Модельные методы обучения с подкреплением
часть 2

Осминин Константин

13 августа 2019

Tinkoff.ru

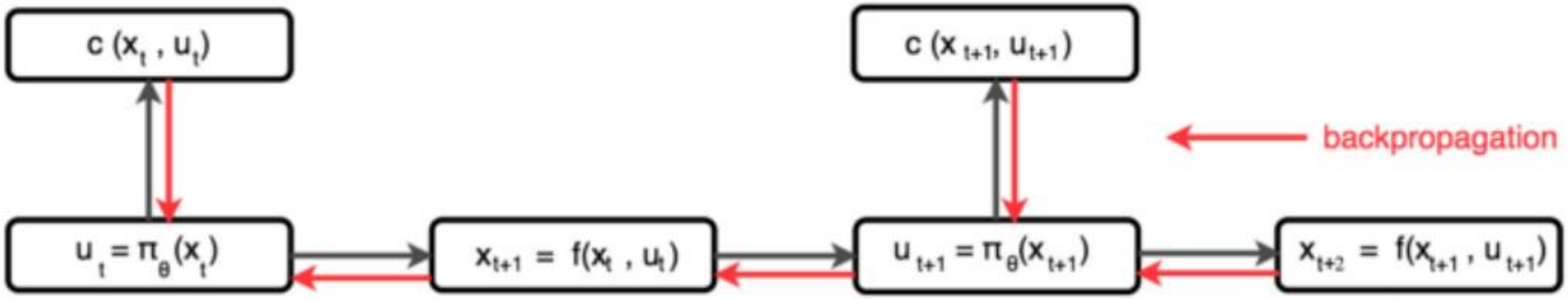
План

- Выучивание динамики среды
- Гауссовский процесс
- PILCO

RL vs control

Обучение с подкреплением	Теория управления
Состояние **s**	Состояние **x**
Действие **a**	Действие **u**
Вознаграждение **r**	Кост **c**
Агент	Контроллер

Выучивание динамики среды



1. run base policy $\pi_0(\mathbf{a}_t|\mathbf{s}_t)$ (e.g., random policy) to collect $\mathcal{D} = \{(\mathbf{s}, \mathbf{a}, \mathbf{s}')_i\}$
2. learn dynamics model $f(\mathbf{s}, \mathbf{a})$ to minimize $\sum_i \|f(\mathbf{s}_i, \mathbf{a}_i) - \mathbf{s}'_i\|^2$
3. backpropagate through $f(\mathbf{s}, \mathbf{a})$ into the policy to optimize $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$
4. run $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$, appending the visited tuples $(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ to $\mathcal{D}$

Гауссовский процесс

- **Определение**: Случайный процесс $\{f_x\}, x \in \mathbb{R}^n$ является гауссовским тогда и только тогда, когда для любого конечного множества индексов $(x_1, .. x_m)$:

  $(f_{x_1}, \dots, f_{x_m})$ есть многомерная гауссовская случайная величина.

- Если $\forall x$: $\mathbb{E} f_x = 0$, то ГП полностью определяется $\mathrm{cov}(f_x, f_y)$.

- Для аппроксимации ковариации используются ядра:

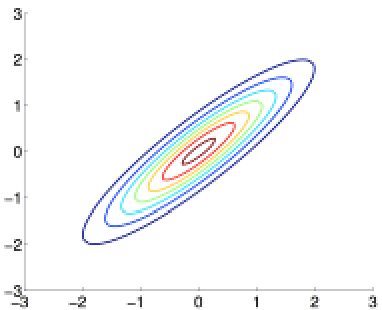
  - Константа: $K_C(x, x') = C$
  - Линейная функция: $K_L(x, x') = x^T x'$
  - Гауссовский шум: $K_{GN}(x, x') = \sigma^2 \delta_{x,x'}$
  - Квадратичная экспоненциальная функция: $K_{SE}(x, x') = \exp\left(-\frac{||d||^2}{2\ell^2}\right)$
  - Функция Орнштейна-Уленбека: $K_{OU}(x, x') = \exp\left(-\frac{|d|}{\ell}\right)$
  - Периодическая функция: $K_P(x, x') = \exp\left(-\frac{2\sin^2\left(\frac{d}{2}\right)}{\ell^2}\right)$

Свойства гауссовского распределения

- **Факт**: Маржинальное и условное распределения гауссовского распределения также гауссовы.

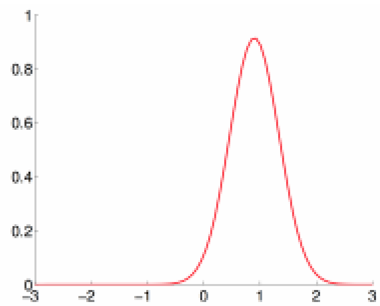
$$\begin{pmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^{\mathrm{T}} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right), \quad p(\mathbf{f}_1) = \int p(\mathbf{f}_1, \mathbf{f}_2) d\mathbf{f}_2 \;= \mathcal{N}(\mathbf{f}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

$$p(\mathbf{f}_1|\mathbf{f}_2) = \mathcal{N}\left(\mathbf{f}_1 | \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{f}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^{\mathrm{T}}\right)$$



$p(\mathbf{f}_1, \mathbf{f}_2) \sim \mathcal{N}(\mathbf{f}_1, \mathbf{f}_2 | \boldsymbol{\mu}, \boldsymbol{\Sigma})$

$p(\mathbf{f}_1)$



$p(\mathbf{f}_1|\mathbf{f}_2)$

Source: DeepBayes2018 Burnaev GP lecture

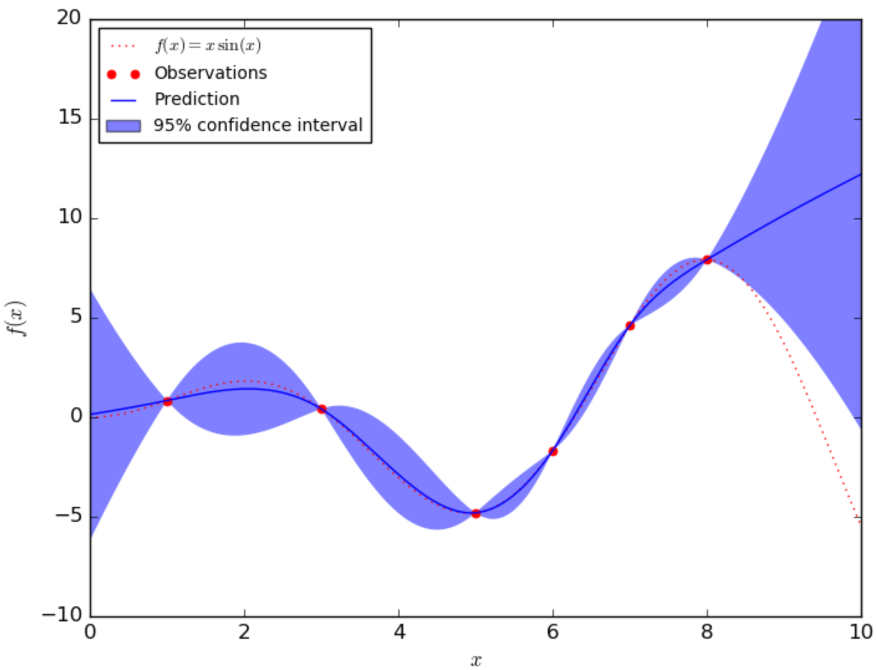
Регрессия гауссовского процесса

- **Задача регрессии:** Дано $(\mathbf{X}, \mathbf{y}) = \{(\mathrm{x_i}, \mathrm{y_i})\}_{\mathrm{i=1}}^{\mathrm{N}}$, $\mathrm{x_i} \in \mathbb{R}^d, y_i \in \mathbb{R}$.

  Нужно найти y^* по x^*.

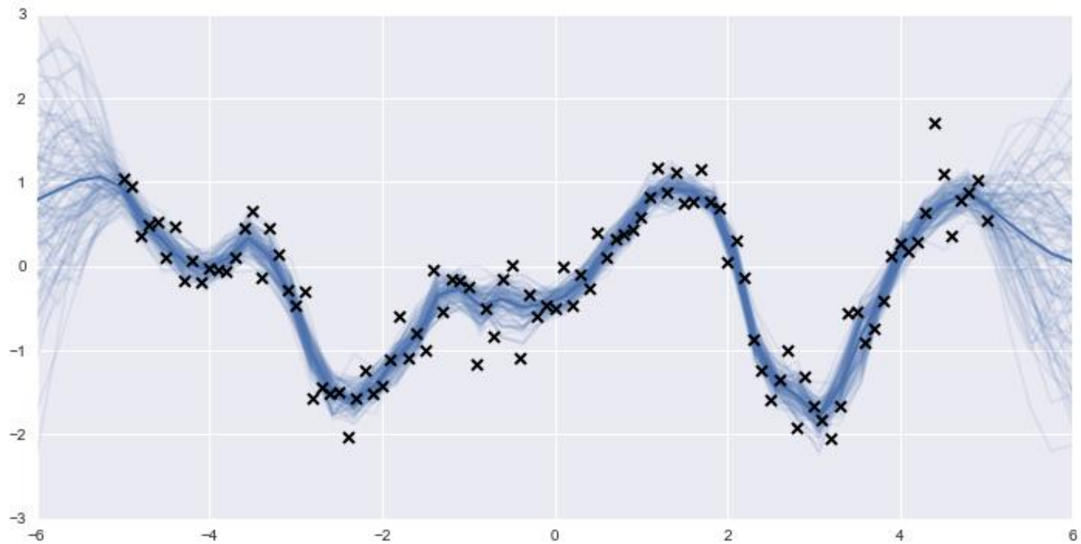
- **Предположение**: $\mathrm{y} = \mathrm{f(x)} + \varepsilon$, где $\mathrm{f(x)}$ - гауссовский процесс со средним 0 и вариацией $k(x\ \ , x')$. $\varepsilon \sim \mathcal{N}\left(o, \sigma_n\ \right)$.

- **Тогда** $\mathrm{y}^*|\mathbf{X}, \mathbf{y}, x^* \sim N\left(m(x^*), \sigma(x^*)\right)$.
  - $m(x^*) = \boldsymbol{k^T K_y^{-1} y},$
  - $\sigma(x^*) = k(x^*, x^*) - \boldsymbol{k^T K_y^{-1} k}$
  - $\boldsymbol{k} \quad = (k(x^*, x_1), \dots, k(x^*, x_N)),$
  - $\boldsymbol{K_y} = \left\|k\left(x_i, x_j\right)\right\|_{\boldsymbol{i,j=1}}^{\boldsymbol{N}} + \sigma_n^2 \boldsymbol{I}$



Свойства регрессии ГП

- Линейна по **y**.
- Условна – нет предположений о распределении **X**.
- Эффективна при малых N и неэффективна при больших.
- Дает не только оценку значения y^*, но и степень неуверенности модели $\sigma(x^*)$.
- (Почти) непараметрична.



PILCO

- PILCO (2011) – probabilistic inference for learning control.
- Моделирует динамику среды гауссовским процессом.
    - Пусть $x_{t-1}, u_{t-1}, x_t, \ldots$ - траектория.
    - Тогда $\Delta_t = x_t$-$x_{t-1} + \varepsilon$ $-$ приращение с шумом $\varepsilon \sim \mathcal{N}\left(o, \sigma_n\right)$.
    - Обозначим $\tilde{x} = [x_t\ u_t]$.
    - Считаем, что Δ_t - гауссовский процесс от $\tilde{x}$ с

$$k(\tilde{x}, \tilde{x}') = \alpha^2 \exp(-0.5\ (\tilde{x} - \tilde{x}')^T \Lambda^{-1}(\tilde{x} - \tilde{x}')),\ \Lambda = \mathrm{diag}\left(\mathrm{l}_1^2, \ldots, \mathrm{l}_\mathrm{D}^2\right).$$

    - Параметры гауссовского процесса $\phi = [\sigma_i, \alpha, l_i]$ определяются из максимизации правдоподобия на обучающей выборке $\widetilde{\boldsymbol{X}}$:

$$\log p(f(x)|\phi\ x) = -\frac{1}{2} f(x)^T K(\phi\ x, x')^{-1} f(x) - \frac{1}{2} \log\det(K(\phi\ x, x')) - \frac{\mathrm{N}}{2} \log 2\pi$$

Probabilistic inference

Дано $\tilde{x}_{t-1} \sim \mathcal{N}\left(\mu'_{t-1}, \Sigma'_{t-1}\right)$. Требуется найти аппроксимацию $x_t \sim \mathcal{N}\left(\mu_t, \Sigma_t\right)$.

$$\mu_\Delta = \mathbb{E}_{\tilde{x}_{t-1}} m(\tilde{x}_{t-1}) = \int m(\tilde{x}_{t-1}) \cdot \boldsymbol{f}_{\mathcal{N}\left(\mu'_{t-1}, \Sigma'_{t-1}\right)}(\tilde{x}_{t-1})\, d\tilde{x}_{t-1} =$$

$$= \left[\boldsymbol{k} = k\left(\widetilde{\boldsymbol{X}}, \tilde{x}_{t-1}\right), \qquad \widetilde{\boldsymbol{X}} - \text{обуч. в} - \text{ка}, \qquad \boldsymbol{K_y} = k\left(\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{X}}\right) + \sigma_n^2 \boldsymbol{I}, \qquad \boldsymbol{y} = \Delta_{\boldsymbol{X}}\right] =$$

$$= \int \boldsymbol{k}^T \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y} \cdot \boldsymbol{f}_{\mathcal{N}\left(\mu'_{t-1}, \Sigma'_{t-1}\right)}(\tilde{x}_{t-1})\, d\tilde{x}_{t-1} = \beta^T \cdot \boldsymbol{q}$$

$$\beta = \boldsymbol{K}_{\boldsymbol{y}}^{-1} \boldsymbol{y}, \quad \boldsymbol{q_i} = \int k(\widetilde{x_i}, \tilde{x}_{t-1}) \cdot \boldsymbol{f}_{\mathcal{N}\left(\mu'_{t-1}, \Sigma'_{t-1}\right)}(\tilde{x}_{t-1})\, d\tilde{x}_{t-1} =$$

$$= \frac{\boldsymbol{\alpha^2}}{\sqrt{\boldsymbol{\Sigma_{t-1} \Lambda_i^{-1} + I}}} \exp\left(-0.5(\widetilde{x_i} - \mu'_{t-1})^T (\Sigma_{t-1} + \boldsymbol{\Lambda_i})^{-1} (\widetilde{x_i} - \mu'_{t-1})\right)$$

$\mu_t = \mu_{t-1} + \mu_\Delta$

Σ_t найти посложнее, но тоже можно.

Целевая функция

- $J(\pi) = \sum_t \mathbb{E}_{x_t}\ c(x_t)$, при условии

  $$x_t - x_{t-1} \sim\ Gauss\ Proc\ (x_{t-1}, \pi_\theta(x_{t-1})), \quad x_0 \sim \mathcal{N}\left(\mu_0, \sigma_0\ \right)$$

- Считаем $c(x) = 1 - \exp(- {|x - x_{target}|^2}/{\sigma_c^2})$

- Контроллер $\pi_\theta(x)$ - произвольный. $\quad \theta = argmin\, J(\pi)$.

- $\mathbb{E}_{x_t} c(x_t) = \int c(x_t) \cdot \boldsymbol{f}_{\mathcal{N}\left(\mu_t, \Sigma_t\ \right)}(x_t)\, dx_t$

- $$\frac{d\mathbb{E}_{x_t} c(x_t)}{d\theta} = \frac{\partial \mathbb{E}_{x_t} c(x_t)}{\partial \mu_t} \cdot \frac{d\mu_t}{d\theta} + \frac{\partial \mathbb{E}_{x_t} c(x_t)}{\partial \Sigma_t} \cdot \frac{d\Sigma_t}{d\theta}$$

- $$\frac{d\mu_t}{d\theta} = \frac{\partial \mu_t}{\partial \mu_{t-1}} \frac{d\mu_{t-1}}{\partial\theta} + \frac{\partial \mu_t}{\partial \Sigma_{t-1}}\ \frac{d\Sigma_{t-1}}{\partial\theta} + \frac{\partial \mu_t}{\partial\theta}$$

- $$\frac{\partial \mu_t}{\partial\theta} = \frac{\partial \mu_\Delta}{\partial \mu_u} \frac{d\mu_u}{\partial\theta} + \frac{\partial \mu_\Delta}{\partial \Sigma_u}\ \frac{d\Sigma_u}{\partial\theta}$$

PILCO алгоритм

Algorithm 1 PILCO

1: **init:** Sample controller parameters $\theta \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
Apply random control signals and record data.
2: **repeat**
3: Learn probabilistic (GP) dynamics model,
4: Model-based policy search,
5: **repeat**
6: Approximate inference for policy evaluation,
get $J^{\pi}(\theta)$, get $\mathrm{d}J^{\pi}(\theta)/\mathrm{d}\theta$
7: Gradient-based policy improvement,
8: Update parameters θ (e.g., CG).
9: **until** convergence; **return** θ^*
10: Set $\pi^* \leftarrow \pi(\theta^*)$.
11: Apply π^* to system (single trial/episode) and
record data.
12: **until** task learned

Эффективность по данным

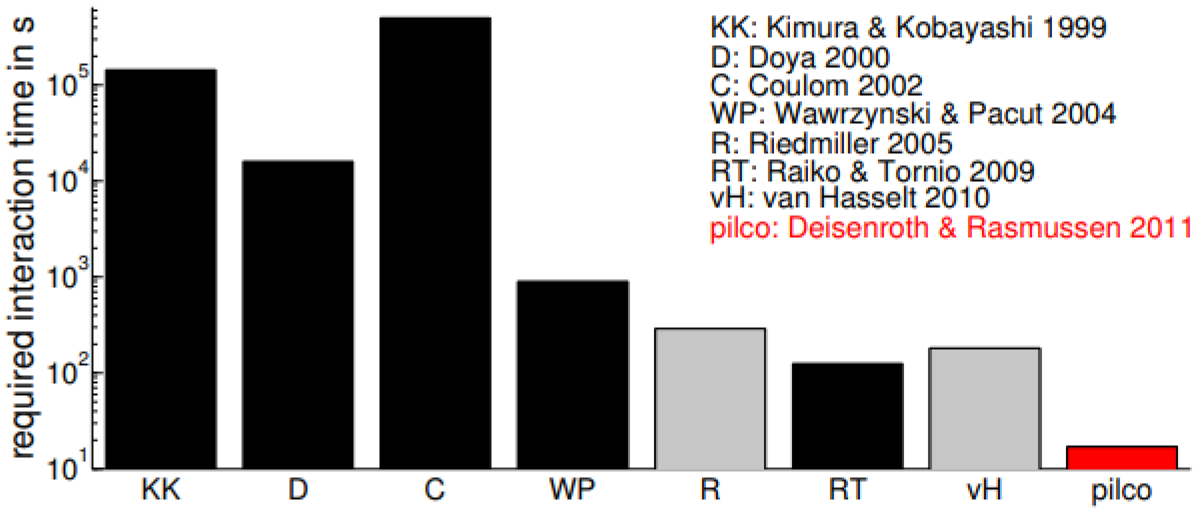


Figure 5. Data efficiency for learning the cart-pole task in the absence of expert knowledge. The horizontal axis chronologically orders the references according to their publication date. The vertical axis shows the required interaction time with the cart-pole system on a log-scale.

Материалы

- Гауссовский процесс
  - вики
  - Лекция DeepBayes2018
- PILCO
  - статья 2011
  - сайт

Семинар

Регрессия гауссовского процесса

https://github.com/bayesgroup/deepbayes-2018/blob/master/day5_gp/gp_basic.ipynb

Спасибо