Inteligencia Artificial

Tema 4 – Procesos de Decisión de Markov

1. Introducción

Vamos a tratar secuencias de acciones cuyos efectos son inciertos: similares a espacios de estados, pero el efecto de una acción está descrito como una distribución de probabilidad. El resultado de una acción es determinista. Además, se introduce el concepto de "recompensa" para un estado.

Se busca cual es la mejor acción a realizar en cada momento, ¿cómo se decide? → sabiendo de antemano sus posibles efectos y con que probabilidad, además de la recompensa en cada caso.

2. El modelo matemático

Un proceso de decisión de Markov viene definido por:

a. Un conjunto *S* de estados (con un estado inicial *S0*)
b. Para cada estado, un conjunto *A(s)* de acciones aplicables al estado
c. Un modelo de transición, dado por una distribución de probabilidad *P(s' | s,a)* para cada par de estados "*s'*", "*s*" y acción aplicable a "*s*" "*a*", que indica la probabilidad de que aplicando "*a*" a "*s*" se obtenga "*s'*".
d. Una función de recompensa *R(s)*.

Propiedad de Markov: el efecto (incierto) de una acción sobre un estado solo depende de la acción y de propio estado (no de estados anteriores). En consecuencia, como hemos dicho, una solución nunca puede ser una secuencia de acciones ya que efecto de cada una es incierto. Lo que buscamos más bien es una "política" que en cada posible estado recomiende que acción aplicar: formalmente una política es una función π definida sobre el conjunto de estados *S* tal que $\pi(s) \in A(s)$.

Una política puede generar secuencias de acciones distintas (unas con más probabilidad que otras) pero buscamos siempre la política "óptima" que maximice la recompensa media esperada para las posibles secuencias de acciones que se generan.

Valoración de una secuencia de estados en el tiempo: suponiendo que mediante una aplicación se genera una secuencia de estados *q0, q1, q2*, …. ¿Cómo valoramos una secuencia de estados? Usamos la valoración mediante recompensa con descuento:

$$V([q0, q1, q2, \cdots]) = R(q0) + \gamma * R(q1) + \gamma^2 * R(q2) + \cdots$$

donde γ es el llamado "factor de descuento".

Las valoraciones en general no son infinitas en tanto que, si hay estados terminales, los asimilamos a estados a partir de los cuales las recompensas son cero. Aún así, si las recompensas están acotadas por una cantidad *Rmax* e $\gamma < 1$, entonces la valoración de una secuencia no puede ser mayor a $Rmax/(1 - \gamma)$.

Dado una política π y un estado *s*, podemos valorar *s* respecto de π teniendo en cuenta la valoración de las secuencias de estados que se generan si se sigue dicha política a partir de *s*. La valoración de un estado respecto de una política π la denotamos como *V^π(s).*

Por suerte, para calcular *V^π(s)* no necesitamos calcular todas las posibles secuencias que podrían iniciarse en *s*. Calcular *V^π* es resolver el siguiente sistema de ecuaciones lineales:

$$V^{\pi}(s) = R(s) + \gamma \cdot \sum_{s'}(P(s'|s, \pi(s)) \cdot V^{\pi}(s'))$$

donde las incógnitas son los *V^π*, una por cada estado *s*.

Llamamos muestreo a generar secuencias de estados aplicando las acciones que indica la política teniendo en cuenta la distribución de probabilidad y calcular la media de sus valoraciones:

a. Comenzamos por un valor arbitrario $V_0^{\pi}(S)$ para cada estado *S*.
b. En cada iteración se aplica la fórmula que relaciona la valoración de un estado con la de sus vecinos:

$$V_{i+1}^{\pi}(s) = R(s) + \gamma \cdot \sum_{s'}(P(s'|s, \pi(s)) \cdot V_i^{\pi}(s'))$$

Políticas óptimas y valoraciones de estados:

- La valoración de un estado s denotada como *V(s)* se define como: *V(S) = max(π) V^π(S)*
- La política óptima se define como *π** tal que *π*(s) = argmax(sumatorio(P(s'|s,a) * V(s'))* donde *a* pertenece a *A(s)*
- V^π* = V y nuestro objetivo es calcular V y π*

3. <u>Ecuaciones de Bellman</u>

De manera análoga a V^π, describimos V(s) en función de las valoraciones de los estados vecinos. Se conoce esto como ecuaciones de Bellman:

$$V(s) = R(s) + \gamma \cdot \max_{a \in A(s)} \sum_{s'}(P(s'|s, a) \cdot V(s'))$$

La solución a este sistema (una ec. por estado) nos da la valoración de cada estado, y en consecuencia, la política óptima. Para resolver las ecuaciones de Bellman plantemos un método iterativo, el muestreo visto arriba. Se puede demostrar que los valores que se obtienen convergen asintóticamente hacia la solución (única si γ < 1) de las ecuaciones. La convergencia es rápida para valores de γ pequeños y viceversa.

Algoritmo de iteración de valores para calcular las ecuaciones de Bellman:

Entrada:

- Un proceso de decisión de Markov: conjunto de estados S, $A(s)$, modelo de transición $P(s'|s,a)$, recompensa $R(s)$ y descuento γ
- $\epsilon > 0$, cota de error máximo permitido

Salida:

- Una valoración $V(s)$ para cada estado

Procedimiento:

- Inicio: Sea V_0 una función sobre los estados, con valor 0 para cada estado, $\delta = \infty$ e i igual a 0
- Repetir
  - Hacer $i = i + 1$ y $\delta = 0$
  - Para cada $s \in S$ hacer:
    - $V_i(s) \longleftarrow R(s) + \gamma \cdot \max_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V_{i-1}(s'))$
    - Si $|V_i(s) - V_{i-1}(s)| > \delta$ entonces $\delta = |V_i(s) - V_{i-1}(s)|$

  hasta que $\delta < \epsilon \cdot (1-\gamma)/\gamma$
- Devolver V_i

4. <u>Iteración de Políticas</u>

Existe un método alternativo de obtener V, comenzando con una política inicial π0:

a. Calcular V^πi.
b. A partir de V^πi calcular una nueva política πi+1 que en cada estado recomiende la acción que mejor valoración espera, respecto de V^πi.
c. Los dos pasos anteriores se iteran hasta que πi y πi+1 coinciden, en cuyo caso hemos alcanzado la política óptima y su valoración asociada, es V.

El proceso anterior termina en tanto que hay un nº finito de políticas distintas. El algoritmo de iteración de políticas:

Entrada:

- Un proceso de decisión de Markov: conjunto de estados $\boldsymbol{S}$, $\boldsymbol{A(s)}$, modelo de transición $\boldsymbol{P(s'|s,a)}$, recompensa $\boldsymbol{R(s)}$ y descuento γ
- Un número $\boldsymbol{k}$ de iteraciones

Salida: Una política óptima π y su valoración asociada

Procedimiento:

- Inicio: π una política aleatoria, asignando una acción a cada estado
- Repetir
  - Calcular $\boldsymbol{V^{\pi}}$
  - Hacer ***actualizada*** igual a ***Falso***
  - Para cada estado $\boldsymbol{s}$:
    - Si $\max_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V^{\pi}(s')) > \sum_{s'} (P(s'|s,\pi(s)) \cdot V^{\pi}(s'))$, entonces hacer $\pi(s)$ igual a $\operatorname{argmax}_{a \in A(s)} \sum_{s'} (P(s'|s,a) \cdot V^{\pi}(s'))$ y ***actualizada*** igual ***Verdad***

    hasta que ***actualizada*** sea ***Falso***
- Devolver π y $\boldsymbol{V^{\pi}}$