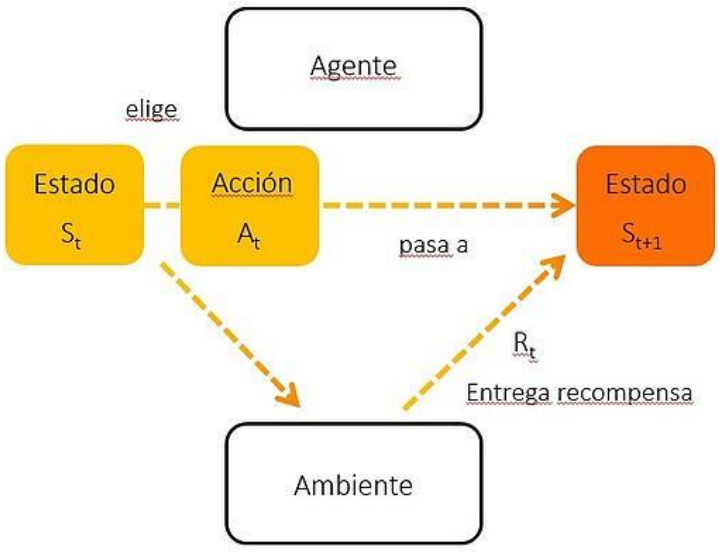
Tarea 6 – Introducción al Aprendizaje Automático
Humberto Gerardo Peña Páez.

Instrucciones:
Objetivo: Que el alumno tenga un punto de partida para consultar temas relacionados al aprendizaje automático.
Entregable 1: Resumen en pdf.
Alcance: No se espera un desarrollo detallado de los conceptos, sino un RESUMEN DE LAS IDEAS EN PALABRAS DEL ALUMNO acompañado con FUENTES DE CONSULTA para ampliar el contexto cuando sea necesario. El resumen debe contener (pero no necesariamente limitarse a) los siguientes temas:

- Aprendizaje reforzado
- Proceso de markov
- Funcion de valor
- Política óptima
- Q - learning
- Policy - Learning
- Aplicaciones

Aprendizaje Reforzado

Es un área del aprendizaje automático que está inspirado en la psicología conductista, cuya ocupación es determinar que acciones debe escoger un agente de software en un entorno dado con el fin de maximizar alguna noción de recompensa o premio acumulado. Es comúnmente utilizado cuando los datos históricos no son suficiente para modelar nuestro problema, ya que esta área coloquialmente es donde las maquinas exploran su ambiente y van aprendiendo de cero en base a recompensas y penalizaciones. A grandes rasgos el siguiente diagrama muestra cómo funciona un algoritmo de aprendizaje reforzado.



Elementos del aprendizaje reforzado

- **Políticas.** Es una tabla (aunque puede tener n-dimensiones) que le indicará al modelo “como actuar” en cada estado.
- **Acciones.** Son las diversas elecciones que puede hacer el agente en cada estado.
- **Recompensas.** Si sumamos o restamos puntaje con la acción tomada.
- **Comportamiento “avaro”**(greedy) **del agente.** Si se dejará llevar por grandes recompensas inmediatas o irá explorando y valorando las riquezas a largo plazo.

Proceso de Markov

Es una serie de experimentos en que cada uno tiene m posibles resultado y la probabilidad de cada resultado depende exclusivamente del que se haya obtenido en los experimentos previos. Es un proceso estocastico con la propiedad de Markov, o sin memoria, es uno para el cual la probabilidad condicional sobre el estado presente, futuro y pasado del sistema son independientes. Surgen en una de las dos siguientes maneras:

- Un proceso estocástico que se define a través de un argumento separado puede demostrarse (matemáticamente) que tiene la propiedad de Markov y como consecuencia tiene las propiedades que se pueden deducir de esta para todos los procesos de Markov.
- De más importancia practica es el uso de la suposición que la propiedad de Markov es válida para un proceso aleatorio con el fin de construir un modelo estocástico para este proceso.

Procesos de primer orden

Algunos procesos de primer orden son:

- Estados: Las condiciones en las cuales se encuentra un ente o sucesos posibles.
- Ensayos: Las ocurrencias repetidas de un evento que se estudia.
- Probabilidad de transición: La probabilidad de pasar de un estado actual al siguiente en un periodo o tiempo y se denota por p_{ij} (la probabilidad de pasar del estado i al estado j en una transición o periodo).

Función de valor

Es la recompensa total que un agente puede acumular empezando en ese estado (predicciones de recompensas). Esta función sirve de base para escoger la acción a realizar (aquella que conduzca al estado con mayor valor) ya que se buscan hacer acciones que den los valores más altos, no la recompensa mayor. El objetivo de los algoritmos de aprendizaje por refuerzo es construir esta función y aprender de ellas mientras interactúa con el ambiente.

Política Óptima

Es definida como la politica de maximizar el valor de todos los estados al mismo tiempo. Si existe una política óptima, entonces la política que maximiza el valor del estado s es la misma que la política que maximiza el valor del estado s'.

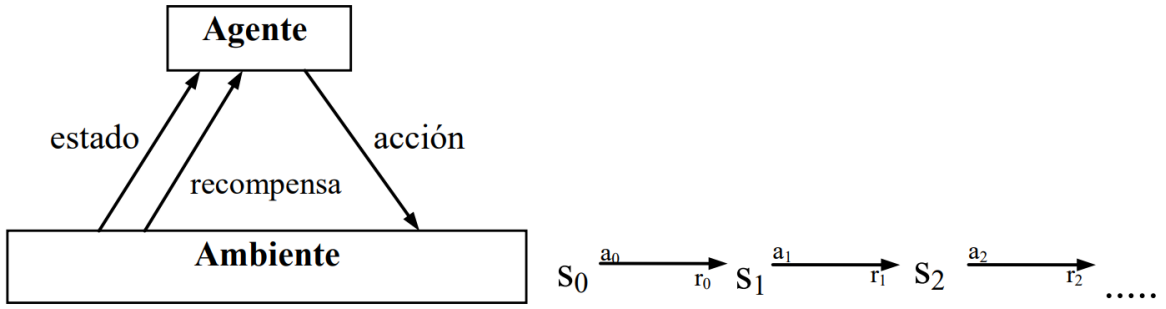
La política π^* es una póliza óptima si y solo si tenemos

$$V_{\pi^*}(S) \geq V_{\pi}(S)$$

para cualquier estado s y cualquier otra política π.

Q – learning

Es un método del aprendizaje reforzado que permite resolver problemas de decisión secuencial y también tareas que pueden ser modeladas como procesos de decisión Markoviano (Markov Decision Processes). La idea general del modelo es tener un agente conectado a algo que denominaremos ambiente vía percepción y acción. La manera en la que interactúa el algoritmo se puede visualizar gráficamente con la siguiente figura.



A medida que el agente se mueve hacia adelante desde un viejo estado a uno nuevo, el algoritmo propaga las estimaciones de Q hacia atrás desde el nuevo estado al viejo. Dado que el agente es responsable tanto de explotar la información disponible, como de explorar nuevos estados y acciones para mejorar sus estimaciones de las funciones de valor, este dilema es conocido como el dilema de exploración/explotación y es común en este método y en la mayoría de los métodos adaptivos.

Aplicaciones

El aprendizaje reforzado se utiliza cuando la información pasada no es tan certera y por lo tanto no se nos permite obtener una predicción tan buena como para tomarla. Aunque también muchas aplicaciones son utilizadas para juegos y control, ya que aquí siguen aprendiendo y interactuando con el ambiente en cuestión.

La primera aplicación en aprendizaje reforzado fue el programa para jugar damas de Samuel. Este utilizó una función lineal de evaluación con pesos usando hasta 16

términos. Su programa era parecido a la ecuación de actualización de pesos, pero no usaba recompensa en los estados terminales. Esto hace que pueda o no converger y puede aprender a perder. En el año 2000 se estaba desarrollando un algoritmo de aprendizaje reforzada que actualizaba las funciones de evaluación en un árbol de búsqueda de juegos.

Referencias

- “Aprendizaje por refuerzo, ¿cómo las máquinas aprenden desde cero?”, Heather Andrwes; consultado en: www.kudaw.com.
- “Aprendizaje por refuerzo”; consultado en: www.wikipedia.com.
- “Proceso de Márkov”; consultado en: www.wikipedia.com.
- “Aprendizaje por refuerzo”; consultado en: http://cayetanoguerra.github.io
- “El proceso de Markov”, Mario Guido Pérez; consultado en: www.dinamica-de-sistemas.com.
- “Learning an Optimal Policy: Model-free Methods”, Leslie Pack Kaelbling; consultado en: www.cs.cmu.edu.
- “Aprendizaje por Refuerzo”, Juan Ignacio Bagnato; consultado en www.aprendemachinelearning.com.
- “Why does the optimal policy exist?”, Alireza Modirshanechi; consultado en: http://towardsdatasciencia.com.
- “Aprendizaje por Refuerzo”; consultado en: http://inaoep.mx
- “Una implementación paralela del algoritmo de Q-Learning basada en un esquema de comunicación con caché”, Alicia Printista, Marcel Errecalde; consultado en: http://sedici.unlp.edu.ar.

Aprendizaje por refuerzo ¿Cómo las máquinas aprenden desde cero? (kudaw.com)

Aprendizaje por refuerzo - Wikipedia, la enciclopedia libre

Proceso de Márkov - Wikipedia, la enciclopedia libre

El proceso de Markov (dinamica-de-sistemas.com)

Microsoft PowerPoint - Tema 6.ppt (cayetanoguerra.github.io)

http://sedici.unlp.edu.ar/bitstream/handle/10915/23363/Documento_completo.pdf?sequence=1&isAllowed=y#:~:text=Q-Learning%20es%20un%20método,que%20está%20situado%20el%20agente.

https://www.aprendemachinelearning.com/aprendizaje-por-refuerzo/

https://ccc.inaoep.mx/~emorales/Cursos/Busqueda/refuerzo.pdf

https://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a-html/node23.html

https://towardsdatascience.com/why-does-the-optimal-policy-exist-29f30fd51f8c