Bootstrap w kontekście metody największej wiarogodności

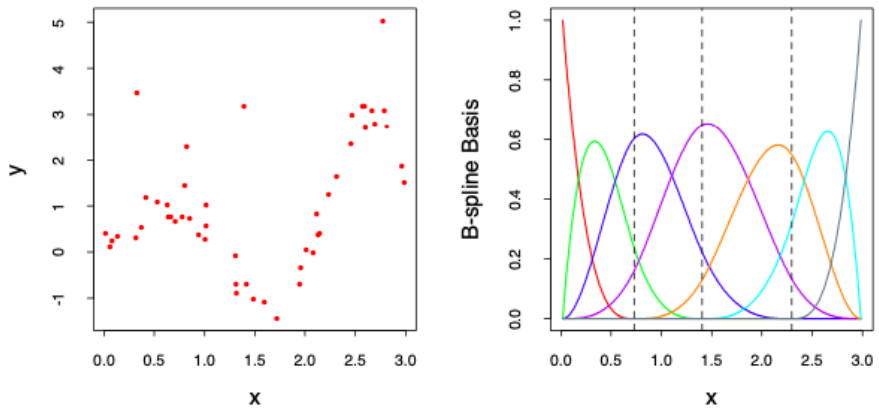
Aleksandra Steiner

1 Bootstrap

Bootstrap to metoda, dzięki której możliwe jest oszacowanie estymatorów parametrów zmiennych losowych o nieznanym rozkładzie. Jest to nieklasyczna metoda statystyczna używająca wielokrotnego losowania ze zwracaniem z próby. Poniżej, zilustrujemy działanie bootstrapu w prostym jednowymiarowym problemie wygładzania oraz pokażemy jego połączenie z metodą największej wiarogodności [1].

1.1 Wprowadzenie - przykład

Oznaczmy dane treningowe przez $Z = \{z_1, z_2, \ldots, z_N\}$, gdzie $z_i = (x_i, y_i)$, $i = 1, 2, \ldots, N$. x_i będzie tutaj jednowymiarową obserwacją pewnej cechy oraz y_i będzie odpowiedzią, ciągłą lub kategoryczną. Przykładem, który rozważymy będzie liczba obserwacji $N = 50$, które znajdują się na ilustracji po lewej poniżej.



Rysunek 1: Przykładowe dane dla przypadku z wygładzaniem (rys. po lewej). Siedem funkcji sklejanych (splajnów). Trzy przerywane proste oznaczają tutaj węzły krzywej (rys. po prawej)

Załóżmy, że decydujemy się dopasować kubiczną funkcję sklejaną z trzema węzłami znajdującymi się w kwartylach wartości X. Otrzymujemy tutaj 7-wymiarową liniową przestrzeń funkcji, która może być przedstawiona przez

liniowe rozwinięcie funkcji sklejanych:

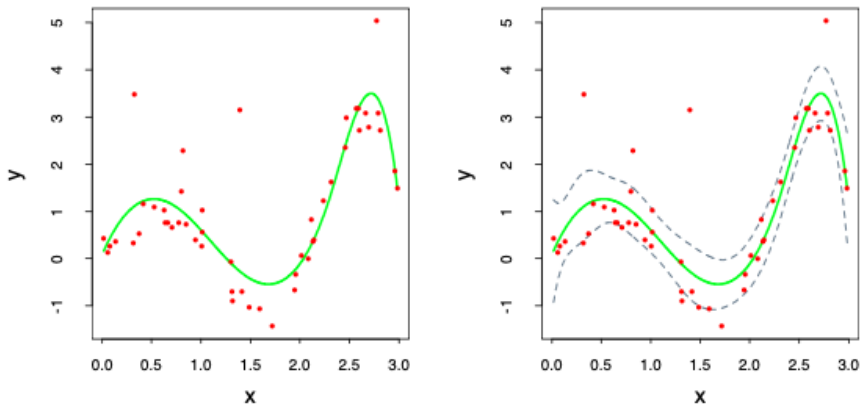
$$\mu(x) = \sum_{j=1}^{7} \beta_j h_j(x). \tag{1}$$

W powyższym równaniu, $h_j(x)$, $j = 1, 2, \ldots, 7$ to siedem funkcji pokazanych na rysunku 1. Będziemy tutaj myśleć o $\mu(x)$ jako odpowiedniku warunkowej wartości oczekiwanej $\mathbb{E}(Y|X = x)$.

Niech H będzie macierzą wymiaru $N \times 7$ z ij-tym elementem $h_j(x_i)$. Klasyczny estymator β w takiej sytuacji, otrzymany przez minimalizację błędu kwadratowego na zbiorze treningowym dany jest przez:

$$\hat{\beta} = (H^T H)^{-1} H^T y. \tag{2}$$

Takie dopasowanie: $\hat{\mu}(x) = \sum_{j=1}^{7} \hat{\beta}_j h_j(x)$ możemy zobaczyć na poniższej ilustracji (wykres po lewej stronie).



Rysunek 2: Funkcja sklejana dopasowana do danych (rys. po lewej) oraz ta sama funkcja wraz z zaznaczonymi przedziałem wyznaczonym przez $+/-$ $1.96\times$ błąd standardowy.

Estymowana macierz kowariancji wektora $\hat{\beta}$ to

$$\widehat{Var}(\hat{\beta}) = (H^T H)^{-1} \hat{\sigma}^2, \tag{3}$$

gdzie $\hat{\sigma}^2$ estymowany jest przez $\sum_{i=1}^{N} (y_i - \hat{\mu}(x_i))^2 / N$.

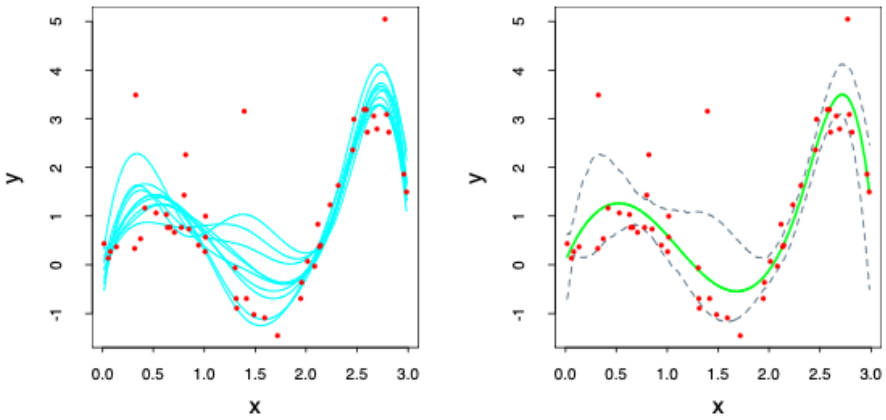
Przyjmując, że $h(x)^T = (h_1(x), h_2(x), \ldots, h_7(x))$, błąd standardowy predykcji $\hat{\mu}(x) = h(x)^T \hat{\beta}$ to:

$$\widehat{se}[\hat{\mu}(x)] = [h(x)^T (H^T H)^{-1} h(x)]^{\frac{1}{2}} \hat{\sigma}. \tag{4}$$

Ponieważ 1.96 to kwantyl rzędu 0.975 standardowego rozkładu normalnego, na rysunku 2 widzimy przybliżony $100 - 2 \cdot 2.5\% = 95\%$ punktowy przedział ufności dla $\mu(x)$.

1.2 Na czym polega metoda bootstrap?

Opierając się na założeniach przykładu opisanego powyżej, losujemy ze zwracaniem B zbiorów danych, każdy rozmiaru $N = 50$ z danych treningowych, gdzie losowaną jednostką jest $z_i = (x_i, y_i)$. Dla każdego takiego bootstrapowego wylosowanego zbioru danych Z^* dopasowujemy kubiczną funkcję sklejaną $\hat{\mu}^*(x)$. Przykład takiego działania (dla dziesięciu wylosowanych prób) możemy zobaczyć na ilustracji poniżej:



Rysunek 3: Dziesięć dopasowanych funkcji sklejanych opartych na losowanych próbach z danych treningowych (rys. po lewej) oraz funkcja sklejana z 95% pasmem ufności błędu standardowego.

Używając $B = 200$ bootstrapowych prób, możemy skonstruować 95% punktowe pasmo ufności dla $\mu(x)$. Dla każdej z tych 200 prób wyliczamy estymator $\hat{\mu^*}(x)$ oraz dla każdego x znajdujemy percentyle $0.025 \cdot 100\% = 2.5\%$ i $0.975 \cdot 100\% = 97.5\%$ estymacji bootstrapowych próbek.

Okazuje się, że istnieje ścisłe powiązanie pomiędzy estymacjami najmniejszych kwadratów (*least squares*) z 2 i 3, metodami bootstrap i największej wiarogodności. Załóżmy, że błędy w modelu, który rozważamy są gaussowskie, to znaczy:

$$\begin{aligned} Y &= \mu(X) + \varepsilon, \text{ gdzie } \varepsilon \sim N(0, \sigma^2), \\ \mu(x) &= \sum_{j=1}^{7} \beta_j h_j(x). \end{aligned} \tag{5}$$

Metoda bootstrap, opisana powyżej, w której pobieramy próby ze zwracaniem z danych treningowych nazywana jest bootstrapem nieparametrycznym. Oznacza to metodę tzw. bez modelu, ponieważ korzysta ona z nieprzetworzonych danych aby generować nowe zbiory danych. Nie używa do tego konkretnego modelu.

Rozważmy pewną odmianę bootstrapu, czyli jego parametryczną wersję. Będziemy w niej symulować nowe zmienne odpowiedzi dodając błędy z rozkładu normalnego do estymowanych wartości, to znaczy:

$$y_i^* = \hat{\mu}(x_i) + \varepsilon_i^*,\ \varepsilon_i^* \sim N(0, \hat{\sigma}^2),\ i = 1, 2, \ldots, N. \tag{6}$$

Powyższy proces powtarzany jest B razy, gdzie w naszym przypadku, powiedzmy, że ponownie $B = 200$. Wynikowe bootstrapowe zbiory danych są postaci $(x_1, y_1^*), \ldots, (x_N, y_N^*)$, a dla każdego z nich wyliczamy tak jak poprzednio funkcję sklejaną. Pasmo ufności w tej metodzie wraz ze wzrostem do nieskończoności liczby prób bootstrapowych będzie równe dokładnie pasmu ufności uzyskanemu z metodą najmniejszych kwadratów (rysunek 2 po prawej).

Estymator μ z próby y^* dany jest przez

$$\hat{\mu^*}(x) = h(x)^T (H^T H)^{-1} H^T y^*, \tag{7}$$

a jego rozkład to

$$\hat{\mu^*}(x) \sim N(\hat{\mu}(x), h(x)^T (H^T H)^{-1} h(x) \hat{\sigma}^2). \tag{8}$$

Zauważmy, że wartość oczekiwana tego rozkładu to po prostu estymator najmniejszych kwadratów, a odchylenie standardowe jest takie samo jak przybliżona formuła z (4).

2 Metoda największej wiarogodności

Parametryczny bootstrap zgadza się z metodą najmniejszych kwadratów w poprzednim przykładzie, ponieważ model zadany w (5) posiada gaussowskie błędy. Ogólnie jednak, parametryczny bootstrap nie tyle zgadza się z metodą najmniejszych kwadratów, ale z metodą największej wiarogodności, co teraz omówimy.

Zacznijmy od zdefiniowania funkcji gęstości prawdopodobieństwa (w przypadku ciągłej zmiennej) czy też funkcji masy prawdopodobieństwa (w przypadku zmiennej dyskretnej).

$$z_i \sim g_\theta(z). \tag{9}$$

W powyższym wyrażeniu, θ oznacza jeden lub więcej nieznanych parametrów, które wyznaczają rozkład Z. Nazywa się to parametrycznym modelem Z.

Jako przykład, jeśli Z pochodzi z rozkładu normalnego z wartością oczekiwaną μ i wariancją σ^2, wtedy

$$\theta = (\mu, \sigma^2)$$

oraz

$$g_\theta(z) = \frac{1}{\sqrt{(2\pi)\sigma}} e^{-\frac{1}{2}(z-\mu)^2/\sigma^2}.$$

Metoda największej wiarogodności opiera się na funkcji wiarogodności danej przez prawdopodobieństwo zaobserwowanych danych zgodnie z modelem g_θ:

$$L(\theta; Z) = \prod_{i=1}^{N} g_\theta(z_i). \tag{10}$$

Będziemy myśleć o powyższej funkcji jako funkcji θ z ustalonymi danymi Z. Oznaczymy również logarytm $L(\theta, Z)$ (nazywany log-likelihood) przez

$$l(\theta, Z) = \sum_{i=1}^{N} l(\theta, z_i) = \sum_{i=1}^{N} \log g_\theta(z_i). \tag{11}$$

Metoda największej wiarogodności wybiera wartość parametru $\theta = \hat{\theta}$ tak, aby maksymalizować $l(\theta, Z)$.

Zdefiniujmy tutaj również funkcję *score* jako

$$\dot{l}(\theta, Z) = \sum_{i=1}^{N} \dot{l}(\theta, z_i), \tag{12}$$

gdzie $\dot{l}(\theta, z_i) = \delta l(\theta, z_i)/\delta\theta$. Zakładając, że funkcja wiarogodności przyjmuje maksimum we wnętrzu przestrzeni parametrów, $\dot{l}(\hat{\theta}, Z) = 0$. Macierz informacji dana jest przez:

$$I(\theta) = -\sum_{i=1}^{N} \frac{\delta^2 l(\theta, z_i)}{\delta\theta\delta\theta^T}. \tag{13}$$

Wartość macierzy informacji w punkcie θ=$\hat{\theta}$ nazywana jest często *zaobserwowaną informacją*. Informacja Fishera (lub też oczekiwana informacja) to $i(\theta) = \mathbb{E}_\theta[I(\theta)]$. Dodatkowo oznaczmy jeszcze prawdziwą wartość θ przez θ_0.

Podstawowe wyniki, które znamy mówią o tym, że próbkowy estymator największej wiarogodności ma asymptotyczny rozkład normalny

$$\hat{\theta} \to N(\theta_0, i(\theta_0)^{-1}) \text{ gdy } N \to \infty. \tag{14}$$

W naszym przypadku losujemy niezależnie z $g_{\theta_0}(z)$, co sugeruje, że próbkowy rozkład $\hat{\theta}$ możemy przybliżyć przez

$$N(\hat{\theta}, i(\hat{\theta})^{-1}) \text{ lub } N(\hat{\theta}, I(\hat{\theta})^{-1}), \tag{15}$$

gdzie $\hat{\theta}$ jest tutaj estymatorem ML obserwowanych danych.

Odpowiadające estymacje błędów standardowych $\hat{\theta}_j$ otrzymujemy z

$$\sqrt{(i(\hat{\theta})_{jj}^{-1})} \text{ i } \sqrt{(I(\hat{\theta})_{jj}^{-1})}. \tag{16}$$

Punkty ufności dla θ_j mogą być skonstruowane z przybliżenia próbkowego rozkładu $\hat{\theta}$ podanego powyżej. Taki punkt ma postać:

$$\hat{\theta_j} - z^{(1-\alpha)} \cdot \sqrt{i(\hat{\theta})_{jj}^{-1}} \text{ lub } \hat{\theta_j} - z^{(1-\alpha)} \cdot \sqrt{I(\hat{\theta})_{jj}^{-1}}, \tag{17}$$

gdzie $z^{(1-\alpha)}$ jest kwantylem rzędu $1-\alpha$ standardowego rozkładu normalnego. Bardziej dokładne przedziały ufności możemy uzyskać z funkcji wiarogodności używając przybliżenia rozkładem chi-kwadrat:

$$2[l(\hat{\theta}) - l(\theta_0)] \sim \chi_p^2, \tag{18}$$

gdzie p to liczba elementów parametru θ. Zatem otrzymany $1 - 2\alpha$ przedział ufności jest zbiorem wszystkich takich θ_0, że $2[l(\hat{\theta}) - l(\theta_0)] \leq \chi_p^{2^{(1-2\alpha)}}$, gdzie $\chi_p^{2^{(1-2\alpha)}}$ to kwantyl rzędu $1 - 2\alpha$ rozkładu chi-kwadrat z p stopniami swobody.

Wróćmy zatem do naszego przykładu wygładzania, aby zobaczyć co daje nam metoda ML. Rozważane parametry to $\theta = (\beta, \sigma^2)$. Funkcja log-wiarogodności to

$$l(\theta) = -\frac{N}{2} \log \sigma^2 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - h(x_i)^T \beta)^2. \tag{19}$$

Estymatory ML otrzymujemy obliczając $\delta l / \delta \beta = 0$ oraz $\delta l / \delta \sigma^2 = 0$, co daje nam:

$$\begin{aligned} \hat{\beta} &= (H^T H)^{-1} H^T y, \\ \hat{\sigma}^2 &= \frac{1}{N} \sum (y_i - \hat{\mu}(x_i))^2. \end{aligned} \tag{20}$$

Zatem estymacje otrzymane powyżej zgadzają się z tymi otrzymanymi w (2) i (3).

Macierz informacji dla $\theta = (\beta, \sigma^2)$ to macierz klatkowa, a blok odpowiadający β to

$$I(\beta) = (H^T H)/\sigma^2. \tag{21}$$

Widać więc, że estymowana wariancja:$(H^T H)^{-1}\sigma^2$ pokrywa się z estymacją LS.

3 Bootstrap versus Maximum Likelihood

Zasadniczo, bootstrap jest komputerową implementacją nieparametrycznego lub parametrycznego ML. Jednak w przeciwieństwie do ML, ma on tę zaletę, że pozwala nam policzyć estymatory ML błędów standardowych i innych wielkości w przypadku, gdy nie mamy danej informacji o rozkładzie.

Literatura

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2 edition, 2009.