Rafael Manzano

13 de enero de 2018

Practica Q-Learning

1. Argumentación de las diferencias entre la versión clásica e inversa de Q-Learning

En la versión clásica de Q-Learning los valores de la tabla se actualizan cada vez que se transcurre de un estado a otro. Mientras que en la versión inversa se almacenan los pares estado - acción hasta llegar a la meta, y entonces se actualizan todos los valores de la tabla para todos los pares estado - acción transcurridos hasta ese momento mediante un método recursivo.

En las pruebas se puede comprobar claramente como la versión recursiva converge más rápido que la estándar ya que una vez que hemos alcanzado el objetivo actualizamos en la tabla un conjunto de valores Q que nos indican la serie de acciones que tenemos que seguir para alcanzar la meta, mientras que en la versión estándar, solo se actualiza el valor q de la celda contigua a la meta en la primera iteración. Teniendo que alcanzar dicha celda para actualizar el valor de la anterior y así sucesivamente hasta obtener todos los valores Q de un camino.

2. Selección y justificación de los valores gamma y k óptimos para la resolución del problema.

El valor de k determina que las probabilidades de escoger una acción u otra.

Para valores de k mayores que uno, favorecemos la explotación. Es decir, daremos mas probabilidad de ejecutarse a la acción que mas valor tenga en nuestra tabla Q para un determinado estado.

Si damos un valor k demasiado alto, corremos el riesgo de llegar a una solución subóptima, ya que ejecutaríamos la primera serie de acciones que nos llevase a la meta sin explorar si hay una solución mejor.

K = 1,5
[[0 0 2 10]]
[0.01615116 0.01615116 0.03634012 0.93135755]

Para k = 1, independientemente de los valores de la tabla les daríamos a todos las mismas posibilidades, es decir seria completamente aleatorio. Por tanto este valor tampoco nos conviene.

K = 1
[[0 0 2 10]]
[0.25 0.25 0.25 0.25]

Dar valores menores a uno no tendría ningún sentido ya que daríamos más importancia a las acciones con menos valor en nuestra tabla Q, y nuestro agente no conseguiría llegar a la meta.

K = 0.9
[[0 0 2 10]]
[0.3165881   0.3165881   0.25643636 0.11038744]

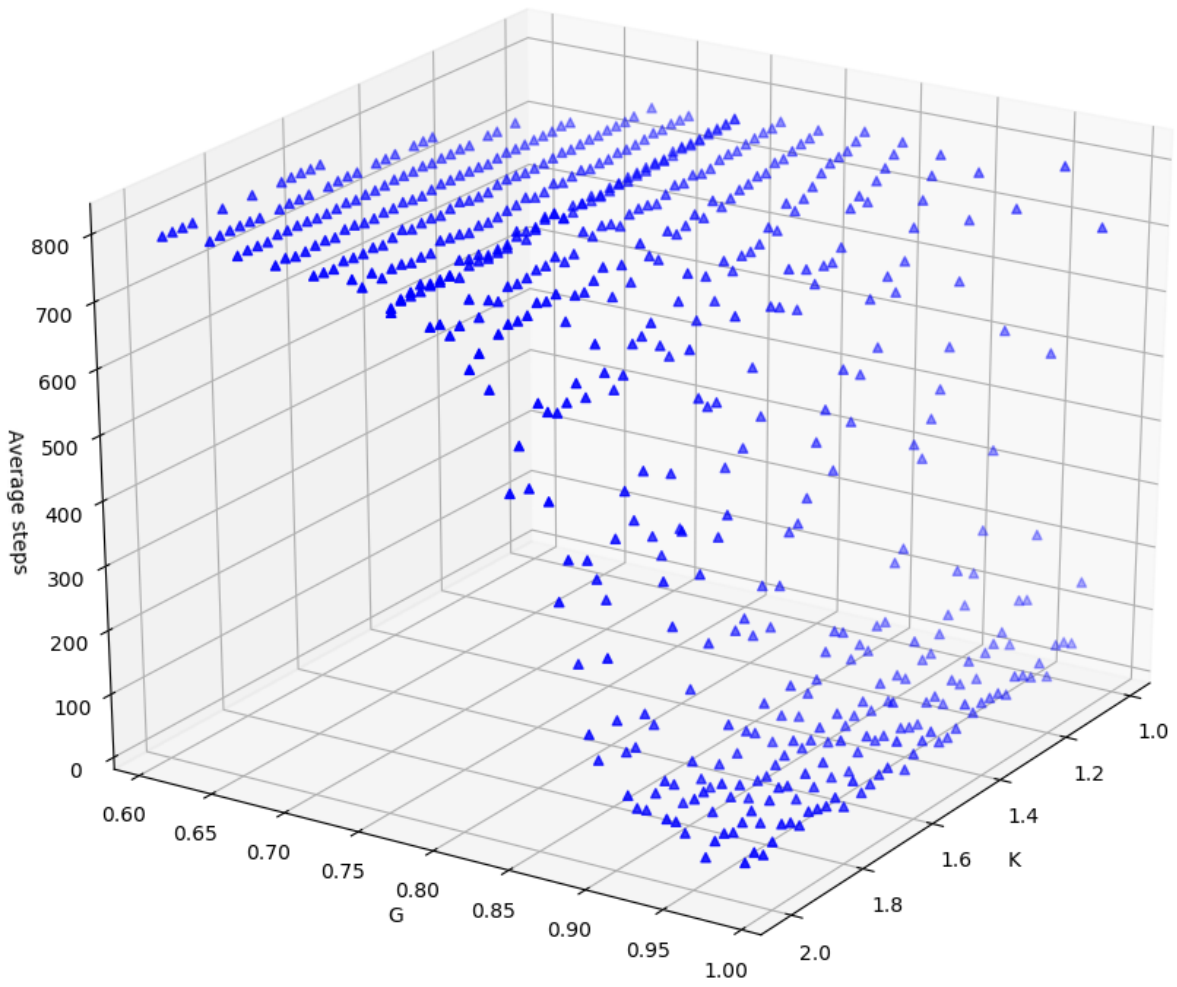
Podemos intuir que el mejor valor de k se encuentra en valores en el intervalo [1.05 , 1.3]. De esta forma nos aseguraríamos de encontrar las solución optima mediante las acciones tomadas para exploración y que una vez encontrada esta solución, tomásemos las acciones que nos llevan a ella.

El valor gamma es el factor de descuento. Representa el valor de las acciones futuras en función de lo lejos que se encuentran de el estado actual.

Un valor de 0 implicaría que solo nos importan las recompensas inmediatas, lo que en este caso no nos conviene ya que tenemos que dar valores a las acciones que nos llevan hasta la meta.

El valor de gamma no puede ser mayor que uno ya que eso implicaría que le daríamos mas importancia a la acción que mas lejos esta de la meta que a las mas próximas a ella, lo cual no tendría sentido ya que no permitiría que el agente llegase a la solución.

He hecho una búsqueda empírica de los valores óptimos para k y gamma en un laberinto de 25x25 y el resultado ha sido el siguiente:



Cada punto del gráfico representa la mediana de cuatro medias finales de cuatro ejecuciones diferentes para los mismos valores de k y gamma. He decidido hacerlo de esta manera para poder ver el gráfico con mas facilidad haciéndolo menos sensible a ejecuciones con resultados muy dispares. Sin embargo, hay que tener en cuenta que esto no nos permite visualizar el problema de escoger valores de k demasiado altos, que pueden dar lugar a resultados subóptimos al no explorar suficiente.

Por tanto concluyo que como se puede ver en el gráfico los valores óptimos para gamma son cercanos al uno, de 0.9 en adelante. Y para k, creo que lo mejor seria una solución dinámica que fuera incrementando k a medida que vamos obteniendo resultados. De esta manera nos aseguraríamos de encontrar la solución optima mediante mucha exploración al inicio, y no perderíamos el tiempo explorando una vez hubiésemos alcanzado la política óptima.