SASE 2013
Simposio Argentino de Sistemas Embebidos

www.sase.com.ar
14/16 de agosto 2013
Fiuba, Buenos Aires, Argentina

Aprendizaje por refuerzo

Juan Carlos Gómez
Claudio Verrastro
juanca@inti.gob.ar
cverra@cae.cnea.gov.ar

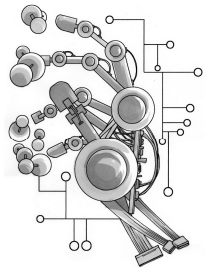


Ilustración: Hernán Juárez

GiAR

UTN.BA
UNIVERSIDAD TECNOLÓGICA NACIONAL
FACULTAD REGIONAL BUENOS AIRES

www.secyt.frba.utn.edu.ar/gia/

Índice

- Contexto
- Aprendizaje, comportamiento, experiencia y tipos de aprendizaje
- Dilema exploración - explotación
- QLearning y SARSAλ
- Experimentos
- Resultados y conclusiones.

Problema a resolver

Sistemas de control para robots

- Multivariable
- Varios parámetros a optimizar
- Modelos complejos
- Sistemas variantes en el tiempo

¿Qué es el aprendizaje?

- RAE: Adquisición de conocimiento por medio de estudio o la experiencia.
- RAE psicología:Adquisición por la práctica de una conducta duradera.

Se pone de manifiesto cuando la experiencia hace que se modifique el comportamiento.

¿Qué es el aprendizaje?
En términos algorítmicos

- Un programa de coputadora APRENDE
  a partir de una experiencia **E**
  a realizar una tarea **T**
  (de acuerdo con una medida de rendimiento **P),**
  si su rendimiento al realizar **T,**
  medido con **P,**
  mejora gracias a la experiencia **E**
  Entonces éxito.
  [Mitchell, 97]

¿Cómo se aprende?

- Aprendizaje inductivo: modelos de conceptos a partir de generalizar ejemplos simples. Patrones comunes (principio de inducción completa)
- Aprendizaje analítico o deductivo: Deducción lógica para obtener soluciones particulares a partir una definición. (modus ponens… Sistemas Expertos)
- Aprendizaje genético: Busqueda estocástica por combinación de soluciones parciales (teorema del esquema).
- Aprendizaje conexionista: aproxima una función multiparamétrica mediante ejemplos (algoritmo de backpropagation)

Objetivo del Aprendizaje Automatico

- Producir programas capaces de mejorar su rendimiento a través de la experiencia
  - Mejoran al realizar una tarea T
  - Respecto a una medida de rendimiento P
  - Gracias a la utilización de la experiencia E

Construir sistemas capaces de adquirir el conocimiento necesario para realizar tareas, usando la experiencia acumulada.

Disciplinas relacionadas

- • Inteligencia Artificial: representación simbólica de conceptos, búsquedas, sistemas expertos,…
- • Estadística: caracterización de errores, muestras, intervalos de confianza, test estadísticos,…
- • Métodos bayesianos: el teorema de Bayes, cálculo de la probabilidad de las hipótesis, …
- • Teoría de Control: procedimientos para el control de procesos
- • Teoría de la Información: medidas de entropía, codificación de hipótesis, MDL, …

Disciplinas relacionadas

- • Teoría Computacional: límites teóricos de la complejidad de las tareas de aprendizaje, …
- • Psicología: ley de la energía de la práctica, más experto, más tiempo, más difícil de mejorar, …
- • Neurología: inspira las redes neuronales artificiales, …
- • Filosofía: la navaja de Ocamm, la hipótesis más simple es la mejor, ….

Aprendizaje por refuerzo

Metodo Inductivo

- Razonamiento Inductivo
- Obtiene conclusiones generales de información específica
- El conocimiento obtenido es nuevo
- No preserva la verdad (nuevo conocimiento puede invalidar lo obtenido)
- No tiene una base teórica bien fundamentada

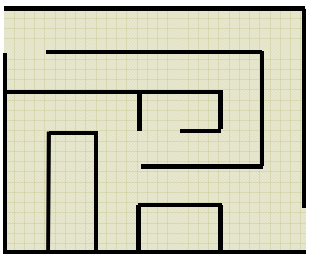
Aprendizaje por refuerzo

Metodo Inductivo

- Desde un punto de vista formal sus resultados no son válidos
- Suponemos que un número limitado de ejemplos representan “Todas” las características que queremos aprender
- Un solo contraejemplo puede invalidar el resultado
- ¡Gran parte del aprendizaje humano es inductivo!

¿Qué observamos?

Homero



Kang y Kodos

Rosquilla

Comportamiento
Toma secuencial de decisiones

Desconocimiento parcial o total del ambiente

Acciones con consecuencias inmediatas o demoradas

Ambiente dinámico

Toma secuencial de decisiones

Maximización de beneficios

Exploración para adquisición de información

Minimización de costos

Sistemas que aprenden

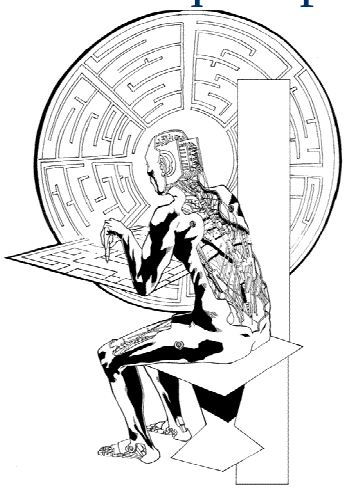


Ilustración Hernán Juárez

Reinforcement learning (RL)

Kaebling, Littman & Moore

RL is the problem faced by an agent that learns behavior through trial and error interactions with a dynamic environment...

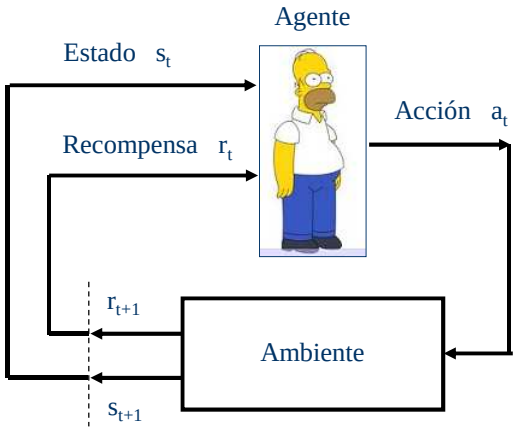
Its promise is beguiling: A way of programming agents by reward and punishment without needing to specify *how* the task is to be achieved...

Russell y Norvig

RL son técnicas para aprender qué hacer a partir de las propias percepciones

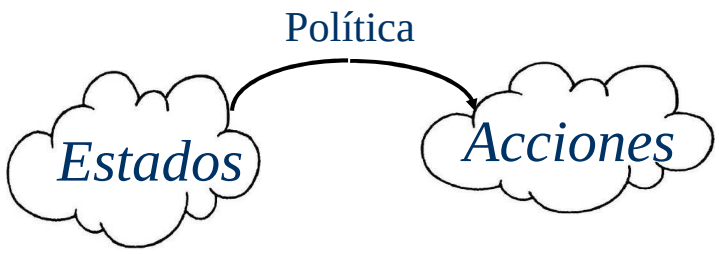
Placer y dolor. Felicidad o infelicidad.
¿robots hedonistas?

Aprendizaje por refuerzo

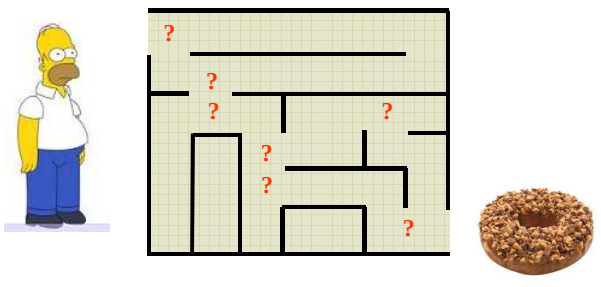


Objetivo

El agente debe encontrar una política que maximice alguna medida de la recompensa esperada a largo plazo.



Recompensa demorada 1



Recompensa demorada 2

Se modelan bien usando procesos de decisión de Markov (MDP)

Procesos de Decisión de Markov

Se definen mediante una tupla

M = (S,A,T,R,γ) con

S: Conjunto de estados en el que se puede encontrar el ambiente y el agente.

A: Conjunto de acciones que puede realizar el agente.

T(*s*,*a*,*s´*): Función de transición, que es la probabilidad de pasar a *s´*, estando en el estado *s* y ejecutando la acción *a*.

R(*s*,*a*): Función de refuerzo, refuerzo esperado al tomar la acción *a* desde el estado *s*.

γ : Factor de descuento.

Procesos de Decisión de Markov

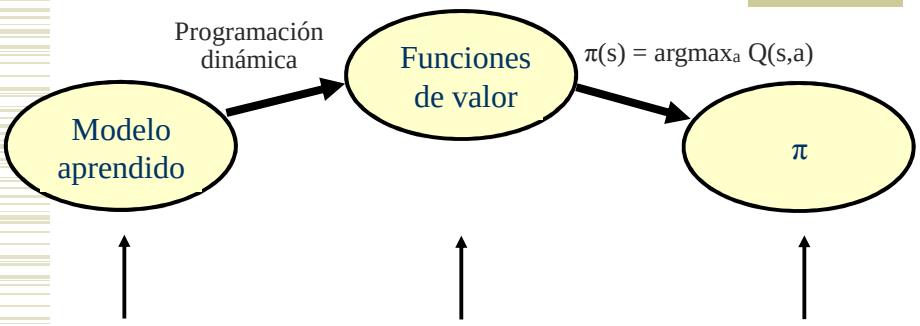
En un MDP las transiciones sólo dependen del estado en que se encuentra el agente.

NO dependen de ningún otro estado anterior, sólo del estado actual.

Andréi Andréyevich Márkov
14/06/1856 – 20/07/1922

Aproximaciones al aprendizaje

Gracias a Diuk



Basados en modelo

Sin modelo

Aprenden Q y V

Búsqueda de la política

Métodos de gradiente sobre π

Dilema

¿cuánto explorar y cuánto explotar?

Explotación vs. Exploración

- Greedy (voraz)
- ε-Greedy
- Softmax / Boltzman

Q – Learning 1

Función de valor de la acción:

$Q(s,a)$

Regla de aprendizaje:

Error de predicción

$$Q'(s,a) := Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

mejor alternativa futura

Q – learning 2

Inicializar $Q(s,a)$

Repetir (para cada episodio)

Inicializar *s := estado inicial*

Repetir (para cada paso del episodio)

se elige una acción *a* considerando el **método de exploración**,

se efectúa la acción *a* en el mundo real (o modelo) y se obtiene la recompensa inmediata *r* y el estado destino *s'*.

se actualiza el valor de $Q(s,a)$ con la regla:

$$Q'(s,a) := Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a))$$

s := s'

Hasta que s sea un estado terminal

Hasta condición de terminación o por siempre

SARSA(λ) 1

Como evolución de Q-Learning...

¿Por qué no actualizar más pares estado-acción hacia atrás, considerando que pasar por ellos es lo que lo trajo a recibir la recompensa actual?

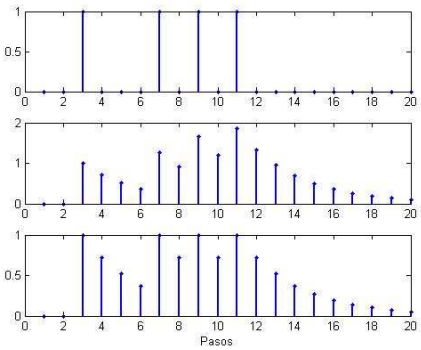
SARSA(λ) 3

Rastro de elegibilidad *(Elegibility trace)*
Medida de cuán recientemente se ha visitado cada par estado-acción.

Visitas

Valor acumulado

Valor de reemplazo



El factor de desvanecimiento de la memoria λ = 0.8 y el de descuento γ = 0.9

SARSA(λ) 4

Se actualiza la función de valor de la acción *Q(s,a)*

$$Q'(s,a) \coloneqq Q(s,a) + \alpha\,\delta\,e(s,a) \qquad \forall(s,a)$$

Donde δ es:

$$\delta(s,a) \coloneqq r + \gamma Q(s',a') - Q(s,a)$$

M

SARSA(λ) 5

Inicializar *Q(s,a) = xx* y, *e(s,a) = 0*

Repetir (para cada episodio):

Inicializar *s*, y luego *a* usando la política de exploración actual (ej: eGreedy)

Repetir (para cada paso del episodio):

Ejecutar acción *a* y observar *r* y *s'*

Elegir acción *a´* (desde s´) con la política de exploración actual (ej:eGreedy)

Calcular: $\delta(s,a) \coloneqq r_t + \gamma Q_{t-1}(s',a') - Q_{t-1}(s,a)$

Actualizar *elegibility trace* según método (acumulativo ó de reemplazo)

Para todo par estado, acción *u*, *a*:

$$Q_t(u,a) \coloneqq Q_{t-1}(u,a) + \alpha\,\delta(s,a)\,e_t(u,a)$$

s:=s' y *a:=a'*

hasta (s = estado terminal)

fin

Ejemplo inicial: La Lumbrí

Modelo de la lumbrí

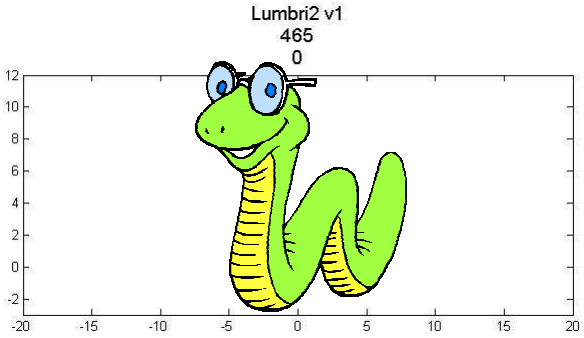
Vínculos: 3 + cabeza y cola

Ángulos entre vínculos: 7

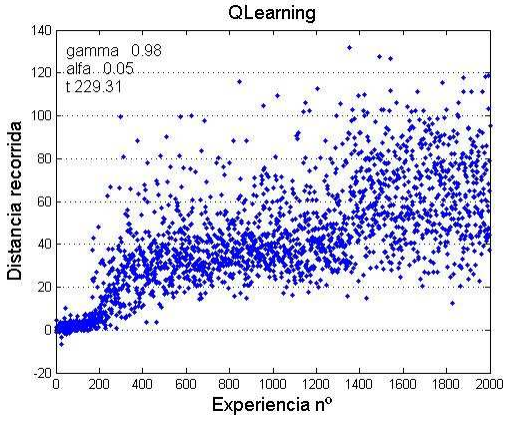
Ángulos del extremo fijo: 5

Cantidad de estados: 3430

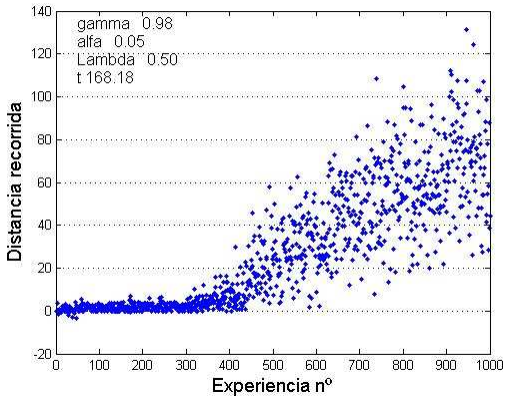
Cantidad de acciones: 27



Experiencia con QLearning



Experiencia con SARSAλ



Resultados

Experiencias en aprendizaje por refuerzo				
Método	Nº de experiencias	Tiempo	Distancia (Promedio 100)	Distancia máxima
QLearning	2000	224.6	42.3	104.5
SARSAλ	1000	167.4	67.3	176.0

Conclusiones y comentarios

- Técnica apropiada para abordar el problema
- SARSAλ tiene mejores resultados
- QLearning es más simple
- Trabajan "*on-line*" (aprendiendo siempre)
- Admiten etapas previas de aprendizaje

Trabajos futuros

Caminar de robots octópodos o hexápodos

Minimizando energía

Maximizando velocidad de desplazamiento

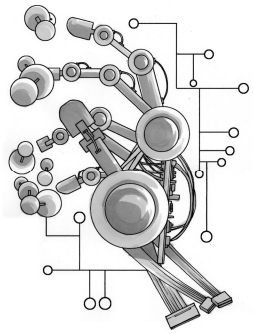
Maximizando confort de la carga

¡ Gracias !

¿Preguntas?

juanca@inti.gob.ar

www.secyt.frba.utn.edu.ar/gia/



Nuestro Homero

Homero Manzi

Periodista, político, poeta, dramaturgo, guionista y director de cine.

SUR
Manzi y Troilo

San Juan y Boedo antigua, y todo el cielo,
Pompeya y más allá la inundación.
Tu melena de novia en el recuerdo
y tu nombre florando en el adiós.
La esquina del herrero, barro y pampa,
tu casa, tu vereda y el zanjón,
y un perfume de yuyos y de alfalfa
que me llena de nuevo el corazón....