Curso de Sistemas Inteligentes
Práctica de Laboratorio No. 3

Temporal Difference Learning

Prof. Francisco Cruz

Ayudante Angel Ayala

Semestre 2 / 2018

1. Fundamentos Teóricos

Temporal Difference (TD) learning es una de las ideas centrales para *reinforcement learning* (RL), este enfoque corresponde a una combinación de las ideas del método de Monte Carlo y de programación dinámica.

Este método permite la estimación de la función de valor para una determinada política, permitiendo converger en una política óptima. El acercamiento más simple para TD corresponde a TD(0), donde la actualización se genera mediante la siguiente ecuación:

$$V(S_t) = V(S_t) + \alpha[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)] \tag{1}$$

1.1. Métodos de predicción

La predicción de valores de acción, permiten el control del aprendizaje para el método de TD, asegurando la convergencia de este. La implementación de estas predicciones se puede realizar mediante dos maneras principalmente, on-policy y off-policy.

1.1.1. Método on-policy: SARSA

Para aprender un valor de acción, se debe estimar el valor de pares estado-acción $q_\pi(s, a)$ para el comportamiento actual de la política, considerando como base la ecuación de TD (1), este método consiste en la reiteración de episodios de una secuencia de pares estados acción como se puede observar en la siguiente ecuación:

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1})) - Q(S_t, A_t)] \tag{2}$$

1.1.2. Método off-policy: Q-learning

La predicción de la acción de valor, para este caso, ocurre considerando el mayor valor en el siguiente estado s_{t+1} sin considerar la política aplicada. Este método genera la convergencia hacia la política óptima $q*$.

$$Q(S_t, A_t) = Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \underset{a}{max} Q(S_{t+1}, a) - Q(S_t, A_t)] \quad (3)$$

2. Actividades

Para demostrar lo indicado anteriormente, se solicita desarrollar un agente que sea capaz de resolver un entorno de grilla cliff walking (véase Figura 1), donde debe ser capaz de obtener una mayor cantidad de recompensa.

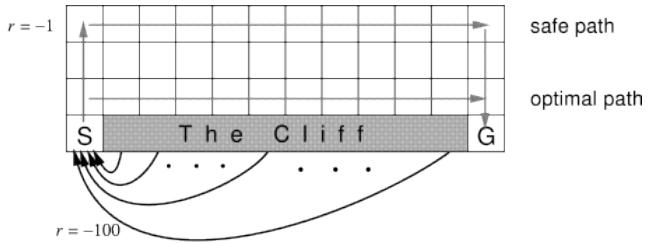

Figura 1: Entorno de Cliff Walking [1]

1. Explicar brevemente el entorno.
2. Diseñar el agente para el MDP del entorno especificado.
3. Implementar los métodos de Q-learning y SARSA.
4. Graficar la recompensa obtenida durante cada episodio.
5. Configurar el valor de ϵ en 0,5, y otro valor a su preferencia.
6. Graficar los resultados de ambos métodos con los nuevos valores de ϵ.

3. Informe

El informe debe ser realizado en formato IEEE de doble columna con un máximo de 5 páginas. Incluir al menos las siguientes secciones:

- Título, autores y filiación.
- Resumen.

- Introducción.
- Fundamentos teóricos.
- Desarrollo y explicación de lo realizado.
    - Descripción del MDP y sus carácteristicas.
    - Especificación del agente y sus métodos de predicción.
    - Entrenamiento y comparación de resultados.
- Conclusiones.
- Referencias científicas en formato IEEE (al menos 5).

3.1. Entrega del Informe

Para la entrega considerar lo siguiente:

- Se debe presentar el código implementado mediante un repositorio Git.
- Solo se aceptarán informes en el formato solicitado.
- La entrega del trabajo como el repositorio es individual.
- El informe debe ser enviado al correo angel4ayala@gmail.com con copia a francisco.cruz@ucentral.cl el día de la entrega hasta las 11:59pm.
- Se aceptarán informes de laboratorio atrasados sujetos a castigo de un punto menos por día (incluido sábado, domingo y feriados).
- No adjuntar el archivo o adjuntar el archivo incorrecto es responsabilidad del alumno.
- El archivo debe ser enviado en formato PDF usando como nombre de archivo el siguiente formato:

  <Nombre><Apellido>LabSI<N° de lab><Semestre><Año>.pdf (4)

  Por ejemplo, IsaacNewtonLabSI222018.pdf correspondería al laboratorio 2 del alumno Isaac Newton.
- El plagio será sancionado con la nota mínima, sin posibilidad de realizar el trabajo nuevamente.
- **Fecha de entrega:** Martes 09 de noviembre.

Referencias

[1] Richard S. Sutton and Andrew G. Barto: *Reinforcement Learning: An Introduction* The MIT Press, Cambridge, Massachusetts, London, England 1998.