Statystyczna analiza danych (molekularnych) – estymacja bayesowska i MLE

Anna Gambin

20 kwietnia 2012

Spis treści

1 Estymacja bayesowska **1**

2 Paradoks więźnia **2**

2.1 Rozkład *a priori* 3

3 Estymacja największej wiarygodności **5**

1 Estymacja bayesowska

Rozważmy następujący przykład (z Wikipedii :). Obserwujemy wejście do szkoły, w której uczą się zarówno chłopcy jak i dziewczynki w proporcjach $6 : 4$. Dysponujemy wiedzą, że wszyscy chłopcy noszą spodnie, oraz że połowa dziewczynek nosi spódnice, a połowa spodnie. Widzimy z daleka postać w spodniach. Jakie jest prawdopodobieństwo, że jest to dziewczyna ? Niech A oznacza zdarzenie bycia dziewczynką, A' – chłopcem, B – zdarzenie noszenia spodni. Chcemy znać prawdopodobieństwo $P(A|B)$.

Wiemy, że $P(A) = 0.4$ oraz $P(A') = 0.6$. Dodatkowo $P(B|A) = 0.5$, ponieważ co druga dziewczyna chodzi w spodniach i $P(B|A') = 1$ (wszyscy chłopcy noszą spodnie). Ze wzoru na prawdopodobieństwo całkowite:

$$P(B) = P(B|A)P(A) + P(B|A')P(A') = 0.5 \cdot 0.4 + 1 \cdot 0.6 = 0.8$$

Czyli ze wzoru Bayesa:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.5 \cdot 0.4}{0.8} = 0.25$$

Pokontemplujmy ten wzór w ogólniejszej postaci:

$$f_{X|Y=y}(x) = \frac{L_{X|Y=y}(x) \cdot f_X(x)}{\int_{-\infty}^{+\infty} L_{X|Y=y}(x) \cdot f_X(x)dx}$$

- $f_{X|Y=y}(x)$ nazywamy gęstością *a posteriori* zmiennej X pod warunkiem danych $Y = y$,
- $L_{X|Y=y}(x)$ nazywamy wiarygodnością,
- $\int_{-\infty}^{+\infty} L_{X|Y=y}(x) \cdot f_X(x)dx$ jest stałą normalizującą,
- a $f_X(x)$ nazywamy gęstością *a priori* zmiennej losowej X.

2 Paradoks więźnia

Na kolejnym przykładzie przedstawimy główne pomysły stojące za bayesowską estymacją parametrów.

Wyobraźmy sobie trzech więźniów, nazwijmy biedaków A, B i C. Wszyscy wiedzą, że następnego dnia dwóch spośród nich wyjdzie na wolność. A zakłada, że ma szanse $\frac{2}{3}$ opuścić mury więzienia, ale postanawia jeszcze dla poprawy nastroju podpytać strażnika. Prosi go o ujawnienie, który z pozostałych więźniów (poza nim) wyjdzie na wolność. Dostaje odpowiedź: B, a następnie zaczyna się zamartwiać, że jego szanse zmalały do $\frac{1}{2}$ (bo albo wyjdzie on, albo C).

Czy to możliwe, że dane (wiedza o tym, że B wychodzi) zmieniły prawdopodobieństwo ?

Podobny paradoks można znaleźć rozważając prostą grę losową: mamy trzy kubeczki, pod jednym znajduje się nagroda. Wskazujemy na jeden z kubeczków, a wtedy prowadzący grę podnosi jeden z pustych kubeczków (ale nie nasz wybrany). Możemy teraz zmienić nasz wybór, albo przy nim pozostać. Co się bardziej opłaca ?

Żeby odpowiedź na to pytanie była całkowicie jasna, wyobraźmy sobie 100 kubeczków (pod jednym znajduje się nagroda). Losujemy kubeczek i z prawdopodobieństwem $\frac{1}{100}$ trafiamy w nagrodę. Następnie prowadzący grę podnosi 98 pustych kubeczków. Teraz jest jasne, że powinniśmy wskazać na drugi z odwróconych kubeczków, żeby nasze szanse na zdobycie nagrody wzrosły do $\frac{1}{2}$.

Powróćmy do naszych więźniów i zastanówmy się, czy i kiedy dane zmieniają prawdopodobieństwo. Załóżmy, że opisana sytuacja powtarza się wielokrotnie (z kolejnymi trójkami więźniów o pseudonimach A, B i C). Są trzy możliwości wyboru pary więźniów wychodzących na wolność. Jeśli jest to para A i B, to strażnik na pytanie A kto wychodzi odpowie B, podobnie jeśli jest to para A i C, to strażnik musi odpowiedzieć C. W sytuacji, gdy A pozostaje w więzieniu odpowiedź strażnika może być B lub C. Oznaczmy przez S_b zdarzenie, że

strażnik wskazuje na B i przyjmijmy, że $P(S_B|BC) = x$, dla $0 < x < 1$. Niech A oznacza zdarzenie, że więzień A wychodzi na wolność, analogicznie BC jest zdarzeniem, że wolność odzyskują więźniowie B i C. Mamy:

$$P(A|S_B) = P(AB|S_B) + P(AC|S_B) =$$

$$= \frac{P(S_B|AB) \cdot P(AB)}{P(S_B|AB) \cdot P(AB) + P(S_B|BC) \cdot P(BC) + P(S_B|AC) \cdot P(AC)} =$$

$$= \frac{\frac{1}{3}}{1 \cdot \frac{1}{3} + x \cdot \frac{1}{3} + 0} = \frac{1}{1+x}$$

Widzimy z powyższych rachunków, że jeśli $x = \frac{1}{2}$ to prawdopodobieństwo wyjścia na wolność więźnia A nie ulega zmianie, mimo znajomości danych. Jednak kiedy $x \neq \frac{1}{2}$ odpowiedź strażnika wnosi nową informację i szanse A rosną lub maleją.

2.1 Rozkład *a priori*

Zastanówmy się co uczynić jak nie znamy x (właściwie skąd mielibyśmy znać:). Często w modelu są takie nieznane (lub nieinteresujące) parametry. Wtedy rozwiązanie polega na "wycałkowaniu"ich, czyli liczymy:

$$P(A|S_B) = \int_x P(A|S_B, x)P(x)dx = \int_x \frac{1}{1+x} P(x)dx$$

gdzie $P(x)$ jest prawdopodobieństwem *a priori*. Zauważmy, że jest to prawdopodobieństwo prawdopodobieństwa – sytuacja dość powszechna we wnioskowaniu bayesowskim.

Jaki rozkład przyjąć może nasz więzień jako rozkład *a priori* ? Jedną z możliwości jest rozkład jednorodny: $P(x) = 1$ dla $0 \leq x \leq 1$. Wtedy:

$$P(A|S_B) = \int_x \frac{1}{1+x} P(x)dx = \ln 2 = 0.693$$

Inny wynik otrzymamy zakładając skoncentrowany w punkcie $x = \frac{1}{2}$ rozkład *a priori*: $P(x) = \delta(x - \frac{1}{2})$ dla $0 \leq x \leq 1$, gdzie $\delta(x)$ jest tak zwaną *funkcją impulsową* zwaną też funkcją *delta Diraca*, zdefiniowaną następująco:

$$\delta(x) = \begin{cases} +\infty \quad \text{dla} \quad x = 0 \\ 0 \quad \text{dla} \quad x \neq 0 \end{cases}$$

Funkcja delta Diraca może być uważana za funkcję rozkładu gęstości ciągłej zmiennej losowej zwanej zmienną pewną. Niech σ^2 oznacza wariancję, a μ wartość oczekiwaną tej zmiennej. Wtedy zachodzi:

$$\delta(x-\mu) = \lim_{\sigma^2 \to 0} f(x)$$

Dla takiego wyboru prawdopodobieństwa *a priori* mamy:

$$P(A|S_B) = \frac{1}{1+\frac{1}{2}} = \frac{2}{3}$$

Naszym zadaniem będzie teraz wyestymować parametr $x = P(S_B|BC)$, czyli prawdopodobieństwo, że strażnik wskaże więźnia B pod warunkiem, że kolejnego dnia zostaną zwolnieni B i C. Do tego zadania niezbędne są dane o historii poprzednich podpowiedzi strażnika w takich sytuacjach. Załóżmy, że te odpowiedzi były następujące (pamiętamy, że zawsze pyta więzień A o innego zwalnianego):

$$BCBCCBCCCBBCBCBCCCCBBCBCCCBCBCBBCCB$$

Mamy $N = 35$, $N_B = 15$, $N_C = 20$. Załóżmy przez chwile, że te dane pochodzą z serii niezależnych prób Bernoulliego. Ponieważ podchodzimy do sprawy "bayesowsko"interesuje nas prawdopodobieństwo *a posteriori*, czyli $P(\text{dane}\,|x)$. W tym momencie wyraźnie rysuje się różnica w podejściu tradycyjnym i bayesowskim. W tym pierwszym nie martwimy się o żadne rozkłady *a priori*, potrzebujemy jedynie wystarczającej statystyki (odpowiednio dużo prób), żeby policzyć:

$$P(\text{dane}\,|x) = \binom{N}{N_B} x^{N_B}(1-x)^{N_C}$$

Natomiast w podejściu bayesowskim mamy następującą proporcjonalność (współczynnik dwumianowy znika w stałej proporcjonalności, gdyż nie zależy od x):

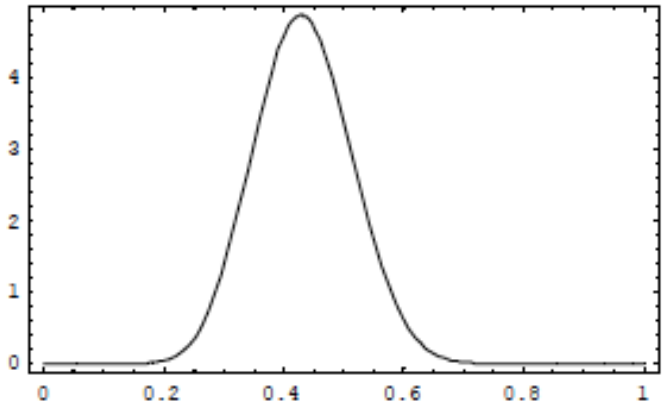
$$P(\text{dane}\,|x) \propto x^{N_B}(1-x)^{N_C} P(x)$$

gdzie $P(x)$ jest rozkładem *a priori*. Za takim podejściem kryje się przesłanie, żeby użyć tylko tych danych, które widzimy, zamiast zakładać coś o **wszystkich** $\binom{N}{N_B}$ permutacjach. Zakładając jednorodny rozkład *a priori* mamy:

$$P(\text{dane}\,|x) = x^{N_B}(1-x)^{N-N_B} \cdot 1$$

Natomiast mianownik (stała normująca) wynosi:

$$\int_0^1 x^{N_B}(1-x)^{N-N_B} dx = \frac{\Gamma(N_B+1)\Gamma(N-N_B+1)}{\Gamma(N+2)}$$



Rysunek 1: Rozkład *a posteriori* dla X, czyli prawdopodobieństwa, że strażnik odpowie B na pytanie więźnia A, dla obserwowanych danych $N = 35$ i $N_B = 15$.

Rysunek 1 przedstawia gęstość dla rozkładu *a posteriori* zmiennej $x = P(S_B|BC)$. Zauważmy, że estymator największej wiarygodności to po prostu procent odpowiedzi typu B:

$$\frac{\partial P(\text{dane}\,|x)}{\partial x} = 0 \implies x = \frac{N_B}{N}$$

Policzmy jeszcze średnią i odchylenie standardowe naszego estymatora dla x.

$$< x >= \int_0^1 xP(\text{dane}\,|x)dx = \frac{N_B + 1}{N + 2}$$

$$\text{var}(x) =< x^2 > - < x >^2 = \frac{(N_B + 1)(N - N_B + 1)}{(N + 2)^2(N + 3)}$$

Widzimy, że rozkład jest coraz bardziej skupiony (ma mniejszą wariancję) wraz z przyrostem danych.

3 Estymacja największej wiarygodności

Sformułujmy wymagania, jakie chcemy, żeby spełniał *dobry* estymator $\hat{\theta}$ dla parametru θ:

- estymator powinien być *nieobciążony*, czyli $E(\hat{\theta}) = \theta$.

- powinien być również *zgodny*, czyli jeśli ξ_n oznacza estymator z n obserwacji dla parametru ξ, to:
$$\forall \epsilon > 0 \quad P(|\xi_n - \xi| > \epsilon) \to_{n\to\infty} 0$$
- powinien mieć również niską wariancję (jeśli jest nieobciążony), lub niski błąd średniokwadratowy (jeśli jest obciążony).[1]
- ostatnia pożądana własność to znana postać rozkładu dla estymatora lub chociaż rozkładu asymptotycznego.

Okazuje się że estymator *największej wiarygodności* (*ang. Maximum Likelihood Estimator – MLE)* spełnia większość tych postulatów.

Rozważmy dyskretne zmienne losowe $Y_1, Y_2, \dots Y_n$, niezależne o identycznym rozkładzie. Ich łączny rozkład wynosi:

$$P(Y_1;\xi) \cdot P(Y_2;\xi) \dots \cdot P(Y_n;\xi)$$

gdzie ξ jest nieznanym parametrem, który chcemy wyestymować. Oznaczmy $Y = (Y_1, Y_2, \dots Y_n)$. Wiarygodność to po prostu rozkład łączny traktowany jako funkcja parametru ξ.

$$L(\xi, Y) = P(Y_1;\xi) \cdot P(Y_2;\xi) \dots \cdot P(Y_n;\xi)$$

Wartość $\hat{\xi}$, która maksymalizuje $L(\xi, Y)$ nazywamy *estymatorem największej wiarygodności* dla parametru ξ. W celu znalezienia tej wartości różniczkujemy funkcję wiarygodności:

$$\frac{\partial}{\partial \xi} L(\xi, Y) = 0 \iff \frac{\partial}{\partial \xi} \log L(\xi, Y) = 0$$

Jeśli chcemy uzyskać konkretne oszacowanie parametru ξ dla danych $y = (y_1, y_2, \dots y_n)$, po prostu podstawiamy dane do funkcji wiarygodności i liczymy maksimum dla $L(\xi, y)$.

Przykład 1 *Niech zmienne $Y_1, Y_2, \dots Y_n$ mają rozkład Poissona z parametrem λ.*

$$L(\lambda, Y) = \frac{e^{-\lambda}\lambda^{Y_1}}{Y_1!} \cdot \frac{e^{-\lambda}\lambda^{Y_2}}{Y_2!} \cdot \dots \cdot \frac{e^{-\lambda}\lambda^{Y_n}}{Y_n!} = \frac{e^{-n\lambda}\lambda^{\sum Y_i}}{\prod (Y_i!)}$$

$$\log L(\lambda, Y) = -n\lambda + (\sum Y_i) \log \lambda - \log(\prod (Y_i!))$$

$$\frac{\partial}{\partial \lambda} \log L(\lambda, Y) = -n + \frac{\sum Y_i}{\lambda} \implies \frac{\partial}{\partial \lambda} \log L(\lambda, Y) = 0 \quad \textit{dla} \quad \hat{\lambda} = \bar{Y}$$

Podsumowując (nieobciążony) estymator MLE dla parametru λ jest równy średniej z obserwacji, a jego wariancja wynosi $\frac{\lambda}{n}$.

[1] W sytuacji kiedy błąd średniokwadratowy (MSE) dla estymatora obciążonego jest mniejszy niż wariancja dla nieobciążonego wariantu, lepiej wybrać estymator obciążony.

Przykład 2 *Rozważny n niezależnych prób Bernouilliego:*

$$Y_i = \begin{cases} 0 & \text{z prawdopodobieństwem } 1-p \\ 1 & \text{z prawdopodobieństwem } p \end{cases}$$

$$L(p, Y) = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^Y(1-p)^{n-Y}$$

gdzie $Y = \sum_{i=1}^{n} Y_i$ jest sumą sukcesów. Załóżmy, że $a \leq Y \leq n-1$, wtedy $\hat{p} = \frac{Y}{n} = \bar{Y}$. Nasz estymator jest oczywiści nieobciążony i łatwo pokazać, że ma wariancję $\frac{p(1-p)}{n}$, a więc jest również zgodny. Powyższą konstrukcje łatwo uogólnić na przypadek rozkładu wielomianowego.

Jednak w przypadkach brzegowych (dla $Y = 0$ lub $Y = n$) musimy postąpić inaczej. Np, dla $Y = n$, funkcja wiarygodności wynosi p^n

$$\frac{\partial}{\partial p} \log L(p, Y) = \frac{n}{p} = 0$$

Widzimy, że powyższe równanie nie ma rozwiązania, czyli musimy szukać estymatora na brzegu obszaru zmienności dla p. Wiemy, że $0 \leq p \leq 1$, czyli estymator MLE dla $Y = n$ jest równy $\hat{p} = 1$.

Rozważmy teraz przypadek zmiennych ciągłych o gęstości f_X zależnej od parametru ξ. Funkcja wiarygodności jest równa:

$$L(\xi, X) = f_X(X_1; \xi) \cdot f_X(X_2; \xi) \cdot \ldots f_X(X_n; \xi)$$

Przykład 3 *Niech $X_1, X_2, \ldots X_n$ mają rozkład normalny o średniej μ i znanej wariancji σ^2. Naszym zadaniem jest znaleźć estymator największej wiarygodności dla μ.*

$$L(\mu, X) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

Po zlogarytmowaniu otrzymujemy:

$$\text{stała niezależna od } \mu - \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2} \implies \hat{\mu} = \bar{X}$$

Estymator jest nieobciążony i ma wiariancję $\frac{\sigma^2}{n}$. Korzystając z nierówności Czebyszewa możemy pokazać zgodność estymatora.

Literatura

[1] Julian J. Faraway Practical Regression and Anova using R, 2002.

[2] K. Seefeld, E. Linder, Statistics using R with biological examples, 2007.

[3] W.P. Krijnen Applied statistics for bioinformatics using R, 2009.

[4] S.K. Mathur, Statistical Bioinformatics with R, Elsevier Academic Press, 2010.

[5] W. J. Ewens, G. R. Grant, Statistical Methods in Bioinformatics Springer-Verlag, 2001.