_3) jsou ve svém účinku zcela nezávislé a ne všichni pacienti nutně absolvují všechny tři. Populační data udávají následující hodnoty: první modalitu absolvuje 60 % pacientů, druhou 40 % a třetí jen 20 % pacientů. Z toho odvodíme, že $P(B_1) = 0{,}6$, $P(B_2) = 0{,}4$ a $P(B_3) = 0{,}2$. Dále jsme z publikovaných klinických studií schopni zjistit, s jakou pravděpodobností jednotlivé modality způsobují sledovanou komplikaci *A*. Půjde o podmíněnou pravděpodobnost $P(A|B_i)$. Nastavme $P(A|B_1) = 0{,}3$, $P(A|B_2) = 0{,}2$ a $P(A|B_3) = 0{,}1$. Klíčová otázka je, jaká je pravděpodobnost, že pacient náhodně vybraný z populace léčených bude mít komplikaci *A*? Tato otázka má velký smysl například za situace, kdy plánujeme určitý výzkum (např. prevalenční studii) a ptáme se, kolik jedinců musíme z dané populace vybrat, abychom jev *A* dobře postihli. Výpočet provedeme podle výše uvedené věty o úplné pravděpodobnosti:

$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots\ldots + P(A|B_k)P(B_k) = 0{,}3 \times 0{,}6 + 0{,}2 \times 0{,}4 + 0{,}1 \times 0{,}2 = 0{,}28.$

Můžeme tedy očekávat, že vybereme-li zcela náhodně z této populace 100 léčených pacientů, 28 z nich bude trpět komplikací *A*. Další aplikace věty o úplné pravděpodobnosti přináší příklad 2.

Zadání: Výskyt onemocnění (*A*) závisí na výskytu určitých genových polymorfizmů v populaci. Celkově rozlišujeme čtyři typy polymorfizmů (B_1, …B_4), u kterých jsou známy podmíněné pravděpodobnosti výskytu nemoci *A*. Naším úkolem je pomocí těchto vstupních informací odhadnout, jaká je pravděpodobnost, že náhodně vybraná osoba z dané populace bude trpět chorobou *A*.

A	Výskyt onemocnění		
B_1	Genetická varianta I	$P(B_1) = 0{,}4$	$P(A \mid B_1) = 0{,}00$
B_2	Genetická varianta II	$P(B_2) = 0{,}2$	$P(A \mid B_2) = 0{,}10$
B_3	Genetická varianta III	$P(B_3) = 0{,}3$	$P(A \mid B_3) = 0{,}01$
B_4	Genetická varianta IV	$P(B_4) = 0{,}1$	$P(A \mid B_4) = 0{,}60$

$$P(A) = P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_4)P(B_4)$$

$$P(A) = 0{,}00 * 0{,}4 + 0{,}10 * 0{,}2 + 0{,}01 * 0{,}3 + 0{,}60 * 0{,}1$$

$$P(A) = 0{,}083$$

Výsledek: Pravděpodobnost výskytu onemocnění *A* u náhodně vybraného jedince z dané populace je 0,083.

Příklad 2. Využití věty o úplné pravděpodobnosti pro odhad výskytu onemocnění ve sledované populaci.

S využitím výše uvedeného příkladu můžeme odvodit tvar Bayesova teorému pro situaci, kdy sledujeme jev *A*, který může být podmíněn více vzájemně nezávislými jevy B_i. Dostáváme tzv. **naivní bayesovský klasifikátor**, který odhaduje pravděpodobnost výskytu jevu *A* ve vazbě na různé podmiňující jevy $B_1,\ldots,B_k$. Tato metoda se nazývá naivní proto, že teoreticky předpokládá absolutní nezávislost všech podmiňujících jevů. Tento předpoklad sice u většiny praktických aplikací splněn není,

Zadání: Máme k dispozici trénovací soubor 12 pacientů popisující výskyt komplikace po léčbě (*A*) a její možné rizikové faktory, věk (B_1) a diabetes (B_2). Pro vytvoření modelu predikce komplikací u pacientů využijeme naivního bayesovského klasifikátoru.

Pacient	Věk (B_1)	DM (B_2)	Komplikace (*A*)
1	vysoký	ano	ano
2	vysoký	ano	ano
3	vysoký	ne	ano
4	vysoký	ne	ano
5	nízký	ano	ano
6	nízký	ne	ano
7	střední	ano	ano
8	střední	ano	ano
9	nízký	ano	ne
10	nízký	ne	ne
11	střední	ne	ne
12	střední	ne	ne

Aposteriorní pravděpodobnost výskytu komplikací (*A*) v závislosti na rizikových faktorech (B_1, B_2) lze získat pomocí naivního bayesovského klasifikátoru.

$$P(A \mid B_1, B_2) = \frac{P(B_1, B_2 \mid A)P(A)}{P(B_1, B_2)}$$

Apriorní pravděpodobnost výskytu komplikací je odvozena ze souboru

A	Komplikace	$P(A) = \frac{8}{12} = 0{,}667$
$\overline{A}$	Bez komplikace	$P(\overline{A}) = \frac{4}{12} = 0{,}333$

Pro jednotlivé rizikové faktory (B_i) je možné spočítat podmíněné pravděpodobnosti $P(A|B_i)$, v příkladu budeme uvažovat pacienta středního věku s diabetem:

$$P(B_1 = \text{stredni} \mid A) = \frac{2}{8} = 0{,}25 \qquad P(B_1 = \text{stredni} \mid \overline{A}) = \frac{2}{4} = 0{,}5 \qquad P(B_2 = \text{ano} \mid A) = \frac{5}{8} = 0{,}625 \qquad P(B_2 = \text{ano} \mid \overline{A}) = \frac{1}{4} = 0{,}25$$

…obdobně lze podmíněné pravděpodobnosti spočítat pro všechny další kombinace rizikových faktorů a výskytu sledované komplikace (*A*). Pro výpočet lze za předpokladu nezávislosti věku (B_1) a přítomnosti diabetu (B_2) spočítat pravděpodobnost výskytu komplikací následovně:

$$P(A \mid B_1, B_2) = \frac{P(B_1, B_2 \mid A)P(A)}{P(B_1, B_2)} = \frac{P(B_1 \mid A)P(B_2 \mid A)P(A)}{P(B_1 \mid A)P(B_2 \mid A)P(A) + P(B_1 \mid \overline{A})P(B_2 \mid \overline{A})P(\overline{A})} = \frac{0{,}1042}{0{,}1042 + 0{,}0417}$$

Pro vybraného pacienta středního věku trpícího diabetem vychází výpočet takto:

$$P(A \mid B_1 = \text{stredni}, B_2 = \text{ano}) = 0{,}7142$$

$$P(\overline{A} \mid B_1 = \text{stredni}, B_2 = \text{ano}) = 0{,}2858$$

⇨ Pacient má vyšší pravděpodobnost, že u něj budeme pozorovat komplikace.

Příklad 3. Využití naivního bayesovského klasifikátoru pro predikci zdravotního stavu pacientů.

ale při dostatečném počtu jevů B_i dosahuje výpočet uspokojivé přesnosti. Jelikož u jevu A v našem případě rozlišujeme pro jednoduchost pouze dva stavy (jev A nastal/nenastal), pak zde klasifikujeme právě do dvou tříd, tedy A a *not* A. Odhadujeme aposteriorní pravděpodobnost nastání jevu A při nastání všech jevů $B_1,\ldots,B_k$:

$$P(A|B_1,B_2,\ldots,B_k) = \frac{P(B_1,B_2,\ldots,B_k|A)P(A)}{P(B_1,B_2,\ldots,B_k)}.$$

anebo v jiném vyjádření odhadujeme aposteriorní pravděpodobnost platnosti hypotézy H při platnosti všech uvažovaných evidencí (E_i):

$$P(H|E_1,E_2,\ldots,E_k) = \frac{P(E_1,E_2,\ldots,E_k|H)P(H)}{P(E_1,E_2,\ldots,E_k)}.$$

Testujeme-li (klasifikujeme) takto více hypotéz (H_1, …H_j) a pro zjednodušení použijeme pouze jednu evidenci, E, pak nejpravděpodobnější je hypotéza s maximální aposteriorní pravděpodobností (značená jako H_{MAP} – **m**aximální **a**posteriorní pravděpodobnost). Tedy podle naivního klasifikátoru odvozeného z Bayesovy věty jde o hypotézu, pro kterou platí:

$$P(H_{MAP}|E) = max_i \frac{P(E|H_i)P(H_i)}{P(E)}, \; i = 1,\ldots,j.$$

přičemž uvedený vztah bychom opět mohli rozvést pro více zvažovaných evidencí $E_1,\ldots,E_k$.

Příklad 3 přináší ukázku použití naivního bayesovského klasifikátoru v klinické praxi, a to pro jednu zvažovanou evidenci a pro více evidencí.

Všimněme si, že využití Bayesovy věty je velice intuitivní a umožňuje i jistou adaptaci na zkoumaný problém a data. Pokud apriorní informace získáváme přímo z experimentálně získaných dat, nazýváme tento soubor trénovací a vlastně na něm nastavujeme parametry Bayesova klasifikátoru pro vlastní využití v neznámém terénu.

Při aplikaci Bayesovy věty nemusíme zkoumat pouze jevy binární (tedy např. výskyt jevu A ano/ne), ale i chování spojitých, a tedy kvantitativních proměnných. Hodnoty těchto náhodných proměnných lze modelovat pomocí známých rozdělení pravděpodobnosti, např. pomocí normálního rozdělení. Odhadujeme tak např. aposteriorní pravděpodobnost výskytu určitého intervalu hodnot náhodné veličiny X při platnosti evidence E, např. $P(X<x_i|E)$. Využíváme přitom stejné vstupní pravděpodobnosti, jako u všech dosud uvedených příkladů, tedy apriorní pravděpodobnost $P(X<x_i)$ a podmíněnou pravděpodobnost $P(E|X<x_i)$. Výpočet lze samozřejmě rozšířit i pro spojité proměnné v roli evidence E, a podmíněné pravděpodobnosti tak zkoumají vzájemné vztahy dvou nebo i více spojitých proměnných. Avšak tato problematika již přesahuje plánovaný rozsah našeho seriálu. V příštím díle tuto část uzavřeme ukázkami aplikace bayesovské statistiky v neurovědách.