Aprendizaje Automático y Data Mining

Bloque V

APRENDIZAJE POR REFUERZO

Índice

- Introdución.
- Modelado mediante estados y acciones.
- Recompensa retrasada.
- Aprendizaje Q (Q-learning)

INTRODUCCIÓN

Introducción (I)

- No hay ejemplos de entrenamiento, no se suministran ejemplos etiquetados (aprendizaje no supervisado).
- Se aprende mediante prueba y error.
- El sistema realiza una determinada tarea repetidamente, para adquirir experiencia y mejorar su comportamiento.
- Se requiere un número de repeticiones muy elevado.
- Aplicación práctica limitada.

Introducción (II)

- Aplicaciones:
  - en procesos que se realizan como una secuencia de acciones:
    - Robots móviles: aprendizaje de la forma de escapar de un laberinto.
    - Juego de ajedrez: aprendizaje de la mejor secuencia de movimientos para ganar un juego.
    - Brazo robot: aprendizaje de la secuencia de pares a aplicar a las articulaciones para conseguir un cierto movimiento.

MODELADO MEDIANTE ESTADOS Y ACCIONES

Estados y acciones (I)

- Se deben modelar dos elementos:
  - Estados: posibles situaciones para el sistema (ej. Posibles situaciones del robot móvil en el laberinto o posibles situaciones de piezas en el tablero de ajedrez.
  - Acciones: posibles acciones que el sistema puede realizar en un momento determinado (ej. Posibles movimientos para el robot: izquierda, derecha, adelante; o movimientos válidos de las fichas en el ajedrez).
- El objetivo es aprender cuál es la mejor acción a ejecutar cuando el sistema se encuentra en un cierto estado.

RECOMPENSA RETRASADA

Recompensa retrasada (I)

- Proceso de aprendizaje:
  - Inicialmente, se ejecutan acciones de forma aleatoria desde cualquier estado.
  - Cuando una de esas acciones produce el resultado deseado, es recompensada.
- Problema: recompensa retrasada. El resultado no se conoce inmediatamente después de una acción, sino después de una larga secuencia de acciones (ej. una partida de ajedrez completa).
- Las recompensas suelen tomar sólo dos valores: 1 o 0 (ej. 1: partida ganada; 0: partida perdida).

Recompensa retrasada (II)

- Recompensa acumulada: recompensa obtenida durante todas las acciones ejecutadas por el sistema.
- La recompensa es tanto mayor cuanto antes se alcance el resultado deseado:
  - menos movimientos de ajedrez para ganar una partida.
  - menor recorrido realizado por el robot para salir del laberinto.
- Recompensa acumulada para todas las acciones realizadas desde el instante t y el estado s_t:

$$V(s_t) = r_t + g \cdot r_{t+1} + g^2 \cdot r_{t+2} + g^3 \cdot r_{t+3} + \ldots$$

$$V(s_t) = \sum_{i=0}^{\infty} g^i \cdot r_{t+i}$$

  - r_{t+i} = recompensa de la acción realizada en el instante $t+i$.
  - $\gamma < 1$

APRENDIZAJE Q

Aprendizaje Q (I)

- Objetivo del aprendizaje: inferir la función Q(s,a)
  - s=estado, a=acción.
- Q(s,a) representa la máxima recompensa que se podría obtener desde el estado s si:
  - La primera acción que se realiza es la acción a.
  - Las acciones posteriores se suponen perfectas (las mejores posibles).
- Conocido Q(s,a), el sistema sabe qué acción tiene que ejecutar en cada estado:
  - La acción a elegir es la que ofrece el valor máximo para Q(s,a).

Aprendizaje Q (II)

- ¿Cómo inferir Q?
  1. Inicialmente, Q(s,a) toma valores aleatorios (A cada estado y a cada acción posible en ese estado se les asigna un valor Q aleatorio).
  2. Partiendo del estado inicial, se realizan acciones aleatoriamente hasta que se cumple una de estas condiciones:
     1. El sistema alcanza el objetivo buscado.
     2. El número de acciones realizadas alcanza un límite máximo.
  3. Se actualizan los valores de Q en función del resultado obtenido:
     1. Si se ha alcanzado el objetivo, se recompensan las acciones proporcionalmente a la distancia al objetivo final.
     2. Si se ha realizado el número límite de acciones sin alcanzar el objetivo final, no hay recompensa.

Aprendizaje Q(III)

- Las etapas 2 y 3 se repiten hasta cumplirse una de estas dos condiciones:
  - Q converge (a un cierta precisión).
  - Se alcanza un número máximo de repeticiones.
- Se ha demostrado que es posible obtener Q con total precisión si se recorren todos los estados y todas las acciones un número infinito de veces.
- En la práctica no es necesario alcanzar una precisión total, pero aun así se requiere un número de iteraciones muy elevado.
- Algunas demostraciones:
  - Pathlearner.exe
  - www.fe.dis.titech.ac.jp/~gen/robot/robodemo.html

Bloque V

APRENDIZAJE POR REFUERZO