**Documento wikipedia**

* ¿Qué es el aprendizaje automático?
* ¿Qué es el aprendizaje por refuerzo?
* Aprendizaje por refuerzo.
  + Un agente.
  + Un conjunto de estados S.
  + Un conjunto de acciones A por cada estado.
  + Ejecutar una acción modifica (o no) el estado del agente y genera una recompensa.
  + Ejemplo de subir al tren -> Exploración aleatorio que, inicialmente, parece peor.
* Algoritmo Q-Learning. (Nota: está bastante bien explicado y resumido)
* Explicación semántica de cada variable del algoritmo Q-Learning.
* *“Empezando con un factor de descuento menor y aumentándolo hacia su valor final se acelera el aprendizaje”.*

**Documento wordpress**

* La recompensa depende de la acción y el estado actual.
* Las recompensas de acciones de victoria o derrota son fáciles de definir, pero necesitamos también las recompensas de las acciones intermedias. De eso se ocupa Q-Learning.
* Recompensas a corto y largo plazo, no solo al finalizar la partida.
* Definición de experiencia: (estado, acción, recompensa, estado siguiente).
* La tabla parte con recompensas directas.
* Para un estado y una acción es posible tener múltiples experiencias y sólo aprendemos un poco de cada una.
* Parámetros del algoritmo: velocidad de aprendizaje y factor de descuento.
* Ejemplo para ajustar parámetros.
* Algoritmo para aprender a partir de una lista de experiencias.
* Concepto de exploración para obtener experiencias.
* Criterio para elegir entre acción o exploración: la probabilidad de una acción se pondera por su recompensa. Así, todas tienen la posibilidad de hacerse.
* Criterio para finalizar la exploración.
* Las celdas con acciones inválidas pueden ser un problema.

**Video 1**

* Se conoce: estados S y acciones A.
* Se desconoce: modelo de transición y función de recompensa.
* Conceptos de exploración, explotación y muestreo.
* Los valores Q del Q-Learning son valores asociados al par Estado-Acción.
* El Q-Learning es aprendizaje activo: la política se calcula probando acciones.
* El Q-Learning funciona aunque el agente no actúe de forma óptima y requiere mucha exploración.

**Video 2 (No)**

**Video 3**

* Aunque actuemos de forma óptima, puede que se cometan errores porque las acciones no son deterministas.
* Requiere de mucha exploración.
* Ejecución del algoritmo: el algoritmo se repite por cada estado y por cada acción
* (1) Esta fórmula hace únicamente explotación de los datos, no hay nada de exploración.
  + La recompensa actual es lo que recibo por ejecutar la acción que acabo de ejecutar. (Duda)
  + Para determinar el máx del estado siguiente, tenemos que ejecutar todas las acciones que se pueden hacer desde el estado siguiente y escoger la que tenga mayor recompensa.
  + y es un parámetro que indica la importancia de las predicciones a futuro.
* (2) El primer sumando es lo que ya sabíamos, y el segundo es lo que hemos aprendido. **El valor que se recomienda para alfa es un valor bajo como 0.1**.
* Tenemos que ir decrementando la tasa de aprendizaje para garantizar la solución óptima con:
* El valor que le demos a la k o la interpretación que hagamos del tiempo puede ser muy diferente. El parámetro n puede ser:
  + El número de ejecuciones del algoritmo.
  + El número de veces que ha procesado ese estado.
  + El tiempo de ejecución del algoritmo.
* Los valores de k y n depende de cada problema, así que, debemos ir probando. (Esto puede ser susceptible a graficar).
* Sin exploración tampoco podemos llegar a la solución óptima. Debemos empezar con mucha exploración e ir reduciéndola con el tiempo para que la solución sea óptima. La exploración podemos incluirla en el algoritmo de dos formas:
  + Método voraz: Utiliza un parámetro que representa una pequeña probabilidad que representa la probabilidad de hacer explotación o exploración. Esta probabilidad debe reducir a medida que se ejecute el algoritmo porque no tiene sentido explorar cuando ya tenemos la política óptima. Podemos hacer k/(k+n).
  + Funciones de exploración: si hemos probado mucho una acción en un estado, tiene menos probabilidad de volver a hacerse. Podemos guardar en una tabla (N) el número de veces que hemos ejecutado una acción. (Esto es factible de graficar)
* Con las funciones de exploración, la exploración desaparece con el tiempo.