Unyleya
EDUCACIONAL

APRENDIZAGEM POR REFORÇO SEGURO

APRENDIZAGEM POR REFORÇO SEGURO

Conceito e objetivo

Conceito: aprendizado por reforço seguro como o processo de aprender políticas que maximizam a expectativa de retorno em problemas nos quais é importante garantir o desempenho razoável do sistema e/ou respeitar as restrições de segurança durante os processos de aprendizado e/ou implantação

Objetivo: Criar um algoritmo de aprendizado que seja seguro durante o teste e o treinamento.

APRENDIZAGEM POR REFORÇO SEGURO

Exemplos de algumas restrições

O caso de resfriamento do centro de dados, onde temperaturas e pressões devem ser mantidas abaixo dos respectivos limites em todos os momentos.

Um robô que não deve exceder os limites de velocidade, ângulos e torques

Um veículo autônomo que deve respeitar suas restrições cinemáticas.

Esses problema pode ser abordado de duas principais maneiras: **alterando os critérios de otimização** ou **alterando o processo de exploração**.

APRENDIZAGEM POR REFORÇO SEGURO

Critérios de otimização

Existem alguns métodos para incorporar o risco ao objetivo de otimização, são eles:

Critérios de pior caso: uma política é considerada ótima se tiver o retorno máximo do pior caso, ou seja, a pior recompensa obtida pela política é maximizada. Toda a tarefa simplifica a resolução do objetivo *min-max* abaixo:

$$max_{\pi \in \Pi}\ min_{\omega \in \Omega^{\pi}}\ E_{\pi,\omega}\ (R) = max_{\pi \in \Pi}\ min_{\omega \in \Omega^{\pi}}\ E_{\pi,\omega}(\sum_{t=0}^{\infty} \gamma^{t}\ r_{t})$$

Onde Ω é um conjunto de trajetórias da forma (s_0 , a_0 , s_1, a_1,…) que ocorre sob a política π .

APRENDIZAGEM POR REFORÇO SEGURO

Critérios de otimização

Critérios sensíveis ao risco: Inclui a notação de “risco” no objetivo de maximização da recompensa ao longo do prazo. Algumas literaturas definem como a variância do retorno. Considerando um exemplo de sensibilidade ao risco com base na função exponencial, uma função objetiva típica pode se parecer com:

$$max_{\pi \epsilon \Pi}\ \beta^{-1}\ logE_{\pi}(exp^{\beta R}) = max_{\pi \epsilon \Pi}\ \beta^{-1}\ logE_{\pi}(exp^{\beta \sum_{t=0}^{\infty} \gamma^{t}\, rt})$$

Uma expansão de Taylor do termo *exp* e *log* nos dá:

$$max_{\pi \epsilon \Pi}\ \beta^{-1}\ logE_{\pi}(exp^{\beta R}) = max_{\pi \epsilon \Pi}\ E_{\pi}(R) + \frac{\beta}{2}\ Var(R) + \vartheta(\beta^{2})$$

Onde β denota o parâmetro sensível ao risco, com o efeito de que β é negativo tem como o objetivo reduzir a variação nas recompensas e, consequentemente o risco.

APRENDIZAGEM POR REFORÇO SEGURO

Critérios de otimização

Critérios restritos : A expectativa de retorno está sujeita a uma ou mais restrições. A forma geral dessas restrições é mostrada abaixo:

$$max_{\pi \epsilon \Pi} E_{\pi}(R) \text{ sujeito a } c_i \in C,\ c_i = \{h_i \leq \alpha_i$$

Onde c_i são as restrições pertencentes ao conjunto C. Uma política é atualizada se for segura com certa confiança, dadas as restrições.

Obrigada!

hulianeufrn@gmail.com