Statystyka Bayesowska i Teoria Decyzji

Wojciech Niemiro [1]

13 lutego 2021

[1]Wydział Wydział Matematyki, Informatyki i Mechaniki, Uniwersytet Warszawski oraz Wydział Matematyki i Informatyki, Uniwersytet Mikołaja Kopernika w Toruniu. wniem@mimuw.edu.pl, wniem@mat.umk.pl

Spis treści

Rozdział 1

Bayesowskie modele statystyczne

1.1 Wstęp

Podejście bayesowskie polega na tym, że wnioskowanie statystyczne interpretujemy w sposób czysto probabilistyczny, jako obliczanie prawdopodobieństwa pewnych zdarzeń. Dane statystyczne traktujemy jako wynik doświadczenia losowego, które już zostało przeprowadzone, czyli jako ustalone wartości zmiennych losowych. Zakładamy, że losowy mechanizm generujący dane jest zależny od nieznanych parametrów. W statystyce bayesowskiej parametry traktujemy jako ukryte, nieobserwowalne zmienne losowe. Zakładamy znajomość rozkładu prawdopodobieństwa *a priori*, który wyraża naszą wiedzę (lub przekonania) o wartościach parametrów przed zaobserwowaniem danych. Obliczamy rozkład prawdopodobieństwa *a posteriori*, który uwzględnia informację zawartą w danych. Z matematycznego punktu widzenia jest to po prostu prawdopodobieństwo warunkowe. Używamy zatem pojęć probabilistycznych do opisu niepewności, braku pełnej informacji. Taka subiektywna interpretacja prawdopodobieństwa jest źródłem pewnych kontrowersji. Przeciwnicy podejścia bayesowskiego wskazują na trudności w określeniu rozkładu *a priori*. Niemniej, liczne zastosowania świadczą o użyteczności statystyki bayesowskiej. Ważną zaletą statystyki bayesowskiej jest, w moim przekonaniu, logiczna prostota.

1.2 Przykład wprowadzający

Rozpatrzmy prosty model, dobrze ilustrujący bayesowski punkt widzenia. Wykorzystamy ten model, aby przedstawić intuicje stojące za podstawowymi pojęciami statystyki bayesowskiej, zanim podamy ogólne definicje.

1.2.1 Przykład. Towarzystwo ubezpieczeniowe chce oszacować prawdopodobieństwo zgłoszenia szkody. Załóżmy, że są dwa typy kierowców: „Ostrożni" i „Ryzykanci". Kierowca „O" w ciągu roku powoduje szkodę z prawdopodobieństwem 0.1 a „R" z prawdopodobieństwem 0.4. Niech $X = 1$ oznacza szkodę, a $X = 0$ jej brak. Napiszmy

$$\mathbb{P}(X = 1|O) = 0.1, \qquad \mathbb{P}(X = 1|R) = 0.4. \tag{a}$$

Towarzystwo ubezpieczeniowe podpisując nową umowę nie wie, jakiego typu jest klient. Szanse natrafienia na „Ostrożnego" i „Ryzykanta" oceniamy przed zawarciem umowy na

$$\mathbb{P}(O) = 0.80, \qquad \mathbb{P}(R) = 0.20. \tag{b}$$

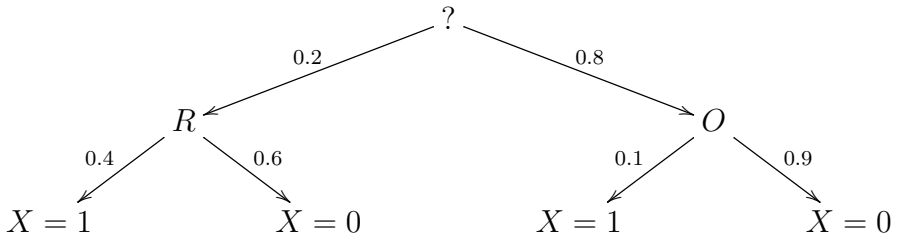
Prawdopodobieństwo zgłoszenia szkody obliczamy ze wzoru na prawdopodobieństwo całkowite:

$$\begin{aligned} \mathbb{P}(X = 1) &= \mathbb{P}(X = 1|O)\mathbb{P}(O) + \mathbb{P}(X = 1|R)\mathbb{P}(R) \\ &= 0.1 \times 0.80 + 0.4 \times 0.20 = 0.16. \end{aligned} \tag{c}$$

Po upływie roku, dysponujemy obserwacją X. Jeśli klient zgłosił szkodę, to elementarny wzór Bayesa pozwala obliczyć, że

$$\mathbb{P}(R|X = 1) = \frac{\mathbb{P}(X = 1|R)\mathbb{P}(R)}{\mathbb{P}(X = 1)} = 0.5. \tag{d}$$

Zbudowaliśmy model dwuetapowego doświadczenia losowego. Pierwszy etap polega na wylosowaniu klienta zgodnie z rozkładem *a priori* (b). Wynik pierwszego etapu losowania jest nieznany. Obserwujemy wynik drugiego etapu, wystąpienie lub brak szkody, czyli zmienną X i na tej podstawie obliczamy prawdopodobieństwo *a posteriori* (d). Prawdopodobieństwa warunkowe we wzorze (a) będziemy nazywać wiarygodnością. Wzór (c) określa brzegowy rozkład obserwacji. Poniższy schemat przedstawia oba etapy doświadczenia.



Przypuśćmy, że kierowca zgłosił szkodę w 1-szym roku ubezpieczenia. Chcemy obliczyć prawdopodobieństwo, że zgłosi szkodę i w 2-gim roku. Rozbudujmy nasz model. Niech zmienne losowe X_1 i X_2 opisują wystąpienie szkody w kolejnych latach (poprzednio rozpatrywaną zmienną X przemianowowaliśmy na X_1). Zakładamy, że

$$\mathbb{P}(X_2 = 1|O, X_1) = 0.1, \qquad \mathbb{P}(X_2 = 1|R, X_1) = 0.4$$

(dla danego „typu", zmienna losowa X_2 jest warunkowo niezależna od X_1 i obie mają taki sam rozkład prawdopodobieństwa). Możemy teraz łatwo wywnioskować, że

$$\begin{aligned}\mathbb{P}(X_2 = 1|X_1 = 1) &= \mathbb{P}(X_2 = 1|O)\mathbb{P}(O|X_1 = 1) + \mathbb{P}(X_2 = 1|R)\mathbb{P}(R|X_1 = 1) \\ &= 0.1 \times 0.5 + 0.4 \times 0.5 = 0.25.\end{aligned} \tag{e}$$

Skomentujmy poczynione powyżej założenia. Oczywiście, przykład ma wyłącznie charakter ilustracyjny i jest skrajnie uproszczony, ale pozwala prześledzić pewne aspekty modelowania statystycznego. Obserwowana zmienna losowa X_1 reprezentuje dane statystyczne, zaś X_2 jest zmienną losową, której wartość chcemy przewidzieć. Wzór (e) podaje tak zwane prawdopodobieństwo *predykcyjne*. Przed otrzymaniem danych możemy obliczyć prawdopodobieństwo „predykcyjne *a priori*", czyli po prostu rozkład brzegowy obserwacji, zgodnie ze wzorem (c). Podkreślmy, że zmienne losowe X_1 i X_2 opisują dwa „realne" doświadczenia losowe, z których jedno już się dokonało, a wynik drugiego jest nieznany. Natomiast zmienna losowa opisująca typ klienta jest z zasady nieobserwowalna. Jest to tylko element modelu matematycznego, który pomaga obliczyć prawdopodobieństwo realnego, przyszłego zdarzenia „$X_2 = 1$" na podstawie informacji „$X_1 = 1$".

Dodajmy, że w naszym przykładzie rozkład *a priori* ma obiektywną interpretację (nie jest to sytuacja typowa w statystyce bayesowskiej). Założenie (b) mówi, że wśród kierowców zawierających nową umowę, 80% należy do typu „O" i 20% do typu „R". Tego rodzaju informacja może pochodzić z przeszłości, podobnie zresztą jak założenie (a), które precyzuje mniej kontrowersyjny element modelu (tak zwaną wiarygodność).

Na koniec wprowadźmy nieco inne oznaczenia, standardowe w statystyce bayesowskiej. Pierwszy etap doświadczenia interpretujemy jako wylosowanie parametru rozkładu prawdopodobieństwa rządzącego drugim etapem. Niech ϑ będzie zmienną losową taką, że $\mathbb{P}(\vartheta = 0.1) = \mathbb{P}(O) = 0.8$ i $\mathbb{P}(\vartheta = 0.4) = \mathbb{P}(R) = 0.8$. Możemy teraz napisać $\mathbb{P}(X = 1|\vartheta = \theta) = \theta$ dla $\theta \in \{0.1, 0.4\}$. Zbiór $\{0.1, 0.4\}$ jest tu przestrzenią parametrów. △

1.3 Rozkłady *a priori* i *a posteriori*

Niech X będzie obserwowaną zmienną losową o wartościach w przestrzeni $\mathcal{X}$. Zakładamy, że rozkład prawdopodobieństwa P_θ tej zmiennej zależy od nieznanego parametru $\theta \in \Theta$. W statystyce bayesowskiej określamy, oprócz rodziny rozkładów $\{P_\theta : \theta \in \Theta\}$ na przestrzeni obserwacji $\mathcal{X}$, również rozkład prawdopodobieństwa Π na przestrzeni parmetrów Θ. Jest on zwany rozkładem *a priori*, ponieważ wyraża on naszą wiedzę (lub przekonania) o parametrze przed zaobserwowaniem zmiennej losowej X (bez brania pod uwagę danych). Nieznany parametr θ traktujemy jako realizację *zmiennej losowej* ϑ. Wnioskowanie statystyczne w

ujęciu bayesowskim sprowadza się do obliczania warunkowego rozkładu prawdopodobieństwa zmiennej ϑ przy ustalonym $X = x$. Mówimy, że jest to rozkład *a posteriori* (po zaobserwowaniu danych).

Prześledźmy najpierw konstrukcję modelu formalnie, z matematycznego punktu widzenia. Założymy (dla uproszczenia), że każdy rozkład P_θ ma gęstość p_θ (względem ustalonej miary oznaczanej w skrócie $\mathrm{d}x$; zwykle jest to miara Lebesgue'a albo miara licząca). Podobnie, niech rozkład *a priori* Π ma gęstość π (względem miary oznaczanej w skrócie $\mathrm{d}\theta$). W dalszych rozważaniach z reguły będziemy utożsamiali rozkłady prawdopodobieństwa z ich gęstościami.

Łączną gęstość zmiennych losowych ϑ i X *definiujemy* wzorem

$$p(\theta, x) = \pi(\theta)p_\theta(x), \qquad (\theta \in \Theta, x \in \mathcal{X}). \tag{1.3.1}$$

W ten sposób określamy *jeden* rozkład prawdopodobieństwa, nazwijmy go $\mathbb{P}$, *na przestrzeni probabilistycznej* $\Omega = \Theta \times \mathcal{X}$. Jeśli zmienne X i ϑ są dyskretne, to przez łączną gęstość rozumiemy $p(\theta, x) = \mathbb{P}(\vartheta = \theta, X = x)$. Będziemy też rozważali modele w których jedna ze zmiennych X i ϑ jest dyskretna a druga ma rozkład absolutnie ciągły. Symbole $\mathbb{E}$, Var, p (bez wskaźnika θ) będą odtąd oznaczały wartość oczekiwaną, wariancję i gęstość względem rozkładu $\mathbb{P}$. Zauważmy, że teraz p_θ staje się *warunkową* gęstością zmiennej losowej X dla $\vartheta = \theta$:

$$p_\theta(x) = \frac{p(\theta, x)}{\pi(\theta)} = p(x|\theta).$$

Symbole P_θ i E_θ oznaczają warunkowe prawdopodobieństwo i wartość oczekiwaną.

Jeśli dana jest wartość naszej obserwacji i wiemy, że $X = x$, to możemy przy pomocy znanego *wzoru Bayesa* obliczyć rozkład *warunkowy* losowego parametru ϑ. Jest to tak zwany rozkład *a posteriori*. Gęstość tego rozkładu oznacza się czasami przez π_x.

1.3.2 Twierdzenie (Wzór Bayesa)**.** *Rozkład* ***a posteriori*** *parametru ϑ, dla danej obserwacji $X = x$, ma gęstość*

$$\pi_x(\theta) = \pi(\theta|x) = \frac{\pi(\theta)p_\theta(x)}{p(x)},$$

gdzie

$$p(x) = \int_\Theta \pi(\theta)p_\theta(x)\mathrm{d}\theta.$$

Gęstość p opisuje *rozkład brzegowy* zmiennej losowej X w modelu bayesowskim. W tym kontekście mówi się p jest *mieszanką* wyjściowych gęstości p_θ. W istocie, możemy traktować p jako „średnią ważoną" funkcji p_θ, z „funkcją wagową" π.

W typowych prostych modelach bayesowskich rozkład *a priori* jest tak dobrany do rozkładu obserwacji, aby wyliczenie rozkładu *a posteriori* było bardzo łatwe. Są to tak zwane sprzężone rodziny rozkładów. Mianownik $p(x)$ we wzorze Bayesa jest dość skomplikowany, ale nie zawsze musimy go obliczać. Interesuje nas zależność gęstości a posteriori od θ, zatem pomijając czynniki nie zależące od θ (choć być może zawierające x) przepiszemy wzór Bayesa w bardziej przyjaznej postaci:

$$\pi_x(\theta) \propto \pi(\theta)p_\theta(x).$$

Symbol $\propto$ oznacza, że lewa i prawa strona są proporcjonalne jako funkcje θ.

1.4 Przykłady

W tym podrozdziale zrobimy przegląd bardzo typowych i ważnych w zastosowaniach modeli bayesowskich. Jak później wyjaśnimy, we wszystkich tych przykładach mamy do czynienia z tak zwanymi sprzężonymi rozkładami *a priori*, które są wygodne do obliczeń.

1.4.1 Przykład (Rozkład dwumianowy i rozkład *a priori* beta)**.** Załóżmy, że X jest liczbą sukcesów w n próbach Bernoulliego z nieznanym prawdopodobieństwem sukcesu θ. Zmienna X ma rozkład dwumianowy $\mathrm{Bin}(n, \theta)$, czyli mamy

$$p_\theta(x) = P_\theta(X = x) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \propto \theta^x(1-\theta)^{n-x} \qquad (x = 0, 1, \ldots, n).$$

Wygodnie przyjąć, że rozkład *a priori* jest rozkładem $\mathrm{Beta}(\alpha, \beta)$:

$$\pi(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \qquad (0 < \theta < 1).$$

Rozkład *a posteriori* ma zatem gęstość

$$\pi_x(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^x(1-\theta)^{n-x} = \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}.$$

Rozkład *a posteriori* jest więc rozkładem $\mathrm{Beta}(\alpha + x, \beta + n - x)$.

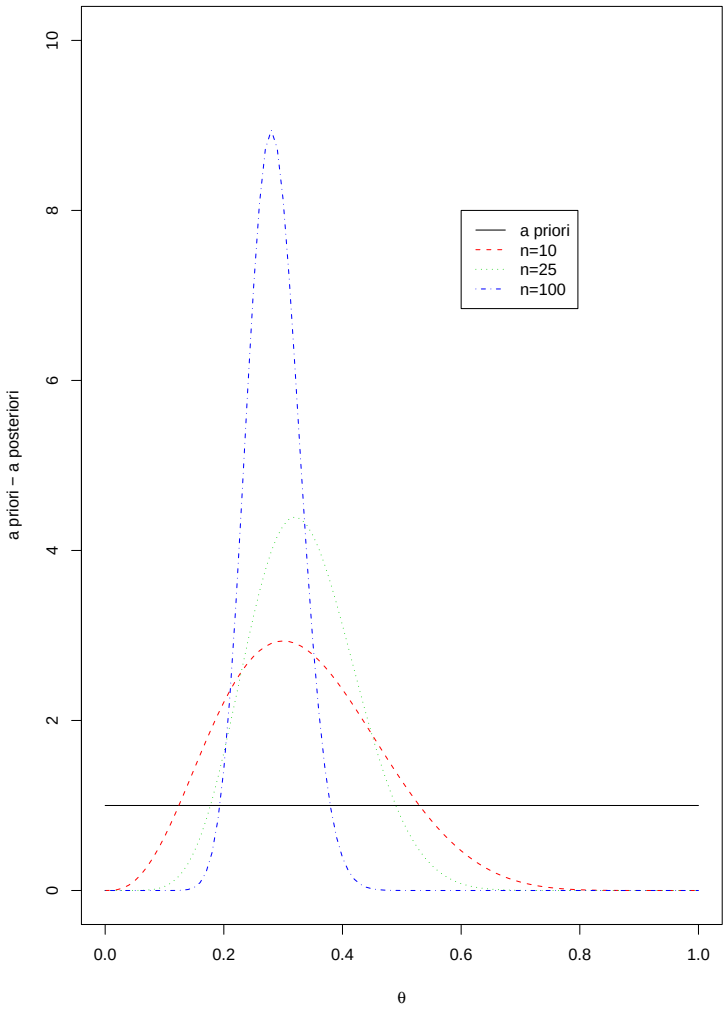
Zauważmy, jak uprościliśmy rachunki pomijając czynniki nie zawierające θ. Musimy jednak uwzględnić stałe normujące, jeśli chcemy obliczyć rozkład brzegowy obserwacji, patrz Zadanie 8. W szczególnym przypadku $n = 1$ obliczenie rozkładu brzegowego jest łatwiejsze:

$$\mathbb{P}(X = 1) = \mathbb{E}(X) = \mathbb{E}\mathbb{E}(X = 1|\vartheta) = \mathbb{E}\vartheta = \frac{\alpha}{\alpha+\beta},$$
$$\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = 1) = \frac{\beta}{\alpha+\beta}.$$

△

W szczególności jednostajny rozkład *a priori* należy do rodziny rozkładów beta: $\mathrm{U}(0,1) = \mathrm{Beta}(\alpha, \beta)$, dla $\alpha = \beta = 1$. Ten rozkład wydaje się dobrze modelować „całkowitą niewiedzę na temat parametru θ", ale jest to interpretacja kontrowersyjna[1].

Rysunek 1.1 przedstawia gęstości jednostajnego rozkładu *a priori* i rozkładów *a posteriori* w Przykładzie 1.4.1 dla kilku wartości n i dla $x = 0.3n$.



Rysunek 1.1: Rozkład *a priori* (jednostajny) i rozkłady *a posteriori* dla $n = 10$, $n = 25$ i $n = 100$.

[1] Jeśli „nic nie wiemy o parametrze θ" to nic nie wiemy o θ^2. A więc $\vartheta \sim \mathrm{U}(0,1)$ czy $\vartheta^2 \sim \mathrm{U}(0,1)$?

1.4.2 Przykład (Rozkład Poissona i rozkład *a priori* gamma)**.** Załóżmy, że X jest zmienną losową o rozkładzie Poissona Poiss$(n\theta)$, gdzie θ jest nieznanym parametrem a $n > 0$ znaną liczbą. Mamy zatem

$$p_\theta(x) = P_\theta(X = x) = \mathrm{e}^{-n\theta}\frac{(n\theta)^x}{x!} \propto \mathrm{e}^{-n\theta}\theta^x \qquad (x = 0, 1, \ldots, n).$$

Rozważmy rozkład *a priori* Gamma(α, λ):

$$\pi(\theta) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\mathrm{e}^{-\lambda\theta} \propto \theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}, \qquad (\theta > 0).$$

Rozkład *a posteriori* ma zatem gęstość

$$\pi_x(\theta) \propto \theta^{\alpha-1}\mathrm{e}^{-\lambda\theta}\mathrm{e}^{-n\theta}\theta^x = \theta^{\alpha+x-1}\mathrm{e}^{-(\lambda+n)\theta}.$$

Jest to więc rozkład Gamma$(\alpha + x, \lambda + n)$.

Rozważany przez nas model znajduje zastosowanie w matematyce ubezpieczeniowej i tak zwanej „teorii zaufania". Interpretacja jest podobna jak w Przykładzie 1.2.1. Wyobraźmy sobie, że zmienna losowa X oznacza liczbę szkód dla konkretnego klienta w ciągu n lat. Parametr θ jest średnią liczbą roszczeń przypadających na rok. Każdemu klientowi odpowiada inna wartość parametru θ. Rozkład *a priori* opisuje „rozrzut" tego parametru w populacji klientów. Zgłaszanie się klientów uznajemy za zjawisko przypadkowe. Każda nowa umowa jest *dwuetapowym* doświadczeniem losowym. W *pierwszym etapie* pojawia się realizacja θ zmiennej losowej ϑ. W *drugim etapie* obserwujemy liczbę szkód $X = x$, powiedzmy w ciągu n lat. Nasz stan wiedzy (lub raczej stopień niewiedzy) o zmiennej θ po zaobserwowaniu liczby x opisuje rozkład *a posteriori*.

Rozkład brzegowy zmiennej losowej X interpretujemy jako rozkład liczby szkód dla pojedynczego klienta pochodzącego z niejednorodnej populacji, opisanej rozkładem gamma.

Aby obliczyć rozkład brzegowy, wykorzystamy znajomość zmiennej normującej w rozkładzie gamma.

$$\begin{aligned}
p(x) = \mathbb{P}(X = x) &= \int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\lambda\theta}\cdot\frac{(n\theta)^x}{x!}e^{-n\theta}\mathrm{d}\theta \\
&= \frac{\lambda^\alpha n^x}{x!\Gamma(\alpha)}\int_0^\infty \theta^{x+\alpha-1}e^{-(\lambda+n)\theta}\mathrm{d}\theta \\
&= \frac{\lambda^\alpha n^x}{x!\Gamma(\alpha)}\cdot\frac{\Gamma(x+\alpha)}{(\lambda+n)^{x+\alpha}} \\
&= \frac{(x+\alpha-1)(x+\alpha-2)\cdots(\alpha+n)\alpha}{x!}\left(\frac{\lambda}{\lambda+n}\right)^\alpha\left(\frac{n}{\lambda+n}\right)^x \\
&= \binom{x+\alpha-1}{x}\Big(\frac{\lambda}{\lambda+n}\Big)^\alpha\Big(\frac{n}{\lambda+n}\Big)^x.
\end{aligned}$$

Przyjmując oznaczenie $\lambda/(\lambda+n) = q$ i $n/(\lambda+n) = 1-q$ możemy napisać

$$\mathbb{P}(X = x) = \binom{-\alpha}{x} q^{\alpha}(q-1)^{x}. \tag{1.4.3}$$

Brzegowo, obserwacje w tym przykładzie mają rozkład *Ujemny Dwumianowy* $\mathrm{Bin}^{-}(\alpha, q)$. △

1.4.4 Przykład (Rozkład normalny, średnia ma rozkład *a priori* normalny)**.** Niech $X = (X_1, \ldots, X_n)$ będzie próbką z rozkładu normalnego $\mathrm{N}(\mu, \sigma^2)$ ze znaną wariancją σ^2, to znaczy

$$p_{\mu}(x_1, \ldots, x_n) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right].$$

Załóżmy, że σ^2 jest znane, zaś nieznany parametr μ ma rozkład *a priori* normalny $\mathrm{N}(m, v^2)$, czyli

$$\pi(\mu) \propto \exp\left[-\frac{1}{2v^2}(\mu-m)^2\right].$$

Możemy napisać gęstość *a posteriori*:

$$\begin{aligned}\pi_{x_1,\ldots,x_n}(\mu) &\propto \exp\left[-\frac{1}{2v^2}(\mu-m)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\ &\propto \exp\left[-\frac{nv^2+\sigma^2}{2\sigma^2v^2}\Big(\mu - \frac{nv^2\bar{x}+\sigma^2 m}{nv^2+\sigma^2}\Big)^2\right].\end{aligned}$$

Oczywiście, $\bar{x}$ oznacza $\sum x_i/n$. Wykonaliśmy tu znaną ze szkoły średniej operację sprowadzania trójmianu kwadratowego do postaci kanonicznej. Wyrazy wolne zostają „pochłonięte" przez symbol $\propto$. Dodanie stałej do argumentu funkcji wykładniczej jest tym samym, co pomnożenie tej funkcji przez stałą. Widać już, że rozkład *a posteriori* jest normalny,

$$\mu|x_1, \ldots, x_n \sim \mathrm{N}\Big(\frac{nv^2\bar{x}+\sigma^2 m}{nv^2+\sigma^2}, \frac{\sigma^2v^2}{nv^2+\sigma^2}\Big). \tag{1.4.5}$$

Aby obliczyć rozkład brzegowy obserwacji w rozpatrywanym modelu, nie musimy całkować. Zamiast tego posłużymy się następującym prostym rozumowaniem. Skoro $X_i|\mu \sim \mathrm{N}(\mu, \sigma^2)$, to $X_i - \mu|\mu \sim \mathrm{N}(0, \sigma^2)$. Ponieważ rozkład warunkowy zmiennej $X_i - \mu$ nie zależy od wartości zmiennej warunkującej μ, to te dwie zmienne są niezależne i $X_i - \mu \sim \mathrm{N}(0, \sigma^2)$. Wiemy z założenia, że $\mu \sim \mathrm{N}(m, v^2)$. Ponieważ $X_i = \mu + (X_i - \mu)$, wnioskujemy że (brzegowo)

$$X_i \sim \mathrm{N}(m, v^2 + \sigma^2).$$

W Zadaniu 9 proponujemy wyznaczenie w podobny sposób łącznego rozkładu brzegowego wektora $(X_1, \ldots, X_n)$. △

1.4.6 Przykład (Rozkład normalny, precyzja ma rozkład *a priori* gamma)**.** Niech $X = (X_1, \ldots, X_n)$ będzie próbką z rozkładu normalnego $\mathrm{N}(\mu, \sigma^2)$. Załóżmy teraz, że μ jest znane, zaś nieznana jest wariancja σ^2. Wygodnie przyjąć, że parametrem jest odwrotność wariancji, $\nu = \sigma^{-2}$. To jest tak zwany parametr *parametr precyzji*. Wiarygodność ma teraz postać

$$p_\nu(x_1, \ldots, x_n) \propto \nu^{n/2} \exp\left[-\frac{\nu}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right]$$

(w odróżnieniu od poprrzedniego przykładu, musieliśmy teraz uwzględnić obecność nieznanego parametru w stałej normującej rozkładu normalnego). Załóżmy, że nieznany parametr ν ma rozkład *a priori* $\mathrm{Gamma}(\alpha, \lambda)$, czyli

$$\pi(\nu) \propto \nu^{\alpha-1}\mathrm{e}^{-\lambda\nu}.$$

Możemy napisać gęstość *a posteriori*:

$$\pi_{x_1,\ldots,x_n}(\nu) \propto \nu^{\alpha+n/2} \exp\left[-\left(\lambda + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right)\nu\right]$$

Widać stąd, że *a posteriori* mamy rozkład gamma,

$$\nu|x_1, \ldots, x_n \sim \mathrm{Gamma}\Big(\alpha + \frac{n}{2}, \lambda + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\Big).$$

Alternatywnie mówi się, że wariancja $\sigma^2 = 1/\nu$ ma rozkład „odwrotny gamma": *a priori* $\sigma^2 \sim \mathrm{IG}(\alpha, \lambda)$. △

1.4.7 Przykład (Rozkład normalny, łączny rozkład *a priori* na średnią i precyzję)**.** Rozważamy rodzinę rozkładów normalnych $\mathrm{N}(\mu, \sigma^2)$ z nieznanymi oboma parameterami. Tak jak w porzednich 2 przykładach, obserwujemy próbkę $X_1, \ldots, X_n$.

Rozkład *a priori* (μ, σ^2) jest opisany w następujący sposób. Brzegowo σ^2 ma rozkład „odwrotny gamma" $\sigma^2 \sim \mathrm{IG}(\alpha, \lambda)$. Warunkowy rozkład μ (*a priori*) jest normalny: $\mu|\sigma^2 \sim \mathrm{N}(m, r\sigma^2)$ dla pewnych hiperparametrów μ and $r > 0$. Rozkład *a posteriori* okazuje się być tej samej postaci, z innymi paramerami α', λ', m' i r' zastępującymi α, λ, m i r (Zadanie 7). Zauważmy, że 2-wymiarowy rozkład *a priori* w tym przykładzie nie jest rozkładem produktowym. △

Przykład 1.4.1 łatwo uogólnia się na przypadek wielowymiarowy. Rozkład wielomianowy jest odpowiednikiem rozkładu dwumianowego. Opisuje „liczbę wyników d-typów" w ciągu n niezależnych doświadczeń o jednakowym rozkładzie. Niech N_j oznacza liczbę otrzymanych wyników typu i, dla $i = 1, \ldots, d$. Wektor losowy $(N_1, \ldots, N_d)$ ma rozkład wielomianowy $\mathrm{Mult}(\theta_1, \ldots, \theta_d)$, dany wzorem

$$\mathbb{P}_\theta(N_1 = n_1, \ldots, N_d = n_d) = p_\theta(n_1, \ldots, n_d) = \frac{n!}{\prod_{i=1}^d n_i!} \prod_{i=1}^d \theta_i^{n_i}.$$

Rozkład Dirichleta jest wielowymiarowym odpowiednikiem rozkładu Beta. Mówimy, że d-wymiarowa zmienna losowa ϑ ma rozkład *Dirichleta*,

$$\vartheta = (\vartheta_1, \ldots, \vartheta_d) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d),$$

jeśli $\vartheta_1 + \cdots + \vartheta_d = 1$ i zmienne $\vartheta_1, \ldots, \vartheta_{d-1}$ mają łączną gęstość

$$p(\theta_1, \ldots, \theta_{d-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_d)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_d)} \theta_1^{\alpha_1 - 1} \ldots \theta_{d-1}^{\alpha_{d-1} - 1} (1 - \theta_1 - \ldots \theta_{d-1})^{\alpha_d - 1}.$$

Parametry $\alpha_1, \ldots, \alpha_r$ mogą być dowolnymi liczbami dodatnimi. Inną definicję rozkładu Dirichleta i związek z rozkładami gamma podaje Zadanie 12 w Dodatku A. Zauważmy, że dla $d = 2$, rozkład Dirichleta sprowadza się do rozkładu beta:

$$\vartheta_1 \sim \mathrm{Beta}(\alpha_1, \alpha_2) \quad \text{wtedy i tylko wtedy, gdy} \quad (\vartheta_1, 1 - \vartheta_1) \sim \mathrm{Dir}(\alpha_1, \alpha_2).$$

1.4.8 Przykład (Rozkłady wielomianowe, rozkład *a priori* Dirichleta)**.** Rozkład Dirichleta jest wygodnym (sprzężonym) rozkładem *a priori*, jeśli obserwacje mają rozkład wielomianowy. Jeśli $(N_1, \ldots, N_d | \vartheta = \theta) \sim \mathrm{Mult}(\theta_1, \ldots, \theta_d)$ i $\vartheta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$ to $(\vartheta | N_1, \ldots, N_d) \sim \mathrm{Dir}(\alpha_1 + N_1, \ldots, \alpha_d + N_d)$. Istotnie, rozkład wielomianowy jest dany wzorem

$$p_\theta(n_1, \ldots, n_d) \propto \prod \theta_i^{n_i}.$$

Jeśli napiszemy gęstość Dirichleta w postaci $\pi(\theta) \propto \prod \theta_i^{\alpha_i - 1}$, to wspomniany wyżej fakt staje się natychmiast widoczny. △

Zreasumujemy nasze rozważania i dokładniej omówimy *interpretację* modelu bayesowskiego. Jeśli rozkład *a priori* wyraża tylko przekonania statystyka i jest wybrany arbitralnie, to obliczone na jego podstawie prawdopodobieństwo ma charakter *subiektywnej* oceny szans. To jest klasyczny punkt widzenia teorii bayesowskiej. Wspomnieliśmy w Przykładzie 1.4.1 o związanych z tym trudnościach. W wielu zastosowaniach rozkład *a priori* ma inną, bardziej obiektywną interpretację, tak jak w Przykładzie 1.4.2. Łączny rozkład prawdopodobieństwa zmiennych losowych ϑ i X jest probabilistycznym modelem dwuetapowego doświadczenia losowego, w którym *obserwowujemy tylko wynik drugiego etapu.*

1.5 Warunkowa niezależność i dostateczność

W modelach bayesowskich ważną rolę gra pojęcie warunkowej niezależności. Rozważmy 3 zmienne losowe X, Y i Z, określone na tej samej przestrzeni probabilistycznej (przestrzenie wartości być być różne). Dla uproszczenia załóżmy, że istnieje łączna gęstość $p(x, y, z)$, albo „w zwykłym sensie" albo dyskretna. Zmienne X i Y są warunkowo niezależne pod warunkiem Z, jeśli

$$p(x, y|z) = p(x|z)p(y|z)$$

dla (prawie)[2] wszystkich x, y i z. Łatwo sprawdzić, że warunkowa niezależność jest równoważna każdemu z dwóch następujących warunków, spełnionemu dla (prawie) wszystkich x, y i z:

$$p(x|y, z) = p(x|z), \qquad p(y|x, z) = p(y|z).$$

Uogólnienie na przypadek większej liczby zmiennych nie przedstawia trudności.

W typowym modelu bayesowskim przyjmuje się założenie o warunkowej niezależności obserwowanych zmiennych losowych $X_1, \ldots, X_n$ dla danej wartości parametru $\vartheta = \theta$. Wiarygodność ma postać

$$p_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_\theta(x_i) \qquad (\theta \in \Theta, x_i \in \mathcal{X}).$$

Przyjmijmy, że w powyższym wzorze symbole $p_\theta(x_i)$ oznaczają tę samą funkcję gęstości $p_\theta(\cdot)$ obliczaną dla różnych argumentów $x_1, \ldots, x_n$. Innymi słowy zakładamy dodatkowo, że zmienne $X_1, \ldots, X_n$ mają ten sam warunkowy rozkład prawdopodobieństwa[3]. Łączny rozkład zmiennych ϑ i $X_1, \ldots, X_n$ ma gęstość

$$p(\theta; x_1, \ldots, x_n) = \pi(\theta) \prod_i p_\theta(x_i). \tag{1.5.1}$$

Zauważmy, że obserwacje $X_1, \ldots, X_n$ w świecie bayesowskim *nie są* niezależne, bo brzegowa gęstość $p(x_1, \ldots, x_n) = \int p(\theta; x_1, \ldots, x_n)\mathrm{d}\theta$ *nie jest* iloczynem $p(x_1) \cdots p(x_n)$. Zmienna X_j zależy od X_i „poprzez zmienną ϑ".

Modele bayesowskie przedstawione w Przykładach 1.4.4 i 1.4.6 mają postać (1.5.1). W dalszej części tego podrozdziału zobaczymy, że w Przykładach 1.4.1 i 1.4.2 też mamy do czynienia z „ukrytym" modelem (1.5.1), zredukowanym przez przejście do statystyki dostatecznej.

[2] Gęstości względem miary Lebesgue'a są jednoznacznie zdefiniowane prawie wszędzie. W przypadku zmiennych dyskretnych można pominąć zastrzeżenie „prawie".

[3] Nasza notacja jest skrócona i przez to niejednoznaczna: symbole $p(\cdot)$ i $p_\theta(\cdot)$ mogą, w zależności od kontekstu, oznaczać różne funkcje gęstości.

Rozkład predykcyjny

Rozważmy nieco rozszerzoną wersję modelu (1.5.1). Oprócz obserwowanych zmiennych $X_1, \ldots, X_n$ dodajmy jeszcze zmienną X_{n+1}, która dotyczy przyszłości. Załóżmy, że wszystkie zmienne $X_1, \ldots, X_n, X_{n+1}$ są warunkowo niezależne i mają ten sam warunkowy rozkład:

$$p(\theta; x_1, \ldots, x_n, x_{n+1}) = \pi(\theta) \prod_{i=1}^{n+1} p_\theta(x_i).$$

Interesuje nas przewidywanie zmiennej X_{n+1} na podstawie danych $X_1 = x_1, \ldots X_n = x_n$. Zgodnie z bayesowską filozofią, zadanie sprowadza się do obliczenia warunkowego rozkładu prawdopodobieństwa

$$X_{n+1}|X_1, \ldots, X_n.$$

Powiemy, że jest to *rozkład predykcyjny*. Obliczamy ten rozkład korzystając z warunkowej niezależności:

$$\begin{aligned} p(x_{n+1}|x_1, \ldots, x_n) &= \int p(x_{n+1}|\theta; x_1, \ldots, x_n) p(\theta|x_1, \ldots, x_n) \mathrm{d}\theta \\ &= \int p(x_{n+1}|\theta) p(\theta|x_1, \ldots, x_n) \mathrm{d}\theta \\ &= \int p_\theta(x_{n+1}) \pi_{x_1, \ldots, x_n}(\theta) \mathrm{d}\theta. \end{aligned}$$

Prypomnijmy sobie, że brzegowy rozkład pojedynczej obserwacji X_i jest dany wzorem

$$p(x_i) = \int p_\theta(x_i) \pi(\theta) \mathrm{d}\theta.$$

Porównanie tych wzorów prowadzi do następującego wniosku.

- Rozkład predykcyjny obliczamy w ten sam sposób co rozkład brzegowy, zastępując w odpowiedniej całce rozkład *a priori* przez rozkład *a posteriori*.

1.5.2 Przykład. Zilustrujmy tę regułę na Przykładzie (1.4.4). Jeśli *a priori* $\mu \sim \mathrm{N}(m, v^2)$ to wiemy, że pojedyncza zmienna X_i ma rozkład brzegowy $\mathrm{N}(m, v^2+\sigma^2)$. Wstawiając w miejsce „hiperparametrów" m i v^2 parametry rozkładu *a posteriori* ze wzoru (1.4.5), natychmiast otrzymujemy rozkład predykcyjny

$$X_{n+1}|x_1, \ldots, x_n \sim \mathrm{N}\Big(\frac{nv^2\bar{x} + \sigma^2 m}{nv^2 + \sigma^2}, \frac{\sigma^2 v^2}{nv^2 + \sigma^2} + \sigma^2\Big).$$

△

Dostateczność

W modelu bayesowskim dostateczność jest (prawie) równoważna warunkowej niezależności parametru i obserwacji pod warunkiem statystyki. Niestety, porządna dyskusja pojęcia dostateczności wymaga użycia teorii miary. W tym wykładzie pominiemy teorio-miarowe subtelności. Zasadnicza idea jest niezwykle prosta. Rozważamy ogólny model bayesowski określony wzorem (1.3.1) i *statystykę* (o wartościach w przestrzeni $\mathcal{T}$), czyli pewną funkcję $T : \mathcal{X} \to \mathcal{T}$.

1.5.3 Stwierdzenie (Bayesowska dostateczność)**.** *W modelu bayesowskim następujące warunki są równoważne:*

(i) zmienne losowe ϑ i X są warunkowo niezależne pod warunkiem $T(X)$,

(ii) zachodzi własność faktoryzacji, to znaczy wiarygodność można przedstawić w postaci $p_\theta(x) = g_\theta(T(x))h(x)$ z wyjątkiem, być może, zbioru parametrów θ o zerowym prawdopodobieństwie a priori[4]*.*

Szkic dowodu. Poniższe rozumowanie jest ścisłym dowodem w przypadku dyskretnej przestrzeni $\mathcal{X}$, gdy gęstość warunkowa jest elementarnie zdefiniowanym prawdopodobieństwem warunkowym. W przypadku ogólnym są kłopoty techniczne związane z definicją rozkładów warunkowych[5].

Jeśli zachodzi warunek (ii) to rozkład *a posteriori* zależy od x tylko poprzez funkcję $T(x)$, bo

$$\pi_x(\theta) \propto p_\theta(x)\pi(\theta) \propto g_\theta(T(x))\pi(\theta).$$

Znaczy to, że rozkład warunkowy ϑ przy danej wartości $X = x$ jest taki sam jak rozkład warunkowy przy danej wartości $T(X) = t$, gdzie $t = T(x)$. Zatem

$$\pi(\theta|x) = \pi(\theta|x,t) = \pi(\theta|t),$$

a więc zmienne losowe ϑ i X są warunkowo niezależne dla danej wartości $T(X) = t$. Otrzymaliśmy (i).

Wynikanie w przeciwną stronę jest równie proste. Jeśli zachodzi (i), to $\pi(\theta|x) = \pi(\theta|t)$ dla $t = T(x)$. Zatem

$$p(x|\theta) = \frac{\pi(\theta|x)p(x)}{\pi(\theta)} \propto \pi(\theta|x)p(x) = \pi(\theta|t)p(x),$$

Otrzymaliśmy faktoryzację (ii) z $g_\theta(t) = \pi(\theta|t)$ i $h(x) = p(x)$. □

[4] Zastrzeżenie o „wyjątkowym zbiorze parametrów" jest niestety konieczne.

[5] Nie istnieje, na ogół, łączna gęstość zmiennych losowych X i $T = T(X)$, nie możemy zatem zastosować uproszczonej „definicji" warunkowego rozkładu $X|T$.

Standardowa (częstościowa) definicja dostateczności mówi, że rozkład warunkowy X przy danym $T(X) = t$ nie zależy od parametru θ. Moim zdaniem, jest dość trudno zrozumieć, dlaczego ta definicja „mówi to, co należy", to znaczy że wartość $T(x) = t$ zawiera tyle samo informacji o nieznanym parametrze θ, co obserwacja $X = x$. W świecie bayesowskim *standardowa* definicja dostateczności sprowadza się (w zasadzie, pomijając subtelności) do stwierdzenia

$$p(x|\theta, t) = p(x|t)$$

(żeby to zauważyć, wystarczy chwila zastanowienia). *Bayesowska* interpretacja dostateczności jest bardziej intuicyjna:

$$\pi(\theta|x) = \pi(\theta|t).$$

Rozkład *a posteriori* (czyli wszystko co wie statystyk po zaobserwowaniu x) zależy tylko od $t = T(x)$.

1.5.4 Przykład (Schemat Bernoulliego i rozkład *a priori* beta)**.** Rozważmy binarne zmienne losowe $X_1, \ldots, X_n$ opisujące schemat Bernoulliego prawdopodobieństwem sukcesu θ. Zakładamy, że są to zmienne warunkowo niezależne i

$$P_\theta(X_i = 1) = \theta, \qquad P_\theta(X_i = 0) = 1 - \theta.$$

Oczywiście,

$$p_\theta(x_1, \ldots, x_n) = \theta^s (1-\theta)^{n-s},$$

gdzie $s = \sum_{i=1}^n x_i$, a więc $S = \sum_{i=1}^n X_i$ jest statystyką dostateczną. Jeśli, tak jak w Przykładzie (1.4.1), rozkładem *a priori* jest Beta(α, β) to *a posteriori* parametr ma rozkład Beta$(\alpha + s, \beta + n - s)$. Wystarczy powołać się na Przykład (1.4.1). Rozkład predykcyjny X_{n+1} (przy standardowych założeniach) jest dany wzorem

$$\begin{aligned} \mathbb{P}(X_{n+1} = 1|x_1, \ldots, x_n) &= \frac{\alpha + s}{\alpha + \beta + n}, \\ \mathbb{P}(X_{n+1} = 0|x_1, \ldots, x_n) &= \frac{\beta + n - s}{\alpha + \beta + n}. \end{aligned} \tag{1.5.5}$$

Wzór (1.5.5) ma interesującą interpretację „urnową", patrz Zadanie 8. Historyczna, dyskretna wersja naszego przykładu przedstawiona jest w Zadaniu 1. △

1.5.6 Przykład (Rozkład Poissona i rozkład *a priori* gamma)**.** Wróćmy do Przykładu 1.5.6 i jego „ubezpieczeniowej" interpretacji. Niech zmienne $X_1, \ldots, X_n$ opisują liczby szkód klienta w kolejnych n latach. Załóżmy, że $X_i|\theta \sim \text{Poiss}(\theta)$. Wtedy $S = \sum_{i=1}^n X_i$ jest statystyką dostateczną i $S|\theta \sim \text{Poiss}(n\theta)$. Jeśli parametr θ ma rozkład *a priori* Gamma(α, λ), to rozkładem *a posteriori* jest Gamma$(\alpha + s, \lambda + n)$. Wykorzystując wzór (1.4.3) dla $n = 1$ i stosując ogólną regułę wyznaczamy bez dalszych rachunków rozkład predykcyjny:

$$\mathbb{P}(X_{n+1} = x|x_1, \ldots, x_n) = \binom{-\alpha - s}{x} \left(\frac{\lambda + n}{\lambda + n + 1}\right)^{\alpha+s} \left(-\frac{1}{\lambda + n + 1}\right)^x, \quad (x = 0, 1, \ldots).$$

Podkreślmy, że powyższy wzór interpretujemy jako przewidywanie liczby szkód x w roku $n+1$ dla klienta, wylosowanego z niejednorodnej populacji, który zgłosił w poprzednich n latach s szkód[6]. △

1.6 Sprzężone rodziny rozkładów

Jeśli wiarygodność należy do wykładniczej rodziny rozkładów prawdopodobieństwa, to pewna rodzina rozkładów *a priori* jest specjalnie wygodna z punktu widzenia obliczeniowego. Przypomnijmy, że *wykładnicza rodzina* gęstości ma postać

$$p_\theta(x) = \exp\left\{\sum_{j=1}^{k} T_j(x)g_j(\theta) + g_0(\theta)\right\}h(x) = \exp\Big\{T(x)^\top g(\theta) + g_0(\theta)\Big\}h(x),$$

gdzie $g(\theta) = (g_1(\theta), \ldots, g_k(\theta))^\top$ and $T(x) = (T_1(x), \ldots, T_k(x))^\top$. Ważną własnością takich rozkładów jest istnienie statystyk dostatecznych o ustalonym wymiarze k, niezależnym od rozmiaru próbki n. Istotnie, wiarygodność dla próbki $X_1, \ldots, X_n \sim_{\text{i.i.d.}} p_\theta$ jest postaci

$$\prod_{i=1}^{n} p_\theta(x_i) = \exp\left\{\sum_{j=1}^{k}\left(\sum_{i=1}^{n} T_j(x_i)\right)g_j(\theta) + ng_0(\theta)\right\}\prod_{i=1}^{n} h(x_i)$$

Kryterium faktoryzacji pokazuje, że $\sum_{i=1}^{n} T(x_i)$ jest k-wymiarową statystyką dostateczną.

Definiujemy *sprzężoną* rodzinę rozkładów a priori wzorem

$$\begin{aligned}\pi(\theta) = \pi_{t_0,n_0}(\theta) &= c(t_0, n_0)\exp\left\{\sum_{j=1}^{k} t_{0j}g_j(\theta) + n_0 g_0(\theta)\right\} \\ &= c(t_0, n_0)\exp\Big\{t_0^\top g(\theta) + n_0 g_0(\theta)\Big\},\end{aligned}$$

przy założeniu, że funkcja w powyższym wzorze (po przestrzeni Θ) jest skończona, czyli istnieje stała normalizująca $c(t_0, n_0)$. Podstawową własnością rodzin sprzężonych jest to, że rozkład *a posteriori* ma tę samą postać co rozkład *a priori*, tylko ma inne parametry. Faktycznie,

$$\begin{aligned}\pi(\theta|x_1, \ldots, x_n) &\propto \prod_{i=1}^{n} p_\theta(x_i)\pi(\theta) \propto \exp\left\{\sum_{i=1}^{n} T(x_i)^\top g(\theta) + ng_0(\theta)\right\}\exp\Big\{t_0^\top g(\theta) + n_0 g_0(\theta)\Big\} \\ &= \exp\left\{\sum_{i=1}^{n}\Big(T(x_i) + t_0\Big)^\top g(\theta) + (n+n_0)g_0(\theta)\right\}.\end{aligned}$$

[6] Przy założeniu, że klient „zachowuje się w kolejnych latach identycznie", niezależnie od wypadków, którym ulega. Adekwatność tego założenia w praktyce jest, oczywiście, mocno wątpliwa.

A więc, rozkład π_{t_0,n_0} po zaobserwowaniu próbki aktualizujemy do $\pi_{t_0+n\bar{t},n_0+n}$, gdzie $n\bar{t} = \sum_{i=1}^n T(x_i)$. Interesujące, że t_0 i n_0 grają rolę podobną jak $\sum T(x_i)$ i n, odpowiednio. Można sobie wyobrażać, że t_0 jest sumą wartości statystyki T dla n_0 „wirtualnych", fikcyjnych obserwacji. To sugeruje wygodny sposób sformułowania wiedzy eksperta w postaci umożliwiającej wybranie rozkładu *a priori*.

Mówimy, że rodzina wykładnicza jest przedstawiona w *naturalnej parametryzacji*, jeśli $\Theta \subseteq \mathbb{R}^k$, $g_j(\theta) = \theta_j$ i $g_0(\theta) = -\psi(\theta)$. Wzór na gęstość p_θ upraszcza się:

$$p_\theta(x) = \exp\left\{\sum_{j=1}^k T_j(x)\theta_j - \psi(\theta)\right\}h(x),$$
$$\psi(\theta) = \log\int_{\mathcal{X}} \exp\left\{\sum_{j=1}^k T_j(x)\theta_j\right\}h(x)\mathrm{d}x.$$

Wiadomo, że $\psi(\theta)$ jest funkcją generującą kumulanty i w szczególności $\mathbb{E}(T(X)|\theta) = \nabla\psi(\theta)$ and $\mathrm{VAR}(T(X)|\theta) = \nabla\nabla^\top\psi(\theta)$ (patrz Zadanie 10). Jeśli $\mu(\theta) = \mathbb{E}(T(X)|\theta)$ to

$$\begin{aligned}\mathbb{E}\mu(\theta) &= \int_\Theta \mu(\theta)\pi_{t_0,n_0}(\theta)\mathrm{d}\theta = \frac{t_0}{n_0},\\ \mathbb{E}[\mu(\theta)|x_1,\ldots,x_n] &= \int_\Theta \mu(\theta)\pi_{t_0+n\bar{t},n_0+n}(\theta)\mathrm{d}\theta = \frac{t_0+n\bar{t}}{n_0+n},\end{aligned} \tag{1.6.1}$$

jeśli Θ jest otwartym wypukłym podzbiorem $\mathbb{R}^k$. Oczywiście, $\mathbb{E}\mu(\theta) = \mathbb{E}T(X)$ (wartość oczekiwana względem $p_\theta(x)\pi(\theta)$.

Szkic dowodu (1.6.1). Załóżmy, że $k = 1$ i $\Theta = \mathbb{R}$. Wtedy mamy $\psi'(\theta) = \mu(\theta)$ i $\pi'(\theta) = (t_0 - n_0\mu(\theta))\pi(\theta)$. Można pokazać (intuicyjnie jest to raczej oczywiste), że $\int \pi'(\theta)\mathrm{d}\theta = 0$, a więc $t_0 = n_0\int \mu(\theta)\pi(\theta)\mathrm{d}\theta$. Pierwsze równanie we wzorze (1.6.1) wynika natychmiast. Drugie jest wnioskiem z pierwszego. □

We wszystkich modelach przedstawionych w tym rozdziale mieliśmy do czynienia ze sprzężonymi rozkładami *a priori*.

Jest kilka przykładów gęstości, które nie należą do rodzin wykładniczych, ale dla których istnieją rozkłady *a priori* o własnościach analogicznych do rozkładów sprzężonych.

1.6.2 Przykład (Jednostajna wiarygodność i rozkłady *a priori* Pareto)**.** Załóżmy, że mamy próbkę $X_1,\ldots,X_n$ z rozkładu jednostajnego $\mathrm{U}(0,\theta)$. Niech rozkład *a priori* będzie rozkładem Pareto $\pi(\theta) \propto \theta^{-n_0}\mathbb{1}(\theta > t_0)$. Ponieważ $p_\theta(x_1,\ldots,x_n) = \theta^{-n}\mathbb{1}(\max(x_1,\ldots,x_n) < \theta)$, rozkład *a posteriori* jest też rozkładem Pareto, mianowicie $\propto \theta^{-n_0-n}\mathbb{1}(\theta > t_0 \vee \max(x_1,\ldots,x_n))$. Chociaż rozkłady jednostajne nie tworzą rodziny wykładniczej, istnieje dla tych rozkładów 1-wymiarowa statystyka dostateczna, mianowicie $T = \max(x_1,\ldots,x_n)$. Konstrukcja „quasi-sprzężonego" rozkładu *a priori* bazuje na tej statystyce. △

Sprężone modelele bayesowskie są wygodne obliczeniowo. Ich oczywistą wadą jest ograniczony wybór rozkładów *a priori*. Pewnym wyjściem może być użycie mieszanek rozkładów sprzężonych.

1.6.3 Przykład (Mieszanki rozkładów sprzężonych)**.** Rozważmy rodzinę wykładniczą rozkładów (z naturalną parametryzacją)

$$p_\theta(x) = \exp\Big\{T(x)^\top\theta - \psi(\theta)\Big\}h(x)$$

i sprzężoną rodzinę rozkładów *a priori*

$$\pi_{t_0,n_0}(\theta) = c(t_0,n_0)\exp\Big\{t_0^\top\theta - n_0\psi(\theta)\Big\}.$$

Jeśli za rozkład a priori wybierzemy mieszankę rozkładów z tej rodziny, czyli

$$\bar\pi(\theta) = \sum_{j=1}^k w_j\pi_{t_{0j},n_{0j}}(\theta),$$

dla pewnych wartości hiper-parametrów (t_{0j},n_{0j}) i wag $w_j > 0$ ($\sum w_j = 1$) to rozkład *a posteriori* też jest mieszanką,

$$\bar\pi_{x_1,\ldots,x_n}(\theta) = \sum_{j=1}^k w'_j\pi_{t_{0j}+n\bar t,n_{0j}+n}(\theta),$$

gdzie wagi *a posteriori* są równe $w'_j \propto w_j c(t_{0j},n_{0j})/c(t_{0j}+n\bar t, n_{0j}+n)$. Weryfikację tego faktu zostawiamy jako ćwiczenie (Zadanie 11). △

1.7 Rozkłady *a priori* Jeffreysa

Dla danej parmetrycznej rodziny $\{p_\theta : \theta\in\Theta\}$ gęstości prawdopodobieństwa na przestrzeni obserwacji, rozważamy gęstość *a priori*, która jest funkcją informacji Fishera, $\pi(\theta) = g(I(\theta))$. Pytanie brzmi, jak wybrać g, aby zapewnić niezmienniczość ze względu na re-parameryzację. Na początek rozważmy przypadek parametru jednowymiarowego. Niech $\Theta,\Psi\subseteq\mathbb{R}$ będą dwoma zbiorami otwartymi. Wprowadźmy nowy parametr $\psi = h^{-1}(\theta)$, gdzie $h:\Psi\to\Theta$ jest funkcją „$1-1$" (dyfeomorfizmem). Oznaczmy przez $\tilde I(\psi)$ informację Fishera w modelu parametryzowanym przez ψ a przez $\tilde\pi(\psi)$ gęstość zmiennej losowej ψ. Ponieważ

$$I(\theta) = \int\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)\right)^2 p_\theta(x)\mathrm{d}x$$

i

$$\frac{\mathrm{d}}{\mathrm{d}\psi}\log p_{h(\psi)}(x) = h'(\psi)\frac{\mathrm{d}}{\mathrm{d}\theta}\log p_\theta(x)_{|\theta=h(\psi)},$$

stąd wynika, że $\tilde{I}(\psi) = I(h(\psi))\,[h'(\psi)]^2$. Z drugiej strony, jeśli π jest gęstością zmiennej losowej θ to zmienna losowa ψ ma gęstość $\tilde{\pi}(\psi) = \pi(h(\psi))|h'(\psi)|$. Zatem powinniśmy mieć

$$g(\tilde{I}(\psi)) = g\left(I(h(\psi))\,[h'(\psi)]^2\right) = g\big(I(h(\psi))\big)|h'(\psi)|.$$

Jeśli to równanie jest spełnione dla dowolnego dyfeomorfizmu h w każdym punkcie ψ to widać, że $g(I) \propto \sqrt{I}$. Tak więc postulat niezmienniczości prowadzi do rozkładów *a priori* postaci

$$\pi(\theta) \propto \sqrt{I(\theta)}. \tag{1.7.1}$$

Mówimy, że ten wzór określa rozkład *a priori Jeffreysa*.

1.7.2 Przykład. Rozważmy rodzinę rozkładów Bernoulliego:

$$p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Ze wzoru na informację Fishera, $I(\theta) = \dfrac{n}{\theta(1-\theta)}$, otrzymujemy

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$

Rozkładem *a priori* Jeffreysa jest Beta$(1/2, 1/2)$. Szczęśliwie, należy on sprzężonej rodziny rozkładów. [7] △

1.7.3 Przykład. Rozważmy rodzinę rozkładów normalych $\mathrm{N}(\theta, \sigma^2)$ ze znanym parametrem σ and i nieznaną średnią θ. Informacja Fishera jest równa $I(\theta) = \sigma^{-2}$, a więc rozkład Jeffreysa powinien mieć gęstość

$$\pi(\theta) \propto 1.$$

To nieprzyjemny wynik, bo nie istnieje rozkład prawdopodobieństwa na prostej rzeczywistej $\Theta = \mathbb{R}$ z gęstością stałą! △

Powyższy negatywny wynik nie jest jednak końcem historii. Musimy w tym miejscu zrobić dygresję.

[7] W przypadku rozkładów dwumianowych często używa się, zamiast θ, parametru „log-iloraz szans", $\psi = \log\dfrac{\theta}{1-\theta}$.

Niewłaściwe rozkłady *a priori*

W praktyce statystycznej dość często używa się funkcji, które nie mają skończonej całki jako „gęstości *a priori*" na przestrzeni parametrów. Odpowiadają one tzw. „niewłaściwym rozkładom prawdopodobieństwa" (są to miary o nieskończonej masie całkowitej). Praktyczne podejście jest brutalne: nawet jeśli $\int \pi(\theta)\mathrm{d}\theta = \infty$., to możemy *formalnie* zastosować regułę Bayesa,

$$\pi_x(\theta) \propto p_\theta(x)\pi(\theta),$$

i może się zdarzyć, że $\int \pi_x(\theta)\mathrm{d}\theta < \infty$. Jeśli mamy właściwy rozkład *a posteriori* wynikający z niewłaściwego rozkładu *a priori*, to praktycznie nastawiony statystyk jest zadowolony, gdyż to właśnie rozkład *a posteriori* jest podstawą wnioskowania. Niestety, takie podejście jest nieścisłe i tracimy interpretację rozkładu *a posteriori* jako rozkładu warunkowego. (Istnieje ścisła teoria uzasadniająca użycie rozkładów niewłaściwych, stworzona przez węgierskiego matematyka Rényi'ego. Jednak ta teoria jest niemal zapomniana i rzadko używana. Dyskusja na ten temat przekracza zakres tego skryptu.)

1.7.4 Przykład (Dalszy ciąg Przykładu 1.7.3)**.** Rozważmy rodzinę rozkładów normalnych $\mathrm{N}(\theta, \sigma^2)$ z nieznaną średnią θ i znaną wariancją. Niewłaściwy rozkład Jeffreysa jest to „rozkład jednostajny na całej prostej" $\mathrm{U}(-\infty, \infty)$ z gęstością $\pi(\theta) \propto 1$. Jeśli $x_1, \ldots, x_n$ jest próbką z rozkładu $\mathrm{N}(\theta, \sigma^2)$ to

$$\pi_{x_1,\ldots,x_n}(\theta) \propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x} - \theta)^2\right).$$

Tak więc *a posteriori* mamy $\theta \sim \mathrm{N}\left(\bar{x}, \sigma^2/n\right)$. Gęstość a posteriori jest równa znormalizowanej wiarygodności i estymator MAP (maximum *a posteriori*) jest równy estymatorowi największej wiarygodności ($\hat{\theta} = \bar{x}$).

Zauważmy, że ten sam rezultat otrzymamy, jeśli zaczniemy od właściwego (sprzężonego) rozkładu *a priori* $\theta \sim \mathrm{N}(\mu, v^2)$, a następnie przejdziemy do granicy z $v \to \infty$. △

1.7.5 Przykład. Rozważmy rodzinę rozkładów normalych $\mathrm{N}(\theta, \sigma^2)$ ze znaną średnią μ i nieznaną wariancją. Informacja Fishera jest równa $I(\sigma) = 2\sigma^{-2}$, a więc rozkład Jeffreysa jest następujący:

$$\pi(\sigma) \propto \sigma^{-1}.$$

To jest rozkłąd niewłaściwy na półprostej $]0, \infty[$. Zauważmy, że rozkładem Jeffreysa dla $\psi = \sigma^{-2}$ (dla parametru *precyzji*) jest $\tilde{\pi}(\psi) \propto \psi^{-1}$. Możemy ten wzór interpretować jako „$\psi \sim \mathrm{Gamma}(0, 0)$ *a priori*". Jeśli $x_1, \ldots, x_n$ jest próbką z $\mathrm{N}(\mu, \sigma^2)$ to *a posteriori* $\psi \sim \mathrm{Gamma}\left(n/2, \sum(x_i - \mu)^2/2\right)$. Ten sam wynik otrzymamy jeśli zaczniemy od właściwego rozkładu *a priori* $\psi \sim \mathrm{Gamma}(\alpha, \lambda)$ i następnie przejdziemy do granicy z $\alpha, \lambda \to 0$. △

Parametr wielowymiarowy

Przypadek parametru wielowymiarowego $\theta \in \mathbb{R}^d$ można potraktować podobnie. Rozkład Jeffreysa jest dany przez

$$\pi(\theta) \propto \sqrt{\det I(\theta)}, \tag{1.7.6}$$

gdzie $I(\theta)$ jest macierzą informacyjną daną wzorem

$$I(\theta) = \int \left(\nabla_\theta \log p_\theta(x) \nabla_\theta^\top \log p_\theta(x)\right) p_\theta(x) \mathrm{d}x.$$

Rozważmy a re-parametryzację $\theta = h(\psi)$ określoną poprzez dyfeomorfizm h pomiędzy dwoma otwartymi podzbiorami $\mathbb{R}^d$. Oznaczmy przez $\tilde{I}(\psi)$ macierz informacyjną Fishera w modelu sparametryzowanym przez ψ i niech $\tilde{\pi}(\psi)$ będzie gęstością zmiennej losowej ψ. Ponieważ

$$\nabla_\psi \log p_{h(\psi)}(x) = h'(\psi) \nabla_\theta \log p_\theta(x)_{|\theta=h(\psi)},$$

wnioskujemy, że $\tilde{I}(\psi) = h'(\psi) I(h(\psi)) h'(\psi)^\top$, gdzie $h'(\psi)$ jest $d \times d$ macierzą nieosobliwą. Jeśli $\pi(\theta) \propto \sqrt{\det I(\theta)}$, to wzór na przekształcenie gęstości implikuje $\tilde{\pi}(\psi) = \pi(h(\psi)) |\det h'(\psi)| = \sqrt{\det I(h(\psi))} |\det h'(\psi)| = \sqrt{\det h'(\psi) \det I(h(\psi)) h'(\psi)^\top} = \sqrt{\det \tilde{I}(\psi)}$. Jeśli więc rozkład *a priori* $\pi(\theta)$ jest wyznaczony zgodnie z regułą Jeffreysa, to to samo jest spełnione dla $\tilde{\pi}(\psi)$.

1.7.7 Przykład. Dla rodziny rozkładów normalych $\mathrm{N}(\theta, \sigma^2)$ z oboma parameterami θ and σ nieznanymi, macierz informacyjna jest dana wzorem

$$I(\theta, \sigma) = \begin{pmatrix} 1/\sigma^2 & 0 \\ 0 & 2/\sigma^2 \end{pmatrix}.$$

A zatem otrzymujemy rozkład Jeffreysa $\pi(\theta, \sigma) \propto \sigma^{-2}$ (w szczególności to znaczy, że θ jest *a priori* niezależna od σ i ma niewłaściwy rozkład $\mathrm{U}(-\infty, \infty)$). △

Zadania

1. (Przykład Laplace'a). Mamy $r+1$ urn o numerach $0, 1, \ldots, r-1$. W każdej urnie znajduje się r kul, przy tym urna numer i zawiera i kul białych oraz $r-i$ kul czarnych. Losowanie przebiega w następujący spoób.

   - Wybieramy jedną z urn z jednakowym prawdopodobieństwem $\frac{1}{r+1}$.
   - Losujemy $n+1$ razy ze zwracaniem z wybranej uprzednio urny.

   Wiadomo, że w n losowaniach wybraliśmy same kule białe.

   (a) Obliczyć prawdopodobieństwo, że w kolejnym $(n+1)$-szym losowaniu też wyjdzie biała?

   (b) Przejść do granicy z $r \to \infty$ (liczba urn jest bardzo duża) i $n = \text{const}$. Porównać z prawdopodobieństwem predykcyjnym w modelu „Rozkład dwumianowy i rozkład *a priori* jednostajny $\mathrm{U}(0,1)$", Przykład 1.4.1.

2. (Wykładniczy/Gamma) Załóżmy, że zmienne losowe $X_1, \ldots, X_n, X_{n+1}$ są, przy danym $\vartheta = \theta$, warunkowo niezależne i każda z nich ma wykładniczy rozkład prawdopodobieństwa $\mathrm{Ex}(\theta)$. Przyjmijmy rozkład a priori Gamma: $\vartheta \sim \mathrm{Gamma}(\alpha, \lambda)$.

   (a) Oblicz rozkład *a posteriori* $\vartheta | x_1, \ldots, x_n$.

   (b) Oblicz rozkład brzegowy X_1.

   (c) Podaj rozkład predykcyjny $X_{n+1} | x_1, \ldots, x_n$.

3. (Pareto/Gamma) Zmienne losowe $X_1, \ldots, X_n$ są, przy danym $\vartheta = \theta$, warunkowo niezależne i każda z nich ma rozkład prawdopodobieństwa o gęstości

$$p(x|\theta) = \begin{cases} \theta(1+x)^{-\theta-1} & \text{dla } x \geqslant 0; \\ 0 & \text{dla } x < 0 \end{cases}$$

   (tzw. rozkład *Pareto*).

   Parametr ϑ jest zmienną losową o rozkładzie *a priori* $\mathrm{Gamma}(\alpha, \lambda)$.

   (a) Obliczyć rozkład *a posteriori* $\pi(\theta | x_1, \ldots, x_n)$.

   (b) Podać jednowymiarową statystykę dostateczną.

   (c) Obliczyć $\mathbb{E}(\vartheta | x_1, \ldots, x_n)$.

4. (Potęgowy/Gamma) Zmienne losowe $X_1, \ldots, X_n$ są, przy danym $\vartheta = \theta$, warunkowo niezależne i każda z nich ma rozkład prawdopodobieństwa o gęstości

$$p(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{dla } 0 < x < 1; \\ 0 & \text{w pozostałych przypadkach.} \end{cases}$$

   Parametr ϑ jest zmienną losową o rozkładzie *a priori* $\mathrm{Gamma}(\alpha, \lambda)$.

(a) Obliczyć rozkład *a posteriori* $\pi(\theta|x_1,\ldots,x_n)$.

(b) Podać jednowymiarową statystykę dostateczną.

(c) Obliczyć $\mathbb{E}(\vartheta|x_1,\ldots,x_n)$.

5. (Ujemny dwumianowy). Zadanie jest kontynuacją Przykładu 1.4.2. Oblicz $\mathbb{E}X$ i $\mathrm{Var}X$ dla zmiennej losowej o rozkładzie ujemnym dwumianowym $\mathrm{Bin}^-(\alpha,q)$, wykorzystując reprezentację tego rozkładu jako mieszanki rozkładów Poissona (oczywiście, można to zrobić inaczej, bezpośrednio wykorzystując wzór (1.4.3)).

6. (Normalny/Odwrotny Gamma) Oblicz rozkład brzegowy w Przykładzie 1.4.6 (jest to przesunięty i przeskalowany rozkład t Studenta).

7. Obliczyć rozkład *a posteriori* w Przykładzie 1.4.7, to znaczy podać parametry α', λ', m' i r'.

8. (Schemat urnowy Polya). Zadanie jest kontynuacją Przykładu 1.5.4.

   (a) Oblicz łączny rozkład brzegowy zmiennych $(X_1,\ldots,X_n)$.

   (b) Zinterpretuj wzory (1.5.5) w terminach losowania z urny (zakładając, że α i β są liczbami całkowitymi).

9. (Normalny/Normalny). Zadanie jest kontynuacją Przykładu 1.4.4. Oblicz łączny rozkład brzegowy zmiennych $(X_1,\ldots,X_n)$. *Wskazówka:* Jest to wielowymiarowy rozkład normalny. Można skorzystać z przedstawienia $X_i = \vartheta + (X_i - \vartheta)$.

10. Rozważmy rodzinę wykładniczą postaci (1.6.1) (naturalna parametryzacja).

    (a) Obliczyć $\mathbb{E}\left(\exp\{r^\top T(X)\}|\theta\right)$, gdzie $r \in \mathbb{R}^k$ w terminach funkcji ψ.

    (b) Pokazać, że $\mathbb{E}(T(X)|\theta) = \nabla\psi(\theta)$ i $\mathrm{VAR}(T(X)|\theta) = \nabla\nabla^\top\psi(\theta)$.

11. W Przykładzie 1.6.3 obliczyć rozkład *a posteriori*.

Rozdział 2

Teoria decyzji statystycznych

Teoria decyzji statystycznych została stworzona, głównie przez Abrahama Walda, w latach 1940-ch. Najkrócej mówiąc, jest to interpretacja wnioskowania statystycznego z punktu widzenia teorii gier. Teoria gier powstała w tymże czasie. Wśród jej twórców wymienić należy Johna von Neumanna.

2.1 Optymalne decyzje

Zajmiemy się zadaniem wyznaczania optymalnych decyzji w sytuacji niepewności. Dostępne nam informacje (lub brak informacji) opisujemy przy pomocy (znanego) rozkładu prawdopodobieństwa zmienej losowej Z o wartościach w przestrzeni $\mathcal{Z}$. Określamy przestrzeń akcji lub inaczej *przestrzeń decyzji* $\mathcal{A}$ oraz *funkcję straty* $\ell : \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$. Liczba $\ell(z, a)$ jest stratą jaką poniesiemy podejmując akcję a, jeśli okaże się, że $Z = z$. Jak ktoś nie lubi myśleć o stratach, to może uznać, że $-\ell$ jest wypłatą lub „użytecznością". Będziemy co prawda stale zakładać, że $\ell \geqslant 0$ ale to nie ma większego znaczenia, bo dodanie stałej do funkcji straty nie wpływa na wybór decyzji. Mamy zminimalizować wartość oczekiwaną:

$$r(a) = \mathbb{E}\ell(Z, a) \to \min_a . \tag{2.1.1}$$

Będziemy funkcję r nazywać ryzykiem. Jeśli $\ell(z, a) \geqslant 0$ to ryzyko $r(a) \in [0, +\infty]$ jest dobrze określone (wartość oczekiwana istnieje, ale może być nieskończona). Załóżmy, że istnieje $a_* \in \mathcal{A}$ takie, że

$$r(a_*) \leqslant r(a) \quad \text{dla wszystkich } a \in \mathcal{A}$$

i $r(a_*) < \infty$. Tak będzie we wszystkich niżej przedstawionych przykładach. Nie musimy zakładać, że a_* jest jedynym minimalizatorem.

Pokażemy, że popularne „charakterystyki rozkładów prawdopodobieństwa" są w istocie rozwiązaniami zadania (2.1.1) dla odpowiednio dobranych funkcji strat.

2.1.2 Przykład (Wartość oczekiwana)**.** Niech $\mathcal{Z} = \mathcal{A} = \mathbb{R}$ i

$$\ell(z, a) = (z - a)^2.$$

Wtedy $a_* = \mathbb{E}Z$ jest jedynym minimalizatorem $r(a) = \mathbb{E}(Z-a)^2$. Interpretujemy zatem wartość oczekiwaną jako „liczbę najlepiej przybliżającą zmienną losową" w sensie błędu średniokwadratowego. Przy tym minimalna wartość ryzyka jest równa wariancji: $r(a_*) = \operatorname{Var} Z$. Patrz Zadanie 1. △

2.1.3 Przykład (Ważona wartość oczekiwana)**.** Uogólnimy poprzedni przykład. Wprowadzimy nieujemną funkcję „wagową" w i przyjmiemy

$$\ell(z, a) = (z - a)^2 w(z).$$

Wtedy minimalizatorem ryzyka jest liczba $a_* = \mathbb{E}Zw(Z)/\mathbb{E}w(Z)$. Patrz Zadanie 2.

Ciekawym szczególnym przypadkiem jest względny błąd kwadratowy,

$$\ell(z, a) = \left(\frac{z - a}{z}\right)^2.$$

Jeśli zmienna losowa $Z > 0$ przybiera wartości „bardzo małe" lub „bardzo duże", to powyższa funkcja straty wydaje się rozsądnie formalizować zadanie przybliżenia Z przez liczbę a. △

2.1.4 Przykład (Kwantyl)**.** Niech $\mathcal{Z} = \mathcal{A} = \mathbb{R}$. Ustalmy $p \in]0, 1[$ i rozważmy funkcję straty

$$\ell(z, a) = a + \frac{z - a}{1 - p}\mathbb{1}(z > a).$$

Liczba a_* minimalizuje $r(a)$ wtedy i tylko wtedy, gdy $\mathbb{P}(Z < a_*) \leqslant p \leqslant \mathbb{P}(Z \leqslant a_*)$. Tak więc, a_* jest p-tym kwantylem Z. Na ogół, kwantyl nie jest jednoznacznie wyznaczony. Jeśli założymy dodatkowo, że dystrybuanta $F(a) = \mathbb{P}(Z \leqslant a)$ jest ciągła, to po prostu mamy $F(a_*) = p$. Co więcej, minimalna wartość ryzyka ma ładną interpretację, mianowicie $r(a_*) = \mathbb{E}(Z|Z > a_*)$. W świecie finansów, kwantyl a_* jest znany jako VaR (Value at Risk), a $\mathbb{E}(Z|Z > a_*)$ nosi nazwę CVaR (Conditional Value at Risk). Patrz Zadanie 3. △

Uwaga. Jeśli rozważymy afiniczną transformację funkcji strat, to znaczy zdefiniujemy $\tilde{\ell}(z, a) = c + b\ell(z, a)$, gdzie $b > 0$ to, oczywiście, funkcje ryzyka r and $\tilde{r}$ odpowiadające ℓ and $\tilde{\ell}$ mają ten sam minimalizator. Na przykład, funkcja straty

$$\tilde{\ell}(z, a) = (a - z)(1 - p) + (z - a)\mathbb{1}(z > a) = \begin{cases} (a - z)(1 - p) & \text{for } z \leqslant a; \\ (z - a)p & \text{for } z > a, \end{cases}$$

może być użyta w Przykładzie 2.1.4. W szczególności dla $p = 1/2$ widzimy, że mediana jest minimalizatorem $\tilde{r}(a) = \mathbb{E}|a - Z|$.

Ogólniej, możemy rozważyć przekształconą stratę postaci $\tilde{\ell}(z, a) = c(z) + b\ell(z, a)$ i nadal otrzymać te same minimalizatory r i $\tilde{r}$ (czyli wyraz wolny transformacji afinicznej może zależeć od z) Ten fakt jest przydatny, aby rozluźnić założenia dotyczące momentów. Na przykład, jeśli $\ell(z, a) = |a - z|$ to musimy założyć że $\mathbb{E}|Z| < \infty$, aby zdefiniować $r(a) = \mathbb{E}|a - Z|$. Minimalizator r jest medianą Z, pod warunkiem, że Z ma skończoną wartość oczekiwaną. Jeśli zmienimy ℓ na $\tilde{\ell}(z, a) = |a-z|-|z|$, to minimalizator $\tilde{r}(a) = \mathbb{E}(|a-Z|-|Z|)$ jest medianą, bez żadnych ograniczeń na momenty.

2.1.5 Przykład (Funkcja Tworząca Kumulanty)**.** Niech Z będzie jednowymiarową zmienną losową i rozważmy następującą funkcję straty 'LINEX ':

$$\ell(z, a) = \exp\{\kappa(z - a)\} - \kappa(z - a) - 1,$$

gdzie $\kappa > 0$. Minimalizatorem ryzyka jest $a_* = \dfrac{1}{\kappa} \log \mathbb{E} \exp\{\kappa Z\}$. Patrz Zadanie 4. △

2.1.6 Przykład (Moda)**.** Załóżmy, że $\mathcal{Z}$ jest zbiorem skończonym, $\mathcal{A} = \mathcal{Z}$ i rozważmy zero-jedynkową funkcję straty

$$\ell(z, a) = \mathbb{1}(z \neq a).$$

Minimalizatorem ryzka jest $a_* = \arg\max_a \mathbb{P}(Z = a)$. △

2.2 Optymalne reguły decyzyjne

Rozważmy model nieco bardziej rozbudowany niż w poprzednim podrozdziale. Załóżmy, że znamy łączny rozkład pary zmiennych losowych (X, Y) o wartościach w przestrzeni $\mathcal{X} \times \mathcal{Y}$ i mamy podjąć decyzję $a \in \mathcal{A}$ na podstawie obserwowanej wartości $X = x$. Zmienna losowa Y jest nieobserwowana. Strata zależy od decyzji a i nieznanej wartości $Y = y$. Innymi słowy $\ell : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$. Zadanie polega na znalezieniu reguły decyzyjnej, czyli funkcji $\delta : \mathcal{X} \to \mathcal{A}$, która minimalizuje ryzyko zdefiniowane jako wartość oczekiwana straty:

$$r(\delta) = \mathbb{E}\ell(Y, \delta(X)) \to \min_{\delta}. \tag{2.2.1}$$

Mówimy, że δ_* jest **regułą bayesowską**[1], jeśli

$$r(\delta_*) \leqslant r(\delta)$$

[1]Używamy tu terminów nawiązujących do statystyki bayesowskiej, ale rozważania w tym podrozdziale mają zastosowanie w ogólniejszej sytuacji.

dla każdej reguły δ. Sposób obliczania reguł bayesowskich jest „w zasadzie łatwy". *Ryzyko a posteriori* definiujemy, dla $a \in \mathcal{A}$, wzorem

$$r(a|x) = \mathbb{E}(\ell(Y,a)|X=x) = \int_{\mathcal{Y}} \ell(y,a)p(y|x)\mathrm{d}y. \tag{2.2.2}$$

Statystyk, poszukując najlepszej decyzji a ma do dyspozycji ustaloną obserwację x, może więc w zasadzie obliczyć $r(a|x)$ dla różnych wartości a i wybrać tę decyzję, ktora minimalizuje ryzyko *a posteriori*. W ten sposób określa regułę decyzyjną, która okazuje się optymalna (bayesowska). Mówi o tym następujące, niemal oczywiste ale przy tym bardzo ważne twierdzenie.

2.2.3 Twierdzenie. *Jeżeli reguła δ_* jest taka, że dla każdego x mamy $r(\delta_*(x)|x) \leqslant r(a|x)$ dla każdego a, to δ_* jest regułą bayesowską.*

Dowód. Zamiana kolejności całkowania prowadzi do wniosku, że

$$\begin{aligned} r(\delta) &= \iint_{\mathcal{X}\times\mathcal{Y}} \ell(y,\delta(x))p(x,y)\mathrm{d}x\mathrm{d}y \\ &= \int_{\mathcal{X}}\int_{\mathcal{Y}} \ell(y,\delta(x))p(y|x)\mathrm{d}y p(x)\mathrm{d}x \\ &= \int_{\mathcal{X}} r(\delta(x)|x)p(x)\mathrm{d}x. \end{aligned}$$

Dla reguły δ_* i dowolnej (innej) reguły decyzyjnej δ mamy zatem

$$r(\delta_*) = \int_{\mathcal{X}} r(\delta_*(x)|x)p(x)\mathrm{d}x \leqslant \int_{\mathcal{X}} r(\delta(x)|x)p(x)\mathrm{d}x = r(\delta),$$

ponieważ z założenia $r(\delta_*(x)|x) \leqslant r(\delta(x)|x)$ dla każdego x. □

Przepis na skonstruowanie reguły bayesowskiej jest zatem prosty. Pozostają „tylko" trudności obliczeniowe. Rozkład warunkowy we wzorze (2.2.2) może być bardzo trudny do obliczenia.

2.3 Gra statystyczna

Opiszemy teraz model „gry statystycznej" w klasycznym ujęciu, obejmującym zarówno statystykę częstościową jak i bayesowską. Niech $(\mathcal{X}, \{P_\theta : \theta \in \Theta\})$ będzie przestrzenią statystyczną. Rozważamy dodatkowo przestrzeń akcji lub inaczej przestrzeń decyzji $\mathcal{A}$ oraz funkcję straty $\ell : \Theta \times \mathcal{A} \to \mathbb{R}$. Interpretacja jest następująca. Wyobrażamy sobie grę, której uczestnikami są „Statystyk" i „Natura". Z jednej strony uznajemy, że parametr θ opisuje interesujące nas aspekty rzeczywistości. Obrazowo mówiąc, Natura wybiera element $\theta \in \Theta$.

Statystyk podejmuje akcję $a \in \mathcal{A}$. Liczba $\ell(\theta, a)$ jest stratą, jaką ponosi Statystyk, który wybrał akcję a, gdy stanem Natury jest θ. Wstępny model gry schematycznie wygląda tak:

$$\ell(\theta, a), \qquad \theta \in \Theta, a \in \mathcal{A}.$$

W teorii decyzji statystycznych rozpatruje się model bardziej rozbudowany. Statystyk przy podejmowaniu decyzji wykorzystuje dane, czyli obserwuje zmienną losową X wylosowaną z rozkładu prawdopodobieństwa P_θ. Odpowiada to następującym regułom gry: Natura po wybraniu θ „jest zobowiązana" przeprowadzić losowanie $X \sim P_\theta$ i pokazać statystykowi wynik, czyli element $x \in \mathcal{X}$. Statystyk może uzależnić wybór decyzji od od obserwacji. Niech $\delta(x) \in \mathcal{A}$ będzie decyzją podjętą po zaobserwowaniu $x \in \mathcal{X}$. tak więc Statystyk w istocie wybiera *regułę decyzyjną*, czyli funkcję $\delta : \mathcal{X} \to \mathcal{A}$. Za wynik gry przyjmuje się średnią (oczekiwaną) wartość straty, którą nazywamy *ryzykiem*. Jeśli założymy, że każdy z rozkładów prawdopodobieństwa P_θ ma gęstość względem ustalonej miary $\mathrm{d}x$ to ryzyko wyraża się wzorem

$$R(\theta, \delta) = E_\theta \ell(\theta, \delta(X)) = \int_{\mathcal{X}} \ell(\theta, \delta(x)) p_\theta(x) \mathrm{d}x. \tag{2.3.1}$$

Jak zwykle, $\mathrm{d}x$ może oznaczać w zasadzie dowolną ustaloną miarę. Zbiór wszystkich reguł decyzyjnych oznaczmy $\mathcal{D}$. Otrzymujemy rozszerzony model gry statystycznej:

$$R(\theta, \delta), \qquad \theta \in \Theta, \delta \in \mathcal{D}.$$

Kolejne rozszerzenie gry wiąże się z podejściem bayesowskim. Niech π będzie rozkładem *a priori* na przestrzeni Θ (utożsamiamy rozkład prawdopodobieństwa z gęstością). *Ryzyko bayesowskie* określamy jako średnią (oczekiwaną) wartość ryzyka względem rozkładu π:

$$r(\pi, \delta) = \int_{\Theta} R(\theta, \delta) \pi(\theta) \mathrm{d}\theta. \tag{2.3.2}$$

Niech $\mathcal{P}$ oznacza zbiór wszystkich rozkładów *a priori*. Bayesowską grę statystyczną schematycznie możemy przedstawić w następujący sposób:

$$r(\pi, \delta), \qquad \pi \in \mathcal{P}, \delta \in \mathcal{D}.$$

W statystyce bayesowskiej rozkład *a priori* jest ustalony, możemy zatem pomijać zależność od π, pisząc $r(\pi, \delta) = r(\delta)$. Ryzyko bayesowskie jest wartością oczekiwaną względem względem łącznego rozkładu zmiennych (X, ϑ):

$$r(\pi, \delta) = \iint_{\Theta \times \mathcal{X}} \ell(\theta, \delta(x)) p_\theta(x) \pi(\theta) \mathrm{d}x \mathrm{d}\theta = \mathbb{E}\ell(\vartheta, \delta(X)). \tag{2.3.3}$$

Znaleźliśmy się w sytuacji przedstawionej w poprzednim podrozdziale, porównaj z równaniem (2.2.1). Dla ustalonego (i znanego) rozkładu *a priori*[2] można wskazać reguły najlepsze w sensie minimalnego ryzyka bayesowskiego. Otrzymujemy je, zgodnie z Twierdzeniem 2.2.3, minimalizując ryzyko a posteriori $r(a|x) = \mathbb{E}(\ell(\vartheta, a)|X = x)$.

[2]Dla nas, bayesistów, rozkład *a priori* jest znany i ustalony. Statystycy „częstościowi" traktują rozkład *a priori* jako „zrandomizowaną strategię Natury" (wybór pojedynczej wartości parametru θ jest w tym ujęciu „czystą strategią Natury").

Dostateczność w teorii decyzji

Rolę dostateczności w bayesowskiej teorii decyzji wyjaśnia następujący fakt.

2.3.4 Stwierdzenie. *Rozpatrzmy bayesowski model statystyczny. Jeśli T jest statystyką dostateczną to dla dowolnej funkcji strat ℓ, bayesowska reguła decyzyjna δ^* zależy od x tylko poprzez $T(x)$.*

Dowód. To jest oczywiste, ponieważ ryzyko bayesowskie a posteriori $r(\delta|x)$ jest całką względem rozkładu a posteriori π_x. Wiemy, że ten rozkład zależy od x tylko poprzez $T(x)$. □

W statystyce częstościowej sprawa jest bardziej skomplikowana. Potrzebujemy dodatkowo pewnych założeń o wypukłości.

2.3.5 Stwierdzenie (Blackwell-Rao)**.** *Rozpatrzmy częstościowy model gry statystycznej. Załóżmy, że przestrzeń akcji $\mathcal{A}$ jest wypukłym podzbiorem $\mathbb{R}^d$ i $\ell(\theta, a)$ jest wypukłą funkcją argumentu a dla dowolnego θ. Niech T będzie statystyką dostateczną. Dla dowolnej reguły decyzyjnej δ, zdefiniujmy*

$$\delta'(t) = E_\theta(\delta(X)|T = t).$$

Reguła δ' jest jednostajnie niegorsza niż δ w następującym sensie: $R(\theta, \delta') \leqslant R(\theta, \delta)$ dla wszystkich θ.

Dowód. Najważniejsze jest spostrzeżenie, że δ' jest dobrze określoną regułą decyzyjną. Po pierwsze, jeśli $\delta(X) \in \mathcal{A}$ to wartość oczekiwana $\delta(X)$ też należy do $\mathcal{A}$, ponieważ $\mathcal{A}$ jest zbiorem wypukłym. Po drugie, warunkowa wartość oczekiwana $E_\theta(\delta(X)|T = t)$ nie zależy od θ, ponieważ T jest statystyką dostateczną (korzystamy tu z częstościowej definicji dostateczności).

Dalsza część dowodu jest łatwa. Na mocy nierówności Jensena,

$$\begin{aligned} R(\theta, \delta') = E_\theta \ell(\theta, \delta'(T)) &= E_\theta \ell(\theta, E_\theta(\delta(X)|T)) \\ &\leqslant E_\theta E_\theta \left(\ell(\theta, \delta(X))\middle|T\right) \\ &= E_\theta \ell(\theta, \delta(X)) = R(\theta, \delta). \end{aligned}$$

□

Reguły minimaksowe

Zasada „minimaksu" jest pewną alternatywą podejścia bayesowskiego. Pozwala wprowadzić liniowy porządek w przestrzeni reguł decyzyjnych bez odwoływania się do rozkładu *a priori*.

Mówimy, że reguła decyzyjna δ^* jest **minimaksowa**, jeśli dla każdej reguły δ,

$$\sup_\theta R(\theta, \delta^*) \leqslant \sup_\theta R(\theta, \delta).$$

Przypomnijmy, że reguła decyzyjna δ^* jest **bayesowska**, jeśli dla każdej reguły δ,

$$\int R(\theta, \delta^*)\pi(\theta)\mathrm{d}\theta \leqslant \int R(\theta, \delta^*)\pi(\theta)\mathrm{d}\theta.$$

Tak więc, reguła bayesowska minimalizuje „średnie ryzyko", podczas gdy reguła minimaksowa minimalizuje „ryzyko w najgorszym przypadku".

Zasada minimaksu wydaje się zakładać, że w grze Satatystyka z Naturą, ta ostatnia „wybiera strategię najbardziej niekorzystną dla Statystyka". To założenie, oczywiście, jest bardzo kontrowersyjne! Niemniej, idea minimaksu pojawia się w różnych kontekstach w statystyce i warto jej poświęcić uwagę.

Poniższe stwierdzenie przedstawia najprostszą metodę znajdowania reguł minimaksowych.

2.3.6 Stwierdzenie. *Jeżeli reguła δ^* ma stałą funkcję ryzyka i jednocześnie jest bayesowska dla pewnego rozkładu a priori π, to δ^* jest minimaksowa.*

Dowód. Zakładamy, że $R(\theta, \delta^*) \equiv c$. Gdyby istniała reguła δ taka, że $\sup_\theta R(\theta, \delta) < \sup_\theta R(\theta, \delta^*) = c$, to mielibyśmy $\int R(\theta, \delta)\pi(\theta)\mathrm{d}\theta < \int R(\theta, \delta^*)\pi(\theta)\mathrm{d}\theta = c$, co byłoby sprzeczne z założeniem, że δ^* jest π-bayesowska. □

Zadanie 4 w Rozdziale 3 ilustruje wykorzystanie tego stwierdzenia. Okazuje się, że można udowodnić stwierdzenie podobne do 2.3.6 przy nieco słabszych założeniach.

2.3.7 Stwierdzenie. *Załóżmy, że reguła δ^* ma stałą funkcję ryzyka i jednocześnie istnieje ciąg rozkładów a priori π_k taki, że $r(\pi_k, \delta^*) - r(\pi_k, \delta_k) \to 0$ dla $k \to \infty$, gdzie δ_k oznacza regułę π_k-bayesowską. Wtedy reguła δ^* jest minimaksowa.*

Dowód pozostawiam Czytelnikowi: Zadanie 5. Ważny wniosek jest sformułowany w Zadaniu 5 w Rozdziale 3: dla próbki z rozkładu normalnego, średnia z próbki okazuje się być estymatorem minimaksowym.

Zadania

1. Uzasadnić fakty wymienione w Przykładzie 2.1.2.
2. Uzasadnić wynik w Przykładzie 2.1.3.
3. Uzasadnić fakty wymienione w Przykładzie 2.1.4.
   Wskazówka: Można skorzystać z równości $\mathbb{E}(Z-a)\mathbb{1}(Z>a) = \int_a^\infty \mathbb{P}(Z>x)\mathrm{d}x$.
4. Uzasadnić wynik w Przykładzie 2.1.5.
5. Udowodnić Stwierdzenie 2.3.7.

Rozdział 3

Estymacja i predykcja w modelu bayesowskim

Dla prawdziwego bayesisty, obliczenie rozkładu *a posteriori* jest ostatecznym wynikiem wnioskowania statystycznego. Z bardziej pragmatycznego punktu widzenia, rozkład *a posteriori* jest jedynie podstawą wyboru optymalnych decyzji. Problemy statystyczne różnią się „typem decyzji" i odpowiednim wyborem funkcji straty. W świecie bayesowskim zadania estymacji i predykcji sprowadzają się do przybliżenia jednej zmiennej losowej funkcją innej zmiennej losowej.

3.1 Estymacja

Mamy przybliżyć nieobserwowany parametr θ funkcją obserwacji x (danych). W modelu bayesowskim zarówno $\theta \in \Theta$, jak i $x \in \mathcal{X}$ są wartościami zmiennych losowych ϑ i X. W zadaniu estymacji zazwyczaj przestrzenią decyzji jest $\mathcal{A} = \Theta$ (czasami zdarza się, że $\mathcal{A} \supset \Theta$). Nieco ogólniej, zadanie może polegać na przybliżeniu pewnej funkcji parametru, powiedzmy $g(\theta)$[1]. Jeśli $g : \Theta \to \mathbb{R}$, to za przestrzeń decyzji przyjmujemy $\mathcal{A} = \mathbb{R}$. Regułę decyzyjną nazywamy estymatorem. Często, dla podkreślenia, estymator funkcji $g(\theta)$ oznaczamy symbolem $\hat{g}$ (lub $\hat{\theta}$, jeśli $g(\theta) = \theta$). Estymatory bayesowskie wyznaczamy na podstawie Twierdzenia 2.2.3, minimalizując ryzyko *a posteriori*

$$r(a|x) = \mathbb{E}\left(\ell(\vartheta, a)|X = x\right) = \int_{\Theta} \ell(\theta, a)\pi_x(\theta)\mathrm{d}\theta.$$

[1]Powiedzmy, że $g : \Theta \to \mathbb{R}$. Funkcja g jest znana, nieznany jest jej argument θ. Typowe przykłady to estymacja $g(\theta) = \theta$, $g(\theta) = 1/\theta$ albo $g(\theta_1, \theta_2) = \theta_1$.

W Podrozdziale 2.1 wymieniliśmy kilka typowych funkcji strat, które można uznać za „karę za błąd przybliżenia" i które są stosowane w estymacji. Reguły bayesowskie odpowiadające funkcjom strat z Przykładów 2.1.2, 2.1.3, 2.1.4, 2.1.5 są to odpowiednie charakterystyki rozkładu *a posteriori*. Wniosek z Przykładu 2.1.2 jest taki:

- Dla kwadratowej funkcji straty $\ell(\theta, a) = (g(\theta) - a)^2$, estymatorem bayesowskim funkcji $g(\theta)$ jest wartość oczekiwana *a posteriori*,

$$\hat{g}(x) = \mathbb{E}(g(\vartheta)|x) = \int g(\theta)\pi_x(\theta)\mathrm{d}\theta.$$

Analogiczne wnioski dotyczą Przykładów 2.1.3, 2.1.4, 2.1.5. Dodajmy, że w standardowych modelach bayesowskich (ze sprzężonymi rozkładami *a priori*) wartość oczekiwana, kwantyle, funkcja tworząca kumulanty i tym podobne charakterystyki rozkładu *a posteriori* są dane jawnymi wzorami albo łatwo je obliczać numerycznie.

3.2 Predykcja

Zadanie predykcji z punktu widzenia matematycznego nie różni się znacząco od zadania estymacji. Rozważamy obserwowaną wartość zmiennej losowej $X = x$ i zmienną Y, która dotyczy przyszłości. Zakładamy, że zmienne X i Y są warunkowo niezależne przy danym $\vartheta = \theta$. Mamy przybliżyć nieobserwowaną zmienną Y funkcją x (danych). Predyktor jest funkcją $\delta : \mathcal{X} \to \mathcal{Y}$, gdzie $\mathcal{X}$ i $\mathcal{Y}$ są przestrzeniami wartości zmiennych losowych X i Y. Dla podkreślenia, czasami oznacza się predyktor zmiennej Y symbolem $\hat{Y}$. Ryzyko bayesowskie jest dane wzorem (2.2.1). Jasne, że optymalny predyktor δ_* otrzymujemy minimalizując wartość oczekiwaną straty *względem rozkładu predykcyjnego*. Przypomnijmy wzór (2.2.2):

$$r(a|x) = \mathbb{E}\left(\ell(Y, a)|X = x\right) = \int_{\mathcal{Y}} \ell(y, a)p(y|x)\mathrm{d}y,$$

gdzie $p(y|x) = \int p_\theta(y)\pi_x(\theta)\mathrm{d}\theta$. Sposób obliczania rozkładu predykcyjnego w modelu bayesowskim omówiliśmy w poprzednim rozdziale. Analogicznie jak w przypadku estymacji, odwołujemy się do Przykładów 2.1.2, 2.1.3, 2.1.4, 2.1.5 aby wyznaczyć optymalne predyktory dla typowych funkcji strat. Na przykład:

- Dla kwadratowej funkcji straty $\ell(y, a) = (y - a)^2$, predyktorem bayesowskim zmiennej Y jest wartość oczekiwana rozkładu *predykcyjnego*,

$$\hat{Y}(x) = \mathbb{E}(Y|x) = \int yp(y|x)\mathrm{d}y.$$

W moim przekonaniu, pomiędzy estymacją i predykcją jest zasadnicza różnica metodologiczna. Rozwiązanie zadania *predykcji* można poddać weryfikacji empirycznej. W przyszłości pojawi się wartość $Y = y$ i będziemy mogli porównać tę wartość z naszym przewidywaniem $\delta(x)$ (jakkolwiek wyznaczymy $\delta(x)$). Z drugiej strony, wynik *estymacji* dotyczy zmiennej ϑ, która jest z zasady nieobserwowalna, jest tylko elementem modelu matematycznego.

Uwaga. Dla kwadratowej funkcji straty, predyktor bayesowski $\hat{Y}$ zmiennej losowej Y jest identyczny z estymatorem bayesowskim $\hat{\mu}$ funkcji $\mu(\theta) = \mathbb{E}(Y|\theta)$. To wynika natychmiast z założenia o warunkowej niezależności X i Y przy danym $\vartheta = \theta$ i własności warunkowej wartości oczekiwanej: $\hat{Y} = \mathbb{E}(Y|x) = \mathbb{E}(\mathbb{E}(Y|x,\vartheta)|x) = \mathbb{E}(\mathbb{E}(Y|\vartheta)|x) = \mathbb{E}(\mu(\vartheta)|x) = \hat{\mu}$.

Z drugiej strony, błąd (średniokwadratowy) predykcji Y jest większy, niż błąd estymacji $\mu(\vartheta)$. Łatwo zauważyć, że

$$\begin{aligned}\mathbb{E}(\hat{\mu} - \mu(\vartheta))^2 &= \mathbb{E}\mathrm{Var}(\mu(\vartheta)|X),\\ \mathbb{E}(\hat{Y} - Y)^2 &= \mathbb{E}\mathrm{Var}(\mu(\vartheta)|X) + \mathbb{E}\mathrm{Var}(Y|\vartheta).\end{aligned}$$

3.2.1 Przykład. Rozważmy model Normalny/Normalny z Przykładu 1.4.4. Zakładamy, że $(X_1, \ldots, X_n, X_{n+1}|\mu) \sim_{\text{i.i.d.}} \mathrm{N}(\mu, \sigma^2)$ i $\mu \sim \mathrm{N}(m, v^2)$. Dla kwadratowej funkcji straty, predyktor przyszłej obserwacji X_{n+1} jest identyczny z estymatorem bayesowskim parametru μ. Ze wzoru (1.4.5) wynika, że

$$\hat{X}_{n+1} = \hat{\mu} = z\bar{X} + (1-z)m, \quad \text{gdzie} \quad z = \frac{nv^2}{nv^2 + \sigma^2}.$$

Błąd średniokwadratowy estymacji i predykcji jest równy, odpowiednio,

$$\begin{aligned}\mathbb{E}(\hat{\mu} - \mu)^2 &= \frac{\sigma^2 v^2}{nv^2 + \sigma^2},\\ \mathbb{E}(\hat{\mu} - X_{n+1})^2 &= \frac{\sigma^2 v^2}{nv^2 + \sigma^2} + \sigma^2.\end{aligned}$$

△

3.2.2 Przykład (Ciąg dalszy)**.** Rozpatrzmy kwantylową funkcję straty z Przykładu 2.1.4 w tym samym modelu Normalny/Normalny. Bayesowski estymator μ i predyktor X_{n+1} minimalizują, odpowiednio,

$$\begin{aligned}&\mathbb{E}q_p(\hat{\mu} - \mu)\\ &\mathbb{E}q_p(\hat{X}_{n+1} - X_{n+1})^2,\end{aligned}$$

gdzie $q_p(t) = t(1-p) + t\mathbb{1}(t < 0)$. Łatwo widzieć, że

$$\begin{aligned}\hat{\mu} &= z\bar{X} + (1-z)m + \Phi^{-1}(p)\sqrt{\frac{\sigma^2 v^2}{nv^2 + \sigma^2}},\\ \hat{X}_{n+1} &= z\bar{X} + (1-z)m + \Phi^{-1}(p)\sqrt{\frac{\sigma^2 v^2}{nv^2 + \sigma^2} + \sigma^2},\end{aligned}$$

gdzie Φ^{-1} jest funkcją kwantylową standardowego rozkładu normalnego $\mathrm{N}(0,1)$. △

Podobny przykład dotyczący funkcji straty LINEX jest zamieszczony w Zadaniu 2.

3.3 Klasyfikacja statystyczna

Załóżmy, że „obiekt" należy do jednej z „klas" oznaczanych $1, \ldots, k$, ale nie wiemy do której z nich. Obserwujemy wektor $x = (x_1, \ldots, x_d) \in \mathcal{X} \subseteq \mathbb{R}^d$, którego składowe opisują „cechy" obiektu. Na podstawie x mamy „zaklasyfikować" obiekt, czyli odgadnąć jego przynależność do klasy. Na przykład, obiekt może być pacjentem, klasom odpowiadają (wzajemnie wykluczające się) jednostki chorobowe, składowe wektora x mogą być objawami, wynikami diagnostycznych badań medycznych itp.

Jeśli założymy, że (nieobserwowana) etykieta klasy c i (zaobserwowany) wektor cechy x są wartościami zmiennych losowych C i X, wtedy problem klasyfikacji możemy rozważać w ujęciu bayesowskim. Zbiór etykiet klasyfikacyjnych $\mathcal{C} = \{1, \ldots, k\}$ może być traktowany jako (szczególnie prostą, bo skończoną) przestrzeń parametrów. Rolę wiarygodności odgrywają „wewnątrzklasowe" gęstości $p(x|c)$, $c = 1, \ldots, k$, które opisują warunkowe rozkłady prawdopodobieństwa wektora cech X, dla $C = c$. Prawdopodobieństwa *a priori* $\pi(c)$ są względnymi częstościami występowania klas. Zwróćmy uwagę, że rozkład *a priori* ma tu obiektywną interpretację. W naszym przykładzie diagnostyki medycznej, $\pi(c)$ może oznaczać odsetek pacjentów cierpiących na chorobę c (w hipotetycznej populacji potencjalnych pacjentów).

Naturalną przestrzenią decyzji w problemie klasyfikacji jest albo $\mathcal{A} = \mathcal{C}$ albo $\mathcal{A} = \mathcal{C} \cup \{0\}$. Po zaobserwowaniu x zgadujemy, że obiekt należy do klasy $\hat{c}(x) \in \mathcal{C}$ albo, być może, mówimy „nie wiem". Ta ostatnia akcja – zawieszona decyzja – jest zakodowana jako $\hat{c}(x) = 0$. Reguła decyzyjna $\hat{c} : \mathcal{X} \to \mathcal{A}$ jest nazywana *klasyfikatorem*. Regułę $\hat{c}$ można równoważnie opisać jako podział przestrzeni cech $\mathcal{X}$ na sumę obszarów decyzyjnych $\mathcal{D}_a = \{x : \hat{c}(x) = a\}$. Funkcję straty utożsamiamy z macierzą $(\ell(a, c), a \in \mathcal{A}, c \in \mathcal{C})$. Wybór funkcji strat zależy, oczywiście od specyfiki problemu. Zawsze zakładamy, że

- $\ell(a, c) \geqslant 0$ dla wszystkich a and c,
- $\ell(c, c) = 0$ dla wszystkich c (jeśli zgadujemy poprawnie, to nic nie tracimy),
- dla wszystkich $a \neq 0$ mamy $\ell(a, c) > \ell(0, c)$ dla przynajmniej jednego c (tracimy mniej odraczając decyzję, niż podejmując błędną decyzję).

Z Twierdzenia 2.2.3 wynika, że najlepsza reguła decyzyjna, *klasyfikator bayesowski* $c^* : \mathcal{X} \to \mathcal{A}$, jest dana wzorem

$$c^*(x) = \arg\min_a r(a|x) = \arg\min_a \sum_{c=1}^{k} \ell(a, c)\pi(c|x),$$

gdzie prawdopodobieństwa *a posteriori* są obliczone zgodnie ze wzorem Bayesa

$$\pi(c|x) = \frac{p(x|c)\pi(c)}{p(x)}.$$

Równoważnie,

$$c^*(x) = \arg\min_a \sum_{c=1}^{k} \ell(a,c)p(x|c)\pi(c).$$

Jeśli minimum jest osiągane dla więcej niż jednego a, możemy wybrać dowolny spośród minimalizatorów.

W dlaszych rozważaniach skupiamy się na specjalnych funkcjach straty. Jeśli uwzględniamy możliwość zawieszenia decyzji „0" to możemy przyjąć

$$\ell(a,c) = \begin{cases} 0 & \text{jeśli } a = c; \\ 1 & \text{jeśli } a \neq c \text{ i } a \neq 0; \\ \lambda & \text{jeśli } a = 0. \end{cases}$$

Ryzyko bayesowskie jest równe

$$r(\hat{c}) = \mathbb{P}(\hat{c}(X) \neq C, \hat{c}(X) \neq 0) + \lambda \mathbb{P}(\hat{c}(X) = 0).$$

Bayesowski klasyfikator $c^* : \mathcal{X} \to \mathcal{C} \cup \{0\}$ jest dany wzorem

$$c^*(x) = \begin{cases} \arg\max_c \pi(c|x) & \text{jeśli } \max_c \pi(c|x) \geqslant 1-\lambda; \\ 0 & \text{jeśli } \max_c \pi(c|x) < 1-\lambda. \end{cases}$$

W przypadku $\mathcal{A} = \mathcal{C}$ (jeśli zawieszenie decyzji nie jest dozwolone) i dla zero-jedynkowej funkcji straty $\ell(a,c) = \mathbb{1}(a \neq c)$ ryzyko bayesowskie jest prawdopodobieństwem błędnej klasyfikacji,

$$r(\hat{c}) = \mathbb{P}(\hat{c}(X) \neq C).$$

Bayesowski klasyfikator $c^* : \mathcal{X} \to \mathcal{C}$ jest dany wzorem

$$c^*(x) = \arg\max_c \pi(c|x) = \arg\max_c p(x|c)\pi(c) = \arg\max_c \left[\log p(x|c) + \log \pi(c)\right]$$

(najczęściej używa się wersji zlogarytmowanej).

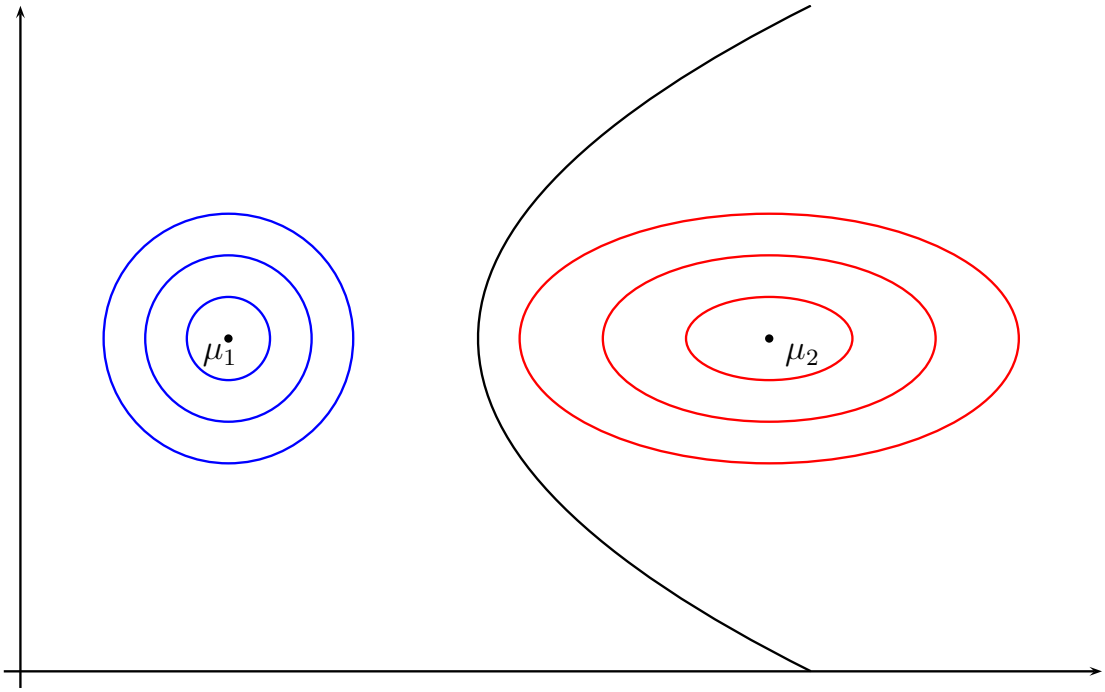
3.3.1 Przykład (Rozkłady normalne z nierównymi macierzami kowariancji)**.** Załóżmy, że w każdej z klas wektor cech ma wielowymiarowy rozkład normalny $(X|C = c) \sim \mathrm{N}(\mu_c, \Sigma_c)$. Załóżmy, że macierze Σ_c są nieosobliwe. Wtedy

$$p(x|c) = (2\pi)^{-\frac{d}{2}} (\det \Sigma_c)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(x-\mu_c)^\top \Sigma_c^{-1}(x-\mu_c)\Big).$$

Rozważmy klasyfikację bez zawieszenia decyzji, $\mathcal{A} = \mathcal{C}$. Skoro mamy

$$\log p(x|c) = -\frac{1}{2}(x-\mu_c)^\top \Sigma_c^{-1}(x-\mu_c) - \frac{1}{2}|\det \Sigma_c| + \text{const},$$

więc klasyfikator bayesowski wybiera maksimum spośród k kwadratowych funkcji $d_c(x) = \log p(x|c) + \log \pi(c)$. Nazywamy je *kwadratowymi funkcjami dyskryminacyjnymi*. W rzeczywistości potrzebujemy tylko $k-1$ funkcji, na przykład $d_c(x) - d_1(x)$ dla $c = 2, \ldots, k$. Granice bayesowskich obszarów decyzyjnych są rozmaitościami kwadratowymi, jak widać na Rysunku 3.1. △



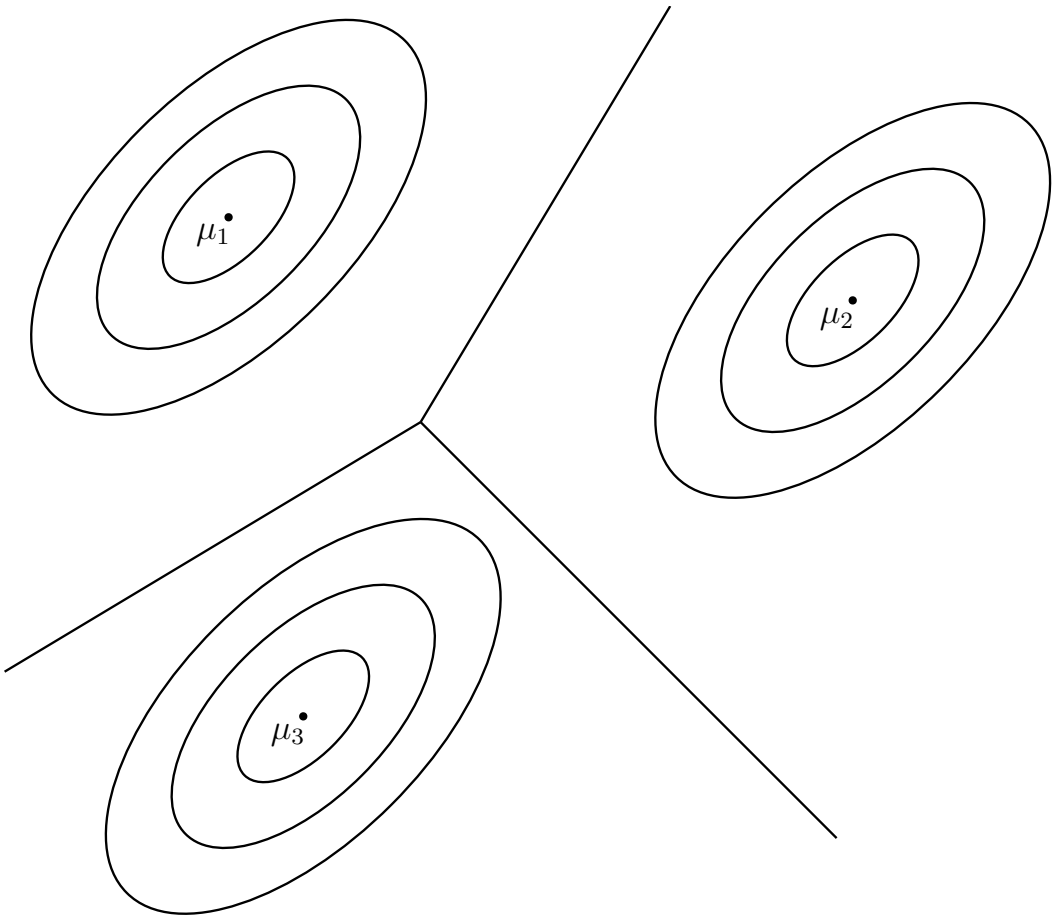
Rysunek 3.1: Kwadratowe obszary decyzyjne w Przykładzie 3.3.1.

3.3.2 Przykład (Rozkłady normalne z równymi macierzami kowariancji)**.** Załóżmy dodatkowo, że macierze kowariancji wewnątrzklasowych rozkładów są jednakowe, $(X|C=c) \sim \mathrm{N}(\mu_c, \Sigma)$. Mamy teraz

$$\log p(x|c) - \log p(x|1) = \left(x - \frac{\mu_c + \mu_1}{2}\right)^\top \Sigma^{-1}(\mu_c - \mu_1),$$

a więc mamy $k-1$ *liniowych funkcji dyskryminacyjnych* $d_c(x) = \log p(x|c) - \log p(x|c) + \log(\pi(c)/\pi(1))$. Granice bayesowskich obszarów decyzyjnych są hiperpłaszczyznami, jak to pokazano na Rysunku 3.2.

Przykłady 3.3.1 i 3.3.1 przedstawiają, z punktu widzenia bayesowskiej teorii decyzji, podstawy obszernego działu statystyki, który nosi nazwę analizy dyskryminacyjnej. Opisaliśmy sytuację, w której znamy rozkłady prawdopodobieństwa i możemy wyznaczyć klasyfikatory bayesowskie. W istocie analiza dyskryminacyjna zajmuje się bardziej realistyczną sytuacją, kiedy rozkłady wewnątrzklasowe są nieznane i estymuje się parametry tych rozkładów (na przykład wektory μ_c i macierze Σ_c lub macierz Σ). Najczęściej w analizie dyskryminacyjnej stosuje się estymatory „częstościowe", nie-bayesowskie. To jest tematyka, która wykracza poza zakres naszych rozważań. △



Rysunek 3.2: Granice obszarów decyzyjnych w Przykładzie 3.3.2.

3.4 Parametryczne klasy reguł decyzyjnych

Twierdzenie 2.2.3 pozwala wyznaczać optymalne reguły decyzyjne w klasie wszystkich reguł (funkcji $\delta : \mathcal{X} \to \mathcal{A}$). Bardzo często interesuje nas znalezienie najlepszej reguły w pewnej

ograniczonej klasie funkcji, na przykład w klasie funkcji liniowych. Motywacje stojące za takim postawieniem zadania są dwojakiego rodzaju. Po pierwsze, użytkownik może być zainteresowany rozwiązaniem prostej postaci i łatwym do interpretacji Po drugie, wyznaczenie optymalnej reguły decyzyjnej wymaga znajomości rozkładów prawdopodobieństwa, a w praktyce prawie nigdy nie mamy do czynienia z taką sytuacją.

Przedstawimy kilka typowych zagadnień, w których chodzi o przybliżenie nieobserwowanej zmiennej losowej Y przez funkcję $\delta_a(X)$, gdzie X jest obserwowaną zmienną losową i poszukiwana funkcja należy do parametrycznej rodziny $\{\delta_a : a \in \mathcal{A}\}$. Znajdujemy się w sytuacji przedstawionej w Podrozdziale 2.1, przy tym rolę zmiennej losowej Z odgrywa para (X, Y).

3.4.1 Przykład (Regresja liniowa, kwadratowa funkcja straty)**.** Niech $Z = (X, Y)$, gdzie X i Y są dwiema 1-wymiarowymi zmiennymi losowymi. Rozważamy problem predykcji Y na podstawie obserwowanej wartości X. Ograniczamy się do liniowych (afinicznych) predyktorów postaci $\hat{Y} = \beta_0 + \beta_1 X$. Decyzja ma postać $a = (\beta_0, \beta_1)$, a więc przestrzenią decyzji jest $\mathcal{A} = \mathbb{R}^2$. Przyjmijmy klasyczną kwadratową funkcję straty:

$$\ell(\beta_0, \beta_1, x, y) = (y - \beta_0 - \beta_1 x)^2.$$

Musimy założyć, że $\mathbb{E}X^2, \mathbb{E}Y^2 < \infty$. Najlepszy liniowy predyktor otrzymujemy dla $\beta_1^* = \mathrm{Cov}(X, Y)/\mathrm{Var}(X)$, $\beta_0^* = \mathbb{E}Y - \beta_1^* \mathbb{E}X$. △

3.4.2 Przykład (Wielowymiarowa regresja liniowa)**.** Rozważamy parę (X, Y), podobnie jak w przykładzie poprzednim. Załóżmy teraz, że $X = (X_1, \dots X_d)^\top$ jest wektorem losowym, a Y jest jednowymiarową zmienną losową. Liniowe predyktory Y na podstawie X możemy zapisać w postaci $\hat{Y} = \beta_0 + \sum_{j=1}^d \beta_j X_j = \beta_0 + X^\top \beta$, gdzie $\beta = (\beta_1, \dots, \beta_d)^\top$. Dla kwadratowej funkcji straty, $\ell(\beta_0, \beta, x, y) = (y - \beta_0 - x^\top \beta)^2$, współczynniki najlepszego liniowego predyktora są dane wzorami

$$\beta^* = \mathrm{VAR}(X)^{-1}\mathrm{COV}(X, Y), \qquad \beta_0^* = \mathbb{E}Y - \mathbb{E}X^\top \beta^*.$$

Musimy założyć, że $\mathbb{E}X_i^2 < \infty$ for $i = 1, \dots, d$, $\mathbb{E}Y^2 < \infty$ i że macierz wariancji-kowariancji $\mathrm{VAR}(X) = (\mathrm{Cov}(X_i, X_j), i, j = 1, \dots, d)$ jest dodatnio określona (nieosobliwa). △

3.4.3 Przykład (Regresja kwantylowa)**.** Zamiast kwadratowej funkcji straty, w modelu regresji możemy użyć straty zdefiniowanej w Przykładzie (2.1.4). Rozważamy parę (X, Y), gdzie X jest d-wymiarowym wektorem losowym a Y jest jednowymiarową zmienną losową. Mówimy o *regresji kwantylowej*, jeśli przyjmiemy funkcję straty

$$\ell(\beta_0, \beta, x, y) = \begin{cases} (\beta_0 + x^\top \beta - y)(1 - p) & \text{for } y \leqslant \beta_0 + x^\top \beta; \\ (y - \beta_0 - x^\top \beta)p & \text{for } y > \beta_0 + x^\top \beta. \end{cases}$$

Nazwę „regresji kwantylowej" można objaśnić w następujący sposób. Jeśli założymy, że $Y = \beta_0^* + X^\top \beta^* + W$, gdzie W jest zmienną niezależną od X, to $\beta_0^* + x^\top \beta^*$ jest p-tym kwantylem warunkowego rozkładu Y dla danego $X = x$. W szczególności, dla $p = 1/2$ otrzymujemy tak zwane estymatory „najmniejszych wartości bezwzględnych" (*Least Absolute Deviations*, LAD lub *Minimum Absolute Deviations*, MAD). W rzeczy samej, kwantylowa funkcja straty dla $p = 1/2$ jest równoważna funkcji

$$\ell(\beta_0, \beta, x, y) = |y - \beta_0 - x^\top \beta|.$$

W ogólnej sytuacji, nie ma jawnego wzoru na minimalizatory (β_0^*, β^*), ale istnieją efektywne metody numeryczne dla ich obliczania. △

3.4.4 Przykład (Liniowa Klasyfikacja)**.** Podobnie jak w Podrozdziale 3.3, rozważamy parę $Z = (X, C)$, gdzie $X = (X_1, \dots X_d)^\top$ jest losowym wektorem cech, C jest etykietą klasy. Załóżmy, że mamy tylko dwie klasy. Wygodnie przyjąć, że etykiety należą do zbioru $\{-1, 1\}$ (stosujemy tu inne kodowanie klas, niż w Podrozdziale 3.3). Problem kalsyfikacji jest analogiczny jak w modelach regresji: mamy przewidzieć wartość C, mając daną wartość X: na tym polega „zaklasyfikowanie" obiektu. Tak, jak w poprzednich przykładach dotyczących regresji, ograniczymy się do klasyfikatorów oparych na funkcjach liniowych $\beta_0 + x^\top \beta$. Załóżmy, że przewidywaną etykietą jest $\hat{C} = \operatorname{sign}(\beta_0 + X^\top \beta)$. Najbardziej naturalną funkcją straty jest indykator of błednej klasyfikacji:

$$\ell(\beta_0, \beta, x, c) = \mathbb{1}(\operatorname{sign}(\beta_0 + x^\top \beta) \neq c).$$

Dla tej $0-1$ funkcji straty, ryzyko jest prawdopodobieństwem błednej klasyfikacji, $\mathbb{P}(\hat{C} \neq C)$. Niestety, ta $0-1$ funkcja straty ma poważne wady. Ryzyko jest funkcją niewypukłą, może mieć wiele lokalnych minimów, w praktyce zadanie minimalizacji jest beznadziejnie trudne. Istnieje funkcja, która do pewnego stopnia „zastępuje" stratę $0-1$ ale jest bardziej przyjazna obliczeniowo bo jest wypukłą funkcją argumentu (β_0, β). Ta funkcja ma angielską nazwę „*hinge loss*" („zawiasowa?"). W moich pracach używałem terminu „perceptronowa funkcja kryterialna". Ta funkcja jest dana wzorem

$$\begin{aligned} \ell(\beta_0, \beta, x, c) &= \big(1 - c(\beta_0 + x^\top \beta)\big)_+ \\ &= \begin{cases} 1 - \beta_0 - x^\top \beta & \text{jeśli } c = 1 \text{ i } \beta_0 + x^\top \beta < 1; \\ 1 + \beta_0 + x^\top \beta & \text{jeśli } c = -1 \text{ i } \beta_0 + x^\top \beta > -1; \\ 0 & \text{w.p.p.} \end{cases} \end{aligned} \tag{3.4.5}$$

Ta funkcja jest wygodna obliczeniowo i ma przyjemne własności (Zadanie 10). △

Zadania

1. Rozważmy model Poisson/Gamma (Przykład 1.4.2).

   (a) Obliczyć estymator bayesowski dla funkcji straty $\ell(\theta, a) = (\theta - a)^2/\theta^2$ (względny błąd kwadratowy).

   (b) Porównać z estymatorem bayesowskim dla kwadratowej funkcji straty.

   Wskazówka: Wykorzystać wynik z Przykładiu 2.1.3.

2. Rozważmy model Normalny/Normalny (Przykład 1.4.4).

   (a) Oblicz estymator bayesowski parametru μ i predyktor bayesowski przyszłej obserwacji X_{n+1} dla funkcji straty LINEX (Przykład 2.1.5).

   (b) Porównaj z estymatorem i predyktorem dla kwadratowej funkcji straty.

3. Załóżmy, że zmienne losowe N_0 i N_1 są, dla danego θ, warunkowo niezależne i mają rozkłady Poissona: $N_0 \sim \text{Poiss}(t_0\theta)$ i $N_1 \sim \text{Poiss}(t_1\theta)$, gdzie t_0 i t_1 są znane i nielosowe. Parametr θ jest realizacją zmiennej losowej $\vartheta \sim \text{Gamma}(\alpha, \lambda)$. Niech

$$S = \sum_{i=1}^{N_1} X_i,$$

   gdzie $X_1, \ldots, X_i, \ldots$ są zmiennymi i.i.d. niezależnymi od N_0 i N_1 i ϑ. Rozkład prawdopodobieństwa zmiennych X_i jest znany i opisany funkcją tworzącą momenty $M(r) = \mathbb{E}e^{rX_i}$ ($\mathbb{E}X_i = \mu$). Obserwujemy $N_0 = n_0$ i mamy przewidzieć wartość S.

   (a) Oblicz bayesowski predyktor S dla kwadratowej funkcji straty.

   (b) Oblicz bayesowski predyktor S dla funkcji straty LINEX.

4. Niech $X \sim \text{Bin}(n, \theta)$. Znajdź minimaksowy estymator parametru θ dla kwadratowej funkcji straty.

   Wskazówka: Skorzystaj ze Stwierdzenia 2.3.6. Poszukaj odpowiedniego rozkładu *a priori* w rodzinie sprzężonej, wśród rozkładów beta.

5. Niech $X_1, \ldots, X_n \sim_{\text{i.i.d.}} \text{N}(\mu, \sigma^2)$. Zakładamy, że σ jest znana i rozważamy zadanie estymacji μ z kwadratową funkcją straty. Udowodnij, że średnia z próbki, $\hat{\mu} = \bar{X}$ jest estymatorem minimaksowym.

   Wskazówka: Zauważmy, że $\bar{X}$ nie jest estymatorem bayesowskim, nie można tu zastosować Stwierdzenia 2.3.6, ale można zastosować 2.3.7. Poszukaj odpowiedniego rozkładu *a priori* w rodzinie sprzężonej, wśród rozkładów normalnych.

6. (Regresja Logistyczna) Rozważmy model klasyfikacji z Przykładu 3.3.2 dla dwóch klas (dwa rozkłady normalne z jednakowymi macierzami kowariancji). Pokaż, że prawdopodobieństwo *a posteriori* (powiedzmy, klasy drugiej) jest logistyczną funkcją liniowej funkcji x:
$$p(2|x) = \frac{\exp(\alpha + \beta^\top x)}{1 + \exp(\alpha + \beta^\top x)}$$
dla pewnych $\alpha \in \mathbb{R}$ i $\beta \in \mathbb{R}^d$.

7. (Odległość Mahalanobisa) W modelu z Przykładu 3.3.2(rozkłady normalne z jednakowymi macierzami kowariancji), zdefiniujmy odległość Mahalanobisa pomiędzy punktami $\xi_1, \xi_2 \in \mathbb{R}^d$ wzorem
$$|\xi_1 - \xi_2|_{\Sigma^{-1}} = \sqrt{(\xi_1 - \xi_2)^\top \Sigma^{-1} (\xi_1 - \xi_2)}.$$
Pokaż, że dla równych prawdopodobieństw *a priori*, klasyfikator bayesowski przypisuje x do tej spośród klasy c, dla której odległość od „środka" $|x - \mu_c|_{\Sigma^{-1}}$ jest najmniejsza.

8. (Prawdopodobieństwo błędnej klasyfikacji) Rozważmy model klasyfikacji z Przykładu 3.3.2 dla dwóch klas (dwa rozkłady normalne z jednakowymi macierzami kowariancji). Napisz wzór na minimalne prawdopodobieństwo błędnej klasyfikacji w terminach odległości Mahalanobisa $|\mu_2 - \mu_1|_{\Sigma^{-1}}$ i dystrybuanty Φ standardowego rozkładu normalnego.

9. (Naiwny klasyfikator bayesowski dla cech binarnych) Załóżmy, że $\mathcal{X} = \{0,1\}^d$ (obserwujemy d binarnych cech) i
$$p(x|c) = \prod_{i=1}^{d} \theta_{ci}^{x_i} (1 - \theta_{ci})^{1 - x_i}.$$
Pokaż, że klasyfikator bayesowski jest liniowy (granice obszarów decyzyjnych są hiperpłaszczyznami). Nazwa „naiwny" odnosi się do uproszczonego założenia, że w każdej klasie cechy są niezależne. Niemniej, to „naiwne" założenie dramatycznie zmniejsza liczbę parametrów, które są w praktyce nieznane i muszą być oszacowane z próbki uczącej (mamy kd parametrów θ_{ci}).

10. (*Hinge loss*) Udowodnij następujące własności liniowego klasyfikatora minimalizującego funkcję straty zdefiniowaną wzorem (3.4.5) w Przykładzie 3.4.4.

    (a) Rozważmy afiniczną nieosobliwą transformację $\mathbb{R}^d \to \mathbb{R}^d$ wektora cech: $\tilde{X} = \mathrm{T} \cdot X + \alpha$, gdzie T jest $d \times d$ nieosobliwą macierzą i $\alpha \in \mathbb{R}^d$. Pokaż, że minimalizator ryzyka $r(\beta_0, \beta) = \mathbb{E}\ell(\beta_0, \beta, X, C)$ jest ekwiwariantny: $\tilde{\beta}_0^* + \tilde{X}^\top \tilde{\beta}^* = \beta_0^* + X^\top \beta^*$.

    (b) Pokaż, że wybranie liczby 1 jako „szerokości marginesu" we wzorze (3.4.5) jest nieistotne. Jeśli zmienimy definicję funkcji straty na $\ell(\beta_0, \beta, x, c) = \big(h - c(\beta_0 + x^\top \beta)\big)_+$ z dowolnie wybranym parametrem $h > 0$, to wyniki klasyfikacji nie zmienią się.

(c) Załóżmy, że dwie klasy są liniowo rozdzielne „z pewnym marginesem", to znaczy istnieją współczynniki $\beta_0^\dagger, \beta^\dagger$ takie, że

$$\mathbb{P}(\beta_0^\dagger + X^\top \beta^\dagger > 1 | C = 1) = 1 = \mathbb{P}(\beta_0^\dagger + X^\top \beta^\dagger < -1 | C = -1).$$

Niech β_0^*, β^* będą współczynnikami minimalizującymi wartość oczekiwaną *hinge loss* (funkcji strat zdefiniowanej w Przykładzie 3.4.4). Pokaż, że funkcja liniowa $\beta_0^* + X^\top \beta^*$ ma tę samą własność rozdzielania klas:

$$\mathbb{P}(\beta_0^* + X^\top \beta^* > 1 | C = 1) = 1 = \mathbb{P}(\beta_0^* + X^\top \beta^* < -1 | C = -1).$$

Wskazówka: Mamy $\mathbb{E}\ell(\beta_0^\dagger, \beta^\dagger, X, C) = 0$, a zatem $(\beta_0^\dagger, \beta^\dagger)$ minimalizuje *hinge loss*. W rozpatrywanej sytuacji nie jest to na ogół jedyny minimalizator; punkt (β_0^*, β^*) oznacza jakikolwiek inny minimalizator.

11. (Regresja) Uzasadnij wynik w Przykładzie 3.4.2.

(a) Wyprowadź podobny wzór na optymalny liniowy predyktor bez wyrazu wolnego, innymi słowy postaci $\hat{Y} = \sum_{j=1}^d \beta_j X_j = X^\top \beta$.

(b) Rozważ przypadek, kiedy wektor losowy X ma pierwszą współrzędną równą 1, to znaczy $X = (1, X_1, \ldots, X_d)$. Zastosujmy wynik otrzymany w poprzedniej części zadania do wyznaczenia współczynników $\beta_0^*, \beta_1^*, \ldots, \beta_d^*$. Pokaż, że dostaniemy to samo rozwiązanie co w Przykładzie 3.4.2 (regresja z wyrazem wolnym).

3.5 Testowanie hipotez statystycznych

Dwie hipotezy proste

Problem testowania hipotezy zerowej przeciw hipotezie alternatywnej przypomina na pierwszy rzut oka klasyfikację z dwiema klasami. W rzeczywistości, abstrakcyjne sformułowanie obu zadań jest prawie identyczne, ale kontekst i interpretacja jest zupełnie inna. Na początek rozważmy dwie proste hipotezy

$$H_0 : X \sim p_0,$$
$$H_1 : X \sim p_1.$$

Rozważamy zatem przestrzeń parametrów $\{0,1\}$ i przestrzeń decyzji $\{0,1\}$ (gdzie 1 jest interpretowane jako odrzucenie H_0 na rzecz H_1). Reguła decyzyjna $\delta : \{0,1\} \to \{0,1\}$ jest (niezrandomizowanym) testem. Rozkład *a priori* precyzuje prawdopodobieństwa $\pi(0) = \mathbb{P}(H_0)$ i $\pi(1) = \mathbb{P}(H_1)$. Ryzyko bayesowskie jest dane wzorem

$$r(\delta) = \alpha(\delta)\ell_0\pi(0) + \beta(\delta)\ell_1\pi(1),$$

gdzie $\alpha(\delta)$ i $\beta(\delta)$ są to prawdopodobieństwa dwóch typów błędu:

$$\alpha(\delta) = \mathbb{P}(\delta(X) = 1|H_0) = \int \mathbb{1}(\delta(x) = 1)p_0(x)\mathrm{d}x \quad \text{(błąd I rodzaju)},$$
$$\beta(\delta) = \mathbb{P}(\delta(X) = 0|H_1) = \int \mathbb{1}(\delta(x) = 0)p_1(x)\mathrm{d}x \quad \text{(błąd II rodzaju)},$$

zaś $\ell_0 = \ell(1,0) > 0$ i $\ell_1 = \ell(0,1) > 0$ są wagami przypisanymi obu typom błędu.

Test δ^* jest bayesowski (optymalny w sensie bayesowskim) jeśli jest postaci

$$\begin{cases} \delta^*(x) = 1 & \text{if} \quad \dfrac{p_1(x)}{p_0(x)} > \dfrac{\ell_0\pi(0)}{\ell_1\pi(1)}; \\ \\ \delta^*(x) = 0 & \text{if} \quad \dfrac{p_1(x)}{p_0(x)} < \dfrac{\ell_0\pi(0)}{\ell_1\pi(1)}. \end{cases}$$

Jeśli $p_1(x)/p_0(x) = (\ell_0\pi(0))/(\ell_1\pi(1))$ to decyzja $\delta^*(x)$ może być wybrana dowolnie. W tym sensie test bayesowski nie jest jednoznacznie wyznaczony.

Interesujące jest porównanie testu bayesowskiego ze stosowanym w statystyce częstościowej TNM (testem najmocniejszym) na danym poziomie istotności α_0. Lemat Neymana-Pearsona mówi, że TNM jest testem ilorazu wiarygodności, $p_1(x)/p_0(x)$. Test bayesowski δ^* jest zawsze

TNM na poziomie istotności $\alpha_0 = \alpha(\delta^*)$, o ile $\alpha(\delta^*) > 0$. Z drugiej strony nierandomizowany TNM niekoniecznie musi być bayesowski i może być całkowicie głupi. Przykład jest podany w Zadaniu 1. Środkiem zaradczym jest losowanie decyzji. Test zrandomizowany to funkcja $\delta_{\text{rand}} : \{0,1\} \to [0,1]$. Jeśli $\delta(x) = p$, to wykonujemy losowy eksperyment (próbę Bernoulliego z prawdopodobieństwem sukcesu p). „Sukces" jest interpretowany jako decyzja „1", czyli odrzucenie hipotezy zerowej. „Porażka" to decyzja „0", czyli brak podstaw do odrzucenia hipotezy zerowej. Prawdopodobieństwa błędu dla testu zrandomizowanego są następujące:

$$\alpha(\delta_{\text{rand}}) = \mathbb{E}(\delta_{\text{rand}}(X)|H_0) = \int \delta_{\text{rand}}(x)p_0(x)\mathrm{d}x \quad \text{(błąd I rodzaju)},$$
$$\beta(\delta_{\text{rand}}) = 1 - \mathbb{E}(\delta_{\text{rand}}(X)|H_1) = 1 - \int \delta_{\text{rand}}(x)p_1(x)\mathrm{d}x \quad \text{(błąd I rodzaju)}.$$

Zdefiniujmy „zbiór ryzyka" $\mathcal{R} \subseteq [0,1]^2$ jako

$$\mathcal{R} = \{(\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) : \delta_{\text{rand}} \text{ jest testem zrandomizowanym. }\}$$

$\mathcal{R}$ ma kilka przyjemnych własności.

- $\mathcal{R}$ jest wypukły i zwarty,
- $(0,1) \in \mathcal{R}$ i $(1,0) \in \mathcal{R}$,
- jeśli $(\alpha, \beta) \in \mathcal{R}$ to $(1-\alpha, 1-\beta) \in \mathcal{R}$.

Korzystając z tych faktów, łatwo jest wykazać, że każdy randomizowany test δ_{rand}, który jest TNM na jakimś poziomie istotności $\alpha_0 > 0$, jest testem bayesowskim dla pewnych stałych $\ell_0, \ell_1 > 0$, pod warunkiem, że $\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) > 0$. (Ponieważ jesteśmy bayesistami, traktujemy prawdopodobieństwa $\pi(0), \pi(1) > 0$ jako ustalone.) Zobacz Zadania 2 i 3. Zwróć uwagę na analogię ze Stwierdzeniami 2.3.4 i 2.3.5.

Zadania

1. Rozważ dwa rozkłady prawdopodobieństwa na przestrzeni $\mathcal{X} = \{1,2,3\}$:

x	1	2	3
$p_0(x)$	0.2	0.5	0.3
$p_1(x)$	0.1	0.3	0.6

   Rozważ problem testowania $H_0 : X \sim p_0$ vs $H_1 : X \sim p_1$.

   (a) Naszkicuj zbiór $\mathcal{R}_{\text{non-rand}} = \{(\alpha(\delta), \beta(\delta)) : \delta \text{ jest testem niezrandomizowanym}\}$. Znajdź niezrandomizowany TNM na poziomie istotności $\alpha_0 = 0.25$.

(b) Naszkicuj zbiór $\mathcal{R} = \{(\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) : \delta_{\text{rand}} \text{ jest testem zrandomizowanym}\}$. Znajdź zrandomizowany TNM na poziomie istotności $\alpha_0 = 0.25$.

2. Pokaż, że zbiór ryzyka $\mathcal{R}$ jest wypukły i zwarty.

   Uwaga: Sprawdzenie zwartości *nie* jest łatwe.

3. Pokaż, że każdy zrandomizowany TNM δ_{rand} na poziomie istotności $\alpha_0 > 0$ jest bayesowski dla pewnych $\ell_0, \ell_1 > 0$, o ile $\alpha(\delta_{\text{rand}}), \beta(\delta_{\text{rand}}) > 0$. Znajdź kontrprzykład pokazujący, że to stwierdzenie nie zachodzi, jeśli rozpatrujemy tylko testy niezrandomizowane.

4. Rozważ problem testowania $H_0 : X \sim \text{Ex}(1)$ vs $H_1 : X \sim \text{Ex}(1) + 1$, gdzie Ex(1) jest rozkładem wykładniczym z parameterem 1 a „Ex(1)+1" jest rozkładem wykładniczym przesuniętym na prawo $p_1(x) = \mathrm{e}^{-(x-1)}\mathbb{1}(x > 1)$. Naszkicuj zbiór ryzyka $\mathcal{R}$.

5. Rozważ problem testowania $H_0 : X \sim \text{N}(0, 1)$ vs $H_1 : X \sim \text{N}(\mu_1, 1)$, gdzie μ_1 jest ustalone. Naszkicuj zbiór ryzyka $\mathcal{R}$. Dla TNM na poziomie $\alpha > 0$, znajdź współczynniki $\ell_0, \ell_1 > 0$, dla których ten test jest bayesowski (porównaj z Zadaniem 3).

Hipotezy złożone i czynniki Bayesa

Przeformułujmy problem testowania hipotez statystycznych jako problem *wybór modelu*. Najpierw rozważmy dwa modele bayesowskie, $M = 0$ i $M = 1$, z dwiema funkcjami wiarygodności i dwoma rozkładami *a priori* (być może na dwóch różnych przestrzeniach parametrów). Jeśli jesteśmy konsekwentnymi bayesistami, powinniśmy również zdefiniować prawdopodobieństwa *a priori* obu modeli. Podsumowując, mamy następujące elementy:

$$\begin{array}{lll} p(x|\theta_0, M = 0), & \pi_0(\theta_0), & \pi(0) = \mathbb{P}(M = 0), \\ p(x|\theta_1, M = 1), & \pi_1(\theta_1), & \pi(1) = \mathbb{P}(M = 0). \end{array}$$

W większości ciekawych przykładów przestrzeń parametrów w modelu 0 jest podzbiorem przestrzeni w modelu 1; w takiej sytuacji symbol θ jest używany zamiast θ_0 i θ_1.) Po zaobserwowaniu $X = x$ mamy zdecydować, co jest prawdą:

$$H_0 : M = 0 \quad \text{czy} \quad H_1 : M = 1.$$

Wzór Bayesa pozwala obliczyć $\mathbb{P}(M = i|x)$ for $i = 0, 1$:

$$\begin{aligned} \mathbb{P}(M = 1|x) &= \frac{\pi(1)p(x|M = 1)}{\pi(1)p(x|M = 1) + \pi(0)p(x|M = 0)} \\ &= \frac{\pi(1)p(x|M = 1)/p(x|M = 0)}{\pi(1)p(x|M = 1)/p(x|M = 0) + \pi(0)}. \end{aligned}$$

Prawdopodobieństwa *a posteriori* hipotez (modeli) zależą od $\pi(i)$ i od ilorazu

$$B_{10} = \frac{p(x|M=1)}{p(x|M=0)},$$

który nazywamy czynnikiem Bayesa (*Bayes Factor*, BF). Jest on interpretowany jako „stopień, w jakim dane świadczą na korzyść hipotezy H_1 w porównaniu z H_0. Standardowe postępowanie polega na obliczeniu BF bez specifikowania $\pi(i) = \mathbb{P}(H_i)$. (Tradycyjna, heurystyczna reguła każe wybrać próg odrzuceń $B_{10} = 1$. W istocie, reguła „odrzuć H_0 na rzecz H_1 jeśli $B_{10} > 1$" jest regułą bayesowską dla $\pi(0) = \pi(1) = 0.5$ i dla symetrycznej funkcji straty $\ell(0,1) = \ell(1,0)$. Zazwyczaj jakikolwiek sensowny wybór zarówno $\pi(i)$ jak i ℓ byłby trudny do uzasadnienia. Umowny próg $B_{10} = 1$ gra w statystyce bayesowskiej podobną rolę, jak „standardowy poziom istotności $\alpha = 0.05$" w statystyce częstościowej.

Typowo, konkurencyjne modele mają różne „wymiary" i mamy zdecydować, czy wybrać „mniejszy" czy „większy " model (tj. model z mniejszą lub większą liczbą parametrów). Najprostszym przypadkiem jest testowanie prostej hipotezy H_0 przeciw złożonej alternatywie H_1 (mamy albo wybrać w pełni określony model albo model o nieznanych parametrach, czyli $\dim = 0$ vs $\dim > 0$).

Podejście oparte na czynnikach Bayesa pozwala uogólnić problem wyboru na przypadek więcej niż dwóch konkurujących modeli. Jeśli mamy k modeli ($M = 0, 1, \ldots, k-1$), to

$$\mathbb{P}(M = i|x) = \frac{\pi(i)B_{i0}}{\pi(0) + \sum_{l=1}^{k-1} \pi(l)B_{l0}}.$$

Czynnik Bayesa jest ilorazem *gęstości brzegowych* w konkurujących modelach:

$$p(x|M=i) = \int p(x|\theta_i, M=i)\pi_i(\theta_i)\mathrm{d}\theta_i.$$

Poza prostymi modelami sprzężonymi rozkłady brzegowe, a co za tym idzie czynniki Bayesa, są trudne do obliczenia. Poniższy przykład opisuje raczej rzadką sytuację, w której istnieje jawny wzór na BF.

3.5.1 Przykład. Niech $X_1, \ldots, X_n$ będzie próbką z rozkładu $\mathrm{N}(\theta, \sigma^2)$. Załóżmy, że σ^2 jest zanana, i rozważmy dwie konkurujące ze sobą hipotezy dotyczące θ. Hipoteza zerowa w pełni specyfikuje warość parameteru, $H_0 : \theta = \mu$, gdzie μ jest dane. Alternatywa ma postać $H_1 : \theta \sim \mathrm{N}(\mu, v^2)$. (Note that Zauważ, że to samo μ występuje w hipotezie zerowej i w alternatywie.) Żeby obliczyć czynnik Bayesa, użyjemy brzegowych gęstości statystyki dostatecznej $\bar{x} = \sum x_i/n$. Ponieważ

$$p(\bar{x}|H_0) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-\frac{1}{2\sigma^2/n}(\bar{x}-\mu)^2\right),$$
$$p(\bar{x}|H_1) = \frac{1}{\sqrt{2\pi\left(\sigma^2/n + v^2\right)}} \exp\left(-\frac{1}{2\left(\sigma^2/n + v^2\right)}(\bar{x}-\mu)^2\right),$$

więc

$$B_{10} = \frac{p(\bar{x}|H_1)}{p(\bar{x}|H_0)} = \sqrt{\frac{\sigma^2/n}{\sigma^2/n + v^2}} \exp\left(\frac{v^2}{2\left(\sigma^2/n + v^2\right)\sigma^2/n}(\bar{x} - \mu)^2\right).$$

Test bayesowski odrzuca H_0 na rzecz H_1 jeśli statystyka $z = \sqrt{n}|\bar{x} - \mu|/\sigma$ przekracza pewną wartość progową c, podobnie jak „częstościowy" test na zadanym poziomie istotności. Różnica polega na wyborze c dla obu testów. Ta różnica prowadzi do zjawiska znanego jako „paradoks Jeffreysa-Lindleya", patrz Zadanie 1. △

W powyższym przykładzie użyliśmy następującego stwierdzenia. Rozważmy problem testowania prostej hipotezy zerowej

3.5.2 Stwierdzenie. *Rozważmy problem testowania prostej hipotezy zerowej $H_0 : \theta = \theta_0$ przeciw „rozmytej" alternatywie $H_1 : \theta \sim \pi(\cdot)$, gdzie $\theta_0 \in \Theta$ jest pewną ustaloną wartością, π jest rozkładem a priori na Θ. Załóżmy, że $T = T(X)$ jest statystyką dostateczną w modelu odpowiadającym H_1 i $p(t|\theta)$ jest gęstością statystyki w punkcie $t = T(x)$. Wtedy*

$$B_{10} = \frac{\int_\Theta p(t|\theta)\pi(\theta)\mathrm{d}\theta}{p(t|\theta_0)}.$$

Dowód. Jeżeli $T(x) = t$ to $p(x|\theta) = p(x|t,\theta)p(t|\theta) = p(x|t)p(t|\theta)$, ponieważ X jest warunkowo niezależne od θ dla danego T. Stąd,

$$B_{10} = \frac{\int_\Theta p(x|\theta)\pi(\theta)\mathrm{d}\theta}{p(x|\theta_0)} = \frac{p(x|t)\int p(t|\theta)\pi(\theta)\mathrm{d}\theta}{p(x|t)p(t|\theta_0)}.$$

□

Zadania

1. (Paradoks Jeffreysa-Lindleya) Powróćmy do Przykładu 3.5.1. Niech $z_n^2 = n(\bar{x} - \mu)^2/\sigma^2$ będzie standardową statystyką używaną do weryfikacji $H_0 : \theta = \mu$ przeciw alternatywie $\theta \neq \mu$ w teorii częstościowej. Zauważmy, że czynnik Bayesa w Przykładzie 3.5.1 jest też rosnącą funkcją z_n^2. Porównajmy test częstościowy z odpowiednikiem bayesowskim opartym na B_{01}. Załóżmy, $z_n^2 = 2$ pozostaje stałe podczas gdy $n \to \infty$. Oczywiście, używana przez „częstościowca" p-wartość jest wciąż taka sama i prowadzi do odrzucenia H_0 (na typowym poziomie istotności $\alpha = 0.05$). Pokaż, że z drugiej strony, $B_{10} \to 0$ dla $n \to \infty$, a więc „bayesista" nie widzi żadnych powodów do odrzucenia H_0!

2. Rozważmy rodzinę rozkładów dwumianowych $\mathrm{Bin}(n, \theta)$ i problem testowania hipotezy $H_0 : \theta = \gamma$ przeciw $H_1 : \theta \sim \mathrm{Beta}(s\gamma, s(1-\gamma))$. Oblicz czynnik Bayesa B_{10} (nie oczekuj prostego jawnego wyrażenia).

   Uwaga: Zauważ, że dla H_1 mamy $\mathbb{E}\theta = \gamma$, a więc alternatywa jest rodzajem „rozmytej hipotezy zerowej", podobnie jak w Przykładzie 3.5.1.

3.6 Zbiory ufności

Bayesowski odpowiednik „estymacji przedziałowej" jest szczególnie prosty i łatwy w interpretacji. Bayesowski obszar ufności[2] na poziomie $1-\alpha$ jest to, z definicji, taki zbiór $C_x \subseteq \Theta$ że

$$\mathbb{P}(\vartheta \in C_x|x) = \int_{C_x} \pi_x(\theta)\mathrm{d}\theta \geqslant 1-\alpha.$$

Jak zwykle w teorii bayesowskiej, nietrudno zdefiniować pewną własność optymalności.

3.6.1 Lemat. *Niech $\pi(\cdot)$ będzie gęstością prawdopodobieństwa na Θ. Niech $h > 0$ i $C^h = \{\theta : \pi(\theta) \geqslant h\}$. Jeśli dla pewnego $C \subseteq \Theta$ mamy*

$$\int_{C^h} \pi(\theta)\mathrm{d}\theta \leqslant \int_C \pi(\theta)\mathrm{d}\theta$$

to

$$\int_{C^h} \mathrm{d}\theta \leqslant \int_C \mathrm{d}\theta.$$

Dowód. Z założenia

$$\int_{C^h\setminus C} \pi(\theta)\mathrm{d}\theta \leqslant \int_{C\setminus C^h} \pi(\theta)\mathrm{d}\theta.$$

Na zbiorze $C^h \setminus C$ mamy $\pi \geqslant h$, podczas gdy na $C \setminus C^h$ zachodzi nierówność przeciwna, $\pi < h$. Zatem

$$\int_{C^h} h\mathrm{d}\theta \leqslant \int_C h\mathrm{d}\theta.$$

Wystarczy podzielić obie strony nierówności przez h i dodać $\int_{C^h\cap C} \mathrm{d}\theta$ do obu stron. Całka $\int_C \mathrm{d}\theta$ jest po prostu „objętością" zbioru C. □

Oczywiście, możemy ten lemat zastosować nie tylko do gęstości *a priori*, ale również do gęstości *a posteriori*. W ten sposób otrzymujemy następujący fakt:

3.6.2 Stwierdzenie (Obszar największej gęstości *a posteriori*)**.** *Jeśli dla pewnego $h > 0$, zbiór $C_x^h = \{\theta : \pi_x(\theta) \geqslant h\}$ jest taki, że $\mathbb{P}(\vartheta \in C_x^h|x) = 1-\alpha$ i dla pewnego innego zbioru C_x mamy $\mathbb{P}(\vartheta \in C_x|x) \geqslant 1-\alpha$ to objętość C_x^h jest niewiększa, niż C_x.*

Innymi słowy, C_x^h jest „najmniejszym" obszarem ufności (w sensie najmniejszej objętości względem miary $\mathrm{d}\theta$; zazwyczaj mówimy tu o d-wymiarowej mierze Lebesgue'a, jeśli $\Theta \subseteq \mathbb{R}^d$).

[2] W statystyce bayesowskiej obowiązuje angielski termin *credible set/region*, podczas gdy w teorii częstościowej mamy *confidence set/region*. Po polsku używamy tego samego określenia „obszar ufności" w obu przypadkach

Rozdział 4

Metody obliczeniowe Monte Carlo (MCMC)

Algorytmy MCMC (Markowowskie Monte Carlo) zrewolucjonizowały statystykę bayesowską. Stworzyły możliwość obliczania (w przybliżeniu) rozkładów *a posteriori* w sytuacji, gdy dokładne, analityczne wyrażenia są niedostępne. W ten sposób statystycy uwolnili się od konieczności używania nadmiernie uproszczonych modeli. Zaczęli śmiało budować modele coraz bardziej realistyczne, zwykle o strukturze hierarchicznej.

4.1 Algorytmy MCMC

.

Algorytmy Monte Carlo wykorzystują komputerową symulację zmiennych losowych (pseudolosowych) do badania rozkładów prawdopodobieństwa i obliczania w przybliżeniu pewnych wielkości. W statystyce bayesowskiej interesuje nas rozkład prawdopodobieństwa *a posteriori* π_y[1] na przestrzeni Θ, który może być trudny do obliczenia. Naszym celem może być na przykład obliczenie wartości oczekiwanej, $\hat{\theta} = \int_\Theta \theta \pi_y(\theta) \mathrm{d}\theta$ (jest to typowy estymator bayesowski, dla kwadratowej funkcji straty). W wielu modelach całka jest trudna lub niemożliwa do obliczenia. Jeśli umiemy generować niezależne zmienne losowe $\vartheta(0), \vartheta(1), \ldots, \vartheta(m), \ldots \sim \pi_y$, to Prawo Wielkich Liczb gwarantuje, że prawie na pewno, dla $m \to \infty$,

$$\bar{\theta}_m = \frac{1}{m} \sum_{t=1}^{m} \vartheta(t) \to \hat{\theta}. \tag{4.1.1}$$

[1] W tym rozdziale obserwowaną przez statystyka zmienną losową i jej wartość oznaczamy symbolem $Y = y$ (zamiast $X = x$, jak poprzednio).

Możemy użyć „średniej empirycznej" $\bar{\theta}_m$ jako przybliżenia $\hat{\theta}$, używając możliwie dużej liczby m (rozmiaru próbki Monte Carlo). W podobny sposób możemy przybliżać medianę i inne wielkości dotyczące rozkładu *a posteriori*.

Skrót MCMC oznacza *Markov Chain Monte Carlo*, czyli po polsku algorytmy MC wykorzystujące łańcuchy Markowa. Podstawowa idea jest taka: jeśli nie umiemy generować zmiennych losowych o rozkładzie π_y to zadowolimy się generowaniem łańcucha Markowa, czyli ciągu zmiennych losowych $\vartheta(0), \vartheta(1), \ldots, \vartheta(m), \ldots$, który w pewnym sensie zbliża się, zmierza do rozkładu π_y. Przy pewnych założeniach dla takiego ciągu zachodzi wzór (4.1.1) pomimo tego, że składniki sumy nie są niezależne i nie mają dokładnie rozkładu π_y.

Warto podkreślić, że w tym rozdziale mówimy *wyłącznie* o losowości generowanej komputerowo, w celach obliczeniowych. Wartości obserwowanych przez statystyka zmiennych, x, są ustalone. W dalszym ciągu będziemy pomijać indeks y pisząc $\pi = \pi_y$ (ale pamiętajmy, że chodzi o rozkład *a posteriori*; rozkład *a priori* zazwyczaj jest „łatwy" i nie wymaga metod Monte Carlo).

Podstawowym algorytmem MCMC w statystyce bayesowskiej jest próbnik Gibbsa (PG) (*Gibbs Sampler*, GS). Załóżmy, że przestrzeń na której żyje docelowy rozkład π ma strukturę produktową: $\Theta = \prod_{i=1}^{d} \Theta_i$. Rozpatrujemy łańcuch Markowa $\vartheta(0), \vartheta(1), \ldots, \vartheta(m), \ldots$ na przestrzeni Θ, $\vartheta(m)$ jest wektorem $(\vartheta_1(m), \ldots, \vartheta_d(m))$. Ponadto przyjmijmy następujące oznaczenia:

- Jeśli $\Theta \ni (\theta_i)_{i=1}^{d}$ to $\theta_{-i} = (\theta_j)_{j \neq i}$: wektor z pominiętą i-tą współrzędną.
- Rozkład docelowy (gęstość): $\pi(\theta)$.
- Pełne rozkłady warunkowe (*full conditionals*):

$$\pi(\theta_i|\theta_{-i}) = \frac{\pi(\theta)}{\pi(\theta_{-i})}, \text{ gdzie } \pi(\theta_{-i}) = \int_{\Theta_i} \pi(\theta_1, \ldots, \theta_i, \ldots, \theta_d) \mathrm{d}\theta_i.$$

Próbnik Gibbsa generuje łańcuch Markowa zgodnie z następującą regułą przejścia. Współrzędne są zmieniane w porządku cyklicznym. Zmiana i-tej współrzędnej polega na wylosowaniu nowej wartości z pełnego rozkładu warunkowego. Jeśli aktualnym stanem łańcucha (w kroku t) jest $\vartheta = \vartheta(t)$, to następny stan $\vartheta' = \vartheta(t+1)$ jest generowany tak:

`Gen` $\vartheta'_1 \sim \pi(\cdot|\vartheta_2, \ldots, \vartheta_d)$;
`Gen` $\vartheta'_2 \sim \pi(\cdot|\vartheta'_1, \vartheta_3, \ldots, \vartheta_d)$;
...
`Gen` $\vartheta'_d \sim \pi(\cdot|\vartheta'_1, \ldots, \vartheta'_{d-1})$;
`Return` ϑ'

4.2 Przykłady

Przedstawię na dwóch dość typowych przykładach konstrukcję próbnika Gibbsa aproksymującego rozkład *a posteriori* w złożonych modelach bayesowskich.

Hierarchiczny model klasyfikacji

4.2.1 Przykład (Statystyka małych obszarów)**.** Zacznijmy od opisania problemu tak zwanych „małych obszarów", który jest dość ważny w dziedzinie badań reprezentacyjnych, czyli w tak zwanej „statystyce oficjalnej". *Małe obszary* to pod-populacje w których rozmiar próbki nie jest wystarczający, aby zastosować „zwykłe" estymatory (średnie z próbki). Podejście bayesowskie pozwala „pożyczać informację" z innych obszarów. Zakłada się, że z każdym małym obszarem związany jest nieznany parametr, który staramy się estymować. Obserwacje pochodzące z określonego obszaru mają rozkład prawdopodobieństwa zależny od odpowiadającego temu obszarowi parameru. Parametry, zgodnie z filozofią bayesowską, traktuje się jak zmienne losowe. W najprostszej wersji taki model jest zbudowany w sposób opisany poniżej. △

Model bayesowski

- $y_{ij} = Y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$ – badana cecha dla j-tej wylosowanej jednostki i-tego obszaru, $(j = 1, \ldots, n_i)$, $(i = 1, \ldots, k)$,
- $\theta_i \sim \mathrm{N}(\mu, \upsilon^2)$ – interesująca nas średnia w i-tym obszarze,
- μ – średnia w całej populacji.

Ciekawe, że ten sam model pojawia się w różnych innych zastosowanianich, na przykład w matematyce ubezpieczeniowej. Przytoczymy klasyczny rezultat dotyczący tego modelu, aby wyjaśnić na czym polega wspomniane „pożyczanie informacji".

Estymator bayesowski

W modelu przedstawionym powyżej, łatwo obliczyć estymator bayesowski (przy kwadratowej funkcji straty), czyli wartość oczekiwaną *a posteriori*.

$$\hat{\theta}_i = \mathbb{E}(\theta_i|y) = z_i\bar{y}_i + (1 - z_i)\mu, \qquad \bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \qquad z_i = \frac{n_i\upsilon^2}{n_i\upsilon^2 + \sigma^2}.$$

Ten wzór jest bardzo dobrze znany specjalistom od małych obszarów i aktuariuszom. W istocie, wynika on natychmiast z Przykładu 1.4.4 (na obecnie rozważanym poziomie rozbudowy modelu, predyktor $\hat{\theta}_i$ zależy tylko od zmiennych $y_{i1}, \ldots, y_{in_i}$; pozostałe zmienne są warunkowo niezależne).

Estymator bayesowski dla i-tego obszaru jest średnią ważoną $\bar{y}_i$ (estymatora opartego na danych z tego obszaru) i wielkości μ, która opisuje całą populację, a nie tylko i-ty obszar. Niestety, estymator $\hat{\theta}_i$ zależy od parametrów μ, σ i v, które w praktyce są nieznane i które trzeba estymować. Konsekwentnie bayesowskie podejście polega na traktowaniu również tych parametrów jako zmiennych losowych, czyli nałożeniu na nie rozkładów *a priori*. Powstaje w ten sposób model hierarchiczny.

Hierarchiczny model bayesowski

Uzupełnijmy rozpatrywany powyżej model, dobudowując „wyższe piętra" hierarchii. Potraktujemy mianowicie parametry rozkładów *a priori*: μ, σ i v jako zmienne losowe i wyspecyfikujemy ich rozkłady *a priori*.

- $y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$,
- $\theta_i \sim \mathrm{N}(\mu, \upsilon^2)$,
- $\mu \sim \mathrm{N}(m, \tau^2)$,
- $\sigma^{-2} \sim \mathrm{Gamma}(p, \lambda)$,
- $\upsilon^{-2} \sim \mathrm{Gamma}(q, \kappa)$.

Zakładamy przy tym, że μ, σ i v są *a priori* niezależne (niestety, są one zależne *a posteriori*). Na szczycie hierarchii mamy „hiperparametry" m, τ, p, λ, q, κ, o których musimy założyć, że są znanymi liczbami.

Łączny rozkład prawdopodobieństwa wszystkich zmiennych losowych w modelu ma postać

$$p(y, \theta, \mu, \sigma^{-2}, \upsilon^{-2}) = p(y|\theta, \sigma^{-2})\pi(\theta|\mu, \upsilon^{-2})\pi(\mu)\pi(\sigma^{-2})\pi(\upsilon^{-2}).$$

We wzorze powyżej i w dalej traktujemy (trochę nieformalnie) σ^{-2} i υ^{-2} jako pojedyncze symbole nowych zmiennych, żeby nie mnożyć oznaczeń. Rozkład prawdopodobieństwa *a posteriori* jest więc taki:

$$\pi_y(\theta, \mu, \sigma^{-2}, \upsilon^{-2}) = \frac{p(y, \theta, \mu, \sigma^{-2}, \upsilon^{-2})}{p(y)}.$$

To jest rozkład „docelowy” π, na przestrzeni $\Theta = \mathbb{R}^{k+3}$, z nieznaną stałą normującą $1/p(y)$. Choć wygląda na papierze dość prosto, ale obliczenie rozkładów brzegowych, wartości oczekiwanych i innych charakterystyk jest, łagodnie mówiąc, trudne.

Opiszemy teraz jak jest skonstruowany **próbnik Gibbsa w modelu hierarchicznym**. Rozkłady warunkowe poszczególnych współrzędnych są proste i łatwe do generowania. Można te rozkłady „odczytać” uważnie patrząc na rozkład łączny:

$$\begin{aligned}\pi_y(\theta, \mu, \upsilon^{-2}, \sigma^{-2}) \propto (\sigma^{-2})^{n/2} \exp\left\{-\frac{\sigma^{-2}}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2\right\} \\ \cdot (\upsilon^{-2})^{k/2} \exp\left\{-\frac{\upsilon^{-2}}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2\right\} \\ \cdot \exp\left\{-\frac{\tau^{-2}}{2}(\mu-m)^2\right\} \\ \cdot (\sigma^{-2})^{p-1}\exp\{-\lambda\sigma^{-2}\} \\ \cdot (\upsilon^{-2})^{q-1}\exp\{-\kappa\upsilon^{-2}\}.\end{aligned}$$

Dla ustalenia uwagi zajmijmy się rozkładem warunkowym zmiennej υ^{-2}. Kolorem niebieskim oznaczyliśmy te czynniki łącznej gęstości, które zawierają υ^{-2}. Pozostałe, czarne czynniki traktujemy jako stałe. Stąd widać, jak wygląda rozkład warunkowy υ^{-2}„ przynajmniej z dokładnością do proporcjonalności:

$$\begin{aligned}\pi_y(\upsilon^{-2}|y,\theta,\mu,\sigma^{-2}) \propto (\upsilon^{-2})^{k/2+q-1} \\ \cdot \exp\left\{-\left(\frac{1}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2+\kappa\right)\upsilon^{-2}\right\}.\end{aligned}$$

Jest to zatem rozkład Gamma$(k/2+q, \sum_{i=1}^{k}(\theta_i-\mu)^2/2+\kappa)$. Zupełnie podobnie rozpoznajemy inne (pełne) rozkłady warunkowe:

$$\begin{aligned}\upsilon^{-2}|y,\theta,\mu,\sigma^{-2} &\sim \text{Gamma}\left(\frac{k}{2}+q, \frac{1}{2}\sum_{i=1}^{k}(\theta_i-\mu)^2+\kappa\right), \\ \sigma^{-2}|y,\theta,\mu,\upsilon^{-2} &\sim \text{Gamma}\left(\frac{n}{2}+p, \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\theta_i)^2+\lambda\right), \\ \mu|y,\theta,\sigma^{-2},\upsilon^{-2} &\sim \text{N}\left(\frac{k\tau^2}{k\tau^2+\upsilon^2}\bar{\theta}+\frac{\upsilon^2}{k\tau^2+\upsilon^2}m, \frac{\tau^2\upsilon^2}{k\tau^2+\upsilon^2}\right), \\ \theta_i|y,\theta_{-i},\mu,\sigma^{-2},\upsilon^{-2} &\sim \text{N}\left(\frac{n\upsilon^2}{n\upsilon^2+\sigma^2}\bar{y}_i+\frac{\sigma^2}{n\upsilon^2+\sigma^2}\mu, \frac{\upsilon^2\sigma^2}{n\upsilon^2+\sigma^2}\right),\end{aligned}$$

gdzie, rzecz jasna, $n = \sum n_i$, $\bar{\theta} = \sum_i \theta_i/k$ i $\theta_{-i} = (\theta_k)_{k\neq i}$. Zwróćmy uwagę, że współrzędne wektora θ są warunkowo niezależne (pełny rozkład warunkowy θ_i nie zależy od θ_{-i}). Dzięki

temu możemy w próbniku Gibbsa potraktować θ jako cały „blok" współrzędnych i zmieniać „na raz".

Próbnik Gibbsa ma w tym modelu przestrzeń stanów Θ składającą się z punktów $\vartheta = (\theta, \mu, \sigma^{-2}, \upsilon^{-2}) \in \mathbb{R}^{k+3}$. Reguła przejścia próbnika,

$$\underbrace{\left(\theta, \mu, \sigma^{-2}, \upsilon^{-2}\right)}_{\vartheta(t)} \longmapsto \underbrace{\left(\theta, \mu, \sigma^{-2}, \upsilon^{-2}\right)}_{\vartheta(t+1)},$$

jest złożona z następujących „małych kroków":

- Wylosuj $\upsilon^{-2} \sim \pi_y(\upsilon^{-2}|\theta, \mu, \sigma^{-2} \quad) = \text{Gamma}(\ldots)$,
- Wylosuj $\sigma^{-2} \sim \pi_y(\sigma^{-2}|\theta, \mu, \quad \upsilon^{-2}) = \text{Gamma}(\ldots)$,
- Wylosuj $\mu \sim \pi_y(\mu \quad |\theta, \quad \sigma^{-2}, \upsilon^{-2}) = \text{N}(\ldots)$,
- Wylosuj $\theta \sim \pi_y(\theta \quad | \quad \mu, \sigma^{-2}, \upsilon^{-2}) = \text{N}(\ldots)$.

Łańcuch Markowa jest zbieżny do rozkładu *a posteriori*:

$$\vartheta(t) \to \pi_y, \quad (t \to \infty).$$

Najbardziej interesujące są w tym hierarchicznym modelu zmienne θ_i (pozostałe zmienne można uznać za „parametry zakłócające). Dla ustalenia uwagi zajmijmy się zmienną θ_1 (powiedzmy, wartością średnią w pierwszym małym obszarze). *Estymator bayesowski* jest to wartość oczekiwana *a posteriori* tej zmiennej:

$$\mathbb{E}(\theta_1|y) = \int \cdots \int \theta_1 \pi_y(\theta, \mu, \sigma^{-2}, \upsilon^{-2}) \mathrm{d}\theta_2 \cdots \mathrm{d}\theta_k \mathrm{d}\mu \mathrm{d}\sigma^{-2} \mathrm{d}\upsilon^{-2}.$$

Aproksymacją MCMC interesującej nas wielkości są średnie wzdłuż trajektorii łańcucha:

$$\theta_1(0), \theta_1(1), \ldots, \theta_1(t), \ldots,$$
$$\text{gdzie } \vartheta(t) = (\theta_1(t), \ldots, \theta_k(t), \mu(t), \sigma^{-2}(t), \upsilon^{-2}(t)).$$

Na Rysunku 4.1 pokazane są dwie przykładowe trajektorie współrzędnej υ^{-2} dla PG poruszającego się po przestrzeni $k + 3 = 1003$ wymiarowej (model uwzględniający 1000 małych obszarów). Dwie trajektorie odpowiadają dwu różnym punktom startowym. Dla innych zmiennych rysunki wyglądają bardzo podobnie. Uderzające jest to, jak szybko trajektoria zdaje się „osiągać" rozkład stacjonarny, przynajmniej wizualnie. Na Rysunku 4.2 pokazane są kolejne „skumulowane" średnie dla tych samych dwóch trajektorii zmiennej υ^{-2}.

Model mieszanek normalnych

Tak, jak w modelu komponentów wariancyjnych, rozważamy obserwacje podzielone na grupy. Różnica jest taka, że podział jest „ukryty". Dane nie zawierają informacji o tym, które jednostki pochodzą z tej samej grupy, a które z różnych grup. Zakładamy, że mamy próbkę losową z mieszanki rozkładów prawdopodobieństwa $\sum_{j=1}^{k} q_j P_j(\cdot)$, gdzie $\sum_{j=1}^{k} q_j = 1$. Każdy z rozkładów $P_j(\cdot)$ zależy od nieznanych parametrów. Prawdopodobieństwa q_j też są nieznane. W modelu, który rozpatrzymy zakłada się, że liczba komponentów k jest znana. W dalszym ciągu rozpatrujemy mieszanki rozkładów normalnych i następującą hierarchię rozkładów *a priori*:

- $Y_1, \ldots, Y_n \sim_{\text{i.i.d.}} \sum_{j=1}^{k} q_j \mathrm{N}(\mu_j, \sigma_j^2)$,
- $(q_1, \ldots, q_k) \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$,
- $\mu_1, \ldots, \mu_k \sim_{\text{i.i.d.}} \mathrm{N}(m, v^2)$,
- $\sigma_1^{-2}, \ldots, \sigma_k^{-2} \sim_{\text{i.i.d.}} \mathrm{Gamma}(\gamma, \lambda)$.

Hiperparametry $m, v^2, \alpha, \gamma, \lambda$ są ustalone i znane. Rozkładem docelowym jest rozkład *a posteriori* $\pi_y(q, \mu, \sigma^{-2})$, gdzie $\sigma^{-2} = (\sigma_1^{-2}, \ldots, \sigma_k^{-2})$.

Przedstawię poniżej próbnik Gibbsa (PG) wykorzystujący ideę *zmiennych pomocniczych*. W naszym modelu, te zmienne pomocnicze, $c_1, \ldots, c_n$, po prostu wskazją, do którego komponentu mieszanki należą poszczególne obserwacje. Innymi słowy,

- $\mathbb{P}(c_i = j|q) = q_j$ (*a priori*),
- $Y_i|c_i = j \sim \mathrm{N}(\mu_j, \sigma_j^2)$ niezależnie od reszty zmiennych.

Ogólnie mówiąc, zmienne pomocnicze ułatwiają konstrukcję algorytmów MCMC. W modelu mieszanek, są same w sobie interesujące.

Łączna gęstość a posteriori w naszym modelu jest następująca:

$$\pi_y(q, c, \mu, \sigma^{-2}) \propto \prod_{i=1}^{n} q_{c_i} (\sigma_{c_i}^{-2})^{1/2} \exp\left(-\frac{\sigma_{c_i}^{-2}}{2}(y_i - \mu_{c_i})^2\right) \cdot \prod_{j=1}^{k} q_j^{\alpha_j - 1} \cdot \prod_{j=1}^{k} \exp\left(-\frac{1}{2v^2}(\mu_j - m)^2\right) \cdot \prod_{j=1}^{k} (\sigma_j^{-2})^{\gamma-1} \exp\left(-\lambda \sigma_j^{-2}\right).$$

Z tego wzoru łatwo wydobyć postać pełnych rozkładów warunkowych (*full conditionals*).

- Rozkład warunkowy c. Niezależnie dla $i = 1, \ldots, n$ mamy

$$\mathbb{P}(c_i = j | q, \mu, \sigma^{-2}, y) \propto q_j (\sigma_j^{-2})^{1/2} \exp\left(-\frac{\sigma_j^{-2}}{2} (y_i - \mu_j)^2 \right).$$

  Jest to łatwy do symulowania rozkład dyskretny na zbiorze $\{1, \ldots, k\}$.

- Rozkład warunkowy q. Grupując wyrażenia zawierające q_j, dostajemy

$$\pi_y(q | c, \mu, \sigma^{-2}) \propto \prod_{j=1}^{k} q_j^{\alpha_j + n_j - 1},$$

  gdzie $n_j = \sum_{i=1}^{n} \mathbb{1}(c_i = j)$. Rozkładem warunkowym jest więc $\mathrm{Dir}(\alpha_k + n_k, \ldots, \alpha_k + n_k)$.

- Rozkład warunkowy μ. Przepiszmy tę część wzoru na łączną gęstość *a posteriori*, która zawiera μ w następujący sposób:

$$\pi_y(\mu | q, c, \sigma^{-2}) \propto \prod_{j=1}^{k} \exp\left(-\frac{\sigma_j^{-2}}{2} \sum_{i: c_i = j} (y_i - \mu_j)^2 \right) \exp\left(-\frac{1}{2v^2} (\mu_j - m)^2 \right).$$

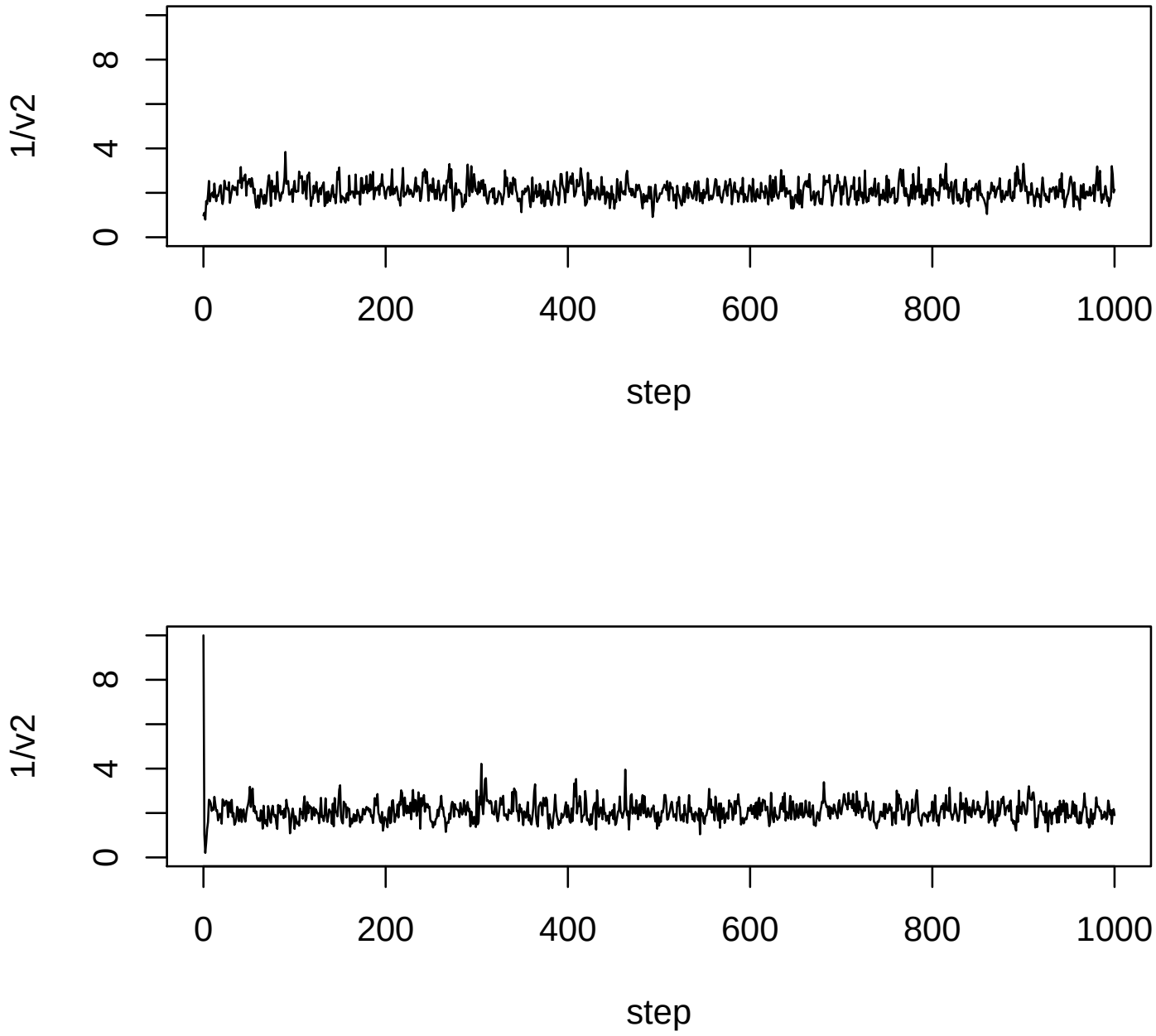
  Niezależnie dla $j = 1, \ldots, k$, obliczamy rozkład μ_j sprowadzając funkcje kwadratowe „pod znakiem exp" do postaci kanonicznej, otrzymując:

$$\mu_j \sim \mathrm{N}\left(z_j \bar{y_j} + (1 - z_j) m, \frac{v^2}{n_j \sigma_j^{-2} + 1} \right),$$
$$z_j = \frac{n_j \sigma_j^{-2} v^2}{n_j \sigma_j^{-2} v^2 + 1}, \quad \bar{y_j} = \frac{1}{n_j} \sum_{i: c_i = j} y_i.$$

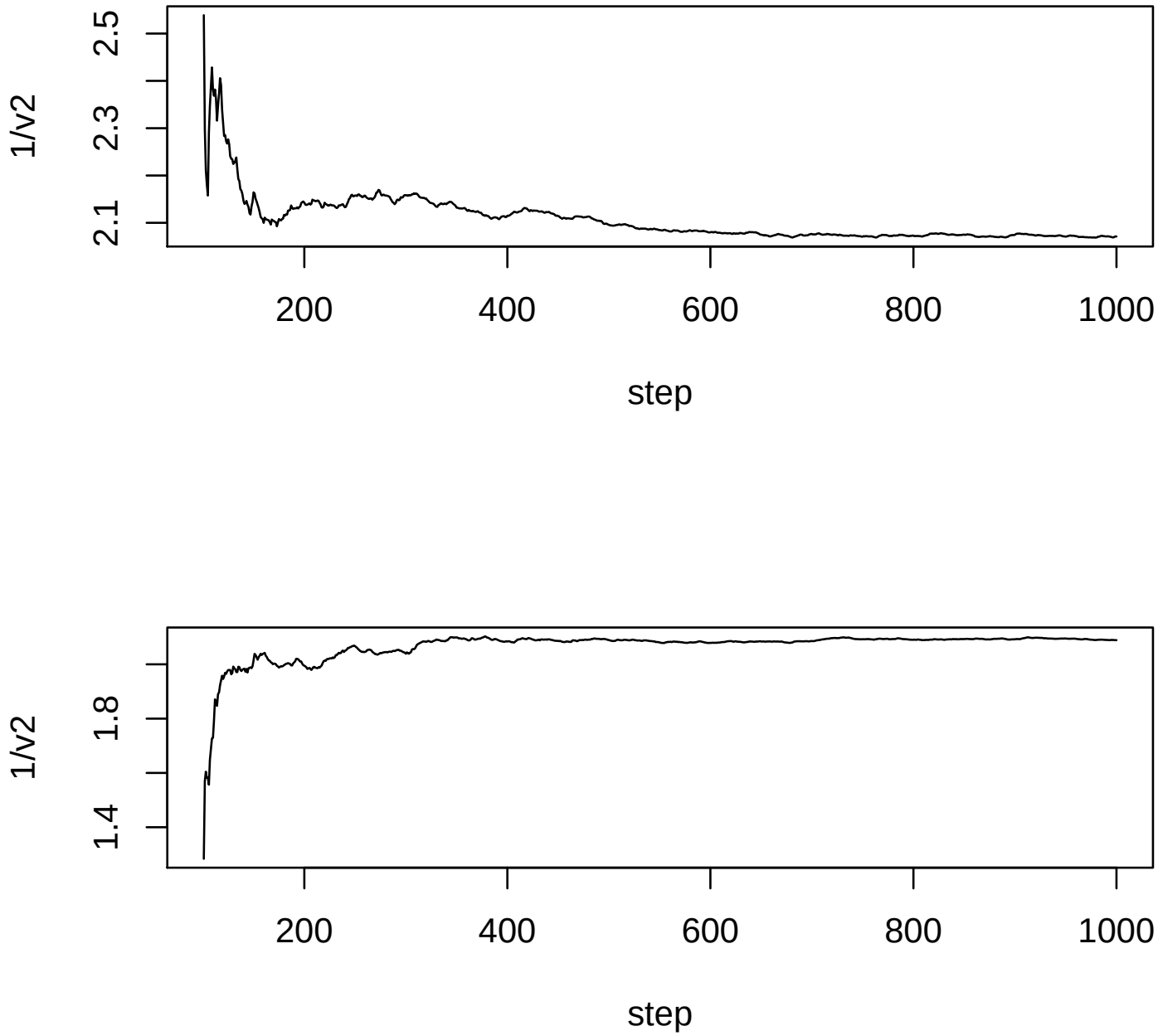
  Zwróćmy uwagę na to, że w każdym kroku PG używamy bieżących wartości zmiennych c_i, czyli aktualizujemy przynależność obserwacji do grup.

- Rozkład warunkowy σ^{-2}. Podobnie jak w poprzednim punkcie, przepisujemy część wzoru na łączną gęstość *a posteriori*, która zawiera σ^{-2} i rozpoznajemy, że pełny rozkład warunkowy jest niezależny dla każdej współrzędnej i równy

$$\sigma_j^{-2} \sim \mathrm{Gamma}\left(\gamma + \frac{n_j}{2}, \lambda + \frac{1}{2} \sum_{i: c_i = j} (y_i - \mu_j)^2 \right).$$



Rysunek 4.1: Trajektorie zmiennej v^{-2} dla dwóch punktów startowych. Hierarchiczny model komponentów wariancyjnych.



Rysunek 4.2: Skumulowane średnie zmiennej v^{-2} dla dwóch punktów startowych. Hierarchiczny model komponentów wariancyjnych.

Rozdział 5

Asymptotyka

5.1 Koncentracja rozkładów *a posteriori*

Zaczniemy od abstrakcyjnego sformułowania, które ma charkter całkowicie deterministyczny.

Rozważamy gęstość prawdopodobieństwa π na przestrzeni Θ i funkcje $f_n : \Theta \to \mathbb{R}$ oraz $f : \Theta \to \mathbb{R}$. Zdefiniujemy ciąg funkcji

$$\pi_n(\theta) = \frac{1}{z_n} \exp\{-n f_n(\theta)\}\pi(\theta), \tag{5.1.1}$$

gdzie

$$z_n = \int_\Theta \exp\{-n f_n(\theta)\}\pi(\theta)\mathrm{d}\theta. \tag{5.1.2}$$

Dla uproszczenia załóżmy, że $\Theta \subseteq \mathbb{R}^d$. Niech θ_0 będzie ustalonym punktem, który należy do wnętrza zbioru Θ.

5.1.3 Lemat. *Załóżmy, że*

1. *Funkcje π i f są ciągłe w punkcie θ_0. Ponadto, $\pi(\theta_0) > 0$.*
2. *Dla punktów θ w pewnym otoczeniu θ_0 mamy zbieżność $f_n(\theta) \to f(\theta)$ $(n \to \infty)$.*
3. *Dla każdego $\varepsilon > 0$, $\liminf_{n\to\infty} \inf_{\{\theta:|\theta-\theta_0|>\varepsilon\}} f_n(\theta) - f(\theta_0) > 0$.*

Wtedy, dla każdego $\varepsilon > 0$,

$$\int_{\{\theta:|\theta-\theta_0|\leqslant\varepsilon\}} \pi_n(\theta)\mathrm{d}\theta \to 1.$$

Zanim przejdziemy do dowodu, skomentujmy założenia i sformułowanie lematu. Można sobie wyobrażać, że π jest gęstością *a priori*. Jeśli $-nf_n$ jest logarytmem wiarygodności[1], to π_n jest gęstością *a posteriori*. Na razie możemy zignorować zależność wiarygodności od obserwacji; to będzie stanowiło kolejny etap naszych rozważań.

Dowód Lematu 5.1.3. Chcemy pokazać, że $\int_{\{\theta:|\theta-\theta_0|>\varepsilon\}} \pi_n(\theta)\mathrm{d}\theta \to 0$. Napiszmy tę całkę w postaci ilorazu

$$\frac{\int_{\{\theta:|\theta-\theta_0|>\varepsilon\}} \exp\left\{-n[f_n(\theta)-f(\theta_0)-\beta]\right\}\pi(\theta)\mathrm{d}\theta}{\int_\Theta \exp\left\{-n[f_n(\theta)-f(\theta_0)-\beta]\right\}\pi(\theta)\mathrm{d}\theta} = \frac{\text{Licznik}}{\text{Mianownik}}.$$

Zajmijmy się najpierw Licznikiem. Korzystając z założenia 3. możemy zakładać, że dla pewnego $\beta > 0$ nierówność $f_n(\theta) - f(\theta_0) > \beta$ zachodzi dla dostatecznie dużych n. Funkcja podcałkowa w Liczniku jest $\leqslant 1$, zatem

$$\text{Licznik} \leqslant \int_{\{\theta:|\theta-\theta_0|>\varepsilon\}} \pi(\theta)\mathrm{d}\theta \leqslant \int_\Theta \pi(\theta)\mathrm{d}\theta = 1.$$

Teraz rozpatrzymy Mianownik. Założenie 1. gwarantuje istnienie liczby $\delta > 0$ takiej, że $\pi(\theta) > 0$ i $f(\theta) - f(\theta_0) \leqslant \beta/2$ dla $|\theta - \theta_0| < \delta$. Stąd wynika, że $\lim_{n\to\infty}[f_n(\theta) - f(\theta_0) - \beta] = [f(\theta) - f(\theta_0) - \beta] \leqslant -\beta/2$, a zatem $\lim \gamma_n(\theta) = +\infty$ dla $|\theta - \theta_0| < \delta$, gdzie $\gamma_n(\theta) = \exp\{-n[f_n(\theta) - f(\theta_0) - \beta]\}$. Na mocy lematu Fatou,

$$\begin{aligned}\liminf \int_\Theta \gamma_n(\theta)\pi(\theta)\mathrm{d}\theta &\geqslant \int_\Theta \liminf \gamma_n(\theta)\pi(\theta)\mathrm{d}\theta \\ &\geqslant \int_{\{\theta:|\theta-\theta_0|<\delta\}} \lim \gamma_n(\theta)\pi(\theta)\mathrm{d}\theta = +\infty.\end{aligned}$$

Znaczy to, że Mianownik dąży to nieskończoności, co kończy dowód. □

Przejdziemy teraz do modelu probabilistycznego. Załóżmy, że $X = X_1, \ldots, X_n, \ldots$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie prawdopodobieństwa p_* na przestrzeni $\mathcal{X}$. Rozpatrzmy funkcję $\rho : \Theta \times \mathcal{X} \to \mathbb{R}$ i połóżmy

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^n \rho(\theta, X_i). \tag{5.1.4}$$

Niech, ponadto

$$f(\theta) = \mathbb{E}_{p_*}\rho(\theta, X). \tag{5.1.5}$$

[1]Znak „minus” wynika stąd, że wolę rozważać minimalizację zamiast maksymalizacji. Czynnik n wprowadziliśmy, aby założenie 2 było spełnione w ważnych przykładach.

Pamiętajmy, że f_n jest funkcją losową, chociaż w zapisie pomijamy argumenty $X_1, \ldots, X_n, \ldots$. W dalszym ciągu będziemy rozważali zachowanie ciągu funkcji f_n dla ustalonego (dla prawie każdego) ciągu wartości $x_1, \ldots, x_n, \ldots$.

Ograniczymy się dla uproszczenia do rozpatrzenia wypukłych funkcji ρ[2]. Założymy, że $\Theta \subseteq \mathbb{R}^d$ jest zbiorem wypukłym.

5.1.6 Twierdzenie. *Załóżmy, że*

1. *Funkcja $\rho(\cdot, x) : \Theta \to \mathbb{R}$ jest wypukła dla każdego $x \in \mathcal{X}$.*
2. *Wartość oczekiwana w definicji $f(\theta)$ jest dobrze określona dla każdego $\theta \in \Theta$.*
3. *Punkt θ_0 jest jedynym minimum funkcji f i należy do wnętrza zbioru Θ.*
4. *Funkcja π jest gęstością prawdopodobieństwa na Θ, jest ciągła w punkcie θ_0 i $\pi(\theta_0) > 0$.*

Jeśli ciąg funkcji π_n zdefiniujemy wzorami (5.1.1) i (5.1.2) dla funkcji f_n danych przez (5.1.4), to prawie na pewno, dla każdego $\varepsilon > 0$,

$$\int_{\{\theta : |\theta - \theta_0| \leqslant \varepsilon\}} \pi_n(\theta) \mathrm{d}\theta \to 1.$$

Dowód. Na mocy Prawa Wielkich Liczb, dla każdego ustalonego θ, z prawdopodobieństwem 1 mamy $f_n(\theta) \to f(\theta)$. Stąd wynika, że z prawdopodobieństwem 1 zbieżność zachodzi dla każdego $\theta \in \Theta_1$, gdzie Θ_1 jest pewnym przeliczalnym, gęstym podzbiorem Θ. Od tego momentu traktujemy wylosowany ciąg $X_1 = x_1, \ldots, X_n = x_n, \ldots$ jak ustalony (zakładając, że dla tego ciągu nie zachodzi „wyjątkowe zdarzenie" o prawdopodobieństwie 0).

Uzasadnimy, że spełnione są założenia Lematu 5.1.3. Założenie o wypukłości bardzo ułatwia argumentację. Ponieważ funkcje f_n są wypukłe, granica f jest też wypukła, a więc ciągła. Spełnione są zatem założenia 1. i 2. Lematu 5.1.3. Pozostaje sprawdzić założenie 3. Skorzystamy z naszego założenia 3. Ze zbieżności funkcji wypukłych $f_n(\theta) \to f(\theta)$ dla gęstego podzbioru wynika zbieżność jednostajna na zbiorach zwartych. Jeśli $\inf_{\{\theta : |\theta - \theta_0| = \varepsilon\}} f(\theta) > f(\theta_0) + 2\beta$ to dla dostatecznie dużych n mamy $\inf_{\{\theta : |\theta - \theta_0| = \varepsilon\}} f_n(\theta) > f(\theta_0) + \beta$. Korzystając jeszcze raz z wypukłości, wnioskujemy, że

$$\inf_{\{\theta : |\theta - \theta_0| \geqslant \varepsilon\}} f_n(\theta) > f(\theta_0) + \beta,$$

co kończy dowód. □

[2]Założenie o wypukłości, może być zastąpione innego typu (dość technicznymi) założeniami, przy których teza naszego Twierdzenia 5.1.6 pozostaje prawdziwa.

Ostatecznie możemy sformułować zasadniczy wniosek z Twierdzenia 5.1.6, dotyczący modelu bayesowskiego. Rozważamy rodzinę gęstości $\{p_\theta : \theta \in \Theta\}$ na przestrzeni $\mathcal{X}$ i gęstość *a priori* π. Jeśli przymiemy

$$\rho(\theta, x) = -\log p_\theta(x), \tag{5.1.7}$$

to ρ jest wiarygodnością, zaś funkcja π_n staje się gęstością *a posteriori*, $\pi_n = \pi_{x_1,\dots,x_n}$.

5.1.8 Wniosek. *Jeśli $X_1, \dots, X_n, \dots \sim_{\text{i.i.d.}} p_*$, funkcja ρ jest dana wzorem (5.1.7) i spełnione są założenia Twierdzenia 5.1.6, to z prawdopodobieństwem 1 mamy zbieżność*

$$\pi_{X_1,\dots,X_n} \to \delta_{\theta_0}, \qquad (n \to \infty),$$

w sensie słabej zbieżności miar probabilistycznych. Punkt θ_0 minimalizuje dywergencję Kullbacka-Leiblera: $\mathcal{D}(p_*\|p_{\theta_0}) = \inf_{\theta\in\Theta} \mathcal{D}(p_*\|p_\theta)$, gdzie

$$\mathcal{D}(p_*\|p_\theta) = \int p_*(x) \log \frac{p_*(x)}{p_\theta(x)} \mathrm{d}x.$$

Podkreślmy, że w tym wniosku *nie zakładamy*, że „prawdziwy rozkład prawdopodobieństwa" o gęstości p_* należy do parametrycznej rodziny $\{p_\theta : \theta \in \Theta\}$. W tym sensie, model statystyczny jest, być może, źle wyspecyfikowany. Gęstość p_{θ_0} jest nazywana rzutem Kullbacka-Leiblera gęstości p_* na rodzinę $\{p_\theta : \theta \in \Theta\}$.

W sytuacji dobrze wyspecyfikowanego modelu otrzymujemy następujący klasyczny wynik.

5.1.9 Wniosek. *Jeśli $X_1, \dots, X_n, \dots \sim_{\text{i.i.d.}} p_{\theta_0}$ dla pewnego $\theta_0 \in \Theta$, funkcja ρ jest dana wzorem (5.1.7) i spełnione są założenia Twierdzenia 5.1.6, to z prawdopodobieństwem 1 mamy zbieżność*

$$\pi_{X_1,\dots,X_n} \to \delta_{\theta_0}, \qquad (n \to \infty),$$

w sensie słabej zbieżności miar probabilistycznych.

5.2 Asymptotyczna normalność

Będziemy podążać podobną ścieżką jak w poprzednim podrozdziale: najpierw udowodnimy wynik sformułowany w deterministyczny sposób, a później pokażemy jego probabilistyczne konsekwencje. Podobnie jak poprzednio, założenie wypukłości będzie odgrywało kluczową rolę. Dla uproszczenia załóżmy, że $\Theta = \mathbb{R}^d$.

5.2.1 Twierdzenie. *Załóżmy, że funkcje $f_n : \Theta \to \mathbb{R}$ są wypukłe i $\theta_n \to \theta_0$ dla pewnego ciągu $\theta_n \in \Theta$ i $\theta_0 \in \Theta$. Ponadto, dla każdego m,*

$$\sup_{|t|\leqslant m} \left| n\left[f_n\left(\theta_n + \frac{t}{\sqrt{n}}\right) - f_n(\theta_n)\right] - \frac{1}{2}t^\top D t \right| \to 0, \tag{5.2.2}$$

gdzie D jest pewną macierzą dodatnio określoną. Niech π będzie gęstością prawdopodobieństwa na przestrzeni Θ, ciągłą i niezerową w punkcie θ_0.

Jeżeli π_n jest zdefiowane przez (5.1.1), to

$$\int_\Theta |\pi_n(\theta) - \gamma_n(\theta)| \mathrm{d}\theta \to 0, \tag{5.2.3}$$

gdzie $\gamma_n(\theta)$ jest gęstością rozkładu normalnego $\mathrm{N}(\theta_n, D^{-1}/n)$.

Dowód. Zdefiniujmy funkcje g_n wzorem

$$\begin{aligned} g_n(t) &= \exp\left\{-n\left[f_n\left(\theta_n + \frac{t}{\sqrt{n}}\right) - f_n(\theta_n)\right]\right\} \pi\left(\theta_n + \frac{t}{\sqrt{n}}\right) \\ &= \pi_n\left(\theta_n + \frac{t}{\sqrt{n}}\right) z_n \mathrm{e}^{n f_n(\theta_n)}. \end{aligned}$$

Łatwo widać, że $g_n(t) \to g_0(t)$, gdzie

$$g_0(t) = \exp\left\{-\frac{1}{2}t^\top D t\right\} \pi(\theta_0).$$

Dowód będzie zakończony, jeśli pokażemy, że

$$\int |g_n(t) - g_0(t)| \mathrm{d}t \to 0. \tag{5.2.4}$$

W rzeczy samej, (5.2.4) implikuje $\int g_n \to \int g_0$. Stąd wynika, że

$$z_n \sim \mathrm{e}^{-n f_n(\theta_n)} c^{-1} n^{-d/2} \pi(\theta_0), \tag{5.2.5}$$

gdzie c jest stałą normującą rozkładu $\mathrm{N}(\theta_n, D^{-1})$[3], gdyż $\int \pi_n(\theta_n + t/\sqrt{n})\mathrm{d}t = n^{d/2}$ i $\int g_0(t)\mathrm{d}t = c^{-1}\pi(\theta_0)$. (Przypomnijmy, że z_n jest zdefiowane wzorem (5.1.2). Symbol $z_n \sim a_n$ znaczy, że $z_n/a_n \to 1$.)

Ponieważ $\pi_n(\theta_n + t/\sqrt{n}) = z_n^{-1}\mathrm{e}^{-n f_n(\theta_n)} g_n(t)$, zaś $\gamma_n(\theta_n + t/\sqrt{n}) = c n^{d/2} \pi(\theta_0)^{-1} g_0(t)$, więc korzystając ze wzoru (5.2.5) możemy przepisać (5.2.4) w następującej, równoważnej postaci:

$$\int \left| \pi_n\left(\theta_n + \frac{t}{\sqrt{n}}\right) - \gamma_n\left(\theta_n + \frac{t}{\sqrt{n}}\right)\right| \frac{\mathrm{d}t}{n^{d/2}} \to 0.$$

[3]Wiadomo, że $c = (2\pi)^{-d/2}(\det D)^{1/2}$, gdzie $\pi = 3.1415\ldots$ oznacza inny obiekt, niż gęstość π.

W ten sposób otrzymujemy tezę (5.2.3)[4].

Dowód będzie zakończony, jeśli uzasadnimy wzór (5.2.4). Napiszmy

$$\int |g_n - g_0| = \int_{|t|\leqslant m} |g_n - g_0| + \int_{|t|\geqslant m} |g_n - g_0|.$$

Pierwszą całkę oszacujemy korzystając z założenia (5.2.2), a drugą korzystając z wypukłości funkcji f_n. Zacznijmy od drugiej całki. Niech $\lambda = \inf_{|t|=1} t^\top Dt$ będzie najmniejszą wartością własną macierzy D. Oczywiście $m^2\lambda = \inf_{|t|=m} t^\top Dt$. Ze wzoru (5.2.2) wynika, że dla dostatecznie dużych n mamy $n[f_n(\theta_n + t/\sqrt{n}) - f_n(\theta_n)] > m^2\lambda/4 = (m\lambda/4)|t|$ dla $|t| = m$. Wypukłość funkcji f_n pozwala wnioskować, że ta ostatnia nierówność zachodzi również dla $|t| \geqslant m$. Wobec tego $g_n(t) \leqslant e_m(t) := \exp\{-(m\lambda)|t|/4\}(\pi(\theta_0) + 1)$ dla dostatecznie dużych n. Ponieważ

$$\int_{\mathbb{R}^d} e_m(t)\mathrm{d}t \to 0, \qquad (m \to \infty),$$

więc możemy znaleźć takie m, że

$$\int_{|t|\geqslant m} g_n(t)\mathrm{d}t \leqslant \int_{|t|\geqslant m} e_m(t)\mathrm{d}t \leqslant \int_{\mathbb{R}^d} e_m(t)\mathrm{d}t < \frac{\varepsilon}{3}.$$

Łatwo widać, że możemy również zapewnić nierówność

$$\int_{|t|\geqslant m} g_0(t)\mathrm{d}t < \frac{\varepsilon}{3}.$$

Wreszcie, z założenia (5.2.2) wynika, że (dla wybranego powyżej m) dla dostatecznie dużych n prawdziwa jest nierówność

$$\int_{|t|\leqslant m} |g_n(t) - g_0(t)|\mathrm{d}t < \frac{\varepsilon}{3}.$$

Stąd $\int |g_n - g_0| < \varepsilon$ dla dostatecznie dużych n, a więc udowodniliśmy (5.2.4). □

Twierdzenie 5.2.1 można zastosować do modelu probabilistycznego opisanego w poprzednim podrozdziale. Przypomnijmy, $X_1, \ldots, X_n, \ldots \sim_{\text{i.i.d.}} p_*$ jest ciągiem niezależnych zmiennych losowych o jednakowym rozkładzie. f_n i f są zdefiniowane wzorami (5.1.4) i (5.1.5), odpowiednio. Dla wypukłej funkcji ρ, jej subgradient oznaczamy $\partial\rho$.

[4] Jeśli zmienna losowa ϑ_n ma rozkład o gęstości $\pi_n(\theta)$ to zmienna losowa $T_n = \sqrt{n}(\vartheta - \theta_n)$ ma gęstość $\pi_n(\theta_n + t/\sqrt{n})n^{-d/2}$.

5.2.6 Twierdzenie. *Załóżmy, że*

1. *Funkcja $\rho(\cdot, x) : \Theta \to \mathbb{R}$ jest wypukła dla każdego $x \in \mathcal{X}$.*
2. *Funkcja f jest dwukrotnie różniczkowalna w punkcie θ_* i macierz $D = \nabla\nabla^\top f(\theta_*)$ jest (ściśle) dodatnio określona.*
3. *Dla punktów θ w pewnym otoczeniu θ_* mamy $\mathbb{E}\,|\partial\rho(\theta, X)|^2 < \infty$.*
4. *Funkcja π jest gęstością prawdopodobieństwa na Θ, jest ciągła w punkcie θ_0 i $\pi(\theta_0) > 0$.*

Ciąg funkcji π_n zdefiniujemy wzorami (5.1.1) i (5.1.2) dla funkcji f_n danych przez (5.1.4). Zakładamy, że θ_n minimalizuje funkcję f_n: $f_n(\theta_n) = \inf_\theta f_n(\theta)$. Niech $\gamma_n(\theta)$ oznacza gęstość rozkładu normalnego $\mathrm{N}(\theta_n, D^{-1}/n)$. Jeśli π_n jest zdefiniowane wzorem (5.1.1) wtedy mamy zbieżność według prawdopodobieństwa (względem rozkładu prawdopodobieństwa ciągu zmiennych $X_1, \ldots, X_n, \ldots$):

$$\int_\Theta |\pi_n(\theta) - \gamma_n(\theta)|\mathrm{d}\theta \to_p 0. \tag{5.2.7}$$

Pominiemy dowód Twierdzenia 5.2.6. Najważniejsza część dowodu polega na sprawdzeniu, że relacja we wzorze (5.2.2)(niemal jednostajna zbieżność do funkcji kwadratowej) zachodzi według prawdopodobieństwa.

Punkty θ_n w Twierdzeniu 5.2.6 noszą nazwę M-estymatorów. Tak więc, nieformalnie podsumowując nasze rozważania możemy powiedzieć, że dla dużych próbek rozkład *a posteriori* zbliża się do rozkładu normalnego, którego średnią jest M-estymator.

Rozważmy na koniec funkcję ρ będącą minus-log-wiarygodnością, to znaczy daną wzorem (5.1.7). W sytuacji dobrze wyspecyfikowanego modelu, to znaczy gdy $p_* \in \{p_\theta : \theta \in \Theta\}$, mamy szczególnie prostą interpretację macierzy kowariancji granicznego rozkładu normalnego. Jeśli we wzorze określającym macierz D (założenie 2. w Twierdzeniu 5.2.6) możemy zamienić kolejność różniczkowania i całkowania, czyli

$$\nabla\nabla^\top \mathbb{E}\rho(\theta, X) = \mathbb{E}\nabla\nabla^\top \rho(\theta, X), \tag{5.2.8}$$

to $-D$ jest macierzą informacyjną. Następujący wniosek jest pewnym wariantem klasycznego twierdzenia Bernsteina-von Misesa[5].

5.2.9 Wniosek. *Załóżmy, że spełnione są założenia Twierdzenia 5.2.6, mamy $p_* = p_{\theta_0}$ dla pewnego $\theta_0 \in \Theta$ i zachodzi wzór (5.2.8). Wtedy we wzorze (5.2.7), γ_n jest gęstością rozkładu $\mathrm{N}(\theta_n, I_n(\theta_0)^{-1})$, gdzie θ_n jest estymatorem największej wiarygodności, zaś $I_n(\theta_0) = nI_1(\theta_0)$ jest macierzą informacyjną Fishera.*

[5]Stale zakładamy, że ρ jest wypukłą funkcją argumentu θ, ale w innych wersjach twierdzenia Bernsteina-von Misesa ten warunek nie jest wymagany.

Zadania

1. Rozważmy model Normalny/Normalny (Przykład 1.4.4). Załóżmy, że $X_1, \ldots, X_n, \ldots$ są i.i.d. próbką z rozkładu $\mathrm{N}(\theta_0, \sigma^2)$ i $n \to \infty$. Pokaż bezpośrednio (nie wykorzystując Wniosku 5.1.8), że prawie na pewno zachodzi $\pi_{x_1,\ldots,x_n} \to \delta_{\theta_0}$ w sensie słabej zbieżności miar probabilistycznych.

   Wskazówka: Użyj PWL i oblicz wartość oczekiwaną i wariancję rozkładu *a posteriori*.

2. Rozważmy model Poisson/Gamma (Przykład 1.4.2). Załóżmy, że $X_1, \ldots, X_n, \ldots$ są i.i.d. próbką z rozkładu $\mathrm{Poiss}(\theta_0)$ i $n \to \infty$. Pokaż bezpośrednio, że $\pi_{x_1,\ldots,x_n} \to \delta_{\theta_0}$ (jak w zadaniu poprzednim).

3. Rozważmy model Bin/Beta (Przykład 1.4.1). Załóżmy, że $X_1, \ldots, X_n, \ldots$ są i.i.d. próbką z rozkładu $\mathrm{Ber}(\theta_0)$ i $n \to \infty$. Pokaż bezpośrednio, że $\pi_{x_1,\ldots,x_n} \to \delta_{\theta_0}$ (jak w zadaniu poprzednim).

4. Niech zmienna losowa ϑ_n ma rozkład o gęstości $\mathrm{Beta}(n+1, n+1)$. Rozważmy zmienną losową $T_n = \sqrt{n}(\vartheta_n - 1/2)$ i oznaczmy przez $p_n(t)$ gęstość rozkładu zmiennej T_n. Pokaż bezpośrednio, że $\lim_{n\to\infty} \pi_n(t)$ jest gęstością rozkładu normalnego. Zidentyfikuj parametry tego rozkładu i porównaj z momentami $\mathbb{E}T_n$ i $\mathrm{Var}T_n$.

   Wskazówka: Można użyć wzoru Stirlinga i rozwinięcia funkcji $\log(1+x)$.

Dodatek A

Rozkłady warunkowe

Warunkowe rozkłady prawdopodobieństwa i warunkowe wartości oczekiwane są definiowane na wykładach z rachunku prawdopodobieństwa w ogólny i abstrakcyjny sposób. W naszym elementarnym wykładzie ograniczymy się do sytuacji, opisanej poniżej.

Załóżmy, że zmienne losowe X i Y mają łączną gęstość $p(x,y)$ względem pewnej miary produktowej (σ-skończonej), oznaczanej $\mathrm{d}x\mathrm{d}y$, na przestrzeni $\mathcal{X}\times\mathcal{Y}$. Przy tym każdy z symboli $\mathrm{d}x$ i $\mathrm{d}y$ może oznaczać albo miarę Lebesgue'a na $\mathbb{R}^d$ albo miarę liczącą na przestrzeni dyskretnej.

- Gęstość warunkowa jest dana wzorem

$$p(y|x)=\frac{p(x,y)}{p(x)},$$

  gdzie $p(x)=\int_{\mathcal{Y}}p(x,y)\mathrm{d}y$ jest gęstością brzegową. Jeśli $A\subset\mathcal{Y}$ to możemy zdefiniować $\mathbb{P}(Y\in A|X=x)=\int_A p(y|x)\mathrm{d}y$. W ten sposób określamy *warunkowy rozkład prawdopodobieństwa* zmiennej Y przy danej wartości $X=x$.

- Ogólniej, niech $h:\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ będzie ograniczoną (lub nieujemną) funkcją. Możemy zdefiniować *warunkową wartość oczekiwaną* jako funkcję $\bar{h}:\mathcal{X}\to\mathbb{R}$ daną wzorem

$$\bar{h}(x)=\mathbb{E}(h(X,Y)|X=x)=\int_{\mathcal{Y}}h(x,y)p(y|x)\mathrm{d}y.$$

  Warunkową wartość oczekiwaną $\mathbb{E}(h(X,Y)|X)$ definiujemy jako zmienną losową $\bar{h}(X)$ (czyli złożenie funkcji $\bar{h}:\mathcal{X}\to\mathbb{R}$ i $X:\Omega\to\mathcal{X}$).

Zadania

1. Uzasadnij zgodność podanych wzorów ze znanymi Ci definicjami $\mathbb{P}(Y \in A|X)$ oraz $\mathbb{E}(h(X,Y)|X)$. Uzupełnij pominięte założenia dotyczące mierzalności. Udowodnij (lub przypomnij sobie) wzór $\mathbb{E}\mathbb{E}(Y|X) = \mathbb{E}Y$.

2. Udowodnij wzór $\mathrm{Var}\mathbb{E}(Y|X) + \mathbb{E}\mathrm{Var}(Y|X) = \mathrm{Var}Y$. Wariancja warunkowa jest zdefiniowana wzorem
$$\mathrm{Var}(Y|X) = \mathbb{E}[(Y - \mathbb{E}(Y|X))^2|X].$$

3. Niech łączna gęstość zmiennych losowych X i Y będzie dana wzorem
$$p(x,y) = \mathrm{e}^{-y} \text{ dla } 0 \leqslant x \leqslant y,$$
zero w pozostałych przypadkach.
   (a) Podaj rozkład brzegowy zmiennej X.
   (b) Podaj rozkład brzegowy zmiennej Y.
   (c) Oblicz rozkład warunkowy zmiennej X dla $Y = y$.
   (d) Oblicz rozkład warunkowy zmiennej Y dla $X = x$.
   (e) Oblicz $\mathbb{E}(X|Y)$. Jaki rozkład prawdopodobieństwa ma ta zmienna losowa?
   (f) Oblicz $\mathbb{E}(Y|X)$. Jaki rozkład prawdopodobieństwa ma ta zmienna losowa?

4. Załóżmy, że zmienne losowe U i V są niezależne i mają jednakowy, jednostajny rozkład prawdopodobieństwa $\mathrm{U}(0,1)$. Niech $X = \min(U,V)$ i $Y = \max(U,V)$.
   (a) Podaj łączną gęstość X i Y.
   (b) Oblicz gęstość brzegową X.
   (c) Podaj rozkład warunkowy zmiennej losowej Y przy danym $X = x$.
   (d) Oblicz $\mathbb{E}(Y|X)$ i $\mathrm{Var}(Y|X)$.
   (e) Podaj rozkład warunkowy zmiennych losowych (U,V) przy danym $(X,Y) = (x,y)$.

5. Załóżmy, że zmienne losowe $U_0, U_1, \ldots, U_n, \ldots$ są niezależne i mają jednakowy, jednostajny rozkład prawdopodobieństwa $\mathrm{U}(0,1)$. Niech $N = \min\{n \geqslant 1 : U_n > U_0\}$.
   (a) Podaj rozkład prawdopodobieństwa zmiennej losowej N.
   Wskazówka: $\mathbb{P}(N > n) = \mathbb{P}(\max(U_0, U_1, \ldots, U_n) = U_0)$.
   (b) Oblicz $\mathbb{E}N$.
   (c) Podaj rozkład warunkowy zmiennej losowej N przy danym $U_0 = u$.
   (d) Oblicz $\mathbb{E}(N|U_0)$ i $\mathbb{E}N$.

(e) Podaj rozkład warunkowy zmiennej losowej U_N przy danym $U_0 = u$.

(f) Oblicz rozkład brzegowy zmiennej losowej U_N.

6. Załóżmy, że zmienne losowe $X_1, \ldots, X_n, \ldots$ oraz N są niezależne, przy tym X_i mają jednakowy rozkład prawdopodobieństwa, znamy $\mathbb{E}X_i = \mu$, $\mathrm{Var}X_i = \sigma^2$ i $\mathbb{E}\mathrm{e}^{tX_i} = M(t)$. Jeśli zmienna losowa N przyjmuje wartości w zbiorze $\{0, 1, 2, \ldots\}$, to możemy zdefiniować $S = S_N = \sum_{i=1}^{N} X_i$. Zakładamy znajomość rozkładu prawdopodobieństwa zmiennej N.

(a) Obliczyć $\mathbb{E}S$, $\mathrm{Var}S$ i $\mathbb{E}\mathrm{e}^{tS}$.

(b) Podaj przykład sytuacji, w której $\mathbb{E}\mathrm{Var}(S|N) = 0$.

(c) Podaj przykład sytuacji, w której $\mathrm{Var}\mathbb{E}(S|N) = 0$.

7. Niech X i Y będą niezależnymi zmiennymi losowymi o rozkładzie dwumianowym, $X \sim \mathrm{Bin}(n, \theta)$, $Y \sim \mathrm{Bin}(m, \theta)$ i $S = X + Y$.

(a) Znajdź rozkład warunkowy zmiennej X dla $S = s$.

(b) Podaj $\mathbb{E}(X|S)$, $\mathrm{Var}(X|S)$.

(c) Oblicz $\mathbb{E}(X|S)$, $\mathrm{Var}\mathbb{E}(X|S)$, $\mathbb{E}\mathrm{Var}(X|S)$.

(d) Znajdź rozkład warunkowy zmiennej S dla $X = x$.

(e) Podaj $\mathbb{E}(S|X)$, $\mathrm{Var}(S|X)$.

8. Niech X i Y będą niezależnymi zmiennymi losowymi o rozkładzie Poissona z parametrami $X \sim \mathrm{Poiss}(\lambda)$, $Y \sim \mathrm{Poiss}(\mu)$ i $S = X + Y$.

(a) Znajdź rozkład warunkowy zmiennej X dla $S = s$.

(b) Podaj $\mathbb{E}(X|S)$, $\mathrm{Var}(X|S)$.

(c) Oblicz $\mathbb{E}(X|S)$, $\mathrm{Var}\mathbb{E}(X|S)$, $\mathbb{E}\mathrm{Var}(X|S)$.

(d) Znajdź rozkład warunkowy zmiennej S dla $X = x$.

(e) Podaj $\mathbb{E}(S|X)$, $\mathrm{Var}(S|X)$.

9. Niech X i Y będą niezależnymi zmiennymi losowymi o jednakowym rozkładzie wykładniczym, $X \sim \mathrm{Ex}(\lambda)$, $Y \sim \mathrm{Ex}(\lambda)$ i $S = X + Y$.

(a) Znajdź rozkład warunkowy zmiennej X dla $S = s$.
Wskazówka: Wzór Bayesa. Ogólniejszy fakt i inną metodę obliczania można znaleźć w Zadaniu 12.

(b) Podaj $\mathbb{E}(X|S)$, $\mathrm{Var}(X|S)$.

(c) Oblicz $\mathbb{E}(X|S)$, $\mathrm{Var}\mathbb{E}(X|S)$, $\mathbb{E}\mathrm{Var}(X|S)$.

(d) Znajdź rozkład warunkowy zmiennej S dla $X = x$.

(e) Podaj $\mathbb{E}(S|X)$, $\mathrm{Var}(S|X)$.

10. Niech $X \sim \mathrm{N}(\mu, v^2)$ i $(Y|X = x) \sim \mathrm{N}(x, \sigma^2)$.

    (a) Podaj rozkład brzegowy zmiennej Y.

    (b) Oblicz rozkład warunkowy zmiennej X dla $Y = y$.

    (c) Oblicz $\mathrm{Cov}(X, Y)$.

11. W urnie znajduje się b białych kul i c kul czerwonych. Losujemy kolejno, jedną po drugiej, bez zwracania n kul.

    (a) Oblicz prawdopodobieństwo, że pierwsza z wylosowanych kul była biała, jeśli wiadomo, że ostatnia, n-ta wylosowana kula jest biała.

    (b) Oblicz $\mathbb{P}(\text{„1-sza biała i } n\text{-ta biała”}) - \mathbb{P}(\text{„1-sza biała”})\mathbb{P}(\text{„}n\text{-ta biała”})$.

    Wskazówka: Prawdopodobieństwo wyniku losowania nie zależy od kolejności, w jakiej kule się pojawiły; zależy tylko od liczby wylosowanych białych/czerwonych kul. Na przykład wynik „BBCCC” ma takie samo prawdopodobieństwo jak „CBCBC” itp.

12. (Rozkłady Dirichleta i Gamma) Niech $X_1, \ldots, X_d$ będą niezależnymi zmiennymi losowymi o rozkładach gamma: $X_i \sim \mathrm{Gamma}(\alpha_i, \lambda)$, $S = X_1 + \cdots + X_d$ i

$$(\vartheta_1, \ldots, \vartheta_d) = \left(\frac{X_1}{S}, \ldots, \frac{X_d}{S} \right).$$

    (a) Pokaż, że $\vartheta \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_d)$.

    (b) Pokaż, że wektor losowy ϑ jest niezależny od S.

    Wskazówka: Oblicz łączną gęstość $(\vartheta_1, \ldots, \vartheta_{d-1}, S)$.