DEBER DE IA

Nombre: Vinicio Veletanga

Realizar una breve Investigación sobre los procesos de decisión de Markov.

Los procesos de decisi´on o control markovianos modelan sistemas din´amicos estoc´asticos controlados, es decir, sistemas cuya evoluci´on est´a sujeta a factores aleatorios y que puede modificarse por medio de la selecci´on de ciertas variables de decisi´on o de control.

3.3 Procesos de Decisión de Markov

3.3.1 Formalización

Un MDP es un modelo matemático de un problema el cual explícitamente considera la incertidumbre en las acciones del sistema. La dinámica del sistema está determinada por una función de transición de probabilidad.

Por otra parte, para cualquier MDP siempre hay una política $\pi : S \rightarrow A$ óptima, que permite decidir en cada estado qué acción tomar de manera que se maximice la utilidad esperada. Esta política π es estacionaria, i. e. no cambia en función del tiempo, y determinista, i. e. siempre se elige la misma acción cuando se está en el mismo estado. Formalmente, un MDP M es una tupla $M = S, A, \Phi, R$, donde: S es un conjunto finito de estados del sistema. A es un conjunto finito de acciones, que se ejecutan en cada estado. $\Phi : AxS \rightarrow \Pi(S)$: es la función de transición de estados dada como una distribución de probabilidades y la cual asocia un conjunto de posibles

estados resultantes de un conjunto de acciones en el estado actual. La probabilidad de alcanzar un estado s' realizando la acción a en el estado s se escribe $\Phi(a,s,s)'$. R: S x A → R es una función de recompensa. R(s, a) es la recompensa que el sistema recibe si lleva a cabo la acción a en el estado s. Esta función define la meta que se quiere alcanzar

3.3.2 Políticas

Puesto que en un MDP el estado actual del entorno se considera completamente observable, i. e. que el agente sabe con exactitud en qué estado se encuentra, el único problema a resolver es determinar la acción a ejecutar en función de dicho estado, para conseguir el objetivo final. No se trata de un problema trivial, puesto que el efecto de las acciones no es determinístico. La única información de la cual se dispone para determinar la acción óptima a ejecutar es: la función de transición de estados T que caracteriza la incertidumbre en el efecto de las acciones, y la función de recompensa R que está relacionada con la utilidad de las acciones en función del objetivo final.

Existen dos tipos de políticas: las estacionarias y las no estacionarias. Una política estacionaria ($a = \pi (s)$) especifica directamente, para cada estado, la acción a realizar independientemente del momento o tiempo en que se encuentra el proceso. Este tipo de políticas se utiliza

principalmente con procesos de horizonte-infinito, puesto que al no existir un límite de pasos, no existe ningún motivo para que el agente cambie su estrategia al elegir las acciones. Una política no estacionaria(a (s)) = π t , sin embargo, asigna una acción u otra al mismo estado en función del momento en que se encuentra 310 el sistema. Este tipo de políticas se utiliza en procesos de horizonte finito, puesto que es conveniente modificar la estrategia a seguir en función del número de pasos que quedan para finalizar el proceso. Así, la política a (s) = π t permite seleccionar la acción a ejecutar en el estado s cuando quedan t pasos para finalizar el proceso. Para un mismo MDP pueden definirse múltiples políticas. Sin embargo, una política será tanto mejor cuanto mayor sea la recompensa que obtiene a largo plazo, siendo éste el criterio que permite seleccionar entre todas ellas, una política óptima.

3.3.3 Función de valor

La función de valor V (s) π determina la utilidad de cada estado s suponiendo que las acciones se escogen según la política π . Se trata de un concepto distinto al de recompensa. La función de recompensa asigna, para cada una de las acciones que pueden ejecutarse en cada estado, un valor numérico que representa la utilidad inmediata de dicha acción. Sin embargo, la función de valor asigna a cada estado un valor numérico que representa la utilidad de dicho estado a largo plazo. Este valor numérico no es la suma exacta de recompensas futuras, que no se puede predecir, sino una aproximación probabilística que se calcula a partir de las funciones de transición ϕ y de recompensa R. Por ejemplo, un agente puede realizar en cierto estado una acción con recompensa positiva, que sin embargo le lleve a un estado que tenga un valor de utilidad negativo. Desde este punto de vista, es mucho más benéfico realizar la acción que prometa una mejor recompensa a largo plazo.

Supóngase una política π en un contexto de horizonte finito con k pasos. La función de valor asociada a esta política, ( ) , V s π k es la recompensa total “esperada” (no real) al ejecutar la política π empezando en el estado s durante k pasos. Obviamente, si k=1, ( ) (, ( )) 1, 1 V s r s π s π = ; es decir, si sólo se dispone de un paso o el proceso se encuentra en el último paso, la función de valor coincide con la recompensa obtenida al ejecutar la acción dada por la política

$$V_{\pi,t}(s) = r(s,\pi_t(s)) + \gamma \sum_{s' \in S} p(s' \mid s,\pi_t(s)) \cdot V_{\pi,t-1}(s')$$

3.3.4 Políticas óptimas

Resolver un MDP consiste en encontrar la política óptima, una directiva de control que maximiza la función de valor sobre los estados. Teóricamente, es posible obtener todas las posibles políticas para un MDP y a continuación escoger entre ellas, aquella que maximiza la función de valor. Sin embargo, este método es computacionalmente costoso, puesto que el

número de políticas crece exponencialmente con el número de estados. Existen, sin embargo, otros métodos que permiten seleccionar la política óptima aprovechando la propiedad de que ésta será también localmente óptima para cada estado individual. Howard (1960) demostró que, para el caso más general de horizonte-infinito, existe una política estacionaria π * que es óptima para cualquier estado inicial del proceso.

3.4 Aprendizaje por refuerzo

El aprendizaje automático propone métodos específicos que permiten a los robots autónomos aprender de su interacción con el entorno, y además, aprender mientras se encuentran inmersos en dicho entorno. 316 En algunos ambientes, muchas veces se puede obtener sólo cierta retroalimentación, recompensa o refuerzo, e. g. valores de ganancia o perdida. El refuerzo puede darse en un estado terminal y/o en estados intermedios. Los refuerzos pueden ser componentes o sugerencias de la utilidad actual a maximizar, e. g. buen movimiento. En aprendizaje por refuerzo (RL, Reinforcement Learning) el objetivo es aprender cómo mapear situaciones a acciones para maximizar una cierta señal de recompensa.

También hay gran flexibilidad en la manera de diseñar los estados. Un estado puede consistir en las percepciones directas del robot recibidas en un momento dado, o puede estar compuesto en parte por datos almacenados Robot Entorno Acción at rt st st+1 rt+1 319 de percepciones anteriores. Incluso pueden existir estados completamente abstractos.