WNIOSKOWANIE BAYESOWSKIE

WPROWADZENIE W MIARĘ BEZBOLESNE :)

Łukasz Kwiatkowski

Katedra Ekonometrii i Badań Operacyjnych

Plan wykładu

1) Prawdopodobieństwo – *revisited*

2) Dwa podejścia do wnioskowania statystycznego

3) Statystyka bayesowska – wyróżniki

4) Wnioskowanie bayesowskie – to „lepsiejsze”

5) Bayesowski model statystyczny – krok po kroku

6) Jeden mem mówi więcej niż tysiąc słów…

Prawdopodobieństwo – *revisited*

- Pytanie 1: Jakie jest prawdopodobieństwo, że jutro między 7:00 a 9:00 spadnie deszcz?

Prawdopodobieństwo – *revisited*

- Pytanie 2: Mam w ręce monetę 5 zł. Jakie są szanse, że jeśli nią rzucę, to wypadnie orzeł?

Prawdopodobieństwo – *revisited*

- Interpretacje prawdopodobieństwa:
  - „**klasyczna**” – prawdopodobieństwo „obiektywne”, częstościowe (definicja Andrieja Kołmogorowa, doprecyzowująca tę zaproponowaną wcześniej przez Richarda von Misesa), rozumiane jako **idealizacja częstości względnej**; zwane także „**fizycznym**”, ale p-stwo **NIE jest jakąś cechą fizyczną**, immanentnie związaną z i określającą jakiś „obiekt” (np. monetę, urnę czy kurs walutowy), zdarzenie czy zdanie logiczne (hipotezę)
  - **bayesowska** – prawdopodobieństwo jest miarą subiektywnego stopnia przekonania/pewności co do zajścia jakiegoś zdarzenia (losowego, tj. obarczonego niepewnością) czy prawdziwości pewnej hipotezy

    → Prawdopodobieństwo jest jak … – każdy ma swoje ;) ALE nie oznacza to jakiegoś dogmatyzmu (*Moje prawdopodobieństwo jest „najmojsze”, i za Chiny Ludowe go nie zmienię!*)

    → P-stwo ma 2 swoje źródła: wychodzimy od **p-stwa *a priori*** (czyli przed doświadczeniem, eksperymentem, przed wglądem w dane, w empirię), a następnie modyfikujemy je wraz z napływającymi danymi = aktualizacja wiedzy („mojej” wiedzy, rozumianej tu jako „moje” wyobrażenie o kształtowaniu się niepewności odnośnie danego zdarzenia/hipotezy itp.)

    → Statystyka bayesowska oferuje w pełni formalny, zgodny z podstawowymi zasadami rachunku prawdopodobieństwa (wychodząc od definicji wg. Pierre Simone de Laplace), wewnętrznie spójny (koherentny) – choć też nie pozbawiony pewnych paradoksów – sposób modyfikacji mojej wiedzy *a priori* w *a posteriori* (tj. po doświadczeniu, po wglądzie w dane/obserwacje), czyli sposób wyznaczania „mojej” końcowej wiedzy, poprzez połączenie wiedzy *a priori* z informacją płynącą z obserwacji (co jest zgodne z neurofizjologią procesu uczenia się :)

Nie zawsze „a priori” oznacza od razu „Bayesa” :)

Prawdopodobieństwo – *revisited*

- Powtórzmy:
  - P-stwo jest matematycznym sposobem wyrażania „mojego” (zatem subiektywnego) stopnia przekonania (a tym samym stopnia *niepewności*)

    → Na gruncie statystyki bayesowskiej prawdopodobieństwo ma interpretację subiektywistyczną
  - P-stwo ulega modyfikacji (w wyniku procesu „uczenia się”, aktualizacji wiedzy) – od wiedzy/informacji wstępnej (tzw. *a priori*, czyli przed wglądem w dane/obserwacje) do wiedzy końcowej (tj. *a posteriori*, czyli uwzględnieniu informacji zawartej w danych; ang. *evidence*)

    → W modelowaniu parametrycznym (np. regresji) owa „wiedza”/niepewność dotyczyć będzie *parametrów modelu*
  - Matematyczny dowód istnienia i jednoznaczności p-stwa jako miary subiektywnego stopnia przekonania nie jest trywialny, ALE JEST :) (XX w.: Morris H. DeGroot, Bruno de Finetti, Leonard Savage)

Dwa podejścia do wnioskowania statystycznego

- Dwa wiodące – istnieją jeszcze inne koncepcje...
- Te dwie wzmiankowane w tym miejscu bazują, odpowiednio, na „klasycznej” i bayesowskiej interpretacji prawdopodobieństwa:
  - **Statystyka „klasyczna”** – inaczej: niebayesowska, częstościowa (ang. *frequentist statistics/inference, frequentist approach to statistics*):
    - Jerzy Spława Neyman (1894-1984) – przedziały ufności
    - Ronald Fisher (1890-1962) – MNW
  - **Statystyka bayesowska** (ang. *Bayesian statistics/inference, Bayesian approach to statistics*):
    - **Thomas Bayes** (1702-1760) – angielski matematyk i duchowny; twórcą jako tako nie jest, ale autorem sławetnego **wzoru Bayesa** (na p-stwo warunkowe; opublikowanego pośmiertnie):

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

    - Pierre Simone de Laplace (1749-1827) – pierwsza formalna definicja p-stwa

→ Chronologicznie podejście bayesowskie jest bardziej „klasyczne” niż podejście „klasyczne” ;)

→ Odnowiciele bayesizmu: Harold Jeffreys, Bruno de Finetti, Leonard Savage, Arnold Zellner

Statystyka bayesowska – wyróżniki

- Podejście bayesowskie do statystyki odróżnia się od ujęcia wnioskowania „klasycznego” w co najmniej kilku aspektach, wśród których za fundamentalne można uznać:
  - Od strony epistemologii (czyli filozofii poznania): subiektywną – w odróżnieniu od częstościowej – interpretację p-stwa

    → p-stwo jako miara „mojego” stopnia niepewności – najbardziej adekwatną metodą do wyrazu/ujęcia tej niepewności okazują się <u>rozkłady prawdopodobieństwa</u>
  - Od strony „technicznej”: wszystkie elementy modelu statystycznego (tj. zarówno dane, jak i – UWAGA – parametry!) są <u>zmiennymi losowymi</u> $y = X\beta + \varepsilon$

    → dla tych, których nie dowierzają w to, co słyszą powtórzmy: na gruncie bayesowskim PARAMETRY są ZMIENNYMI LOSOWYMI

    → przypomnijmy: na gruncie „klasycznym” parametry są… (?)
- Analogia pomiędzy parametrami (bayesowskiego) modelu statystycznego a położeniem elektronu (fizyka kwantowa)
- Tak więc na gruncie bayesowskim parametry modelu mają swoje… rozkłady prawdopodobieństwa

  → Nie ma więc potrzeby konstruowania czegoś takiego jak estymator (=funkcja danych, która ma „ustrzelić” nieznaną, stałą wartość parametru), choć na gruncie bayesowskim także (o ile chcemy) możemy wyznaczać oceny punktowe parametrów (czyli punktowe wskazania jego położenia) – jako wartość oczekiwana/mediana/modalna *rozkładu* – ale już nie estymatora, tylko bezpośrednio parametru (wydaje się bardziej naturalne :)

Wnioskowanie bayesowskie – to „lepsiejsze" ;)

- Zasadniczo, statystyka bayesowska:
    - daje po prostu pełny obraz kształtowania się **niepewności** co do możliwych wartości parametrów (poprzez bezpośrednio *ich* – a nie estymatora – rozkłady prawdopodobieństwa), a o to właśnie chodzi w statystyce (← nauka o niepewności, a nie – by tak rzec – o ocenach punktowych)
    - wewnętrznie spójne (najbardziej ze wszystkich podejść do wnioskowania statystycznego, choć w dalszym ciągu nie jest wolne od pewnych paradoksów)
    - bardzo wymierna i już teraz namacalna dla Państwa korzyść z WB: w końcu naturalnie brzmiące interpretacje przedziałów „ufności" (zwanych w statystyce bayesowskiej *przedziałami wiarygodności*; ang. *credible intervals*)

- Choć „lepsze", to jednak jego zwolennikom/twórcom/propagatorom nie szczędzono batów na przestrzeni dziejów. Te przeplatająco-ścierające się dzieje statystyki bayesowskiej i „klasycznej" to istna „Moda na sukces"...

  → Sharon B. McGrayne, (2011), *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press

Bayesowski model statystyczny – krok po kroku

- Patrz materiał „***Podstawowe pojęcia***.pdf” (Moodle, w sekcji „Materiały pomocnicze”)
- We wnioskowaniu statystycznym (każdym parametrycznym, nie tylko bayesowskim) mamy zasadniczo dwa obiekty: dostępne **dane** (obserwacje) i **parametry** (pomijamy prognozowanie, gdzie mamy jeszcze *obserwacje niedostępne*):
  - **Dane – wektor obserwacji dostępnych**:

$$y = (y_1 \;\; y_2 \;\; \dots \;\; y_T)' \in \underbrace{Y}_{\substack{przestrzeń \\ obserwacji}} \subseteq \mathbb{R}^T$$

- dla y zakładamy, że zostały „wygenerowane” zgodnie z pewnymi założeniami (np. KMNRL), przyjmując dla nich tzw. ***model próbkowy*** (ang. *sampling model*) albo inaczej ***rozkład próbkowy***, czyli rozkład obserwacji, przy założeniu konkretnych (choć o nieznanych wartościach) parametrów, θ – przykładowo w KMNRL:

$$p(y|\beta,\sigma^2) = f_N^T(y|X\beta,\sigma^2 I_T) = (2\pi)^{-\frac{T}{2}}(\sigma^2)^{-\frac{T}{2}}\exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\}$$

- $p(x)$ – **notacja rodzajowa**: funkcja gęstości / masy rozkładu prawdopodobieństwa zmiennej x
- ten „mechanizm powstawania obserwacji” – jeśli spojrzeć na niego tak, jakby argumentem były parametry, a nie obserwacje – to zarazem **funkcja wiarygodności** (ang. *likelihood function*), $L_y(\theta)$:

$$\underbrace{L_y(\theta)}_{\substack{funkcja \\ wiarygodności}} = \underbrace{p(y|\theta)}_{\substack{rozkład/model \\ próbkowy}}$$

- Uwaga: y jako wektor losowy (wektorowa zmienna losowa, wektor zmiennych losowych) ORAZ y jako *realizacja* tego wektora (notacyjnie tego tu nie rozróżniamy)

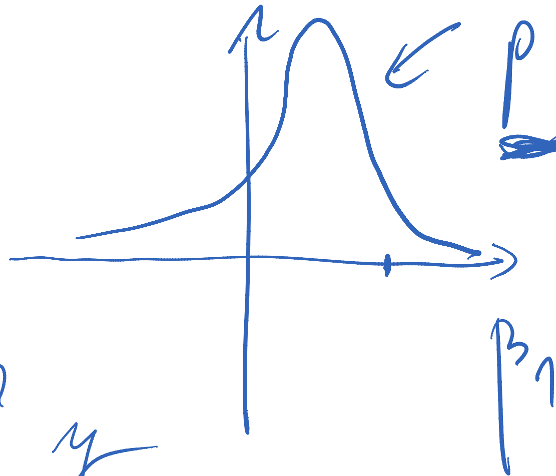
Bayesowski model statystyczny – krok po kroku

- Patrz materiał „***Podstawowe pojęcia***.pdf” (Moodle, w sekcji „Materiały pomocnicze”)
- We wnioskowaniu statystycznym (każdym parametrycznym, nie tylko bayesowskim) mamy zasadniczo dwa obiekty: dostępne **dane** (obserwacje) i **parametry** (pomijamy prognozowanie, gdzie mamy jeszcze *obserwacje niedostępne*):

  - **Parametry** – **wektor** (wszystkich) **parametrów** danego modelu:

$$\theta = (\theta_1 \;\; \theta_2 \;\; \dots \;\; \theta_s) \in \underbrace{\Theta}_{\substack{przestrzeń \\ parametrów}} \subseteq \mathbb{R}^s$$

- przypomnijmy, że „tu”: θ – wektor losowy (a nie nieznanych stałych, jak w ujęciu częstościowym)



$p(\beta_1 | y)$

$p(\theta | y) = ?$

$\theta \quad y$

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} = \frac{Pr(B)\,Pr(A|B)}{Pr(A)}$$

Wzór Bayesa:

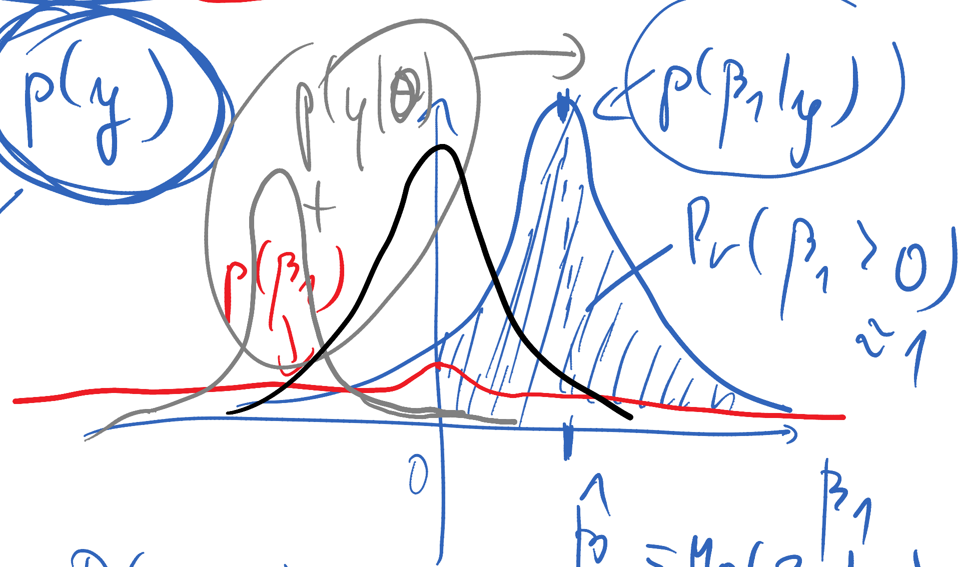
model bayesowski

r. próbkowy / f. wiarygodności

$$\underbrace{p(\theta|y)}_{\text{R. a posteriori}} = \frac{p(y,\theta)}{\underbrace{p(y)}_{\text{brzegowa gęstość obserwacji}}} = \frac{p(y|\theta)\,p(\theta)}{p(y)}$$

$p(\theta)$ → R. a priori

rozkłady łączne,
brzegowe i warunkowe



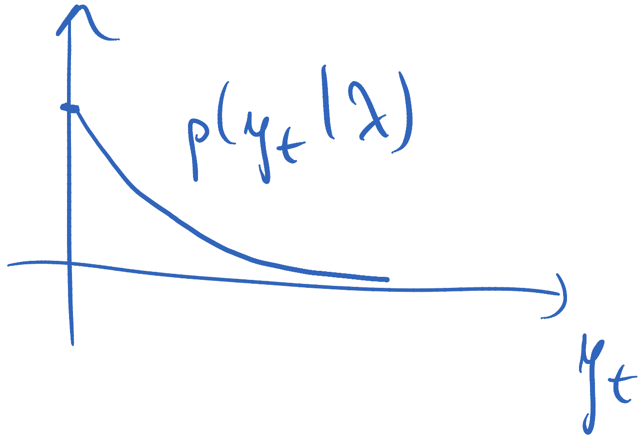
$D(\beta_1|y)$

$\hat{\beta}_1 = Mo(\beta_1|y) \approx E(\beta_1|y) \approx Me(\beta_1|y)$

Przykład 1

Niech $y_t|\lambda \sim iid\ Exp(\lambda)$, $t=1,\dots,T$

r. wykładniczy o par. $\lambda > 0$



→ Wyznaczyć $p(\lambda|y)$.

ingrediencje :)

$$p(\lambda|y) = \frac{p(y,\lambda)}{p(y)} = \frac{\overset{(1)}{p(y|\lambda)}\,\overset{(2)}{p(\lambda)}}{p(y)}$$

$y_t \perp y_s$, $t \neq s$

Ad (1) $$p(y|\lambda) = p(y_1, y_2, \dots, y_T|\lambda) = \prod_{t=1}^{T} p(y_t|\lambda) =$$

$$= \prod_{t=1}^{T}\left(\lambda e^{-\lambda y_t}\right) = \lambda^T \cdot e^{-\lambda \sum_{t=1}^{T} y_t} \Rightarrow \boxed{p(y|\lambda) = \lambda^T e^{-\lambda \sum_{t=1}^{T} y_t}}$$

$e^a \cdot e^b \cdot e^c = e^{a+b+c}$

$\underbrace{a \cdot a \cdot \ldots \cdot a}_{\# T} =$

$$\prod_{t=1}^{T}(a \cdot b_t) = \left(\prod_{t=1}^{T} a\right) \cdot \left(\prod_{t=1}^{T} b_t\right) = a^T \prod_{t=1}^{T} b_t$$

Ad 2) (r. a priori)

$\lambda > 0 \Rightarrow$

$p(\lambda) = ?$

parametry rozkładu a priori = hiperparametry

$\rightarrow$ Sugeruje się by $\boxed{p(\lambda) = f_G(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}}$ „+" dla jakichś zadanych samodzielnie $a, b > 0$

proporcjonalne do

$$\Rightarrow \underbrace{p(\lambda \mid y)} = \frac{p(y \mid \lambda)\, p(\lambda)}{p(y)} \propto p(y \mid \lambda)\, p(\lambda) =$$

stała normująca gęstość r. a posteriori (bo nie zależy od λ)

nie mam, ale... bo nie potrzebuję :)

$$= \lambda^T e^{-\lambda \sum_{t=1}^{T} y_t} \cdot \underbrace{\boxed{\frac{b^a}{\Gamma(a)}}}_{\text{stała}} \lambda^{a-1} e^{-b\lambda} \propto \lambda^{a+T-1} \cdot e^{-\lambda\left(b + \sum_{t=1}^{T} y_t\right)}$$

→ Podsum.: $p(\lambda|y) \propto \underbrace{\lambda^{\alpha+T-1}\, e^{-\lambda(b+\sum_{t=1}^{T} y_t)}}_{\text{jądro rozkładu a posteriori}}$ → Pytanie: Czy jest to jądro jakiegoś znanego rozkładu p-stwa?

→ Odp.: TAK: rozkład <u>gamma</u>

→ Zatem $p(\lambda|y) = f_G(\lambda|\bar{\alpha}, \bar{b})$, gdzie $\bar{\alpha}, \bar{b}$ – „nowe" parametry rozkładu gamma

ta sama rodzina rozkładów, co a priori – gamma (bo r. a priori jest tu – dla rozważanej f. wiarygodności – rozkładem sprzężonym)

$$\bar{\alpha} = \alpha + T$$
$$\bar{b} = b + \sum_{t=1}^{T} y_t$$

Ten r. a posteriori reprezentuje naszą końcową (po wglądzie w dane) „wiedzę"/niepewność dot. param. λ.

- $E(\lambda|y)$
- $M_o(\lambda|y)$
- $M_e(\lambda|y)$
- $D(\lambda|y)$
- $Q_1(\lambda|y)$, $Q_3(\lambda|y)$

i tn.

Charakterystyki r. a posteriori

+

wykres funkcji gęstości