

Model bayesowski i MCMC

Zofia Dzedzic

Uniwersytet Wrocławski

W statystyce bayesowskiej zakłada się, że parametry rozkładu są zmiennymi losowymi.

Rozkład a'priori

$$\theta \sim \pi(\theta)$$

Obserwacje przy znanym θ mają rozkład $X \sim p(X|\theta)$ – wiarygodność. Wnioskujemy o parametrach na podstawie rozkładu θ pod warunkiem zaobserwowanego X

Rozkład a'posteriori

$$\pi(\theta|X) = \frac{\pi(\theta)p(X|\theta)}{\int_X \pi(\theta)p(x|\theta)dx} \propto \pi(\theta)p(X|\theta)$$

Znajomość rozkładu a' posteriori umożliwia nam predykcję wartości przyszłej obserwacji zmiennej X

$$p(x^{new}|X) = \int p(x^{new}|\theta)\pi(\theta|X)d\theta.$$

Porównując z tym podejściem, metoda MLE wyliczyłaby w tym miejscu $p(x^{new}|\hat{\theta})$, nie biorąc pod uwagę niepewności w estymacji parametru.

S – liczba wyrzuconych orłów w N rzutach monetą o prawdopodobieństwie wyrzucenia orła θ .

$$p(S|\theta) \propto \theta^S(1 - \theta)^{N-S}$$

Zakładamy, że rozkład a'priori jest rozkładem Beta(α , β) czyli

$$\pi(\theta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}.$$

Stąd

$$\pi(\theta|S) \propto \theta^{S+\alpha-1}(1 - \theta)^{N-S+\beta-1}$$

czyli $\pi(\theta|S) \sim \beta(S + \alpha, N - S + \beta)$.

Mając zdefiniowany model bayesowski potrzebujemy często losować parametry z zadanych rozkładów a' posteriori. W większości przypadków nie jest to proste, więc potrzebujemy w tym celu dobrych algorytmów. Jedną z takich metod są Monte Carlo Marcov Chains.

Chcemy obliczyć

$$I = \int f(x)\pi(x)dx,$$

gdzie π jest gęstością prawdopodobieństwa. Czyli z MPWL

$$I = Ef(X) \approx \frac{1}{n}f(X_i) =: \hat{I}_n,$$

gdzie $X_1, \dots, X_n \sim i.i.d, \pi$

W prostym MC kolejne iteracje nie zależą od poprzednich. Stąd każdy krok musi być dopasowany globalnie do rozkładu. MCMC pozwalają eksplorować rozkład lokalnie, np. w okolicy wyniku z poprzedniego kroku.

Metoda ta polega na wygenerowaniu ergodycznego łańcucha Markowa X_1, X_2, \dots, X_n o rozkładzie stacjonarnym π . Wówczas

$$\hat{I}_{n_0, n} = \frac{1}{n} \sum_{i=n_0+1}^{n+n_0} f(X_i) \approx \int f(x)\pi(x)dx.$$

Próbnik Gibbsa (Gibbs sampler)

Rozważmy sytuację, w której mamy zmienne losowe X_1, X_2, \dots, X_K i chcemy wygenerować próbę z rozkładu łącznego tych zmiennych. Załóżmy, że trudno to zrobić bezpośrednio, ale znamy rozkłady warunkowe poszczególnych zmiennych w zależności od pozostałych, t.j.

$$\pi(X_i | X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_K)$$

Możemy wtedy do wygenerowania próby z zamierzonego rozkładu zastosować następujący algorytm

Próbnik Gibbsa

- 1 Wybierz indeks $j \in \{1, \dots, K\}$.
- 2 Zmiennym o indeksach i różnych od j przypisz wartości $x_i^{(t+1)} = x_i^{(t)}$, tak że $x_{-j}^{t+1} = x_{-j}^t$.
- 3 Wygeneruj $x_j^{(t+1)} \sim p(x_j | x_{-j}^{(t)})$.

Rozważmy rozkład

$$p(x, y) = \frac{n!}{(n-x)!x!} y^{(x+\alpha-1)}(1-y)^{(n-x+\beta-1)} \quad x \in \{0, \dots, n\}, y \in [0, 1]$$

Trudno generować z niego bezpośrednio, ale łatwo możemy wylosować próbę z rozkładów brzegowych

$$p(x|y) \sim \text{Bin}(n, y)$$

$$p(y|x) \sim \text{Beta}(x + \alpha, n - x + \beta)$$

Niech $Q(x|dy)$ będzie jądrem przejścia o gęstości $q(x, y)$, a π docelowym rozkładem.

Algorytm M-H

- 1 Wygeneruj $Y \sim Q(X_n, \cdot)$
- 2 Oblicz p-stwo akceptacji $\alpha(X_n, Y)$, gdzie

$$\alpha(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} \wedge 1$$

- 3 $X_{n+1} = \begin{cases} Y & \text{z p-stwem } \alpha(X_n, Y) \\ X_n & \text{z p-stwem } 1 - \alpha(X_n, Y) \end{cases}$

Otrzymany z algorytmu M-H łańcuch Markowa jest odwracalny względem rozkładu stacjonarnego π .

- 1 Elements of statistical learning
- 2 Błażej Miasojedow, materiały do wykładu „Algorytmy statystyki praktycznej”