Wyklad7

16 listopada 2012

00:23

3. Podejście Bayesowskie

dotychczas myśleliśmy tak: $p_\theta(x)$

pewien określony nieznany parametr

i mówiliśmy o rodzinie rozkładów prawdopodobieństw

terez myślimy tak $p(x|\theta)$

a θ jest zmienną losową z pewnym rozkładem a priori $p(\theta)$. Zalecane podejście jeśli chcemy polepszyć estymację wykorzystując wiedzę a priori albo gdy nie potrafimy znaleźć opt. estymatorów (nie istnieją ...) w podejściu klasycznym

3.1 Optymalny estymator Bayesowski

w podejściu klasycznym minimalizowaliśmy:

$$\Delta^2\tilde{\theta} = \int dx\, (\tilde{\theta}(x)-\theta)^2 p_\theta(x)$$

i wymagaliśmy nieobciążoności estymatora

W podejściu Bayesowskim minimalizujemy średnią wariancję:

$$\Delta^2\tilde{\theta} = \int dx\, d\theta\, (\tilde{\theta}(x)-\theta)^2 \underbrace{p(x|\theta)\, p(\theta)}_{p(x,\theta)}$$

Korzystając z wzoru Bayesa $p(x,\theta) = p(\theta|x)\,p(x)$:

$$= \int \left[\int d\theta \left[\tilde{\theta}(x)-\theta\right)^2 p(\theta|x)\right] p(x)\, dx$$

Ponieważ $p(x) \geq 0$ szukanie optymalnego estymatora oznacza znalezienie dla każdego x takiego $\tilde{\theta}(x)$ który minimalizuje $\int dx\, (\tilde{\theta}(x)-\theta)^2 p(\theta|x)$

$$\frac{d}{d\tilde{\theta}}\int d\theta\,(\tilde{\theta}-\theta)^2 p(\theta|x) = \int d\theta\, 2(\tilde{\theta}-\theta)\, p(\theta|x) = 0$$

$$\Downarrow$$

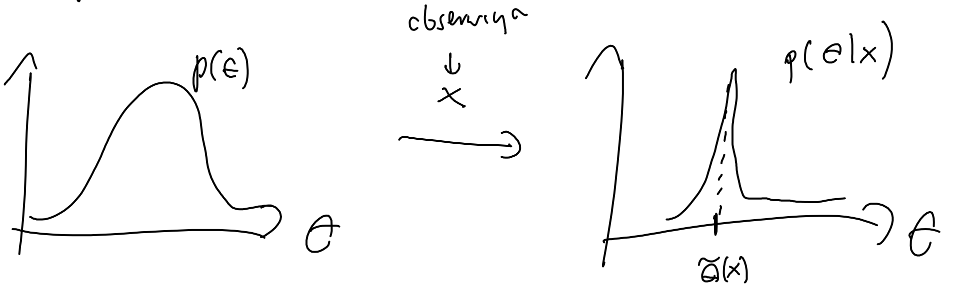
$$\tilde{\theta} = \int d\theta\, \theta\, p(\theta|x)$$

Optymalny estymator: Bayesowski:

$$\boxed{\tilde{\theta}(x) = \langle \theta \rangle_{p(\theta|x)}}$$

wartość oczekiwana rozkładu a posteriori

$$\underbrace{p(\theta|x)}_{\text{a posteriori}} = \frac{p(x,\theta)}{p(x)} = \frac{p(x|\theta)\,\overbrace{p(\theta)}^{\text{a priori}}}{p(x)}$$



optymalny estymator

Dla optymalnego estymatora:

$$\delta^2\tilde{\theta} = \int dx\, d\theta\, (\theta - \tilde{\theta}(x))^2 p(\theta|x)\, p(x)\, dx =$$

$$= \int dx \left(\int d\theta \left(\theta - \langle\theta\rangle_{p(\theta|x)} \right)^2 p(\theta|x) \right) p(x)\, dx$$

$$= \int dx\, \delta^2\theta\big|_{p(\theta|x)}\, p(x)\, dx$$

Intuicja:

W ogólności $\delta\tilde{\theta}$ będzie zależeć od wiedzy a priori ale im więcej danych tym wpływ wiedzy a priori słabszy:

$$p(\theta|x) \propto \underset{\text{szybko zmienne}}{p(x|\theta)}\ \underset{\text{wolno zmienne z } \theta}{p(\theta)} \approx p(x|\theta)$$

2 σ

Jeśli danych mało estymator będzie dociągany w kierunku średniej rozkładu apriori, największy problem podejścia Bayesowskiego: to wybór rozkładu apriori!

Uwaga:

ponieważ mamy dane $p(x|\theta)$, $p(\theta)$, żeby obliczyć $p(\theta|x)$ musimy obliczyć $p(x) = \int d\theta\, p(x|\theta) p(\theta)$

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{\int d\theta'\, p(x|\theta') p(\theta')}$$

Nie zawsze łatwo wykonać całkę, ale wyrażenie w mianowniku to po prostu normalizacja rozkładu $p(\theta|x)$, bo nie zależy od θ.

3.2 Modele gaussowskie

Przykład

$$x_i \sim N(\theta, \sigma^2) \qquad i = 1 \dots N$$

$$p(x|\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N e^{-\frac{1}{2\sigma^2}\sum_i (x_i - \theta)^2}$$

Weźmy gaussowski rozkład a priori

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{1}{2\sigma_\theta^2}(\theta - \mu_\theta)^2}$$

średnia a priori

Zdefiniujmy
$$\sigma_{\theta|x}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$
$$\mu_{\theta|x} = \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_\theta}{\sigma_\theta^2}\right)\sigma_{\theta|x}^2$$

Czyli:
$$p(\theta|x) = \frac{1}{\sqrt{2\pi\sigma_{\theta|x}^2}}\, e^{-\frac{1}{2\sigma_{\theta|x}^2}\left(\theta - \mu_{\theta|x}\right)^2}$$

Rozkład a posteriori też gaussowski (mieliśmy szczęście nie było problemów)

Widać że jego szerokość
$$\sigma_{\theta|x}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}$$

będzie tym węższa im więcej prób a wpływ rozkładu a priori będzie maleć im większe N, średnia z kolei
$$\mu_{\theta|x} = \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_\theta}{\sigma_\theta^2}\right)\sigma_{\theta|x}^2 =$$
$$= \underbrace{\frac{\frac{N}{\sigma^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}}_{\alpha}\bar{x} + \underbrace{\frac{\frac{1}{\sigma_\theta^2}}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\theta^2}}}_{1-\alpha}\mu_\theta =$$
$$= \alpha \cdot \bar{x} + (1-\alpha)\,\mu_\theta$$

wkład od danych

obciążenie ze średnią a priori

Dla dużych N rozkład a priori nie ma znaczenia: $\alpha \to 1$, im α bliższe 1 tym mniejszy wpływ rozkładu a priori.

Dla tego problemu optymalny estymator jest tym który dąży do zmierzenia (średnia a posteriori)

$$\tilde{\Theta}(x) = \mu_{\Theta|x}$$

A średnia wariancja estymatora

$$\Delta^2\tilde{\Theta} = \int dx\, \Delta^2\Theta|_{p(\Theta|x)} p(x) = \int dx\, \sigma^2_{\Theta|x}\, p(x) = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_\Theta^2}}$$

W granicy $N \to \infty$ $\quad \tilde{\Theta}(x) = \bar{x} \quad \Delta^2\tilde{\Theta} = \frac{\sigma^2}{N}$

Przyjemniej jest pracować z rozkładami a priori dla których a posteriori łatwe do liczenia całkowania itp. Tak jest dla gaussowskich rozkładów a priori i gaussowskich procesów

Bayesowski model liniowy

$$\vec{x} = H\vec{\Theta} + \vec{w} \qquad \vec{x} = [x_1, \ldots, x_N] \qquad \vec{\Theta} = [\Theta_1, \ldots, \Theta_K]$$

$$\vec{w} \sim N(0, C_w)$$

macierz kowariancji szumu

$$p(\vec{x}|\vec{\Theta}) = \frac{1}{\sqrt{(2\pi)^K \det C_w}} e^{-\frac{1}{2}(\vec{x} - H\vec{\Theta})^T C_w^{-1}(\vec{x} - H\vec{\Theta})}$$

ale teraz mamy również $\vec{\Theta}$ traktowane jako zmienna losowa $\quad \vec{\Theta} \sim N(\vec{\mu}_\Theta, C_\Theta)$ – rozkład a priori

Czyli $\quad p(\vec{\Theta}) = \frac{1}{\sqrt{(2\pi)^N \det C_\Theta}} e^{-\frac{1}{2}(\vec{\Theta} - \vec{\mu}_\Theta)^T C_\Theta^{-1}(\vec{\Theta} - \vec{\mu}_\Theta)}$

Czyli $p(\vec\theta) = \frac{1}{\sqrt{(2\pi)^N \det C_\theta}} e^{-\frac{1}{2}(\ldots)}$

Chcemy obliczyć $p(\vec\theta \mid \vec x)$

Liczymy (z dokładnością do stałych i wyrazów nie zależnych od θ):

$$p(\vec x \mid \vec\theta)\, p(\vec\theta) \approx e^{-\frac{1}{2}\left[(\vec x - H\vec\theta)^T C_w^{-1} (\vec x - H\vec\theta) + (\vec\theta - \vec\mu_\theta)^T C_\theta^{-1} (\vec\theta - \mu_\theta)\right]}$$

$$\approx e^{-\frac{1}{2}\left[\vec\theta^T H^T C_w^{-1} H \vec\theta - 2\vec\theta^T H^T C_w^{-1} \vec x + \vec\theta^T C_\theta^{-1} \vec\theta - 2\vec\theta^T C_\theta^{-1} \mu_\theta\right]}$$

Zdefiniujmy $C^{-1}_{\vec\theta|\vec x} = H^T C_w^{-1} H + C_\theta^{-1}$

$$\mu_{\vec\theta|\vec x} = C_{\vec\theta|\vec x} \cdot \left(H^T C_w^{-1} \vec x + C_\theta^{-1} \mu_\theta\right)$$

$$\approx e^{-\frac{1}{2}\left((\vec\theta - \mu_{\vec\theta|x})^T C^{-1}_{\vec\theta|\vec x} (\vec\theta - \mu_{\vec\theta|\vec x})\right)}$$

Czyli rozkład a posteriori:

$$p(\vec\theta \mid \vec x) = \frac{1}{\sqrt{(2\pi)^k \det C_{\vec\theta|\vec x}}} e^{-\frac{1}{2}\left((\vec\theta - \mu_{\vec\theta|\vec x})^T C^{-1}_{\vec\theta|\vec x} (\vec\theta - \mu_{\vec\theta|\vec x})\right)}$$

Optymalny estymator: $\tilde\theta = \mu_{\vec\theta|\vec x}$

$\sigma^2_{\tilde\theta} = C_{\vec\theta|\vec x}$

Uwaga: Jeśli $\vec w, \vec\theta$ nie są gaussowskie to wciąż można użyć estymatora

$$\boxed{\begin{array}{l} \tilde\theta(\vec x) = \mu_{\vec\theta|\vec x} = C_{\vec\theta|\vec x} \cdot \left(H^T C_w^{-1} \vec x + C_\theta^{-1} \mu_\theta\right) \\ \text{[illegible] dla kosztu [illegible]} \\ \sigma^2_{\tilde\theta} = C_{\vec\theta|\vec x} = H^T C_w^{-1} H + C_\theta^{-1} \end{array}}$$

Nie koniecznie będzie to najlepszy estymator, ale można pokazać że będzie najlepszy spośród estymatorów liniowych tzn.:

$$\hat{\theta}(\vec{x}) = \sum_n a_n x_n + a_0$$

Często dobre rozwiązanie i efektywne numerycznie

W granicy $C_\theta^{-1} \to 0$ wracamy do najlepszego liniowego estymatora w estymacji klasycznej!

Można skorzystać z Woodbury identity:
{A, C nieosobliwe

$$(A + UCV^{+})^{-1} =$$
$$= A^{-1} - A^{-1}U(C^{-1} + V^{+}A^{-1}U)^{-1}V^{+}A^{-1}$$

Wtedy:

$$C_{\theta|x} = (C_\theta^{-1} + H^T C_w^{-1} H)^{-1} =$$
$$= C_\theta - C_\theta H^T (C_w + H C_\theta H^T)^{-1} H C_\theta$$

Dowód:

$$(A(1 + A^{-1}UCV^{+}))^{-1} =$$
$$= (1 + A^{-1}UCV^{+})^{-1} \cdot A^{-1} =$$
$$= \left(\sum_n^{\infty} (-1)^n (A^{-1}UCV^{+})^n\right) A^{-1} =$$
$$= \left(1 - A^{-1}UCV^{+} \sum_{n=0}^{\infty} (A^{-1}UCV^{+})^n (-1)^n\right) A^{-1} =$$
$$= \left(1 - A^{-1}U \sum_{n=0}^{\infty} (CV^{+}A^{-1}U)^n CV^{+} (-1)^n\right) A^{-1} =$$

$$\{ (C^{-1} + V^{+}A^{-1}U)^{-1} = (1 + CV^{+}A^{-1}U)^{-1} C =$$

$$(C^{-1} + V^{\dagger} A^{-1} U)^{-1} = (I + C V^{\dagger} A^{-1} U)^{-1} C =$$

$$= \sum_{m} (-1)^{m} (C V^{\dagger} A^{-1} U)^{m} \cdot C =$$

$$= \left(I - A^{-1} U (C^{-1} + V^{\dagger} A^{-1} U)^{-1} V^{\dagger}\right) A^{-1} =$$

$$= A^{-1} - A^{-1} U (C^{-1} + V^{\dagger} A^{-1} U)^{-1} V^{\dagger} A^{-1}$$