

APRENDIZAJE REFORZADO

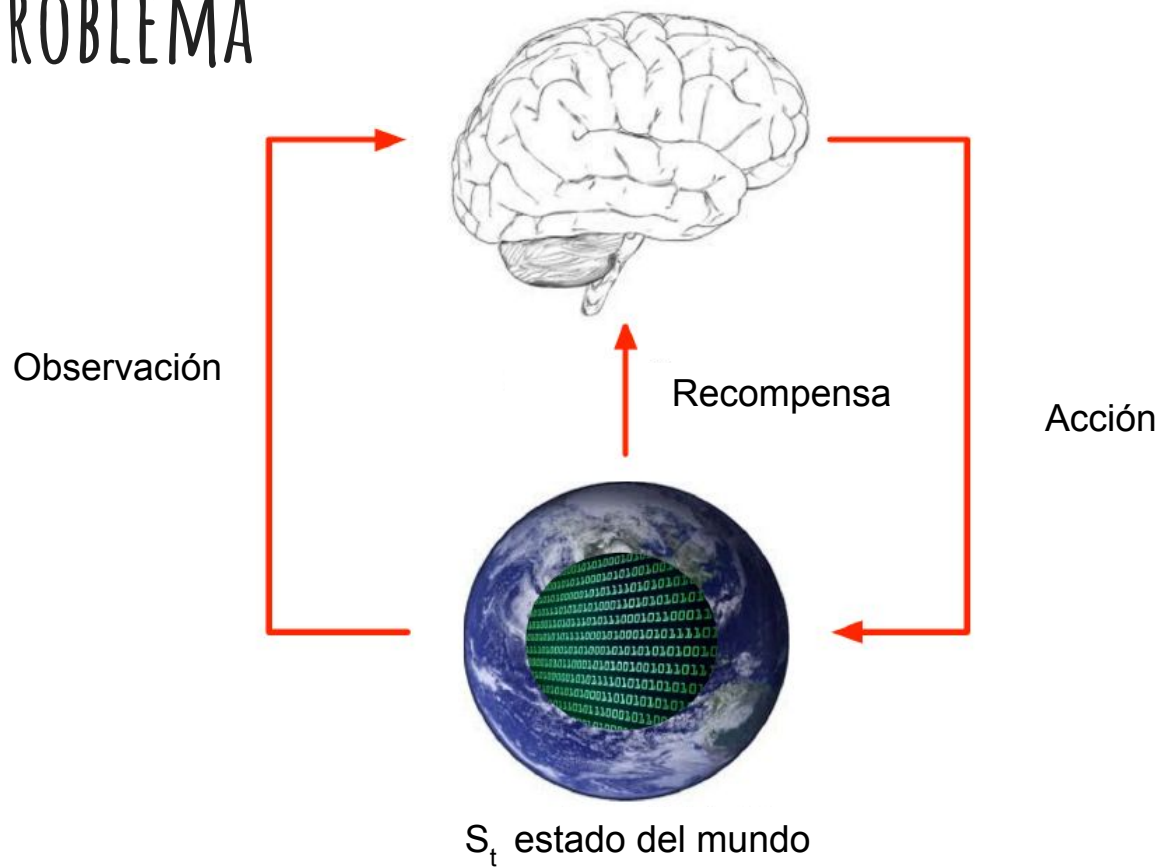
CLASE 1

Julían Martínez, FIUBA

EL CONTENIDO DE ESTE
CURSO FUE DESARROLLADO
EN GRAN PARTE CON LA
AYUDA DE JAVIER KREINER



MARCO DEL PROBLEMA





Download from
Dreamstime.com



DIFERENCIAS CON OTROS PARADIGMAS DE ML, RL VS APRENDIZAJE SUPERVISADO

- No viene dado un dataset con inputs y targets
- No hay un 'supervisor', o sea input y targets, hay una señal de recompensa
- El feedback **se recibe con retraso**, no es instantáneo
- Las decisiones son secuenciales, los datos no son i.i.d.
- Las acciones del agente **modifican** los datos que va recibiendo (el ambiente).

EJEMPLOS

(EN CADA UNO DE ESTOS, RECOMPENSA, ACCIONES, OBSERVACIONES?)

- Un jugador de ajedrez, Teg, go, Backgammon, etc.
- `Un helicóptero debe realizar piruetas
- Diseñar landing page para maximizar retención
- Chatbots con diversos objetivos: psicólogos, servicio al cliente
- Tratamiento médico personalizado
- Administración de una cartera de acciones
- Robots
- Asistentes de navegación

VIDEOS DE ALGUNOS EJEMPLOS

- robot humanoide:

<https://www.youtube.com/watch?v=No-JwwPbSLA>

- helicoptero: <https://www.youtube.com/watch?v=0JL04JJjocc>

- blackout: <https://www.youtube.com/watch?v=eG1Ed8PTJ18>

- space invaders:

<https://www.youtube.com/watch?v=W2CAghUiofY>

- arquero robotico:

<https://www.youtube.com/watch?v=CIF2SBVY-J0>

ÉXITOS

- TD-Gammon (1992)
- Atari Games (DQN, 2015)
- AlphaGo(2015/2016)/AlphaGo Zero(2017)/AlphaZero(2017)
- Dota 2 (2018)
- Starcraft 2 (2019)
- Manipulación Robótica (2018)

(<https://ai.googleblog.com/2018/06/scalable-deep-reinforcement-learning.html>)

El campo no ha hecho un impacto económico significativo aún, pero está comenzando a ser usado en diferentes industrias (grandes oportunidades). Tal vez el problema más importante: necesita ingentes cantidades de datos. Leer

<https://www.oreilly.com/ideas/practical-applications-of-reinforcement-learning-in-industry>

COMPAÑÍAS UTILIZANDO APRENDIZAJE REFORZADO

- Deepmind: AlphaGo, AlphaZero, Atari Games, <https://deepmind.com/>
- Trading algorítmico: Hihedge, <https://www.hihedge.com/>, <https://pit.ai/>
- Ambientes de cultivo controlables: Optimal Labs: <http://optimal.ag/>
- Aprendizaje de robots/vehículos autónomos: <http://covariant.ai/>, <https://www.latentlogic.com/>, <https://www.osaro.com/>, <http://prowler.io/>, <https://www.fanuc.com/>
- Análisis de datos: <http://intelligentlayer.com/>
- Chatbots: <https://rasa.com/>

REPASO DE PROBABILIDAD Y PROCESOS DE MARKOV

$\mathcal{L} = \phi \frac{L}{t} \frac{t}{\phi}$

$f(\omega) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \omega} dx \quad \frac{d}{dt} \frac{d}{d\phi}$

$\nabla \cdot \mathbf{E} = 0 \quad \nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t} \quad \nabla \cdot \mathbf{H} = 0 \quad \nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}$

$(-i\hbar \frac{\partial}{\partial t} \Psi = H \Psi)$

$\rho \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \nabla \cdot \mathbf{T} + \mathbf{f}$

$H = -\sum p(x) \log p(x)$

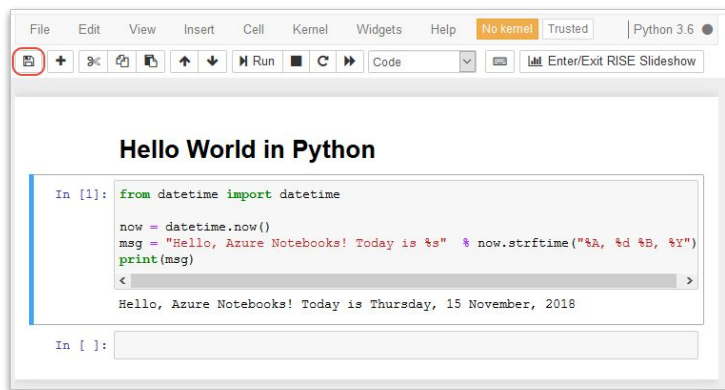
$\frac{1}{2} G^2 S^2 \frac{\partial^2 V}{\partial S^2} + r S \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - r \cdot V = 0$

$(Q, q_i, m_i) = \sum_{i=1}^n \left[\frac{D_i}{m_i q_i} S_i + c_i \cdot D_i + \frac{q_i H_i}{2} \left(m_i \left(1 - \frac{D_i}{P_i} \right) - 1 + 2 \right) \right]$

$\left[\frac{d \Delta p(s, \phi)}{d \phi} \right] = \begin{bmatrix} \gamma & -\mathcal{L} \\ -\beta & 0 \end{bmatrix} \begin{bmatrix} \Delta p(s, \phi) \\ \Delta M(s, \phi) \end{bmatrix}$

$\int_0^{\frac{\pi}{2}} (\log \sin x)^2 dx = \int_0^{\frac{\pi}{2}} (\log \cos x)^2 dx = \frac{\pi}{2} \left\{ \frac{\pi^2}{12} + (\log 2)^2 \right\}$

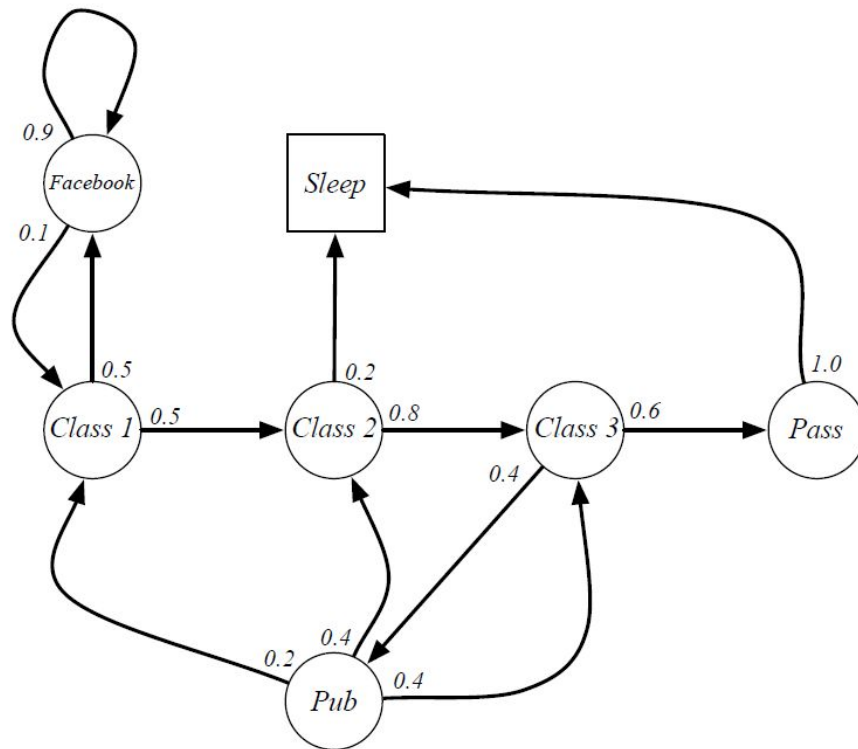
EJEMPLO DEL ESTUDIANTE



The screenshot shows the Azure Notebook interface. The top menu bar includes File, Edit, View, Insert, Cell, Kernel, Widgets, Help, No kernel, Trusted, and Python 3.6. Below the menu is a toolbar with icons for file operations and execution. The main area displays a code cell titled "Hello World in Python". The code in the cell is:

```
In [1]: from datetime import datetime
now = datetime.now()
msg = "Hello, Azure Notebooks! Today is %s" % now.strftime("%A, %d %B, %Y")
print(msg)
<
Hello, Azure Notebooks! Today is Thursday, 15 November, 2018
```

The output of the code is displayed below the cell.



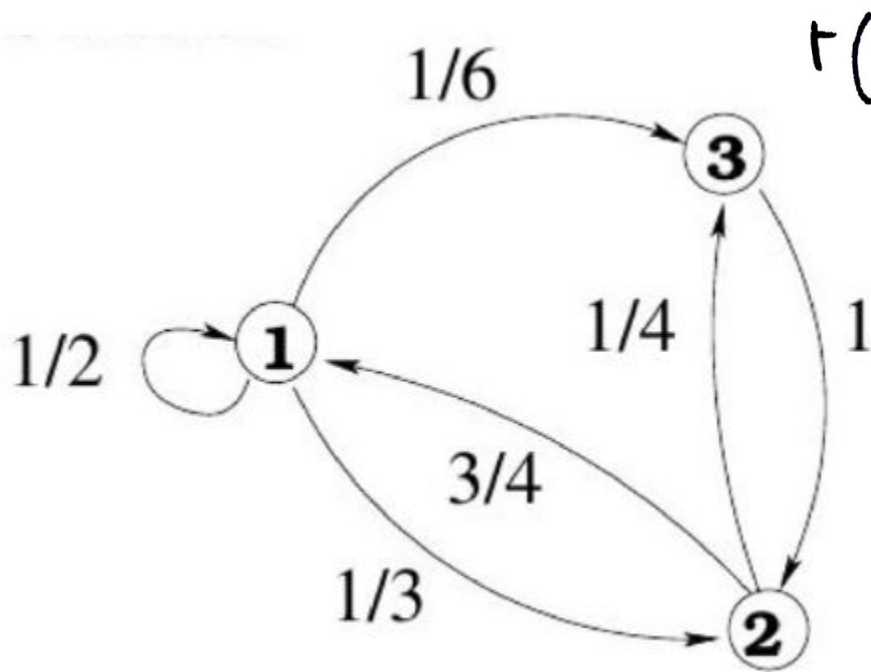
Tomado de los slides de David Silver

EJERCICIO 1.1 - ESTUDIANTE - PYTHON

Usando el código hacer los siguientes ejercicios:

1. simular 100 episodios del estudiante
2. calcular un estimado del tiempo de visita de cada estado
3. calcular un estimado de la longitud media de la cadena

EJERCICIO 1.2 - MC SIMPLE CON REWARD - PAPEL Y PYTHON



$$r(1) = -2, \quad r(2) = 3, \quad r(3) = 5$$

$$g(s_1, s_2) = r(s_1) + r(s_2)$$

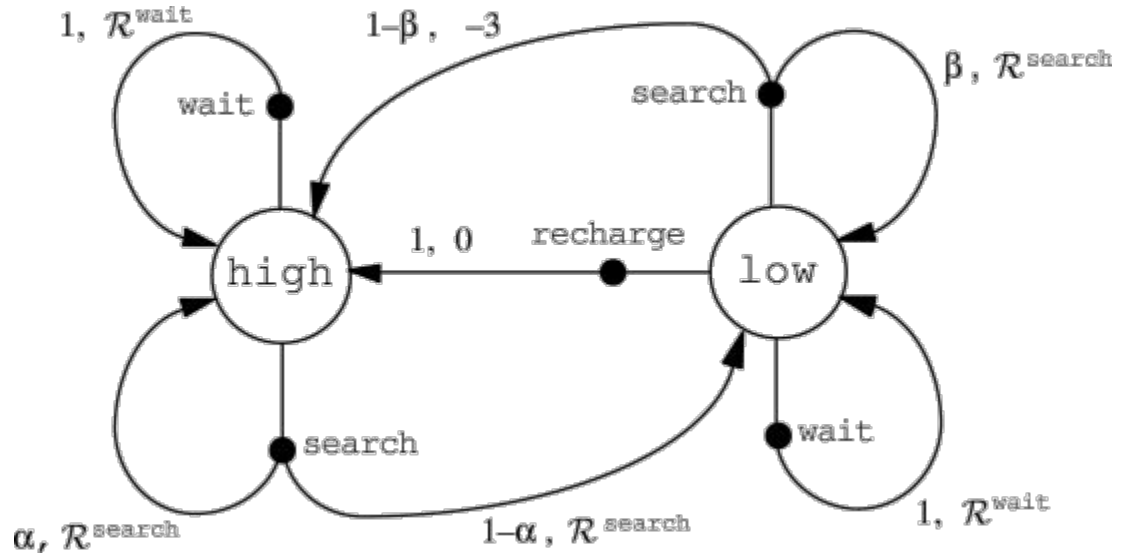
Calcular esto de manera analítica y vía simulación:

$$E_1[g(s_1, s_2)]$$

PROCESOS MARKOVIANOS DE DECISIÓN

$\mathcal{L} = \phi \frac{\partial}{\partial t}$
 $\nabla \cdot \mathbf{E} = 0 \quad \nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t}$
 $\nabla \cdot \mathbf{H} = 0 \quad \nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}$
 $-\hbar \frac{\partial}{\partial t} \Psi = H \Psi$
 $f(w) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x w} dx \quad \frac{d}{dt}$
 $\rho \left(\frac{\partial v}{\partial t} + v \cdot \nabla v \right) = -\nabla p + \nabla \cdot \mathbf{T} + \mathbf{f}$
 $H = -\sum p(x) \log p(x)$
 $\frac{1}{2} G^2 S^2 \frac{\partial^2 V}{\partial S^2} + r S \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - r \cdot V = 0$
 $(Q, q_i, m_i) = \sum_{i=1}^n \left[\frac{D_i}{m_i q_i} S_i + c_i v D_i + \frac{q_i H_i}{2} \left(m_i \left(1 - \frac{D_i}{P_i} \right) - 1 + 2 \right) \right]$
 $\left[\frac{d \Delta p(s, \phi)}{d \phi} \right] = \begin{bmatrix} \gamma & -\mathcal{L} \\ -\beta & 0 \end{bmatrix} \begin{bmatrix} \Delta p(s, \phi) \\ \Delta M(s, \phi) \end{bmatrix}$
 $\int_0^{\frac{\pi}{2}} (\log \sin x)^2 dx = \int_0^{\frac{\pi}{2}} (\log \cos x)^2 dx = \frac{\pi}{2} \left\{ \frac{\pi^2}{12} + (\log 2)^2 \right\}$

EJEMPLO DEL ROBOT




Tomado de los slides del libro de Sutton

EJEMPLO MÁS COMPLEJO - ALQUILER DE AUTOS

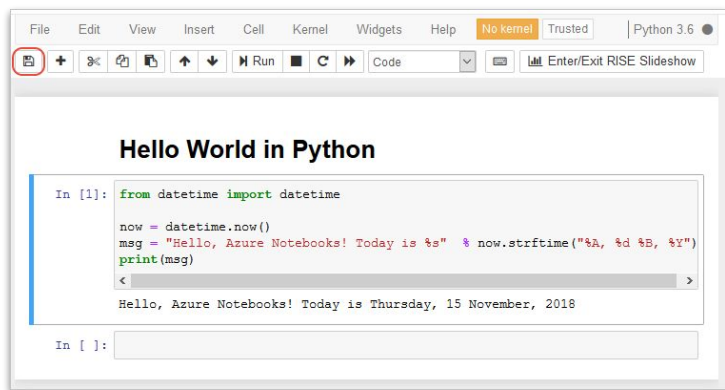
- Dos terminales, A y B.
- Un máximo de 20 autos por terminal.
- Puedo mover máximo en cada noche 5 autos de una terminal a otra. Cada auto cuesta 2\$ moverlo.
- La cantidad de autos *demandados* en cada una de las terminales sigue una distribución de Poisson de medias 3 y 4 respectivamente.
- La cantidad de autos *retornados* en cada una de las terminales sigue una distribución de Poisson de medias 2 y 3 respectivamente.
- Cada auto alquilado da una ganancia de 10\$.
- *Si alguna de las dos terminales se queda sin autos se acaba el negocio.*

EJERCICIO 1.3 - RATA 5.14 - PAPEL

5.13  Una **rata** está atrapada en un laberinto. Inicialmente puede elegir una de tres sendas. Si elige la primera se perderá en el laberinto y luego de 12 minutos volverá a su posición inicial; si elige la segunda volverá a su posición inicial luego de 14 minutos; si elige la tercera saldrá del laberinto luego de 9 minutos. En cada intento, la rata elige con igual probabilidad cualquiera de las tres sendas. Calcular la esperanza del tiempo que demora en salir del laberinto.

5.14 Una rata está atrapada en un laberinto. Inicialmente elige al azar una de tres sendas. Cada vez que vuelve a su posición inicial elige al azar entre las dos sendas que no eligió la vez anterior. Por la primera senda, retorna a la posición inicial en 8 horas, por la segunda retorna a la posición inicial en 13 horas, por la tercera sale del laberinto en 5 horas. Calcular la esperanza del tiempo que tardará en salir del laberinto.

INTRODUCCIÓN A OPENAI GYM



- Introducción General
- Para que “jueguen”: Mountain
- Ejemplo del Robot (Batería)

INTRODUCCIÓN A OPENAI GYM

- ¿Cómo instalarlo? ubuntu 18.04, Python, jupyter, open ai gym
 - Linux:
 - `sudo apt-get update`
 - `sudo apt-get install python3 python3-pip ipython3 python3-fontconfig`
 - `sudo apt-get install libglu1-mesa-dev freeglut3-dev mesa-common-dev python-opengl`
 - `pip3 install numpy pandas matplotlib jupyter gym`
 - Windows:
 - Virtual Box:
 - instalar ubuntu 16.04 y usar el instructivo de la parte de Linux
 - WSL:(inspirado en <https://github.com/openai/gym/issues/11#issuecomment-242950165>)
 - instalar Windows Subsystem for Linux (WSL): <https://docs.microsoft.com/en-us/windows/wsl/install-win10>
 - instalar ubuntu 16.04 LTS para WSL yendo a Microsoft Store (barra de búsqueda de Windows) y buscando Ubuntu 16.04
 - correr una consola WSL (buscar Ubuntu en la barra de búsqueda de Windows)
 - realizar los mismos pasos que en el instructivo de linux en esa consola
 - instalar vcXsrv/xming;
 - correr vcXsrv (elegir one large window); tipear en la consola de comandos de WSL: `export DISPLAY=:0`
 - Correr jupyter: `jupyter notebook --no-browser`
 - Google colab: ir a google colab, <https://colab.research.google.com/notebooks/welcome.ipynb#recent=true>, elegir la solapa Github y buscar en <https://github.com/javkrei/aprendizaje-reforzado-austral>

LECTURAS RECOMENDADAS

- AlphaGo paper: <https://ai.google/research/pubs/pub44806>
- Brief Survey of Deep RL: <https://arxiv.org/pdf/1708.05866.pdf>
- Sutton capítulo 1 para una introducción, capítulo 16 para aplicaciones, 14 y 15 para relación con psicología y neurociencia.