Práctica 06: Reinforcement Learning: Q-Learning

Luis Tong Chabes
Ciencia de la Computación

Junio 2019

1. Modelo Q-Learning

Ees un modelo independiente de política y basado en TD que utiliza una función de valor de la forma par estado-acción $Q(s, a)$ y cuya funcion de actualización es dada por la siguiente expresión:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[r_{t+1} + \gamma maxQ(s_t, a) - Q(s_t, a_t)]$$

y la selección de la accion sera tal que se maximice el valor $Q(s, a)$.

$$a_t = argmaxQ(s_t, a)$$

2. Juego

El juego consiste en el clásico *Snake* donde se tiene un *grid* de 2 dimensiones x, y entonces el *snake* siempre esta en movimiento para los 4 direcciones posible y tiene el objetivo de encontrar su punto objetivo o también conocido como comida. Cabe resaltar que la posición del *snake* es central y del objetivo es aleatorio.

3. Composicion del Juego

Para el juego se necesita establecer valores que son las siguientes:
Para obtener la acción y actualizar en la clase de Q-learning mantiene el modelo antes planteado.
En la clase de *Snake* esta compuesto con un areglo de pares ordenados (x, y) que lleva desde el inicio del juego hasta que termina.
En la clase principal del juego consta de un bucle mientras el jupador en este caso *snake* no llega al objetivo, es decir que ambas posiciones sean iguales.
Entonces se inicializa el modelo y se guarda la distancia anterior que nos ayudara para establecer el premio, Aqui tenemos otro bucle que seguira ejecutandose mientras el juego

no haya terminado.
Si no es asi la dirreccion que del *snake* es igual a la acción, para después aplicar la actualización del estado con el premio.
Se tiene que definir dos funciones para obtener el estado y obtener el premio; el primero consta de nformacion de las posiciones del jugador y del objetivo, mientras el segundo establece el premio usando la distancia.

Despues de esto se hace una revision para saber si el snake ha llegado al objetivo para fin del juego.

4. Desempeño del juego

Fue medido por la cantida de iteraciones requeridas para llegar al objetivo.

4.1. Parámetros a evaluar: tamaño de *grid*,α , γ ,intentos y número de iteraciones

Para acelerar los experimentos, el tamaño de *grid* son cudarados.
Cabe resaltar que intentos se refiere desde el inicio a fin.
Y iteraciones es desde posición inicial avanzando uno en uno, esto quiere decir que en un *grid* de 10*10 el recorrido máximo es 100.

tamaño	α	γ	intentos	iteraciones
10	0.25	0.25	3	6
10	0.25	0.55	9	52
10	0.55	0.9	13	153
10	0.55	0.25	4	25
10	0.55	0.55	8	79
10	0.55	0.9	11	124
10	0.9	0.25	6	40
10	0.9	0.55	9	43
10	0.9	0.9	8	109
100	0.25	0.25	6	579
100	0.25	0.55	9	265
100	0.55	0.9	4	159
100	0.55	0.25	14	2230
100	0.55	0.55	4	433
100	0.55	0.9	9	526
100	0.9	0.25	11	904
100	0.9	0.55	10	1105
100	0.9	0.9	15	1076
1000	0.25	0.25	9	797
1000	0.25	0.55	11	1201
1000	0.55	0.9	11	1161
1000	0.55	0.25	8	236
1000	0.55	0.55	10	769
1000	0.55	0.9	14	1644
1000	0.9	0.25	8	254
1000	0.9	0.55	8	506
1000	0.9	0.9	15	1268
..	..	..	..	..