Kryteria wyboru rozkładu i testy zgodności

Joanna Czarnowska[1]

[1]Uniwersytet Gdański
Instytut Informatyki

Wybór rozkładu

Mamy próbę $x_1, x_2, \ldots, x_n$ z nieznanego rozkładu.

Z poniższych różnych rodzin rozkładów, chcemy wybrać rozkład najmniej różniący się od tego nieznanego rozkładu

$$M_1 = \{F_{\theta_1};\ \theta_1 \in \Theta_1 \subset R^{p_1}\},$$

$$\ldots$$

$$M_k = \{F_{\theta_k};\ \theta_k \in \Theta_k \subset R^{p_k}\}.$$

Jak to zrobić? I co rozumiemy przez „najmniej różniący się”?

Wybór rozkładu – porównanie dystrybuant

Jedną z metod jest porównanie dystrybuanty empirycznej $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n} 1_{\{X_i \leqslant x\}}$ z rozważaną dystrybuantą teoretyczną F. Najczęściej wykorzystywane są do tego statystyki, mierzące różnice w wartościach między obiema dystrybuantami. Najbardziej znaną jest statystyka Kołmogorowa-Smirnowa

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|, \tag{1}$$

mierząca maksymalną różnicę w wartościach dystrybuant.

Zobacz Przykład 1 i Przykład 2 w skrypcie 05KryteriaWyboru-wyklad.R

Wybór rozkładu – porównanie dystrybuant

Inne statystyki to

1. Cramera-von-Misesa

$$CM = n \int_{-\infty}^{\infty} (\hat{F}_n(x) - F(x))^2 dx, \tag{2}$$

   oparta na kwadracie różnic w wartościach dystrybuant,

2. Andersona-Darlinga

$$AD = n \int_{-\infty}^{\infty} \frac{(\hat{F}_n(x) - F(x))^2}{F(x)(1 - F(x))} dx. \tag{3}$$

   oparta na unormowanym kwadracie różnic w wartościach dystrybuant.

Wybór rozkładu – porównanie dystrybuant

1. Niech $F_1, F_2, \ldots, F_k$ bedą dystrybuantami odpowiednio z rodzin $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$, których parametry zostały wyestymowane z użyciem na przykład estymatora największej wiarygodności (MLE).
2. Dla każdej z dystrybuant $F_1, F_2, \ldots, F_k$ i dystrybuanty empirycznej $\hat{F}_n$, obliczamy wartość statystyki Kołmogorowa-Smirnowa D_n. Wybieramy ten rokład, dla którego wartość statystyki jest najmniejsza.

Podobnie wyboru możemy dokonać wykorzystując statystyki CM i AD.

Zobacz Przykład 3 w skrypcie 05KryteriaWyboru-wyklad.R

Wybór rozkładu – porównanie gęstości

Inną metodą wyboru „najlepszego" spośród rozkładów ciągłych, jest porównanie różnic między gęstościami. Prowadzi to do dwóch kryteriów AIC (Akaike Information Criterion) oraz BIC (Bayesian Information Criterion).

Według kryterium AIC (Akaike, 1974) z kilku modeli wybieramy ten, który ma najmniejszą wartość

$$AIC = -2 \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) + 2p,$$

gdzie $\hat{\theta}$ jest estymatorem MLE nieznanego parametru θ, a p liczbą parametrów rozkładu.

BIC (Bayesian Information Criterion)

Według kryterium BIC wybieramy ten model, dla którego jest najmniejsza wartość

$$BIC = -2 \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) + p \ln n,$$

gdzie $\hat{\theta}$ jest estymatorem MLE parametru θ, a p liczbą parametrów rozkładu.

Logarytm wiarygodności

Do wyboru rozkładu wykorzystuje się też logarytm wiarygodności

$$LL = \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i),$$

gdzie $\hat{\theta}$ jest estymatorem MLE nieznanego parametru θ, a p liczbą parametrów rozkładu. W tym przypadku wybieramy model, dla którego wartość LL jest największa.

Kryteria LL, AIC oraz BIC wykorzystujemy też do wyboru z rozkładów dyskretnych.

Zobacz Przykład 3 w skrypcie 05KryteriaWyboru-wyklad.R

Test zgodności rozkładów

Na podstawie realizacji $x_1, x_2, \ldots, x_n$ z rozkładu o nieznanej dystrybuancie F, testujemy hipotezę zerową o równości dystrybuant

$$\mathrm{H}_0 : F = F_0,$$

przeciwko hipotezie alternatywnej(kontrhipotezie)

$$H_1 : \ F \neq F_0,$$

gdzie F_0 jest ustaloną dystrybuantą.

Jako statystyki testowej użyjemy statystyki Kołmogorowa-Smirnowa. Duża wartość tej statystyki, obliczona dla $x_1, x_2, \ldots, x_n$, świadczy przeciwko hipotezie zerowej. W teście można wykorzystać też statystyki CM i AD.

Test zgodności (metoda Monte-Carlo)

Do przetestowania hipotezy zerowej o równości dystrybuant, wykorzystamy następujący algorytm.

1. Metodą Monte-Carlo wyznaczamy rozkład statystyki D_n, dla próbek liczności n z rozkładu F_0.
2. Obliczamy d_n – wartość statystyki D_n, dla $x_1, x_2, \ldots, x_n$ oraz F_0.
3. Korzystając z 1. obliczamy prawdopodobieństwo $p = P(D_n > d_n)$.
4. Ustalamy α – poziom istotności testu (w praktyce przyjmuje się najczęściej 5% lub 1%).
   - Jeśli wartość p wyliczona w punkcie 3. jest mniejsza lub równa od przyjętego poziomu istotności α, to hipotezę H_0 o równości dystrybuant odrzucamy (na przyjętym poziomie istotności α) na korzyść hipotezy alternatywnej H_1.
   - W przeciwnym razie twierdzimy, że nie ma podstaw do orzucenia hipotezy zerowej (na przyjętym poziomie istotności).

Zobacz Przykład 4 i Przykład 5 w skrypcie 05KryteriaWyboru-wyklad.R

AIC – uzasadnienie

Dalej krótkie uzasadnienie formuły AIC.

Informacja Kullback'a-Leibler'a (Kullback-Leibler Divergence)

Niech f_θ będzie gęstością jednej z wybranych rodzin, którą chcemy wykorzystać do aproksymacji nieznanej gęstości g.

Informacją Kulback'a-Leibler'a nazywamy

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx.$$

Informacja Kullback'a-Leibler'a (Kullback-Leibler Divergence)

Niech f_θ będzie gęstością jednej z wybranych rodzin, którą chcemy wykorzystać do aproksymacji nieznanej gęstości g.

Informacją Kulback'a-Leibler'a nazywamy

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx.$$

Dla rozkładów dyskretnych wykorzystujemy funkcje prawdopodobieństwa ($P(X{=}x) = g(x)$).

$$KL(g, f_\theta) = \sum_{i=1}^{\infty} g(x_i) \ln \frac{g(x_i)}{f_\theta(x_i)}.$$

Informacja Kullback'a-Leibler'a (Kullback-Leibler Divergence)

Niech f_θ będzie gęstością jednej z wybranych rodzin, którą chcemy wykorzystać do aproksymacji nieznanej gęstości g.

Informacją Kulback'a-Leibler'a nazywamy

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx.$$

Dla rozkładów dyskretnych wykorzystujemy funkcje prawdopodobieństwa ($P(X{=}x) = g(x)$).

$$KL(g, f_\theta) = \sum_{i=1}^{\infty} g(x_i) \ln \frac{g(x_i)}{f_\theta(x_i)}.$$

Można wykazać, że

- $KL(g, f_\theta) \geqslant 0$ oraz, że
- $KL(g, f_\theta) = 0$ wtedy i tylko wtedy, gdy $f_\theta = g$.

Informacja Kullback'a-Leibler'a (Kullback-Leibler Divergence)

Niech f_θ będzie gęstością jednej z wybranych rodzin, którą chcemy wykorzystać do aproksymacji nieznanej gęstości g.

Informacją Kulback'a-Leibler'a nazywamy

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx.$$

Dla rozkładów dyskretnych wykorzystujemy funkcje prawdopodobieństwa ($P(X{=}x) = g(x)$).

$$KL(g, f_\theta) = \sum_{i=1}^{\infty} g(x_i) \ln \frac{g(x_i)}{f_\theta(x_i)}.$$

Można wykazać, że

- $KL(g, f_\theta) \geqslant 0$ oraz, że
- $KL(g, f_\theta) = 0$ wtedy i tylko wtedy, gdy $f_\theta = g$.

Im mniejsza jest wartość $KL(g, f_\theta)$, tym mniej tracimy informacji biorąc f_θ w miejsce nieznanego rozkładu g.

Informacja Kullback'a-Leibler'a. Kryterium AIC

Rozpiszmy całkę

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx$$

$$= \underbrace{\int_{-\infty}^{\infty} g(x) \ln g(x) dx}_{\text{stała, nie zależy od wyboru } f_\theta} - \int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx$$

Informacja Kullback'a-Leibler'a. Kryterium AIC

Rozpiszmy całkę

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx$$

$$= \underbrace{\int_{-\infty}^{\infty} g(x) \ln g(x) dx}_{\text{stała, nie zależy od wyboru } f_\theta} - \int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx$$

Przenieśmy stałą na drugą stronę równości

$$KL(g, f_\theta) - \text{stała} = -\int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx \tag{4}$$

Informacja Kullback'a-Leibler'a. Kryterium AIC

Rozpiszmy całkę

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx$$

$$= \underbrace{\int_{-\infty}^{\infty} g(x) \ln g(x) dx}_{\text{stała, nie zależy od wyboru } f_\theta} - \int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx$$

Przenieśmy stałą na drugą stronę równości

$$KL(g, f_\theta) - \text{stała} = -\int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx \tag{4}$$

Estymatorami wielkości (4) są

$$L = -\frac{1}{n} \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) \quad (oraz) \quad A = -\frac{1}{n} \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) + \frac{p}{n}, \tag{5}$$

gdzie p jest liczbą parametrów rozkładu.

Informacja Kullback'a-Leibler'a. Kryterium AIC

Rozpiszmy całkę

$$KL(g, f_\theta) = \int_{-\infty}^{\infty} g(x) \ln \frac{g(x)}{f_\theta(x)} dx$$

$$= \underbrace{\int_{-\infty}^{\infty} g(x) \ln g(x) dx}_{\text{stała, nie zależy od wyboru } f_\theta} - \int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx$$

Przenieśmy stałą na drugą stronę równości

$$KL(g, f_\theta) - \text{stała} = -\int_{-\infty}^{\infty} g(x) \ln f_\theta(x) dx \tag{4}$$

Estymatorami wielkości (4) są

$$L = -\frac{1}{n} \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) \quad (oraz) \quad A = -\frac{1}{n} \sum_{i=1}^{n} \ln f_{\hat{\theta}}(X_i) + \frac{p}{n}, \tag{5}$$

gdzie p jest liczbą parametrów rozkładu.

Wzór AIC otrzymujemy mnożąc A przez $2n$, wzór LL – mnożąc przez $-n$.