Matematika IV – 11. týden
Bayesovská analýza, testování hypotéz

Jan Slovák

Masarykova univerzita
Fakulta informatiky

26. 11. 2013

Obsah přednášky

Plán přednášky

Kde je dobré číst?

- Karel Zvára, Josef Štěpán, Pravděpodobnost a matematická pravděpodobnost statistika, Matfyzpress, 2006, 230pp.
- J. Slovák, M. Panák, M. Bulant, Matematika drsně a svižně, Muni Press, Brno 2013, v+773 s., elektronická edice `www.math.muni.cz/Matematika_drsne_svizne`
- Marie Budíková, Štěpán Mikoláš, Pavel Osecký, Teorie pravděpodobnosti a matematická statistika (sbírka příkladů), Masarykova univerzita, 3. vydání, 2004, 117 stran, ISBN 80-210-3313-4.
- Marie Budíková, Tomáš Lerch, Štěpán Mikoláš, Základní statistické metody, Masarykova univerzita, 2005, 170 stran, ISBN 80-210-3886-1.
- Riley, K.F., Hobson, M.P., Bence, S.J. Mathematical Methods for Physics and Engineering, second edition, Cambridge University Press, Cambridge 2004, ISBN 0 521 89067 5, xxiii + 1232 pp.

Plán přednášky

Bayesův vzorec pro podmíněnou pravděpodobnost (tzv. inverzní pravděpodobnost):

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

Bayesův vzorec pro podmíněnou pravděpodobnost (tzv. inverzní pravděpodobnost):

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

Na úrovni hustot (nebo pravděpodobnostních funkcí) náhodných veličin: máli vektor (X, Θ) hustotu $f(x|\theta)$, pak podmíněná pravděpodobnost komponenty Θ za podmínky $X = x$ hustotu $g(\theta|x)$ danou

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)}.$$

Bayesův vzorec pro podmíněnou pravděpodobnost (tzv. inverzní pravděpodobnost):

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

Na úrovni hustot (nebo pravděpodobnostních funkcí) náhodných veličin: máli vektor (X, Θ) hustotu $f(x|\theta)$, pak podmíněná pravděpodobnost komponenty Θ za podmínky $X = x$ hustotu $g(\theta|x)$ danou

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)}.$$

Mluvíme o **apriorní hustotě** $g(\theta)$ a **aposteriorní hustotě** $g(\theta|x)$.

Bayesův vzorec pro podmíněnou pravděpodobnost (tzv. inverzní pravděpodobnost):

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

Na úrovni hustot (nebo pravděpodobnostních funkcí) náhodných veličin: máli vektor (X, Θ) hustotu $f(x|\theta)$, pak podmíněná pravděpodobnost komponenty Θ za podmínky $X = x$ hustotu $g(\theta|x)$ danou

$$g(\theta|x) = \frac{f(x|\theta)g(\theta)}{f(x)}.$$

Mluvíme o **apriorní hustotě** $g(\theta)$ a **aposteriorní hustotě** $g(\theta|x)$. Protože předem víme, že $g(\theta|x)$ je hustota pravděpodobnosti, nemusí nás konstanta $f(x)$ vůbec zajímat — počítáme prostě až na násobek konstantou.

Předpokládejme, že na univerzitě je spokojenost studentů v jednotlivých předmětech náhodná veličina $X \sim \mathrm{N}(\theta, \sigma^2)$, zatímco parametr θ dosahovaný jednotlivými učiteli je náhodná veličina $\theta \sim \mathrm{N}(a, b)$.
Můžeme tedy počítat (pořád až na konstatní násobky, tj. ignorujeme součinitele, ve kterých nevystupuje θ) a dostaneme

$$\theta \sim \mathrm{N}\left(\frac{b^2}{b^2 + \sigma^2} x + \frac{\sigma^2}{b^2 + \sigma^2} a, \frac{b^2 \sigma^2}{b^2 + \sigma^2} \right).$$

Předpokládejme, že na univerzitě je spokojenost studentů v jednotlivých předmětech náhodná veličina $X \sim \mathrm{N}(\theta, \sigma^2)$, zatímco parametr θ dosahovaný jednotlivými učiteli je náhodná veličina $\theta \sim \mathrm{N}(a, b)$.
Můžeme tedy počítat (pořád až na konstatní násobky, tj. ignorujeme součinitele, ve kterých nevystupuje θ) a dostaneme

$$\theta \sim \mathrm{N}\left(\frac{b^2}{b^2+\sigma^2} x + \frac{\sigma^2}{b^2+\sigma^2} a, \frac{b^2\sigma^2}{b^2+\sigma^2} \right).$$

Když tedy z dlouhodobého vyhodnocování anket známe parametry a, b, σ, můžeme po vyjádření nějakého studenta upřesnit apriorní představu o parametrech pro jeden konkrétní předmět. Ve výsledném odhadu rozložení je pak střední hodnota dána váženým průměrem zjištěné hodnoty x a apriorně předpokládané střední hodnoty a, v závislosti na rozptylech σ a b.

Bayesovská interpretace?

Pro $\sigma \to 0$ je váha jediného názoru stále rostoucí a tomu odpovídá 100% váha u x v případdě $\sigma = 0$. Je to plně v souladu s interpretací, že Bayesovská statistika je pravděpodobnostní rozšíření standardní diskrétní matematické logiky.

Bayesovská interpretace?

Pro $\sigma \to 0$ je váha jediného názoru stále rostoucí a tomu odpovídá 100% váha u x v případdě $\sigma = 0$. Je to plně v souladu s interpretací, že Bayesovská statistika je pravděpodobnostní rozšíření standardní diskrétní matematické logiky.
Místo jednoho studenta použijeme výběrový průměr $\bar{X}$ výsledku šetření. Opět o normální rozdělení, jen budeme místo σ^2 dosazovat σ^2/n. Pišme

$$c_n = \frac{nb^2}{nb^2 + \sigma^2}$$

a aposteriorní odhad pro θ je

$$\theta \sim \mathrm{N}(c_n\bar{X} + (1 - c_n)a, c_n\sigma^2/n).$$

Bayesovská interpretace?

Pro $\sigma \to 0$ je váha jediného názoru stále rostoucí a tomu odpovídá 100% váha u x v případě $\sigma = 0$. Je to plně v souladu s interpretací, že Bayesovská statistika je pravděpodobnostní rozšíření standardní diskrétní matematické logiky.
Místo jednoho studenta použijeme výběrový průměr $\bar{X}$ výsledku šetření. Opět o normální rozdělení, jen budeme místo σ^2 dosazovat σ^2/n. Pišme

$$c_n = \frac{nb^2}{nb^2 + \sigma^2}$$

a aposteriorní odhad pro θ je

$$\theta \sim \mathrm{N}(c_n\bar{X} + (1 - c_n)a, c_n\sigma^2/n).$$

Pro rostoucí n se bude střední hodnota našeho rozdělení pro θ stále více blížit výběrovému průměru a jeho rozptyl půjde k nule. Čím je tedy n větší, tím více se blížíme bodovému odhadu z frekventistického přístupu.

Přínosem Bayesovského přístupu je, že s použitím odhadnutého rozdělení můžeme odpovídat na dotazy typu „s jakou pravděpodobností je nový vyučující horší než předchozí?“
Potřebujeme k tomu apriorní údaje.
Předpokládejme, že máme docela dobře hodnocené učitele: $a = 7{,}5$, $b = 2{,}5$ a a ponecháme směrodatnou odchylku $\sigma = 2$. Pro $n = 15$ a výběrový průměr 5,133 dostaneme aposteriorní odhad pro rozdělení $\theta \sim \mathrm{N}(5{,}230,\ 0{,}256)$.

Přínosem Bayesovského přístupu je, že s použitím odhadnutého rozdělení můžeme odpovídat na dotazy typu „s jakou pravděpodobností je nový vyučující horší než předchozí?“
Potřebujeme k tomu apriorní údaje.
Předpokládejme, že máme docela dobře hodnocené učitele: $a = 7{,}5$, $b = 2{,}5$ a a ponecháme směrodatnou odchylku $\sigma = 2$. Pro $n = 15$ a výběrový průměr 5,133 dostaneme aposteriorní odhad pro rozdělení $\theta \sim \mathrm{N}(5{,}230,\ 0{,}256)$.
Zajímá nás $P(\theta < 6)$. Odpověď získáme dotazem na hodnotu distribuční funkce příslušného normálního rozdělení pro argument 6 – odpověď je cca $93,6\%$. Je tedy podobná, jako jsme viděli v frekventistickém přístupu.

Plán přednášky

Definition

Hypotézou rozumíme nějaké tvrzení o rozdělení určeném sdruženou distribuční funkcí $F_X(x)$ náhodného vektoru $X = (X_1, \ldots, X_n)$. Rozhodujeme mezi tzv. **nulovou hypotézou** H_0 a **alternativní hypotézou** H_A, která bývá negací nulové hypotézy. Možnými rozhodnutími jsou **zamítnutí** nebo **nezamítnutí** nulové hypotézy.

Definition

Hypotézou rozumíme nějaké tvrzení o rozdělení určeném sdruženou distribuční funkcí $F_X(x)$ náhodného vektoru $X = (X_1, \ldots, X_n)$. Rozhodujeme mezi tzv. **nulovou hypotézou** H_0 a **alternativní hypotézou** H_A, která bývá negací nulové hypotézy. Možnými rozhodnutími jsou **zamítnutí** nebo **nezamítnutí** nulové hypotézy.

Když nulovou hypotézu zamítneme, přestože ve skutečnosti platí, nastává **chyba prvního druhu**, když ji nezamítneme v situaci, kdy neplatí, hovoříme o **chybě druhého druhu**.

Statistické rozhodování se opírá o předem určený **kritický obor** W, tj. předem určenou množinu výsledků pokusu, při kterých budeme nulovou hypotézu zamítat.

Statistické rozhodování se opírá o předem určený **kritický obor** W, tj. předem určenou množinu výsledků pokusu, při kterých budeme nulovou hypotézu zamítat.
Tvar kritického oboru oboru volíme tak, abychom platnou hypotézu zamítli s pravděpodobností nejvýše α. Tj. zadáváme předem ohraničení velikosti chyby prvního druhu tzv. hladinou testu α.
Zpravidla volíme $\alpha = 0,05$ nebo $\alpha = 0,01$.
Výpočetní síla dnes umožňuje úkol obrátit a pro daná data se ptát, na jaké **nejmenší** hladině bychom ještě hypotézu zamítli. Hovoříme o **dosažené hladině testu** nebo také p**–hodnotě** (v angličtině P-value nebo Sig. level).

Statistické rozhodování se opírá o předem určený **kritický obor** W, tj. předem určenou množinu výsledků pokusu, při kterých budeme nulovou hypotézu zamítat.
Tvar kritického oboru oboru volíme tak, abychom platnou hypotézu zamítli s pravděpodobností nejvýše α. Tj. zadáváme předem ohraničení velikosti chyby prvního druhu tzv. hladinou testu α.
Zpravidla volíme $\alpha = 0,05$ nebo $\alpha = 0,01$.
Výpočetní síla dnes umožňuje úkol obrátit a pro daná data se ptát, na jaké **nejmenší** hladině bychom ještě hypotézu zamítli. Hovoříme o **dosažené hladině testu** nebo také p**–hodnotě** (v angličtině P-value nebo Sig. level).
Mezi všemi kritickými obory na dané hladině testu ale pochopitelně přitom chceme vybrat ten, který bude minimalizovat chybu druhého druhu.

Předpokládejme, že náhodný vektor X má hustotu rozdělení $f(x, \theta)$ závislou na (vektorovém) parametru. Za nulové hypotézy je to rozdělení s hustotou $f(x, \theta_0)$, za alternativní s hustotou $f(x, \theta_1)$.

Theorem (Neymanovo-Pearsonovo lemma)

Nechť k danému $\alpha \in (0, 1)$ existuje $c > 0$ takové, že pro množinu $W_c = \{x : f(x, \theta_1) \geq cf(x, \theta_0)\}$ platí $\int_{W_c} f(x, \theta_0)dx = \alpha$.
Pak pro každou měřitelnou množinu W takovou, že je $\int_W f(x, \theta_0)dx = \alpha$, platí

$$\int_{W_c} f(x, \theta_1)dx \geq \int_W f(x, \theta_1)dx$$

V případě intervalových odhadů můžeme problém přeformulovat jako hypotézy H_0 – „střední hodnota je μ_0“ a H_A – „střední hodnota je μ_1“. Kritický obor je pak dán požadavkem

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \right| \geq z(\alpha/2)$$

a nezávisí na konkrétní hodnotě μ_1.

Example

Úkol v našem předchozím příkladu o výšce desetiletých chlapců lze formulovat tak, že nulovou hypotézou je nezměněná výška populace, zatímco alternativní je, že se výška změnila (tj. náš kritický obor je symetrický). Hladinu testu pak spočteme na $6,66\%$, takže je přirozené, že jsme nulovou hypotézu na úrovni 5% nezamítli.

V případě intervalových odhadů můžeme problém přeformulovat jako hypotézy H_0 – „střední hodnota je μ_0“ a H_A – „střední hodnota je μ_1“. Kritický obor je pak dán požadavkem

$$|Z| = \left| \frac{\bar{X} - \mu_0}{\sigma} \sqrt{n} \right| \geq z(\alpha/2)$$

a nezávisí na konkrétní hodnotě μ_1.

Example

Úkol v našem předchozím příkladu o výšce desetiletých chlapců lze formulovat tak, že nulovou hypotézou je nezměněná výška populace, zatímco alternativní je, že se výška změnila (tj. náš kritický obor je symetrický). Hladinu testu pak spočteme na $6,66\%$, takže je přirozené, že jsme nulovou hypotézu na úrovni 5% nezamítli.
Když interpretujeme zadání tak, že buď se výška nezměnila, nebo vzrostla, bude náš kritický obor nesymetrický a dojdeme k hladině testu $3,33\%$. Nulovou hypotézu proto na hladině 5% zamítneme.

Plán přednášky

Uvažujme náhodný vektor $Y = (Y_1, \ldots, Y_n)^T$ a předpokládejme, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodností k, β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $\mathrm{N}(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o **lineárním modelu** s úplnou hodností.

Uvažujme náhodný vektor $Y = (Y_1, \ldots, Y_n)^T$ a předpokládejme, že platí

$$Y = X \cdot \beta + \sigma Z,$$

kde $X = (x_{ij})$ je konstantní matice reálných čísel s n řádky a $k < n$ sloupci a hodností k, β je neznámý konstantní vektor k parametrů modelu, Z je náhodný vektor, jehož n komponent má rozdělení $\mathrm{N}(0, 1)$, a $\sigma > 0$ je neznámý kladný parametr modelu. Hovoříme o **lineárním modelu** s úplnou hodností.
V praktických problémech jde často o to, že známe veličiny x_{ij} a snažíme se odhadnout nebo predikovat hodnotu Y.

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).
K odhadu vektoru β se často používá **metoda nejmenších čtverců**. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.

Například x_{ij} může ve vztahu $Y = X \cdot \beta + \sigma Z$ vyjadřovat hodnocení i–tého studenta v j–tém semestru ($j = 1, 2, 3$) z matematiky a chceme vědět, jak tento student asi dopadne ve čtvrtém semestru. K tomu potřebujeme znát vektor β (zatímco σZ vystihuje náhodná vychýlení ve sledovaném modelu). Vektor β odhadneme na základě úplných pozorování, tj. ze znalosti hodnot Y (např. z výsledků v přechozích letech).
K odhadu vektoru β se často používá **metoda nejmenších čtverců**. To znamená, že chceme najít odhad $b \in \mathbb{R}^k$ tak, aby vektor $\hat{Y} = Xb$ minimalizoval druhou mocninu délky vektoru $Y - X\beta$.
To je ale jednoduchá úloha lineární algebry a víme, že jde o nalezení kolmého průmětu vektoru Y do podprostoru $\langle X \rangle \subset \mathbb{R}^n$ generovaném sloupci matice X.

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2.$$

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2.$$

Velikost $\|Y - \hat{Y}\|^2$ nazýváme **reziduální součet čtverců**, zpravidla se značí **RSS**. Definujeme také **reziduální rozptyl** jako

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Minimalizujeme přitom funkci

$$\|Y - X\beta\|^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2.$$

Velikost $\|Y - \hat{Y}\|^2$ nazýváme **reziduální součet čtverců**, zpravidla se značí **RSS**. Definujeme také **reziduální rozptyl** jako

$$S^2 = \frac{\|Y - Xb\|^2}{n - k}.$$

Víme, že $\hat{Y} = Xb$ a že, díky našemu přepokladu o maximální hodnosti X, je matice $X^T X$ invertibilní. Můžeme proto rovnou spočíst $b = (X^T X)^{-1} X^T \hat{Y}$.

Theorem

V lineárním modelu $Y = X\beta + \sigma Z$ *paltí pro vhodné matice* P *a* R*:*
(1) Pro odhad $\hat{Y}$ *platí*

$$\hat{Y} = X\beta + \sigma PP^T Z, \quad \hat{Y} \sim \mathsf{N}(X\beta, \sigma^2 PP^T).$$

(2) Reziduální součet čtverců RSS a normovaný čtverec velikosti rezidua mají rozdělení:

$$Y - \hat{Y} \sim \mathsf{N}(0, \sigma^2 RR^T), \quad \|Y - Y\|^2/\sigma^2 \sim \chi^2_{n-k}\,.$$

(3) Náhodná veličina $b = \beta + \sigma(P^T X)^{-1} P^T Z$ *má rozdělení*

$$b \sim \mathsf{N}(\beta, \sigma^2 (X^T X)^{-1}).$$

(4) Pro reziduální rozptyl platí $(n-k)S^2/\sigma^2 \sim \chi^2_{n-k}$.
(5) Střední hodnota reziduálního rozptylu je $\mathsf{E}\, S^2 = \sigma^2$.
(6) Veličiny b *a* S^2 *jsou nezávislé.*