Bayesovská statistika

Petr Šimeček

Podmiňování

Pravděpodobnost jevu A za podmínky, že nastav jev B

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

Obdobně se zavádí pravděpodobnost, že $X = x$ za podmínky $Y = y$

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)},$$

resp. ve spojitém případě podmíněná hustota

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Bayesova věta – nejjednodušší tvar

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A|B)\,\mathrm{P}(B)+\mathrm{P}(A|not(B))\,\mathrm{P}(not(B))}$$

Schematicky

$$posterior = prior \times verohodnost$$

Příklad – lékařská diagnostika

Máme test na rakovinu, která má specificitu a senzitivitu tj.

$$\mathrm{P}(+|rakovina) = \mathrm{P}(-|zdravy) = 0.95$$

a předpokládáme, že prevalence rakoviny je tzn.

$$\mathrm{P}(rakovina) = 0.005$$

Jaká je pravděpodobnost, že máte rakovinu, pokud vám test vyšel pozitivně?

$$\mathrm{P}(rakovina|+) = ?$$

Bayesova věta pro náhodné jevy

Nechť A je náhodný jev. Nechť $B_1, \ldots B_n$ je soubor disjunktních náhodných jevů takový, že jejich sjednocení je Ω. Nechť pravděpodobnost A i pravděpodobnost všech $B_1, \ldots B_n$ je nenulová, potom pro libovolné $i = 1, \ldots, n$ platí

$$\mathrm{P}(B_k|A) = \frac{\mathrm{P}(A|B_k)\,\mathrm{P}(B_k)}{\sum_{i=1}^{n} \mathrm{P}(A|B_i)\,\mathrm{P}(B_i)}.$$

Klasická (frekventistická) statistika

Ideové základy R.A.Fisher (poč. 20. stol.)

- parametry jsou pevné
- testujeme nulovou hypotézu (inspirace Popperovou filosofií), určíme si s jakou pstí jsme ochotni se mýlit

Problémy

- Kdo věří v nulové hypotézy?
- Interpretace, mnohonásobná porovnávání

Bayesovská statistika

- všechno jsou náhodné veličiny
- netestujeme hypotézy, ale děláme intervalové odhady

Problémy

- Kde vzít apriorní rozdělení?
- Ve složitějších případech výpočetně složité (simulace, numerická intergrace)

Bayesovská věta pro spojité náhodné veličiny

$$f_{Y|X}(y|x) = \frac{f_Y(y) f_{X|Y}(x|y)}{\int f_Y(z) f_{X|Y}(x|z) dz}.$$

Lze interpretovat jako

- f_Y apriorní informace o parametru Y (předpokládané rozdělení před pokusem) – prior
- $f_{X|Y}$ jak X závisí na parametru Y (obvykle známe) – věrohodnost
- $f_{Y|X}$ jak se změní rozdělení Y – posterior

Příklad – beta a alternativní rozdělení

Máme posloupnost (X) k úspěchů a l neúspěchů, přičemž předpokládáme, že pravděpodobnost úspěchu (Y) je náhodně rozdělená s hustotou $Beta(a, b)$

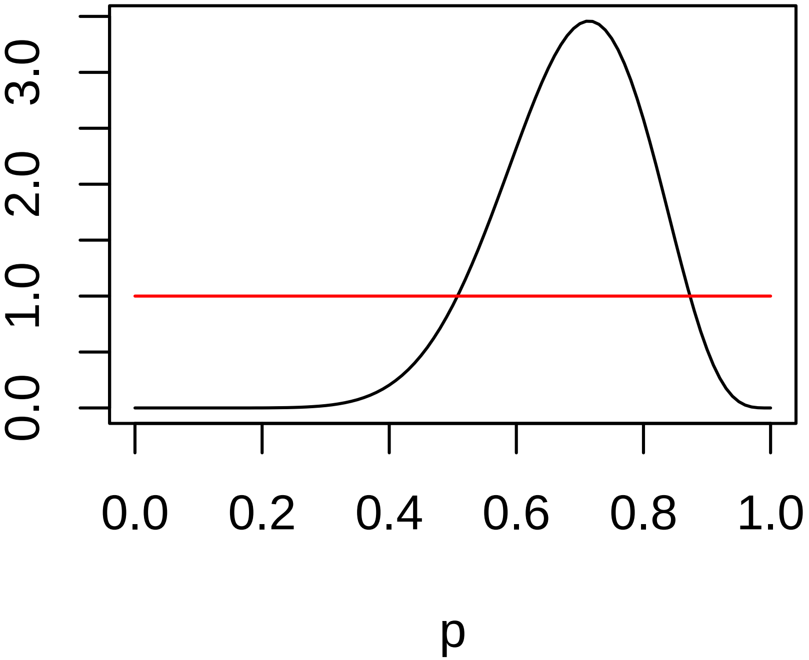
$$f_Y(y) = \frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}$$

Potom aposteriorní rozdělení má hustotu

$$Beta(a + k, b + l)$$

Příklad – beta a alternativní rozdělení

Uvažujme kupříkladu 10 úspěchů a 4 neúspěchy:



Jiný příklad "konjugované" dvojice

Předpokládáme, že máme pozorování z $N(\mu, \sigma^2)$.

Apriorní rozdělení pro μ budiž také normální $N(\mu_0, \sigma_0^2)$.

Pozorování $X_1, \ldots, X_n$ s průměrem m. Potom aposteriorní rozdělení pro μ je opět normální s parametry

$$N\left(\frac{\frac{mn}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}}\right)$$

Gibb's sampler

Problém – často nejsme schopni rozdělení přesně upočítat. Řešením může být numerická integrace, ale častěji je to simulace.

Situace: Máme nějaké známé veličiny (X) a neznámé (skryté, latentní) $Y_1, \dots, Y_n$, jejichž rozdělení bychom rádi spočítali.

- Na počátku všechny neznámé veličiny inicializujeme nějakou (vesměs libovolnou) hodnotou $y_1^{(0)}, \dots, y_n^{(0)}$
- V k-tém kroku simulujeme i-tou náhodnou veličinu Y_i podmíněně na X a na všech ostatních veličinách $Y_1 = y_1^{(k)}, \dots, Y_{i-1} = y_{i-1}^{(k)}, Y_{i+1} = y_{i+1}^{(k-1)}, \dots,$ $Y_n = y_n^{(k-1)}$.

Programová implementace

- **WinBUGS** – pro nekomerční použití zdarma po registraci
- **OpenBUGS** – GNU licence, zatím totožný s WinBUGS
- **BRUGS** – knihovna pod R pro OpenBUGS

K čemu je to dobré v Kalmanově filtraci?

Pokud Y_t závisí na X_t nějak jinak než gaussovsky (např. exponenciální rozdělení, poissonovo rozdělení), potom je těžké upočítat Kalmanův filtr. Naopak nemusí

být tak těžké spočítat to Bayesovsky, resp. ve složitějším případě z této řady simulovat.