Bayesian Reinforcement Learning Methods

Using Bayesian MDPs and GPTD Methods

Vickie Ye and Alexandr Wang

May 12, 2016

Markov Decision Processes

- System described by a known set of states S and actions A, and unknown reward function $R(s, a)$ and transition function $T(s, a, s') = P(X^{(t+1)} = s' | X^{(t)} = s, Y^{(t)} = a)$.
- We define a quality function

$$Q = \sum_{t=0}^{\infty} \gamma^t R^{(t)},$$

which we approximate for each state-action pair as

$$Q(s, a) = \mathbb{E}[R(s, a)] + \gamma \sum_{s'} T(s, a, s') \mathrm{max}_{a'} Q(s', a').$$

- To estimate Q, we need to estimate T and R.

Estimating T with a Bayesian model

We model our observed transition counts for each (s, a) as

$$\mathbf{m}^{(t)} \sim \mathrm{Mult}(\pi(s, a))$$
$$\pi(s, a) \sim \mathrm{Dirichlet}(\alpha)$$

where

$$\pi(s, a) = (T(s, a, s_0), \ldots, T(s, a, s_{N-1})),$$

Our posterior is then

$$\pi^{(t)} | D \sim \mathrm{Dirichlet}(\alpha^{(t)} | \mathbf{m}^{(t)}), \ \alpha_i^{(t)} = \alpha_i + m_i^{(t)}$$

Estimating R with a Bayesian model

We model our reward $R(s, a)$ for each state-action pair as

$$\begin{aligned} r(s, a) &\sim \mathcal{N}(\mu, \tau) \\ \mu &\sim \mathcal{N}(\mu_0, c_0\tau) \\ \tau &\sim \mathrm{Ga}(\beta, \rho) \end{aligned}$$

Our posterior is then

$$\tau \sim \mathrm{Ga}\left(\beta + \frac{k}{2}, \rho + \frac{1}{2}\sum_i (r_i - \bar{r})^2 + \frac{kc_0(\bar{r} - \mu_0)^2}{2(n + c_0)}\right),$$

$$\mu \sim \mathcal{N}\left(\frac{k\bar{r} + c_0\mu_0}{k + c_0}, (k + c_0)\tau\right)$$

Testing Problems

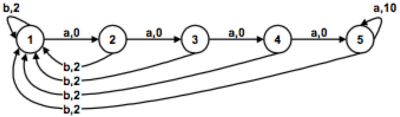


Figure 1: Chain Problem

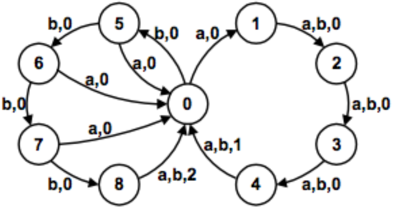


Figure 2: Loop Problem

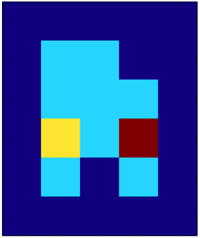


Figure 3: Easy maze

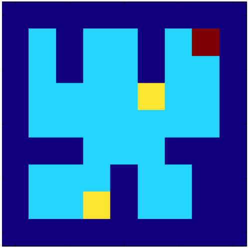
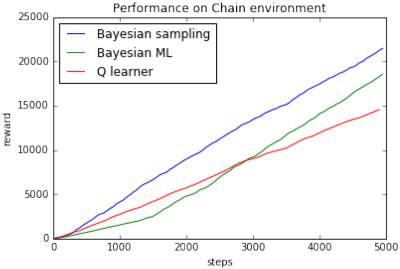
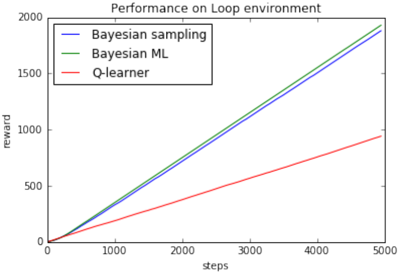


Figure 4: Hard maze

Results on the Chain and Loop problems





Results on the Maze problems