

NMAI059 Pravděpodobnost a statistika 1

13. přednáška

Robert Šámal

Porovnání MAP a ML

- ▶ Pro připomenutí základní metody bodového odhadu klasické a Bayesovské statistiky
- ▶ pro diskrétní veličiny (u spojitých bychom psali f místo p)
- ▶ ML (maximální věrohodnost): naměříme hodnotu x , odhadneme hodnotu parametru ϑ tak, aby byla maximální

$$P(X = x; \vartheta)$$

- ▶ MAP (maximální a posteriori) naměříme hodnotu x , odhadneme hodnotu parametru ϑ (který považujeme za n.v.) tak, aby byla maximální

$$P(\Theta = \vartheta \mid X = x)$$

Přehled

Permutační test

Bootstrap

Generování náhodných veličin

Situace

- ▶ Máme k dispozici dvě sady nezávislých náhodných veličin (náhodné výběry):
- ▶ $X_1, \dots, X_n \sim F_X$ a $Y_1, \dots, Y_m \sim F_Y$
- ▶ Chceme rozhodnout, zda platí $H_0 : F_X = F_Y$ nebo $H_1 : F_X \neq F_Y$
- ▶ Příklady: doba běhu programu před/po vylepšení, hladina cholesterolu u lidí co jedí/nejedí Zázračnou SuperpotravuTM, frekvenci krátkých slov v textu autora X a Y.
- ▶ Nevíme nic o vlastnostech F_X, F_Y (zejména nečekáme, že je normální)

Postup

- ▶ Zvolíme vhodnou statistiku, např.

$$T(X_1, \dots, X_n, Y_1, \dots, Y_m) = |\bar{X}_n - \bar{Y}_m|$$

- ▶ $t_{\text{obs}} := T(X_1, \dots, X_n, Y_1, \dots, Y_m)$
- ▶ Za předpokladu H_0 jsou „všechny permutace stejné“: X_i i Y_j se generovaly ze stejného rozdělení.
- ▶ Náhodně zpermutujeme zadaných $m + n$ čísel a pro každou permutaci výčíslíme T – dostaneme $T_1, T_2, \dots, T_{(m+n)!}$.
- ▶ Jako p -hodnotu vezmeme pravděpodobnost, že $T > t_{\text{obs}}$, neboli

$$p = \frac{1}{(m+n)!} \sum_j I(T_j > t_{\text{obs}}).$$

- ▶ To je pravděpodobnost chyby 1. druhu, neboli H_0 zamítneme, pokud je $p < \alpha$ (pro naši zvolenou hodnotu α , např. $\alpha = 0.05$).

Vylepšení

- ▶ Zkoušet všechny permutace může trvat moc dlouho. Vezmeme tedy jen vhodný počet B nezávisle náhodně vygenerovaných permutací a spočítáme jenom B hodnot T_1, \dots, T_B .
- ▶ Jako p -hodnotu vezmeme odhad pravděpodobnost, že $T > t_{\text{obs}}$, neboli

$$\frac{1}{B} \sum_{j=1}^B I(T_j > t_{\text{obs}}).$$

- ▶ Pro dostatečně velké m, n dává podobné výsledky jako testy založené na CLV, vhodné je tedy zejména pro středně velké počty.

Přehled

Permutační test

Bootstrap

Generování náhodných veličin

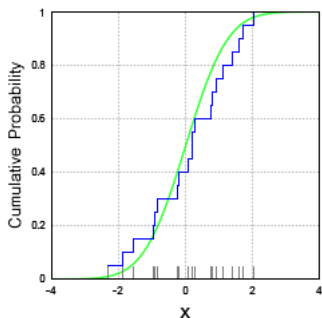
Empirická distribuční funkce – připomenutí

- ▶ $X_1, \dots, X_n \sim F$ n.n.v., F je jejich distribuční funkce
- ▶ **Definice:** *Empirická distribuční funkce (empirical CDF)* je definována

$$\hat{F}_n(x) = \frac{\sum_{i=1}^n I(X_i \leq x)}{n},$$

kde $I(X_i \leq x) = 1$ pokud $X_i \leq x$ a 0 jinak.

(Obrázek vytvořil wiki-editor nagualdesign.)



Bootstrap – základní idea

- ▶ z naměřených dat $X_1 = x_1, \dots, X_n = x_n \sim F$ vytvoříme \hat{F}_n
- ▶ další data můžeme samplovat z \hat{F}_n
- ▶ to se dělá tak, že vybereme uniformně náhodné $i \in \{1, \dots, n\}$ a řekneme x_i

Bootstrap – základní použití

- ▶ $T_n = g(X_1, \dots, X_n)$ nějaká statistika (funkce dat)
- ▶ chceme odhadnout $\text{var } T_n$
- ▶ nasamplujeme $X_1^*, \dots, X_n^* \sim \hat{F}_n$ (viz minulá strana)
- ▶ spočteme $T_n^* = g(X_1^*, \dots, X_n^*)$
- ▶ opakujeme B -krát, dostaneme $T_{n,1}^*, \dots, T_{n,B}^*$
- ▶ odhad rozptylu:

$$\frac{1}{B} \sum_{b=1}^B \left(T_{n,b}^* - \frac{1}{B} \sum_{k=1}^B T_{n,k}^* \right)^2$$

Přehled

Permutační test

Bootstrap

Generování náhodných veličin

Základní metoda (inverse transformation method)

Věta

Nechť F je rostoucí spojitá funkce s $\lim_{x \rightarrow -\infty} F(x) = 0$ a $\lim_{x \rightarrow +\infty} F(x) = 1$.

Nechť $U \sim U(0, 1)$ a $X = F^{-1}(U)$.

Pak X má distribuční funkci F .

- ▶ Funguje dobře, když umíme vyčíslit F^{-1} , třeba pro exponenciální rozdělení.
- ▶ Gamma rozdělení je součet několika exponenciálních – tak ho tak i vygenerujeme.

Zamítací metoda (rejection sampling)

- ▶ Chceme vygenerovat n.v. s hustotou f .
- ▶ Umíme vygenerovat n.v. s hustotou g (která je „podobná“).
- ▶ $\frac{f(y)}{g(y)} \leq c$ pro nějakou konstantu c .
- ▶ Postup
 1. Vygenerujeme Y s hustotou g , a $U \sim U(0, 1)$.
 2. Pokud $U \leq \frac{f(Y)}{cg(Y)}$, tak $X := Y$.
 3. Jinak hodnotu Y , U zamítneme a opakujeme od bodu 1.
- ▶ Zdůvodnění: vygenerovat náhodnou hodnotu X s hustotou f je totéž, jako vygenerovat náhodný bod pod grafem funkce f , jehož vodorovná (x -ová) souřadnice je X (a svislá je uniformně náhodná mezi 0 a X).

Varianta základní metody pro diskrétní proměnné

- ▶ Chceme n.v. X , která nabývá hodnot x_1, x_2, \dots s pravděpodobnostmi p_1, p_2, \dots ($\sum_i p_i = 1$).
 - ▶ Vygenerujeme $U \sim U(0, 1)$.
 - ▶ Najdeme i takové, že $p_1 + \dots + p_{i-1} < U < p_1 + \dots + p_i$.
 - ▶ Položíme $X := x_i$.
-
- ▶ Funguje hezky když máme vzorec pro $p_1 + \dots + p_i$ (např. geometrické rozdělení).
 - ▶ Binomické rozdělení je lepší simulovat jako součet n nezávislých Bernoulliových veličin.
 - ▶ Na další (Poisson) jsou speciální triky).