

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Breve introducción a Reinforcement Learning

A. Atutxa

LSI Bilbao

December 9, 2024

¹Basado en el libro de Sutton y Barto, curso de Adam y Martha White (U. Alberta), curso de UCL D. Silver, curso de Stanford y cursos de de E. BrunSkill, Thomas Simonini, DeepMind y

Overview

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

1 Situando RL

2 RL: El problema de los k-armed bandits

3 Markov Decision Processes

4 Aprendizaje temporal, aprendizaje por episodios

Toma de decisiones secuencial

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

■ **Objetivo de RL:** Seleccionar las acciones que maximicen el futuro premio acumulado

- Cada acción puede tener consecuencias a largo plazo.
- El premio no tiene porque ser inmediato
- El mejor premio a corto plazo no tiene por qué ser el mejor a largo plazo. Actuar de forma Greedy no siempre es la mejor estrategia (p.e. las inversiones)

RL se basa en la siguiente hipótesis (premisa):

Definición (La hipótesis del premio)

Todo Goal puede ser descrito como una maximización del cúmulo de premios esperado

Contexto: Agente y Entorno

Intro RL

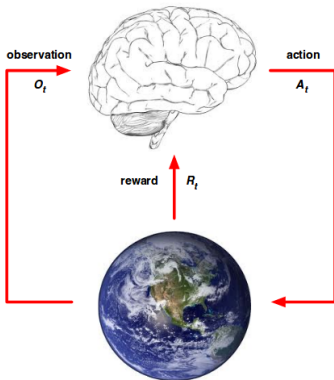
A. Atutxa

Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios



■ En cada paso t el agente:

- Recibe/Percibe una observación
- Ejecuta una acción
- Recibe un premio

■ En cada paso t el entorno:

- Emite una observación
- Recibe una acción
- Emite un premio

El problema de los k-armed bandits (las K máquinas tragaperras)

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

- Ejemplo básico de la *biblia* de RL (libro de Richard S. Sutton y Andrew G. Barto¹)
- Nos va a permitir:
 - Formalizar **la toma de decisiones** bajo **incertidunbre**
 - Entender: **acción, premio, valor de una acción**
- Ejecutar :
<https://mdp.ai/coursera/c01-k-armed-bandit/>

¹<https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf>

K-bandits vs. Aprendizaje Temporal

Intro RL

A. Atutxa

Situando RL

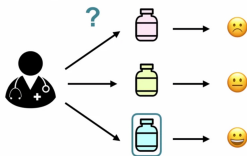
RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios

Consecuencias de cada accion en el entorno: **No** influencia sobre posteriores premios.

Clinical Trials



Consecuencias de cada accion en el entorno: influencia sobre posteriores premios. Aprendizaje temporal.



Aprendizaje temporal.

Intro RL

A. Atutxa

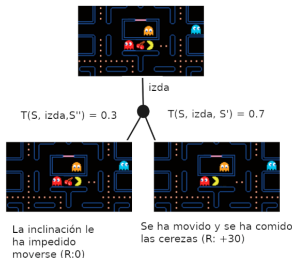
Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios

Conocemos las probabilidades subyacentes en el entorno. Markov Decision Process (Value Iteration)



No conocemos las probabilidades subyacentes en el entorno. Markov Decision Process (Value Iteration)



Formalización (MDP)

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Formalización como Proceso de Decisión de Markov

- $M = \langle S, \gamma, T, R \rangle$
- S : Conjunto finito de estados, $S_t \in S$
- A : Conjunto finito de acciones disponibles. $A_t \in A(S_t)$. A_t es la acción en el instante t que pertenece a las acciones disponibles en el estado S_t .
- T : Función de transición. Cuando se trata de un entorno estocástico $T : S \times A \times S \rightarrow P(S)$.

$$T(s'|s, a) = Pr(S_{t+1} = s' | S_t = s, A_t = a)$$

$$\sum_{s' \in S} T(s'|s, a) = 1$$

- $R : S \times A \times S \rightarrow \mathbb{R}$

MDPs

Intro RL

A. Atutxa

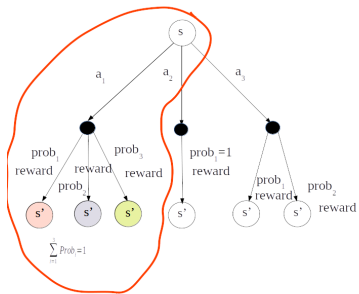
Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

- **Modelo de Transiciones:** función de transición T , es decir, la probabilidad de transicionar el estado actual S realizando una acción A a los siguientes estados s' posibles.



- **Modelo de Premios:** función de los premios R , es decir, el valor del premio que podríamos obtener dado un estado y una acción al pasar a los distintos estados s' posibles.

Concepto: Política

- **La política:** mapeo entre los estados del entorno percibidos por el agente y las acciones que el agente realizará cuando alcance cada uno de esos estados.
- Se suele representar con la letra griega π y habrá tantas como combinaciones de acciones y estados haya
- El aprendizaje consiste en encontrar las **política optima** π^* de entre todas las posibles
- **Value Iteration:** Permite encontrar la política óptima si conocemos la distribución subyacente del entorno.
- Al finalizar el Value Iteration sabemos cuales son los **valores óptimos** V^* de cada estado:

$$\pi^* =_{a \in A} \sum_{s' \in S} T(s, a, s') (R(s, a, s')) + \gamma V^*(s')$$

Problemas con los MDPs

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

No se suele disponer de las **funciones de transición**.

Solo disponemos de los estados y de las acciones y el premio que está asociado a transicionar de un determinado estado a otro.

Dos estrategias posibles:

- **Model based:** Consiste en aprender las funciones de transición y del premio y luego aplico value iteration. No se suele emplear porque es muy costoso.
- **Model free:** Consiste en aprender el valor de cada acción a través de muestras o episodios. **Aprendizaje temporal por episodios.**

Aprendizaje temporal por episodios

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Un **episodio** consiste en exponer al agente a **un ciclo completo** donde hay un estado inicial y un estado final. El episodio está formado por:

- Una lista de estados
- Acciones (posibles acciones a partir de un estado)
- Premios
- Nuevos estados (posibles estados a partir de un estado)

Aprendizaje por episodios: Métodos

Intro RL

A. Atutxa

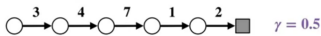
Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios

- **Monte Carlo:** El premio se contabiliza al final del episodio. Los estados se recorren en orden inverso y así los premios se van acumulando en orden inverso.



$$G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 = 7$$

$$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 = 8$$

$$G_2 = R_3 + \gamma R_4 + \gamma^2 R_5 = 8$$

$$G_3 = R_4 + \gamma R_5 = 2$$

$$G_4 = R_5 = 2$$

$$G_5 = 0$$

- **Aprendizaje Temporal:** No se espera hasta el final. En cada paso se va actualizando el valor del estado haciendo una media ponderada entre el valor actual y lo que le propone el "futuro".

$$V(S_{t+1}) = (1 - \alpha)V(S_t) + \alpha[R_{t+1} + \gamma V(S'_t)]$$

$$V(S_{t+1}) = V(S_t) + \alpha[R_{t+1} + \gamma V(S'_t) - V(S_t)]$$

Aprendizaje temporal por episodios

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

- SARSA: Como Q-learning salvo en vez de seleccionar el valor Q de la mejor acción en s' se selecciona el valor Q de una acción seleccionada según la política (epsilon greedy,..)
- Q-Learning²

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\epsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ϵ -greedy)

 Take action A , observe R, S'

$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma \max_a Q(S', a) - Q(S, A)]$

$S \leftarrow S'$

 until S is terminal

² ¡¡¡IMPORTANTEE!! También lo vereis escrito así
 $Q(S, A) \leftarrow (1 - \alpha)Q(S, A) + \alpha[R + \gamma \max_a Q(S', a)]$

Aprendizaje temporal por episodios Q-learning (Q-table). Ejemplo ³

Intro RL

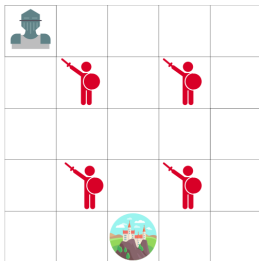
A. Atutxa

Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios



- Cada paso es un premio de -1 (para indicar que el camino más largo es peor).
- Si tocas a un enemigo el premio es -100 y el episodio finaliza.
- Si estás en el castillo el premio es +100.

³Fuente: <https://www.freecodecamp.org/news/diving-deeper-into-reinforcement-learning-with-q-learning-c18d0db58efe/>

Aprendizaje temporal por episodios Q-learning (Q-table). Ejemplo de un gridworld

Intro RL

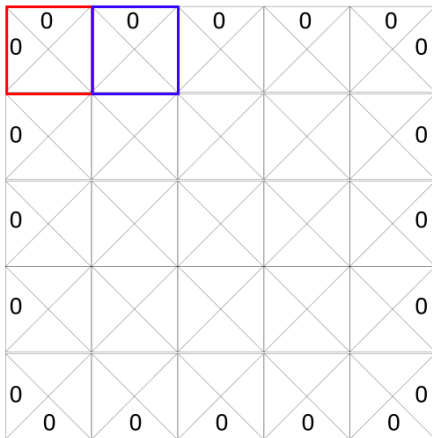
A. Atutxa

Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios



Aprendizaje temporal por episodios Q-learning (Q-table). Ejemplo

Intro RL

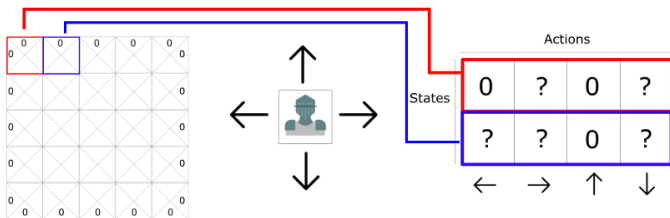
A. Atutxa

Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios



$$Q^{\pi}(s_t, a_t) = \underline{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | s_t, a_t]$$

Aprendizaje temporal por episodios Q-learning (Q-table). Ejemplo

Intro RL

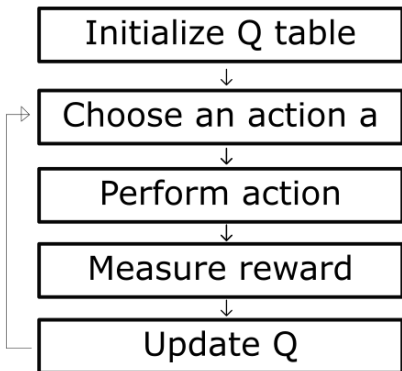
A. Atutxa

Situando RL

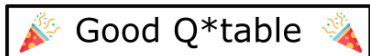
RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios



At the end of the training



Ejercicio de ejemplo

Intro RL

A. Atutxa

Situando RL

RL: El problema de los k-armed bandits

Markov Decision Processes

Aprendizaje temporal, aprendizaje por episodios



	←	→	↑	↓
Start	0	0	0	0
Small <u>cheese</u>	0	0	0	0
Nothing	0	0	0	0
2 small <u>cheese</u>	0	0	0	0
<u>Death</u>	0	0	0	0
Big <u>cheese</u>	0	0	0	0

Aprendizaje temporal por episodios

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$

Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(\text{terminal}, \cdot) = 0$

Loop for each episode:

 Initialize S

 Loop for each step of episode:

 Choose A from S using policy derived from Q (e.g., ε -greedy)

 Take action A , observe R, S'

$Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$

$S \leftarrow S'$

 until S is terminal

¿Cómo equilibramos la exploración versus la explotación?

- Ir decrementando el ε según pasan los episodios
- Añadir una función de exploración que modifica las actualizaciones ligeramente añadiendo un "bias" sobre las acciones aun no experimentadas

$Q(S', a) \rightarrow Q(S', a) + k/(n+1)$ donde $n+1$ es el número de veces que se ha ejecutado esa acción.

Limitaciones del Q-Learning

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Con configuraciones y un número de estados pequeño como los de los ejemplos anteriores, el algoritmo funcionará. Pero,

- ¿qué sucede con casos como el del Pacman en el que el número de estados es enorme? ¿es realista pensar que los vamos a poder explorar todos?
- solución: quizás pueda jugar con la representación de los estados

Q-Learning Aproximado

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Queremos encontrar una representación que generalize y permita agrupar casos similares⁴



⁴Ejemplo de Dan Klein

Q-Learning Aproximado

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

Solución: Representación de estados como vector de rasgos

- Distancia al fantasma más cercano
- Distancia al punto más próximo
- Número de fantasmas
- $\frac{1}{(dist.AIPunto)^2}$
- ...

Q-Learning Aproximado

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

$$Q(S, A) = Q(S, A) + \alpha[(R + \gamma \max_{a'} Q(S', a')) - Q(S, A)]$$

- $Q(S, A) = w_1 f_1(S, A) + w_2 f_2(S, A) + \dots + w_n f_n(S, A)$

- transición (S, A, S', R)

- diferencia entre:

- $R + \gamma \max_{a'} Q(S', a')$: El futuro si se realiza la acción

- $Q(S, A)$: mi estado actual

$$Q(S, A) = Q(S, A) + \alpha[diferencia]$$

$$w_1 = w_1 + \alpha[diferencia] f_1(S, A)$$

Q-Learning Aproximado⁵

Intro RL

A. Atutxa

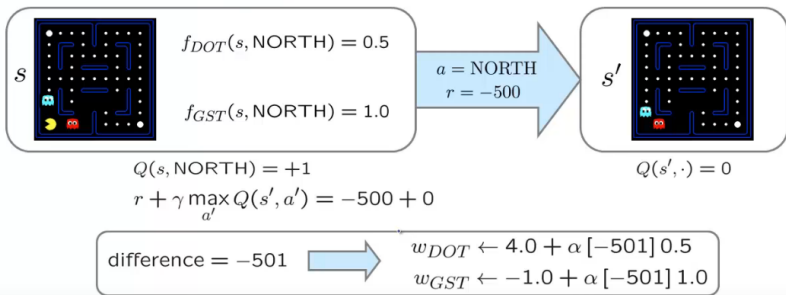
Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

$$Q(s, a) = 4.0 f_{DOT}(s, a) - 1.0 f_{GST}(s, a)$$



⁵Ejemplo de Dan Klein

Bibliografía

Intro RL

A. Atutxa

Situando RL

RL: El
problema de
los k-armed
bandits

Markov
Decision
Processes

Aprendizaje
temporal,
aprendizaje
por episodios

- Reinforcement Learning, An Introduction (Second Edition). By Richard S. Sutton and Andrew G. Barto
<https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf>
- Berkeley curso de Inteligencia Artificial (Dan Klein)
- DLRL2019 (Adam White):
<https://www.youtube.com/watch?v=RancMV1wECg>