Analýza dat v neurologii

XXXIII. Bayesovská statistika v klinických a neurovědních 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

V předchozích dílech seriálu jsme uvedli Bayesův teorém a základy bayesovské statistiky v obecných klinických aplikacích. Zároveň jsme čtenářům slíbili, že tuto poměrně složitou problematiku rozvineme výkladem, jaké má využití v klinickém a neurovědním výzkumu, kde zaujímá významné postavení. Již v díle XXXI jsme se zmínili o samostatné větvi tzv. výpočetních neurověd, která s bayesovskou statistikou pracuje v základním výzkumu i v aplikacích [1]. V tomto díle se pokoušíme přiblížit další užitečné aplikace bayesovského usuzování.

Bayesovská klasifikace na příkladu více jevů

Princip Bayesovy věty jsme vysvětlili v předchozím díle seriálu. V podstatě odhadujeme aposteriorní podmíněnou pravděpodobnost jevu *A* při nastání jevu *B* podle vztahu: $P(A|B) = [P(B|A)P(A)]/P(B)$. Od samotného počátku tedy odhadujeme nikoli pouze pravděpodobnost jevu *A* jako takovou, ale pravděpodobnost jeho nastání při platnosti určité podmínky (nastání jevu *B*). Obecná, nepodmíněná pravděpodobnost jevu *A*, tedy $P(A)$, je jedním ze vstupů výpočtu; hovoříme o tzv. apriorní pravděpodobnosti $P(A)$. Jev *B* v pozici podmínky může nahradit jakkoli specifikovaná evidence *E*, např. platnost určité podmínky v datech. Pomocí znalosti podmíněné pravděpodobnosti $P(B|A)$, nebo alternativně značeno $P(E|A)$, tak zpřesňujeme odhad $P(A|B)$, resp. $P(A|E)$.

Princip **bayesovské klasifikace** vyplývá z výše uvedeného vztahu. Statistickou klasifikaci obecně definujeme jako výpočetní postup zaměřený na konstrukci modelů pro třídění a doplnění dat ve vazbě na jejich informační obsah. Řečeno jednoduše: hledáme pravidla v datech, která nám umožní nějak účelově třídit hodnocené subjekty. Bayesovská klasifikace vychází z Bayesovy věty a tato pravidla definuje pomocí podmíněných pravděpodobností výskytu jevů. Rozhodujeme mezi různými a vzájemně se vylučujícími hypotézami (H_1, H_2, …, H_i) nebo jevy (A_1, A_2, …, A_i), které odpovídají definovaným třídám, a vybíráme nejpravděpodobnější variantu (třídu) s nejvyšší hodnotou $P(H_j|E)$ nebo $P(A_j|E)$. Dle výpočtu se tedy přikloníme k variantě (třídě) s hodnotou

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

V díle XXXII jsme představili jednoduchý příklad pravděpodobnostní klasifikace mezi dvěma jevy (nastání komplikace po léčbě: ano/ne). Příklad 1, který uvádíme zde, rozvíjí téma na složitější situaci s více jevy, mezi nimiž máme rozhodnout. Příklad 1 také dokládá typický postup statistické klasifikace, která by se vždy měla skládat z tzv. **fáze učení** (zahrnuje vlastní vývoj modelu pro klasifikaci dat do daných tříd; učení probíhá na tzv. trénovacích datech nebo **trénovacím souboru dat**), a následně z tzv. **fáze validace**, kdy již dochází k aplikaci modelu na posuzování nových dat a jejich třídění do daných kategorií.

Bayesovské usuzování a klinické aplikace

Poměrně složitý příklad 1 dokládá, že Bayesův vzorec je cenným nástrojem pro klasifikaci subjektů, který může využít i matematický laik. Při pochopení významu vstupních pravděpodobností by dosazení do vztahu neměl být problém. Avšak v této jednoduchosti nesmí zaniknout hlavní přidaná hodnota Bayesova teorému. Jedná se o pravděpodobnostní provádění statistických odhadů, a to i v situacích, kdy vstupní informace nejsou přesné nebo nejsou vůbec dostupné. Samotná apriorní pravděpodobnost jevu *A* v cílové populaci hraje roli jakési váhy, kterou na výstupu zpřesňujeme. Bayesovská koncepce tak v sobě nese značnou adaptabilitu, neboť prováděné odhady je možné lokalizovat dle prevalence uvažovaného onemocnění (jevu) v různých populacích.

V této souvislosti také hovoříme o tzv. **bayesovské dedukci či úsudku** (*Bayesian inference*), nebo obecněji o bayesovském přístupu. V pojetí „klasické" statistiky provádíme úsudky na základě pozorovaných souborů dat, většinou získaných v minulosti. Pracujeme s předem stanovenou hypotézou a také pravděpodobností, s jakou jsme „ochotni" se mýlit, jinými slovy jaká pravděpodobnost chybného závěru je ještě přípustná. Bayesovský přístup otevírá cestu obecnějšímu usuzování, do kterého vstupuje objektivní i subjektivní vážení významu jednotlivých fakt. Na rozdíl od jiných popisných nebo klasifikačních metod nedává bayesovský přístup jen jedinou odpověď, ale nabízí pravděpodobnosti, s jakými jednotlivé hypotézy odpovídají provedeným pozorováním. Proto nachází bayesovská statistika uplatnění především v oborech, kde se definitivní závěry na základě retrospektivních dat potýkají s velkou neurčitostí, tedy v ekonomice, kriminalistice, managementu a samozřejmě také v medicíně.

V medicíně se bayesovský přístup uplatňuje zvláště v situacích, kdy musíme hodnotit pravděpodobnost určité hypotézy nebo nastání jevu, a na vstupu máme z objektivních důvodů pouze neurčité informace. Například pokud nemůžeme násobně opakovat experiment za statisticky stabilních podmínek anebo z etických důvodů, dále při odhadu prevalence chorob, predikce výsledku diagnostických testů při různé prevalenci choroby, zobecnění vý-

Zadání: Máme k dispozici trénovací soubor záznamů 24 pacientů, který obsahuje určení vhodného typu kontaktních čoček (třídy jevu *A*) pro pacienty s různými charakteristikami (jevy $B_1 - B_4$). Pro vytvoření modelu predikujícího správný typ čoček u pacienta využijeme naivního bayesovského klasifikátoru; výsledný model aplikujeme jak na trénovací, tak testovací soubor pacientů.

-	B_1	B_2	B_3	B_4	A
Pacient	Věk	Brýle	Astigma-tizmus	Produkce slz	Vhodné kontaktní čočky
1	nízký	krátkozrakost	ne	redukovaná	žádné
2	nízký	krátkozrakost	ne	normální	měkké
3	nízký	krátkozrakost	ano	redukovaná	žádné
4	nízký	krátkozrakost	ano	normální	tvrdé
5	nízký	dalekozrakost	ne	redukovaná	žádné
6	nízký	dalekozrakost	ne	normální	měkké
7	nízký	dalekozrakost	ano	redukovaná	žádné
8	nízký	dalekozrakost	ano	normální	tvrdé
9	střední	krátkozrakost	ne	redukovaná	žádné
10	střední	krátkozrakost	ne	normální	měkké
11	střední	krátkozrakost	ano	redukovaná	žádné
12	střední	krátkozrakost	ano	normální	tvrdé
13	střední	dalekozrakost	ne	redukovaná	žádné
14	střední	dalekozrakost	ne	normální	měkké
15	střední	dalekozrakost	ano	redukovaná	žádné
16	střední	dalekozrakost	ano	normální	žádné
17	vysoký	krátkozrakost	ne	redukovaná	žádné
18	vysoký	krátkozrakost	ne	normální	žádné
19	vysoký	krátkozrakost	ano	redukovaná	žádné
20	vysoký	krátkozrakost	ano	normální	tvrdé
21	vysoký	dalekozrakost	ne	redukovaná	žádné
22	vysoký	dalekozrakost	ne	normální	měkké
23	vysoký	dalekozrakost	ano	redukovaná	žádné
24	vysoký	dalekozrakost	ano	normální	žádné

Apriorní pravděpodobnost výskytu skupin je odvozena z trénovacího souboru:

$$P(\text{mekke}) = \frac{5}{24} = 0{,}208 \qquad P(\text{tvrde}) = \frac{4}{24} = 0{,}167 \qquad P(\text{zadne}) = \frac{15}{24} = 0{,}625$$

Aposteriorní pravděpodobnost výskytu dané skupiny (*A*) v závislosti na charakteristikách pacientů ($B_1,..,B_4$) lze získat pomocí naivního bayesovského klasifikátoru.

$$P(A \mid B_1,..,B_4) = \frac{P(B_1,..,B_4 \mid A)P(A)}{P(B_1,..,B_4)}$$

Pro vyřešení výše uvedené rovnice je třeba zjistit podmíněné pravděpodobnosti: $P(B_1 \mid A)$ až $P(B_4 \mid A)$

P(nízký | měkké) = 0,400
P(střední | měkké) = 0,400
P(vysoký | měkké) = 0,200
P(dalekozrakost | měkké) = 0,600
P(krátkozrakost | měkké) = 0,400
P(ne | měkké) = 1,000
P(ano | měkké) = 0,000
P(normální | měkké) = 1,000
P(redukovaná | měkké) = 0,000

P(mladý | tvrdé) = 0,500
P(střední | tvrdé) = 0,250
P(vysoký | tvrdé) = 0,250
P(dalekozrakost | tvrdé) = 0,250
P(krátkozrakost | tvrdé) = 0,750
P(ne | tvrdé) = 0,000
P(ano | tvrdé) = 1,000
P(normální | tvrdé) = 1,000
P(redukovaná | tvrdé) = 0,000

P(mladý | žádné) = 0,267
P(střední | žádné) = 0,333
P(vysoký | žádné) = 0,400
P(dalekozrakost | žádné) = 0,533
P(krátkozrakost | žádné) = 0,467
P(ne | žádné) = 0,467
P(ano | žádné) = 0,533
P(normální | žádné) = 0,200
P(redukovaná | žádné) = 0,800

Za použití rovnice pro aposteriorní pravděpodobnosti, zjištěných apriorních pravděpodobností a podmíněných pravděpodobností $P(B_1|A)$ až $P(B_4|A)$ je pro každého pacienta spočtena pravděpodobnost jeho zařazení do třídy s určitým typem kontaktních čoček. Pacient je finálně zařazen do třídy s nejvyšší pravděpodobností. Výsledky modelu jsou uvedeny v následující tabulce.

Pozorované	měkké	tvrdé	žádné
Predikce měkké	5	0	**1**
tvrdé	0	4	0
žádné	0	0	14

Výsledky modelu aplikovaného na trénovací soubor predikují správný typ kontaktních čoček z 95,8 %, tedy s jednou chybnou klasifikací pacienta bez kontaktních čoček do skupiny s měkkými kontaktními čočkami.

Trénovací soubor posloužil k vývoji klasifikačního pravidla, sám o sobě ovšem nemůže sloužit k jeho nezávislému ověření. Aplikace modelu na soubor dat použitý k jeho tvorbě v sobě skrývá riziko nerealisticky vysokého podílu správných predikcí. Pro korektní zhodnocení predikční síly modelu je před jeho nasazením v praxi třeba provést ověření na nezávislém, tzv. testovacím souboru, které je doloženo v druhé části příkladu 1.

Příklad 1a. Využití naivního bayesovského klasifikátoru pro vytvoření modelu predikujícího vhodný typ kontaktních čoček pro pacienty s poruchami zraku.

sledků klinických studií apod. V přehledu literatury níže uvádíme další vybrané práce aplikující bayesovskou statistiku v neurovědách a některé významné aplikace dále popíšeme v následujícím textu.

Bayesovský mozek a bayesovské filtry

Výklad bayesovské pravděpodobnosti má blízko k neurologii a k neurovědám obecně. Řada výzkumů vychází z předpokladu, že nervový systém se při zpracování senzorických signálů řídí pravděpodobnostními modely, které lze reprezentovat pomocí bayesovské statistiky [2]. Tzv. **bayesovský mozek** představuje neurovědní metodický přístup usilující o vysvětlení kognitivních funkcí mozku pomocí statistických principů. Základem je předpoklad, že nervový systém musí data ze senzorických vjemů uspořádat do vlastního interního modelu, odrážejícího realitu vnějšího světa. Mozek je studován jako nástroj generující pravděpodobnostní rozhodnutí na základě podnětů z částečně neznámého vnějšího světa; pravděpodobnost daná bayesovským modelem je takto využívána pro studium behaviorálních i mentálních procesů [3,4].

Bayesovský mozek je modelem intuitivního rozhodování na základě znalosti podmíněných pravděpodobností určitých jevů. Např. jev *A* je vyhodnocen jako nebezpečný na základě toho, s jakou pravděpodobností (dáno přímou nebo předanou zkušeností) je spojen s určitým nebezpečím. Tento způsob kódování vjemů je také nazýván **bayesovský filtr**. Jako metoda je využíván např. při indexaci položek v databázích nebo při boji se spamem či jinou nežádoucí formou elektronické komunikace. Využijme této poměrně aktuální a všeobecně srozumitelné problematiky k vysvětlení funkcí bayesovských filtrů.

K pochopení funkce bayesovského filtru stačí následující princip: slova v těle nebo v záhlaví emailu ukládá počítač do databáze a označuje je podle toho, zda šlo nebo nešlo o spamovou komunikaci. Takto postupně vzniká tzv. **zkušenostní** neboli **kalibrační databáze**. Následně na základě pravděpodobnosti, zda se v historii dané slovo vázalo na spam, filtr vyhodnotí konkrétní zprávu jako spam nebo jako normální sdělení. Abychom mohli odhadnout odpovídající pravděpodobnosti, označme je P_{spam} a $1 - P_{spam}$, potřebujeme vytvořit zdrojovou databázi spamů a normálních emailů. Na základě těchto vstupů (zkušeností) pak vyhodnocujeme slova v nově příchozím emailu a násobením jejich pravděpodobností P_{spam} získáváme hodnotu (výslednou pravděpodobnost), která nám pomůže rozhodnout, zda je email spíše spam, nebo ne [5].

Zdrojová databáze však rychle zastarává nebo se může stát nereprezentativní, a následně tak odchylovat filtr od správných rozhodnutí. To, co tvoří skutečnou přidanou hodnotu bayesovských

Zadání: Pro verifikaci výsledků odvozeného bayesovského klasifikačního pravidla (příklad 1a) využijeme jiný soubor pacientů (n = 30) ze stejné cílové populace, než trénovací soubor využitý k tvorbě modelu. I u tohoto testovacího souboru ovšem známe realitu, tedy nejvhodnější typ čoček pro každého pacienta (třídy jevu A). Pro klasifikaci pacientů zde využijeme apriorní pravděpodobnosti a podmíněné pravděpodobnosti $P(B_1|A)$ až $P(B_4|A)$ zjištěné na trénovacím souboru. Pro každého pacienta z testovacího souboru je takto spočtena pravděpodobnost jeho zařazení do třídy s určitým typem kontaktních čoček.

-	B_1	B_2	B_3	B_4	A	Výsledky modelu			
Pacient	Věk	Brýle	Astigma-tizmus	Produkce slz	Vhodné kontaktní čočky	P (měkké)	P (tvrdé)	P (žádné)	Výsledná klasifikace
1	nízký	krátkozrakost	ano	normální	tvrdé	0,000	0,882	0,117	tvrdé
2	nízký	dalekozrakost	ano	normální	tvrdé	0,002	0,686	0,312	tvrdé
3	střední	dalekozrakost	ano	normální	tvrdé	0,002	0,467	0,531	žádné
4	vysoký	krátkozrakost	ano	normální	tvrdé	0,000	0,715	0,285	tvrdé
5	střední	krátkozrakost	ano	normální	tvrdé	0,001	0,750	0,249	tvrdé
6	nízký	krátkozrakost	ne	normální	měkké	0,820	0,002	0,179	měkké
7	nízký	dalekozrakost	ne	normální	měkké	0,857	0,000	0,142	měkké
8	střední	krátkozrakost	ne	normální	měkké	0,785	0,001	0,214	měkké
9	střední	dalekozrakost	ne	normální	měkké	0,828	0,000	0,172	měkké
10	vysoký	dalekozrakost	ne	normální	měkké	0,667	0,000	0,332	měkké
11	vysoký	krátkozrakost	ne	normální	měkké	0,604	0,001	0,395	měkké
12	nízký	krátkozrakost	ne	redukovaná	žádné	0,001	0,000	0,999	žádné
13	nízký	krátkozrakost	ano	redukovaná	žádné	0,000	0,002	0,998	žádné
14	nízký	dalekozrakost	ne	redukovaná	žádné	0,002	0,000	0,998	žádné
15	nízký	dalekozrakost	ano	redukovaná	žádné	0,000	0,001	0,999	žádné
16	střední	krátkozrakost	ne	redukovaná	žádné	0,001	0,000	0,999	žádné
17	střední	krátkozrakost	ano	redukovaná	žádné	0,000	0,001	0,999	žádné
18	střední	dalekozrakost	ne	redukovaná	žádné	0,001	0,000	0,999	žádné
19	střední	dalekozrakost	ano	redukovaná	žádné	0,000	0,000	1,000	žádné
20	střední	dalekozrakost	ano	normální	žádné	0,002	0,467	0,531	žádné
21	vysoký	krátkozrakost	ne	redukovaná	žádné	0,000	0,000	1,000	žádné
22	vysoký	krátkozrakost	ne	normální	žádné	0,604	0,001	0,395	měkké
23	vysoký	krátkozrakost	ano	redukovaná	žádné	0,000	0,001	0,999	žádné
24	vysoký	dalekozrakost	ne	redukovaná	žádné	0,001	0,000	0,999	žádné
25	vysoký	dalekozrakost	ano	redukovaná	žádné	0,000	0,000	1,000	žádné
26	vysoký	dalekozrakost	ano	normální	žádné	0,001	0,422	0,577	žádné
27	vysoký	krátkozrakost	ano	redukovaná	žádné	0,000	0,001	0,999	žádné
28	vysoký	dalekozrakost	ano	normální	žádné	0,001	0,422	0,577	žádné
29	střední	krátkozrakost	ne	normální	žádné	0,785	0,001	0,214	měkké
30	nízký	dalekozrakost	ano	normální	žádné	0,002	0,686	0,312	tvrdé

Pozorované	měkké	tvrdé	žádné
Predikce měkké	6	0	2
tvrdé	0	4	1
žádné	0	1	16

Model při aplikaci na trénovací soubor určuje správný typ kontaktních čoček z 86,7 %, tedy se čtyřmi chybnými klasifikacemi.

Závěr:
Model vytvořený na trénovacím souboru (n = 24) byl nezávisle ověřen na testovacím souboru pacientů (n = 30) a byla prokázána jeho dobrá predikční (klasifikační) schopnost. Model lze doporučit pro aplikaci do praxe.

Příklad 1b. Validace naivního bayesovského klasifikátoru predikujícího vhodný typ kontaktních čoček pro pacienty s poruchami zraku.

filtrů, je schopnost průběžně pracovat se zpětnou vazbou. Někdy hovoříme o tzv. **bayesovském učení** jako o procesu postupné validace a modifikace modelu na základě nově dostupných hodnot. Opět zde vidíme rozdíl mezi klasickou, frekventistickou statistikou a bayesovským úsudkem, v jehož pojetí je pravděpodobnost určitého sledovaného jevu upravována podle nových podnětů, informací, důkazů (apriorní vstupy).

Další příklady praktického využití bayesovské statistiky

Možnost subjektivního nastavení apriorních pravděpodobností je bayesovskému přístupu často vytýkána, neboť u vědeckých prací se snažíme minimalizovat subjektivní vliv posuzovatele na vstupní informace. Apriorní pravděpodobnosti by tedy měly být co nejlépe podloženy, ideálně nějakými externími a nezávislými zdroji. Pocházejí-li vstupní informace z různých zdrojů, je velmi užitečné použít váženou kombinaci jejich vlivu, např. podle věrohodnosti, podobnosti anebo významu.

Hodnocení relevance vstupních informací je velmi užitečným příkladem využití bayesovské statistiky, který jistě ocení všichni, kdo řeší problém spojování údajů z různých zdrojů. Často totiž v medicíně k studovanému problému získáme informace z randomizovaných prospektivních klinických studií (tzv. *evidence level* A), ale také z retrospektivních observačních studií a případně i z *ad hoc* pozorování. Pokud jsme schopni význam informačních zdrojů určitým způsobem vážit, pak nám bayesovská klasifikace nabízí jednoduchou možnost výsledky kombinovat.

Předpokládejme, že chceme odhadnout pravděpodobnost, že určitá informace I je důvěryhodná, přičemž data čerpáme ze tří různých zdrojů s různou věrohodností. Nastavme věrohodnostní váhy pro jednotlivé zdroje informací jako podmíněnou pravděpodobnost, že informace je spolehlivá, když o ní informuje daný zdroj: $P(I|Z_1)$ = 0,5; $P(I|Z_2)$ = 0,3 a $P(I|Z_3)$ = 0,2. K dispozici máme přehled článků s informací I ze všech tří zdrojů, a to v následujícím zastoupení: $P(Z_1)$ = 0,2; $P(Z_2)$ = 0,3 a nejčastěji zastoupený zdroj č. 3, $P(Z_3)$ = 0,5. Příklad reflektuje reálnou situaci, kdy zdroji 1 věříme nejvíce (jde např. o prospektivní, řádně plánované studie), nicméně prací tohoto typu máme nejméně ze všech. Pravděpodobnost, že z daného přehledu článků získáváme relevantní a důvěryhodnou informaci I, spočítáme následovně:

$$P(I) = P(I|Z_1)P(Z_1) + P(I|Z_2)P(Z_2) + P(I|Z_3)P(Z_3) =$$
$$= 0,5 \times 0,2 + 0,3 \times 0,3 + 0,2 \times 0,5 = 0,29.$$

Pokud bychom tento výpočet srovnali se situací, kde budeme mít k dispozici více zdrojů v nejvěrohodnější třídě, tedy např. $P(Z_1)$ = 0,6; $P(Z_2)$ = 0,3 a $P(Z_3)$ = 0,1, pak získáme $P(I)$ = 0,41. V podstatě takto počítáme pravděpodobnostní skóre důvěryhodnosti heterogenních informačních zdrojů, což můžeme využít při srovnávání různých rešerší nebo metaanalýz.

Kromě hodnocení věrohodnosti vstupních informací lze bayesovský přístup obdobně využít i při skórování důvěryhodnosti nebo relevance výstupů studií (tzv. *Bayesian Credibility Analysis*, např. [6]).

Zadání: Máme k dispozici trénovací soubor n = 8 jedinců, u nichž známe pohlaví (třídy jevu A) a dále tři kvantitativní charakteristiky: výšku, hmotnost a délku chodidla. Trénovací soubor slouží k vývoji klasifikačního pravidla, které bude na základě proměnných $B_1 - B_3$ rozlišovat pohlaví jedince. Příklad ukazuje využitelnost bayesovské klasifikace za situace, kdy prediktory jsou spojité proměnné.

Apriorní pravděpodobnost výskytu kategorií pohlaví (muž/žena) je odhadnuta přímo z datového souboru:

$$P(\text{muz}) = \frac{4}{8} = 0{,}5 \qquad P(\text{zena}) = \frac{4}{8} = 0{,}5$$

Aposteriorní pravděpodobnost výskytu dané třídy (A) v závislosti na prediktorech ($B_1,..,B_n$) lze získat pomocí naivního bayesovského klasifikátoru.

$$P(A \mid B_1,..,B_3) = \frac{P(B_1,..,B_3 \mid A)P(A)}{P(B_1,..,B_3)}$$

-	B_1	B_2	B_3	A
Pacient	**Výška**	**Hmotnost**	**Chodidlo**	**Pohlaví**
1	182,9	81,6	30,5	muž
2	180,4	86,2	27,9	muž
3	170,1	77,1	30,5	muž
4	180,4	74,8	25,4	muž
5	152,4	45,4	15,2	žena
6	167,6	68,0	20,3	žena
7	165,2	59,0	17,8	žena
8	175,3	68,0	22,9	žena

Výpočet podmíněné pravděpodobnosti $P(B_1|A)$ až $P(B_3|A)$ vychází ze znalosti distribuce spojitých proměnných v jednotlivých kategoriích pohlaví. Za předpokladu normálního rozdělení proměnných $B_1 - B_3$ je možné spojitá data popsat pomocí průměru a směrodatné odchylky prediktorů v těchto kategoriích.

Průměr ± ± sm. odch.	Výška (cm)	Hmotnost (kg)	Chodidlo (cm)
muž	178,5 ± 5,7	79,9 ± 5,0	28,6 ± 2,4
žena	165,1 ± 9,5	60,1 ± 10,7	19,1 ± 3,3

Nechť prediktor B_x má průměr μ a rozptyl σ^2. Pravděpodobnost zařazení jedince s hodnotou prediktoru $B_x = v$ do skupiny A_1 (muž) je určena pomocí rovnice pro hustotu pravděpodobnosti normálního rozdělení.

$$P(B_x = v \mid A_1) = \frac{1}{\sqrt{2\pi\sigma_{A_1}^2}} e^{\frac{(v-\mu_{A_1})}{2\sigma_{A_1}^2}}$$

Za použití výše popsaných vstupů lze vypočítat hodnotu pravděpodobnosti $P(A|B_1, \ldots B_3)$, a tedy ke každému jedinci v trénovacím souboru zjistit pravděpodobnost, zda jde o muže či ženu. Výsledky této trénovací fáze klasifikace shrnuje následující tabulka:

-	Výsledky modelu		
Pacient	**P (žena)**	**P (muž)**	**Predikce**
1	0,000	1,000	muž
2	0,000	1,000	muž
3	0,001	0,999	muž
4	0,014	0,986	muž
5	1,000	0,000	žena
6	1,000	0,000	žena
7	1,000	0,000	žena
8	0,933	0,067	žena

Model aplikovaný na trénovací soubor klasifikuje (predikuje) správně pohlaví jedinců ve 100 % případů. Pokud bychom model chtěli dále verifikovat pro použití v praxi, bylo by nutné jej prověřit na nezávislém, testovacím souboru dat.

Příklad 2. Využití naivního bayesovského klasifikátoru pro klasifikaci pohlaví jedinců na základě spojitých charakteristik.

Zde posuzujeme důvěryhodnost výsledků studie (aposteriorní pravděpodobnost) na základě jejích výsledků ve formě intervalů spolehlivosti pro odhadovanou statistiku a dále s pomocí apriorní znalosti těchto intervalů. Těmito aplikacemi se budeme zabývat v některém z dalších dílů seriálu.

Velmi užitečné je využití podmíněné pravděpodobnosti při studiu **vzájemné závislosti výskytu náhodných jevů**, sledovaných např. v 2 × 2 nebo obecněji v $r \times c$ kontingenčních tabulkách. Tyto metody umožňují promítnout do výpočtu apriorní informace o nezávislosti jevů. Jelikož zde výklad již přesahuje rámec našeho seriálu, omezíme se pouze na dva jednodušší příklady pravděpodobnostního hodnocení výskytu dvou jevů:

- Hodnotíme bezpečnost určitého balení léku v plastových lahvičkách. Z rozsáhlejšího auditu vyplynulo, že výskyt balení, které obsahuje problematickou koncentraci škodlivin, je 9 %, a tedy tuto situaci očekáváme u 9 lahviček ze 100. Toxikologické posudky říkají, že problematická expozice by u pacienta nastala v případě, pokud by po sobě zkonzumoval dvě vadná balení. Předpokládejme, že lahvičky při výdeji náhodně vybíráme z větších balení po 100 kusech. Jaká je pravděpodobnost, že pacientovi naráz vydáme dvě balení léku se zvýšenou hladinou škodlivin? K řešení využijeme podmíněnou pravděpodobnost. Aby mohla nastat situace, kdy obě vydané lahvičky budou vadné, musí toto nejprve nastat u první z nich (jev A). $P(A)$ vypočítáme jako 9/100 = 0,09. U druhé vybírané lahvičky již vybíráme z 99 zbylých a teoreticky 8 z nich může být vadných. Tedy pravděpodobnost výběru druhé vadné lahvičky (jev B), když první vybraná byla vadná, je $P(B|A)$ = 8/99 = 0,081. Výsledný výpočet pravděpodobnosti výběru dvou vadných lahviček po sobě je následující: $P(A \cap B) = P(B|A) \times P(A)$ = 0,0073. Pravděpodobnost, že pacienta neúměrně zatížíme škodlivinami, je tedy při daném způsobu vydávání léku relativně nízká.
- Při podání určitého léku hrozí dva typy nebezpečné toxicity. Při sledování n = 100 pacientů jsme pozorovali výskyt jen toxicity typu I u 5 pacientů, jen toxicity typu II u 10 pacientů a oba typy současně nastaly u 20 pacientů. Zajímá nás, zda z těchto dat lze odvodit vzájemnou nezávislost anebo závislost obou typů toxicity. Typ I nastává u 25 pacientů, z čehož vyplývá $P(I)$ = 0,25. Obdobně $P(II)$ = 0,3 a pravděpodobnost nastání obou toxických reakcí současně $P(I \cap II)$ = 0,2. Podmínku pro nezávislost dvou jevů jsme v díle XXXI seriálu definovali jako $P(I \cap II) = P(I) \times P(II)$. V našem případě ovšem zjevně platí $P(I) \times P(II)$ = 0,25 × 0,3 = 0,075 ≠ $P(I \cap II)$ = 0,2. Jinými slovy, oba jevy jsou vzájemně zá-

Zadání: Máme k dispozici soubor pacientů trpících mírnou kognitivní poruchou anebo Alzheimerovou chorobou. Soubor je doplněný i o kontrolní osoby bez potíží, celkem tak pracujeme se vzorkem $n = 833$. Cílem analýzy je na základě dostupných charakteristik pacientů ($B_1 – B_2$) a objemů různých oblastí mozku zjištěných pomocí MR ($B_3 – B_7$) vytvořit model správně klasifikující pacienty do některé ze tří skupin (skupiny A).

-	B_1	B_2	B_3	B_4	B_5	B_6	B_7	A
Pacient	Pohlaví	Věk (roky)	Hippokampus (mm³)	Amygdala (mm³)	Thalamus (mm³)	Pallidum (mm³)	Putamen (mm³)	Skupina
1	M	84	6996	2725	12800	3914	11227	kontrola
2	M	70	6605	2881	12582	3828	11210	kognitivni
3	M	76	6348	2679	12501	3657	11238	Alzheimer
...	...	...	...	...	...	...	...	...
833	M	65	6415	2659	12922	3695	10895	Alzheimer

Apriorní pravděpodobnost výskytu jednotlivých skupin je odvozena ze souboru:

$$P(\text{kontrola}) = \frac{230}{833} = 0{,}276 \qquad P(\text{kognitivní porucha}) = \frac{406}{833} = 0{,}487 \qquad P(\text{Alzheimer}) = \frac{197}{833} = 0{,}236$$

Pro výpočet aposteriorní pravděpodobnosti přiřazení pacienta k některé ze skupin je třeba zjistit apriorní podmíněné pravděpodobnosti výskytu prediktorů v jednotlivých třídách pacientů a kontrol. V případě kategoriálního prediktoru typu pohlaví (B_1) jde přímo o podmíněné pravděpodobnosti *P(mužlA), P(ženalA)*. V případě spojitých proměnných je za předpokladu normálního rozdělení spočítán průměr a rozptyl v rámci skupin pacientů a kontrol a dále dosazen do rovnice pro hustotu normálního rozdělení.

P(mužlkontrola) = 0,522
P(ženalkontrola) = 0,478
P(mužlkognitivní) = 0,640
P(ženalkognitivní) = 0,360
P(mužlAlzheimer) = 0,518
P(ženalAlzheimer) = 0,482

Průměr ± ± sm. odch	Věk (roky)	Hippokampus (mm³)	Amygdala (mm³)	Thalamus (mm³)	Pallidum (mm³)	Putamen (mm³)
kontrola	75,5 ± 5,0	7 054 ± 186	2 994 ± 157	12 661 ± 284	3 745 ± 214	11 275 ± 132
kognitivní	74,3 ± 7,4	6 553 ± 176	2 717 ± 213	12 693 ± 251	3 730 ± 196	11 147 ± 210
Alzheimer	75,2 ± 7,7	6 255 ± 181	2 693 ± 205	12 664 ± 270	3 699 ± 211	11 060 ± 174

Za použití takto zjištěných apriorních pravděpodobností a podmíněných pravděpodobností $P(B_1|A)$ až $P(B_7|A)$ je pro každého pacienta spočtena aposteriorní pravděpodobnost jeho přiřazení k jednotlivým třídám. Pacient je přiřazen ke skupině s nejvyšší aposteriorní pravděpodobností, výsledky modelu jsou uvedeny v následující tabulce:

Skutečnost		kontrola	kognitivní	Alzheimer
Predikce	kontrola	216 (25,9%)	12 (1,4%)	0 (0,0%)
	kognitivní	14 (1,7%)	349 (41,9%)	62 (7,4%)
	Alzheimer	0 (0,0%)	45 (5,4%)	135 (16,2%)

Model poskytuje správnou predikci u 84 % pacientů. Nejčastější chybou predikce je záměna mírné kognitivní poruchy a Alzheimerovy choroby (12,8 %), chybná klasifikace mezi kontrolami a mírnou kognitivní poruchou nastává v 3,1 %. Nedošlo k žádné chybné klasifikaci mezi kontrolami a pacienty s Alzheimerovou chorobou.

Příklad 3. Využití naivního bayesovského klasifikátoru pro identifikaci neurologického postižení na základě spojitých a kategoriálních prediktorů.

vislé. Má zde tedy smysl odhadnout podmíněné pravděpodobnosti vzájemného výskytu obou jevů: $P(I|II) = P(I \cap II)/P(II) = 0{,}2/0{,}3 = 0{,}667$ a obdobně $P(II|I) = 0{,}8$. Je patrné, že v obou případech jsou podmíněné pravděpodobnosti významně vyšší než nepodmíněné pravděpodobnosti výskytu obou jevů $P(I)$ a $P(II)$.

Využití bayesovské statistiky pro spojité proměnné

Jak jsme již konstatovali v díle XXII seriálu, bayesovské odhady a predikce lze v plném rozsahu použít i pro spojité proměnné. Jediným rozdílem je zde způsob, jak kalkulujeme pravděpodobnost výskytu hodnot spojité proměnné. Místo pravděpodobnosti výskytu náhodných jevů A a B do Bayesova teorému dosazujeme hustoty pravděpodobnosti výskytu určitých hodnot spojitých proměnných. U spojité proměnné X tak může jít například o pravděpodobnost výskytu:

- průměrných hodnot ± 2 nebo 3 směrodatné odchylky: $\mu \pm 2\sigma$; $\mu \pm 3\sigma$
- intervalu spolehlivosti pro odhad průměru: $\mu \pm 1{,}96\sigma_e$
- arbitrárně určeného intervalu hodnot daných např. rizikovými hranicemi: $X > x_a$

Jako příklad uveďme odhad aposteriorní pravděpodobnosti výskytu rizikových hodnot koncentrace krevního markeru X za platnosti podmínky B, což může být např. relaps sledovaného onemocnění. Odhadujeme tak aposteriorní pravděpodobnost, že hodnoty markeru překročí hraniční hodnotu označenou x_a:

$$P(X > x_a|B) = [P(B|X > x_a) \times P(X > x_a)] / P(B)$$

Pravděpodobnost překročení koncentrace x_a u daného markeru $P(X > x_a)$ odhadujeme z distribuční funkce neboli z rozdělení pravděpodobnosti tohoto znaku. Z apriorních informací predikujeme aposteriorní pravděpodobnost překročení hranice x_a při nastání relapsu onemocnění, tedy $P(X > x_a|B)$. Příklady 2 a 3 ukazují dva praktické výpočetní postupy pracující s kvantitativními znaky. Příklad 3 dokládá, že bayesovské klasifikátory jsou užitečné i pro složité soubory z klinické praxe obsahující kombinaci různých typů prediktorů s cílem zařazovat pacienty do více než dvou skupin. Výsledná tabulka nám přitom umožňuje velmi přehledně zhodnotit přesnost provedených predikcí a určit, kde nastala případná chyba.

Trochu historie závěrem

Při výkladu Bayesova teorému nemůžeme pominout aspoň krátký náhled do historie této významné kapitoly matematické statistiky. Teorém nese jméno po Thomasu Bayesovi (1701–1761), anglickém matematikovi a teologovi. Ačkoli Thomas Bayes za svého života publikoval teologické i matematické práce, jeho objevy v oblasti teorie pravděpodobnosti zůstaly ve formě poznámek a byly publikovány až po jeho smrti, a to jeho přítelem Richardem Pricem (1793: *An Essay Towards Solving a Problem in the Doctrine of Chances*) [7]. Následné rozpracování matematického systému pravděpodobnostní indukce je zásluhou velmi významného Bayesova následovníka, Pierre Si-

mona Laplace (1749–1827), který v roce 1814 ve slavném spise *Théorie analytique des probabilités* definoval principy teorie pravděpodobnosti. Položil tak základ pojetí pravděpodobnosti jako nástroje pro popis všech problémů s neúplnou vstupní informací. Genialita Laplaceova přínosu spočívá v zobecnění teorie pravděpodobnosti, včetně případů, pro které opakovaný výskyt v experimentu nebo pozorovaný hromadný výskyt nemají smysl.

Je nepochybné, že Thomas Bayes přispěl k velmi užitečnému chápání pravděpodobnosti; tu nevnímal jen jako zobecněnou relativní četnost, ale jako nástroj popisu částečné znalosti systému. Doslova znovuzrození zažívá bayesovská pravděpodobnost od 60. let minulého století, odkdy je díky rozvoji počítačové techniky využívána v řadě vědních oborů. Je až fascinující, jakých rozměrů dosahuje zobecněná interpretace základů starých téměř 250 let. Tento metodický koncept našel významné uplatnění v psychologii, ve výzkumu behaviorálních a motorických funkcí, elektrofyziologii a také v teoretickém výzkumu způsobu kódování informací v centrálním nervovém systému [8,9]. Tento výzkum pak zpětně inicioval moderní metody biostatistiky a bioinformatiky, jako je strojové učení a výpočty na bázi umělé inteligence [10–12].

Literatura

1. Trappenberg TP. Fundamentals of Computational Neuroscience. 2nd ed. Oxford: Oxford University Press 2010.
2. Knill D, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 2004; 27(12): 712–719.
3. Fahlman SE, Hinton GE, Sejnowski TJ. Massively parallel architectures for A.I. Netl, Thistle, and Boltzmann machines. Proceedings of the National Conference on Artificial Intelligence. Washington DC 1983.
4. Jaynes ET. How Does the Brain Do Plausible Reasoning? In: Erickson GJ, Smith CR (eds). Maximum-Entropy and Bayesian Methods in Science and Engineering. Berlin: Springer 1988.
5. Kára M. Jak funguje bayesovský antispamový filtr? (1.). Lupa.cz : server o českém internetu [online] 2005 [cit. 2009-03-11]. Dostupný z URL: http://www.lupa.cz/clanky/jak-funguje-bayesovsky-antispamovy-filtr-1.
6. Matthews RAJ. Methods for assessing the credibility of clinical trial outcomes. Drug Information J 2001; 35(4): 1469–1478.
7. Edwards AWG. Commentary on the Arguments of Thomas Bayes. Scand J Stat 1978; 5(2): 116–118.
8. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 1999; 2(1): 79–87.
9. Koerding KP, Wolpert DM. Bayesian integration in sensorimotor learning. Nature 2004; 427(6971): 244–247.
10. Citro G, Banks G, Cooper G. INKBLOT: a neurological diagnostic decision support system integrating causal and anatomical knowledge. Artif Intell Med 1997; 10(3): 257–267.
11. Lavrac N. Selected techniques for data mining in medicine. Artif Intell Med 1999; 16(1): 3–23.
12. Ghahramani Z. Unsupervised learning. In: Bousquet O, Raetsch G, von Luxburg U (eds). Advanced lectures on machine learning. Berlin: Springer-Verlag 2004.

Další doporučená literatura

Salamon R, Bernadet M, Samson M, Derouesne C, Gremy F. Bayesian method applied to decision-making in neurology – methodological considerations. Methods Inf Med 1976; 15(3): 174–179.

Miller RA. Medical diagnostic decision support systems – past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1994; 1(1): 8–27.