Unyleya
EDUCACIONAL

PROCESSO DE EXPLORAÇÃO

PROCESSO DE EXPLORAÇÃO

Comportamentos exploratórios clássicos na aprendizagem por reforço

Assumem que o agente tem que explorar e aprender a pesar ações diferentes e a agir de forma otimizada.

O agente ignora o risco de ações, potencialmente terminando em estados perigosos. Explorações como a ε-*greedy* podem resultar em situações desastrosas.

Além disso, as políticas de exploração aleatória desperdiçam uma quantidade significativa de tempo explorando as regiões do estado e do espaço de ação onde a política ideal nunca será encontrada.

PROCESSO DE EXPLORAÇÃO

Comportamentos exploratórios clássicos na aprendizagem por reforço

É impossível evitar completamente situações indesejáveis em ambientes de risco sem conhecimento externo, porque o agente precisa visitar o estado perigoso pelo menos uma vez antes de rotulá-lo como “perigoso”. Pode haver duas maneiras de modificar o processo de exploração: **incorporar conhecimento externo** ou **exploração dirigida ao risco**.

PROCESSO DE EXPLORAÇÃO

Incorporar conhecimento externo

Fornecer conhecimento inicial (pode ser considerado como um tipo de procedimento de inicialização) ou derivar uma política usando um conjunto finito de exemplos.

Exemplo: registrar um conjunto finito de demonstrações de um professor humano e fornecer a ele um algoritmo de regressão, para construir uma função Q parcial que pode ser usada para guiar ainda mais a exploração. Essas abordagens de inicialização não são suficientes para evitar situações perigosas que ocorrem na exploração.

Para derivar uma política de um conjunto de demonstrações, um professor, demonstra uma tarefa e as trajetórias de ações do estado são registradas.

Essas tarefas são usadas para derivar um modelo da dinâmica do sistema, e um algoritmo de aprendizagem por reforço encontra a política ótima nesse modelo.

PROCESSO DE EXPLORAÇÃO

Exploração dirigida ao risco

Uma das abordagens define uma métrica de risco sobre a noção de “controlabilidade”.

Intuitivamente, se um estado particular (ou par de ação de estado) produzir muita variabilidade no sinal de erro de diferença temporal, ele será menos controlável. A controlabilidade do par de ação do estado é definida como:

$$C^{\pi}(s, a) = -\boldsymbol{E}_{\pi}[|\delta_t|| \; s_t = s, \; a_t = a]$$

$$C(s_t, a_t) \leftarrow C(s_t, a_t) - \alpha'(|\delta_t| + C(s_t, a_t))$$

Onde δ é o sinal de erro de diferença temporal. O algoritmo de exploração procura usar controlabilidade como uma heurística de exploração em vez de uma exploração geral de Boltzmann. O agente é encorajado a escolher regiões controláveis do ambiente.

Obrigada!

hulianeufrn@gmail.com