Aprendizaje Reforzado

anibal.contreras

June 2022

1 Actividad 1: Comprendiendo los hiperparámetros

- El exploration rate es la probabilidad de que el agente explore el ambiente por sobre *exploit* (explotar lo ya conocido). Con un *exploration rate* de $\epsilon = 1$ se tiene un 100% de certeza que el agente explore. Cuando se inicia el entrenamiento el agente no conoce nada de su ambiente, por lo que lo ideal es que explore por sobre explotar un ambiente que no conoce.
- Si se tiene un *exploration rate* mínimo y un *exploration decay rate* muy alto significa que el agente explorará muy poco, es decir tendrá un *exploit* mayor y al tener un *exploration decay rate* muy alto ese *exploration rate* disminuirá, sin embargo al ser mínimo quedará en 0 y el agente no aprenderá, pues no exploró y en un futuro tampoco explorará, por lo que el desempeño será malo.
- Si el *exploration decay rate* es demasiado bajo, el agente se quedará durante muchas iteraciones dentro de su mismo *exploration rate*, lo cual es perjudicial para el aprendizaje, pues a medida que el agente explora, hay un momento en el que ya encuentra un ambiente en el que está bien recompensando y debe explotarlo mas. Con un *exploration decay rate* bajo esto no ocurrirá y su rendimiento a lo largo del tiempo sería de la forma: bajo $\rightarrow$ *regular* $\rightarrow$ *bajo*.

- Los cambios en los distintos hiperparámetros contextualizados con el juego Pong se pueden reflejar en como el pádel responde ante la proximidad de la pelota. Con un buen ajuste de hiperparámetros, digamos un *exploration rate* medio alto, que sea 0.6, un *exploration decay rate* bajo, que sea 0.01, permitiría que al principio el agente no supiera que significa que la pelota lo pase, pero a medida que van ocurriendo las iteraciones, el agente aprenderá que si la pelota se aproxima, este se debe mover a la posición a la que pelota llegará y ya teniendo mas o menos dominado ese comportamiento dejar de explorar y pasar a explotar, por eso un *exploration decay rate* bajo.

2 Actividad 2: Implementando Q-Learning

Se implementa en *QAgent.py*

3 Actividad 3: Análisis de parámetros del agente

- La tasa de descuento determina que tanto el agente se preocupa de las recompensas en el futuro respecto a las del futuro inmediato. Si la tasa de descuento es 0, el agente será reacio al futuro y pensará exclusivamente en recompensas inmediatas. Si la tasa de descuento es 1, el agente evaluará cada una de sus acciones según la suma total de sus recompensas futuras.
- La tasa de descuento que me dió mejores resultados, manteniendo los otros hiperparámetros constantes, con un número de 5000 episodios fue 0.1, pues se obtuvieron mean scores cada 100 iteraciones de 22 en promedio y record máximo de 148.
- El learning rate es un hiperparámetro que controla como el modelo se adapta al problema. Si se tiene un *learning rate* muy alto, calculará el nuevo Q-value sin considerar los Q-values calculados anteriormente, por tanto siempre tendrá un desempeño aceptable, pero no mejorará con el tiempo. Por el otro lado, si se tiene un *learning rate* muy bajo hará que el entrenamiento se quede estancado y no podrá tener un buen desempeño. El learning rate que me dió mejores resultados, manteniendo todos los hiperparámetros constantes, con un número de 5000 episodios fue 0.4, pues se obtuvieron mean scores cada 100 iteraciones de 23 en promedio y record máximo de 154 puntos.

4 Actividad 4: Nueva política de recompensa

Una política que tal vez haya dado mejores resultados para recompensar al agente es seguir la posición en el eje y en todo momento de la pelota. Si el pádel está en la misma posición que en la posición y de la pelota, se le suma 2 puntos, se le suma 1 punto si se acerca a la posición y de la pelota y se le resta 1 si se aleja de la posición y de la pelota. Es bastante similar a la política ya implementada, pero tiene como incentivo extra el seguir la posición de la pelota en todo momento, para que cuando la pelota se acerque, el agente siempre esté próximo a donde llegará. Como se le suman 2 puntos al estar en la posición de la pelota, intentará siempre estar ahí, si no está en esa posición, igual será recompensado con 1 punto y se verá incentivado a acercarse a donde terminará la pelota y si se aleja se verá perjudicado con -1 punto.

5 Bibliografía

- https://deeplizard.com/learn/video/mo96Nqlo1L8
- https://stats.stackexchange.com/questions/221402/understanding-the-role-of-the-discount-factor-in-reinforcement-learning: :text=The
- https://stackoverflow.com/questions/58266988/learning-rate-gradient-descent-difference: :text=Specifically