Respuestas escritas Tarea 4 IA

Tomás Couso Coddou

Junio 2022

Actividad 1: Comprendiendo los hiperparámetros

- ¿Cuál es la funciónń del exploration rate?

  El exploration rate $\epsilon \in [0, 1]$ tiene como función definir la proporción de exploración/explotación del agente a través de las iteraciones. Esto es posible en tanto la forma en que el agente elige una acción se determina según las siguientes probabilidades:

  $1 - \epsilon$: Probabilidad de explotación; agente toma una acción en base a la Q-table.

  ϵ: Probabilidad de exploración; agente toma una acción aleatoria.

  El exploration rate está diseñado de manera tal que permite proporiciones diferentes de exploración/explotación a medida que transcurren las iteraciones. Esto se ve en el hecho de que actualiza su valor en base a la siguiente expresión:

  $$\epsilon = \epsilon_{min} + (\epsilon_{max} - \epsilon_{min})e^{-decay\_rate * iter\_num}$$

  El decaimiento exponencial de la anterior expresión implica que durante el proceso de entrenamiento, el agente comienza explorando en alta medida, pero en la medida en que ϵ decae, el agente tiende a preferir explotar el conocimiento del que dispone en vez de realizar acciones aleatorias.

  Por último, el exploration rate también tiene influencia en el término de las iteraciones de Q-Learning, pues el criterio de finalización del algoritmo establece que se termina de iterar una vez que $\epsilon = \epsilon_{min}$.

- ¿Cuál es el problema de tener un exploration rate mínimo con un valor muy bajo y un exploration decay rate con un valor muy alto?

  En dicho escenario, el valor del exploration rate al ser actualizado queda de la siguiente manera:

  $$\epsilon = \epsilon_{min} + (\epsilon_{max} - \epsilon_{min})e^{-M * iter\_num}$$

  donde M refiere a una constante de alto valor. Si aproximamos la anterior expresión, llegamos a lo siguiente.

$$\epsilon \approx \epsilon_{max} e^{-\infty} = 0$$

El problema de tener un exploration rate definido de tal manera es que llevaría a un valor de ϵ muy bajo, de manera tal que el agente tendría un sesgo muy marcado hacia la explotación. En la práctica, esto repercutiría en que las acciones llevadas a cabo serán muy repetitivas y poco flexibles.

- ¿Qué pasa si el exploration decay rate es demasiado bajo?

  En dicho escenario, el valor del exploration rate al ser actualizado sería:

  $$\epsilon = \epsilon_{min} + (\epsilon_{max} - \epsilon_{min}) e^{-m * iter\_num}$$

  donde m refiere a una constante de bajo valor. Si aproximamos la anterior expresión , llegamos a los siguiente:

  $$\epsilon \approx \epsilon_{min} + (\epsilon_{max} - \epsilon_{min}) = \epsilon_{max}$$

  Esto indica que para valores muy bajos del exploration decay, el exploration rate va a tender a ser constante. Aquello es problemático, por dos motivos. Primero, dado que el criterio de convergencia es $\epsilon = \epsilon_{min}$, bajo el anterior escenario Q-Learning podría no converger, o demorar muchas iteraciones. Segundo, si el exploration rate es constante y con un valor cercano a ϵ_{max}, el agente tendrá el mismo exploration rate durante todas las iteraciones. Eso podría ser problemático para el desempeño del agente, pues mantener constante la tendencia a explorar implica que no se aprovecha el conocimiento almacenado en la Q-table, lo que en última instancia lleva a que el agente se comporte de manera excesivamente flexible, casi asistemática.

- ¿En qué tipo de situaciones se podrían reflejar los cambios en estos hiperparámetros, contextualizado al juego de Pong?

  El problema del exploration rate mínimo muy bajo y el exploration decay rate muy alto puede llevar a un comportamiento repetitivo en el agente. Para el caso del Pong, esto puede traducirse en un mal desempeño del agente para casos menos típicos, como podría ser una bola lanzada hacia una esquina que rebota con el muro poco antes de alcanzar el extremo opuesto. Un agente con acciones poco flexibles podría verse sesgado a simplemente seguir la dirección de la bola, sin anticipar el cambio de dirección final.

  El problema del exploration decay rate demasiado bajo puede llevar a comportamientos asistemáticos en el agente, que se verían explicados por la baja explotación de las políticas aprendidas. Para el Pong, esto podría verse en un comportamiento sesgado hacia la exploración; por ejemplo, podría pasar que frente a una situación estandar, como una bola en dirección horizontal, el agente decidiera no interceptar la trayectoria, sino que moverse hacia un lado.

Actividad 2: Análisis de parámetros del agente

- ¿Qué rol cumple la tasa de descuento en Q-Learning?

  La tasa de descuento γ es empleada al actualizar los valores de la Q-table:

  $$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r(s,a) + \gamma argmax_{a'}\{Q(succ(s,a)a')\}) \quad (1)$$

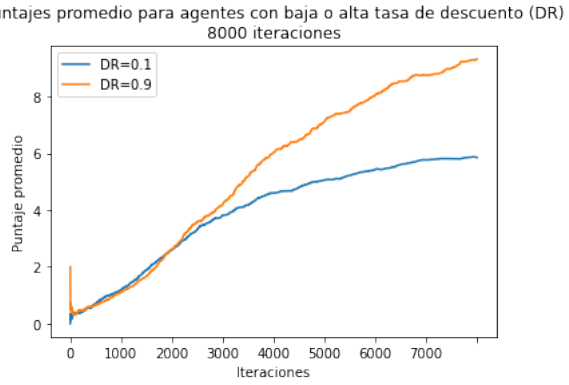
  La anterior expresión indica que cómo acutalizar el valor de ejecutar la acción a desde el estado s. Para efectos de la tasa de descuento, basta que analicemos la siguiente porción:

  $$\gamma argmax_{a'}\{Q(succ(s,a)a')\})$$

  Este componente indica que Q-Learning, al momento de llenar la Q-table, considera la recompensa de las eventuales acciones que el agente puede tomar desde el estado s. No obstante, no existe certeza de que alcancemos el estado sucesor de s óptimo de s, pues el agente podría eventualmente explorar en vez de explotar. De tal modo, γ es un ponderador que cumple la función de tomar en cuenta la incerteza de las recompensas lejanas en el tiempo. Nótese a este respecto que debe cumplirse que $\gamma < 1$, pues de lo contrario, se estaría expresando que las recompensas futuras son una certeza, lo cual no se cumple cuando tenemos un agente que explora.

- ¿Qué tasa de descuento te dio mejores resultados? ¿Por qué crees que fue así?

  A continuación, se detalla el desempeño del agente para una tasa de descuento baja (0.1) y para una alta (0.9).

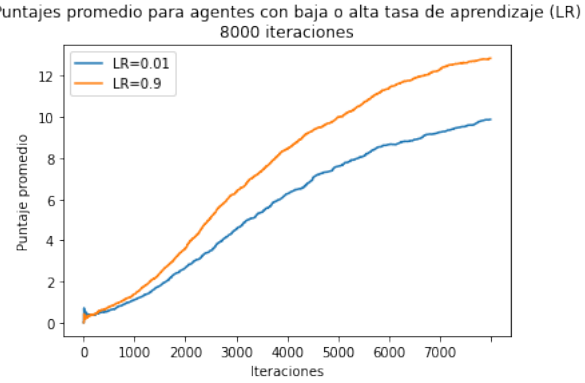


  Puede observarse un desempeño equivalente para ambas tasas de descuento hasta la iteración número 2000, donde la tasa de descuento alta comienza a tener un mejor desempeño. Esto se explica en base a que un valor γ alto se traduce en que consideramos en mayor medida las recompensas futuras al medir la calidad de una accción en un estado en

particular, de modo que se dispone de una fuente de información extra al determinar qué debe hacer el agente. Para tasas de descuento muy bajas, en cambio, la medición de la calidad de una acción estará sesgada hacia las consecuencias inmediatas de dicha acción, sin considerar eventuales beneficios o costos en el futuro.

- ¿Para qué sirve el learning rate? ¿Qué valor te entrego mejores resultados? Comenta los resultados

  La tasa de aprendizaje corresponde al término α de la expresión (1), e indica el sesgo que existe hacia la recompensa o hacia el valor preexistente a la hora de actualizar la Q-table. Para visualizar el efecto de este parámetro, se detalla el desempeño del agente para una tasa de aprendizaje baja (0.01) y para una alta (0.9)



Puede observarse que el agente de tasa de aprendizaje alta tiene un mejor desempeño que el de tasa de aprendizaje baja durante todas las iteraciones. La expliación de esto puede hallarse nuevamente en la expresión (1). Para valores de α cercanos a 1, los valores de la Q-table tendrán tendencia a actualizarse en mayor medida según las recompensas de la acción ejecutada, pues el existirá un sesgo hacia el segundo término de la expresión (1). Esto repercute en que el agente aprenderá de manera más expedita. Por otra parte, para valores cercanos a 0, los valores de la Q-table estarán sesgados a mantener el valor previamente almacenados, repercutiendo en que el aprendizaje sea más lento. Cabe notar que el efecto descrito vale para un número relativamente bajo de iteraciones, pues podría darse el escenario en que, con suficientes iteraciones, el agente con baja tasa de aprendizaje alcance el rendimiento del de alta tasa de aprendizaje.

Actividad 4: Nueva política de recompensas

Como nueva política de recompensas, se propone una extensión de la actual política de recompensas. Para los casos en que el agente se aleja del sitio de impacto, o en que el agente se mantiene estático en el sitio de impacto, se propone mantener la recompensa. Sin embargo, para el caso en que el agente se acerca al sitio de impacto, la recompensa debe ser proporcional a la distancia al sitio de impacto. Así, la idea sería asignar una recompensa más alta por moverse hacia el sitio de impacto desde una distancia larga, y una más baja (pero aún positiva) por moverse hacia el sitio de impacto desde una distancia pequeña.

El fundamento de por qué dicha política de recompensas podría ser efectiva se encuentra en un caso donde el agente entrenado con la política actual rinde de manera deficiente. Cuando una pelota cambia súbditamente de dirección al dar un bote, el agente varias veces es incapaz de corregir el rumbo a tiempo, pues antes del bote seguía la pelota de manera muy cercana. La política propuesta podría ser una buena forma de recompensar al agente en tanto es razonablemente semejante a la política actual (que sí funciona) pero a la vez fomenta movimientos más moderados cuando el agente se encuentra cercano al sitio de impacto, lo que podría eventualmente dar más capacidad de respuesta en los casos en que la pelota sufre de cambios repentinos de dirección.