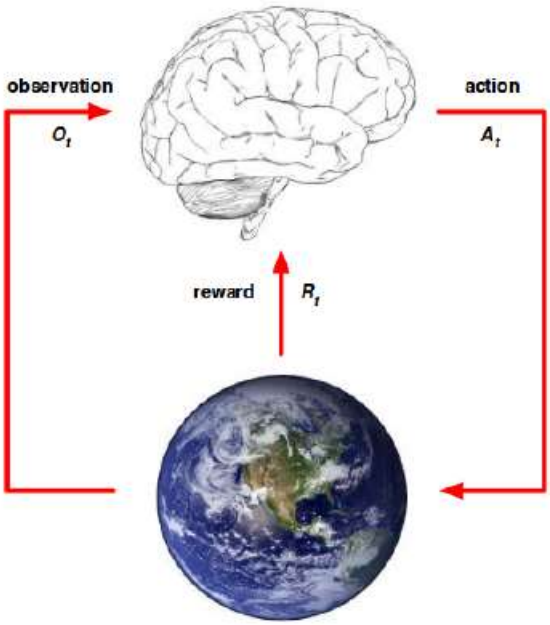
AUTOR: Cuesta Alario David

Aprendizaje Reforzado



Se trata de un proceso de aprendizaje automático basado en un conjunto de recompensas y castigos proporcionas por el entorno que permite al agente obtener información acerca de lo bueno o malo de ejecutar cada acción

Se caracteriza porque:

- No se emplean modelos supervisores que establezcan que acciones son correctas o incorrectas
- El tiempo y la secuencialidad son factores importantes
- Las acciones del agente tienen consecuencias sobre los datos

El agente recibe una observación genera una acción y recibe un premio

El entorno emite una observación recibe una acción y emite un premio

El agente selecciona la acción a realizar basándose en una secuencia de observaciones de las acciones y premios pasados que se representan mediante estados con la finalidad de simplificar el cómputo sin afectar a la información que proporcionan

Función objetivo

La función objetivo puede ser descrita como una maximización del cumulo de los premios esperados

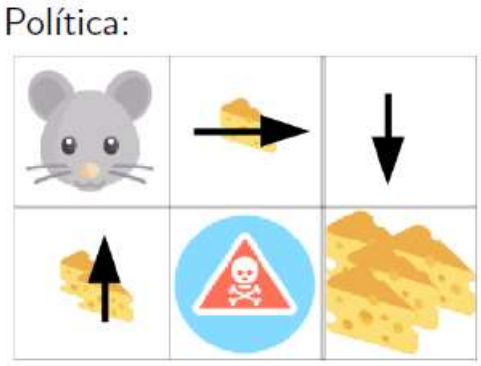
- El premio no tiene por qué ser inmediato:
  - Cada acción puede tener consecuencias a largo plazo.
  - El mejor premio a corto plazo no tiene por qué ser el mejor a largo plazo.
- Asunción de Marcov :
  - El valor futuro de una acción depende únicamente de su valor presente , siendo independiente de la historia de dicha variable, siendo independiente de la historia de dicha variable
  - Para evaluar una acción se le da más peso a los premios presentes que a los premios futuros porque estos son más inciertos

Componentes

Política

Es la estrategia-comportamiento del agente según la cual decide que acción tomar

- **Determinística**: Dado un estado siempre se ejecuta la acción que nos lleve al mejor resultado posible.
  Es más conservador debido a que Solo exploto el conocimiento que he ido almacenando hasta el momento
- **Estocástica**: En ocasiones se permite la exploración ignorando lo aprendido y seleccionando una acción al azar
  Es menos conservador debido a que favorece aprender nuevas relaciones



Función de Valor

Medida de lo bueno o malo que es un estado o acción de cara a los premios que puedo conseguir en un futuro.

- Se suele aprender mediante una tabla de acciones que asocia el estado actual y una acción con otro estado
- Permite deducir la política

Modelo

Es la representación que posee el agente sobre el entorno

- Permite predecir el comportamiento del entorno
- Requiere mucha complejidad para ser aprendido y normalmente no se hace

AUTOR: Cuesta Alario David

Métodos de aprendizaje

Por episodios

Consiste en exponer al agente a un ciclo completo de estados donde hay un estado inicial y un estado final.

Método 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.

Aprendizaje Temporal:

- No se espera hasta el final.
- En cada paso se va actualizando el valor del estado.