PROCESOS DE DESICIÓN DE MÁRKVO

También conocido por sus siglas MDP (Márkov Decisión Process) es un fenómeno aleatorio dependiente del tiempo para el cual se cumple una propiedad específica.
Un MDP modela un problema de decisión secuencial en donde el sistema evoluciona en el tiempo y es controlado por un agente. Esto está determinado por una función de transición de probabilidad que mapea estados y acciones a otros estados.

Los elementos que conforman a un MDP son:

- Un conjunto finito de estados S: {S1,…,Sn}, donde St denota el estado $s \in S$ al tiempo t.
- Un conjunto finito de acciones que pueden depender de cada estado, A(s), donde$a_t \in A(s)$ denota la acción realizada en un estado s en el tiempo t.
- Una función de recompensa ($R^a_{ss'}$) que regresa un número real indicando lo deseado de estar en un estado $s' \in S$ dado que en el estado $s \in S$ se realizó la acción $a \in A(s)$.
- Una función de transición de estados dada como una distribución de probabilidad ($P^{a}{}_{,}$) que denota la probabilidad de llegar al estado $s' \in S$ dado que se
- tomó la acción $a \in A(s)$ en el estado $s \in S$, que también denotaremos como $\Phi(s, a, s')$.

Dado un estado $s_t \in S$ y una acción $a_t \in A(s_t)$, el agente se mueve a un nuevo estado s_{t+1} y recibe una recompensa r_{t+1}

Métodos para resolver

Los principals métodos para resolver los MDPs son:

- Iteración de valor.
- Iteración de política.
- Programación lineal.

Iteración de valor (Bellman)

Se basa en que si sabemos la solución para el subproblema v*(s'), podemos hallar la solución de v*(s), la idea se basa en iniciar desde la recompensa final e ir retrocediendo para encontrar el **valor** óptimo de cada uno de los estados anteriores.
Se puede escribir combinando la mejora en la política y la evaluación de la política truncada como sigue:

$$V_{k+1}(s) = max_a \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V_k(s')].$$

Algoritmo de iteración de valor

Inicializa $V(s) = 0$ para toda $s \in S$
REPEAT
$\Delta \leftarrow 0$
FORALL$(s \in S)$
$v \leftarrow V(s)$
$V(s) \leftarrow max_a \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^*(s')]$
$\Delta \leftarrow max(\Delta, |v - V(s)|)$
ENDFOR
UNTIL$(\Delta < \theta$ (número positivo pequeño))
Regresa una polítíca determinísta
ENSURE $\pi(s) = argmax_a \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V^*(s')]$

Para espacios muy grandes, el ver todos los estados puede ser computacionalmente muy caro. Una opción es hacer estas actualizaciones al momento de estar explorando el espacio, y por lo tanto determinando sobre qué estados se hacen las actualizaciones. El hacer estimaciones en base a otras estimaciones se conoce también como bootstrapping.

Iteración de política (Howards)

para una política dada, se evalúan los valores V hasta que no cambian dentro de una cierta tolerancia θ.

Algoritmo de iteración de política.

Inicializa $V(s) = 0$ para toda $s \in S$
REPEAT
$\Delta \leftarrow 0$
FORALL$(s \in S$)
$v \leftarrow V(s)$
$V(s) \leftarrow \sum_a \pi(s,a) \sum_{s'} \mathcal{P}^a_{ss'}[\mathcal{R}^a_{ss'} + \gamma V(s')$
$\Delta \leftarrow max(\Delta, |v - V(s)|)$
ENDFOR
UNTIL$(\Delta < \theta$ (número positivo pequeño))
regresa $V \approx V^\pi$

Una de las razones para calcular la función de valor de una política es para tratar de encontrar mejores políticas. Dada una función de valor para una política dada, podemos probar una acción a≠ π(s) y ver si su V (s) es mejor o peor que el V $^{\pi}$(s).

Ejemplo de funcionamiento del algoritmo

Inicialización: $V(s) \in \mathcal{R}$ y $\pi(s) \in \mathcal{A}(s)$ arbitrariamente $\forall s \in S$
Evaluación de política:
REPEAT
$\Delta \leftarrow 0$
FORALL$(s \in S)$
$v \leftarrow V(s)$
$V(s) \leftarrow \sum_{s'} \mathcal{P}_{ss'}^{\pi(s)}[\mathcal{R}_{ss'}^{\pi(s)} + \gamma V(s')]$
$\Delta \leftarrow max(\Delta, |v - V(s)|)$
ENDFOR
UNTIL$(\Delta < \theta$ (número positivo pequeño))
Mejora de política:
pol-estable $\leftarrow$ true
FORALL$(s \in S$:)
$b \leftarrow \pi(s)$
$\pi(s) \leftarrow \text{argmax}_a \sum_{s'} \mathcal{P}_{ss'}^{a}[\mathcal{R}_{ss'}^{a} + \gamma V(s')]$
IF$(b \neq \pi(s))$
pol-estable $\leftarrow$ false
ENDIF
ENDFOR
IF$($*pol-estable*$)$
stop
ELSE
go to **Evaluación de política**