Bài thực hành Học Tăng cường:

Tìm đường trong mê cung bằng Q-Learning

1. Mục tiêu

Sinh viên sẽ lập trình một tác nhân (agent) sử dụng thuật toán Q-Learning để tìm đường từ điểm bắt đầu đến mục tiêu trong một mê cung 5x5, tránh các bức tường và tối ưu hóa phần thưởng.

1.1 Mô tả bài toán

Môi trường: Mê cung 5x5 (25 ô), được biểu diễn như sau:

```
S . . . .
. W W . .
. . W . .
W . . . .
. . W . G
```

S: Điểm bắt đầu (Start) tại (0,0).

G: Mục tiêu (Goal) tại (4,4).

W: Tường (Wall) không thể đi qua.

. (dấu .): Ô trống có thể di chuyển.

Tác nhân: Một robot di chuyển trong mê cung.

Hành động: 4 lựa chọn - Lên (0), Xuống (1), Trái (2), Phải (3).

Trạng thái: Tọa độ hiện tại của robot (x, y).

Phần thưởng:

+10 khi đến mục tiêu (4,4).

-1 khi di chuyển đến ô trống (phạt nhỏ để khuyến khích đi nhanh).

-5 khi va vào tường (không thay đổi trạng thái).

Mục tiêu: Tìm đường ngắn nhất từ (0,0) đến (4,4).

1.2 Yêu cầu

a) Xây dựng môi trường:

Tạo một mảng 5x5 đại diện cho mê cung (0: trống, 1: tường, 2: mục tiêu).

Viết hàm kiểm tra hành động hợp lệ (không vượt qua tường, không ra khỏi mê cung) và cập nhật trạng thái.

b) Triển khai Q-Learning:

Khởi tạo Q-table (25 trạng thái x 4 hành động) với giá trị 0.

Sử dụng công thức cập nhật:

$Q(s,a) \leftarrow Q(s,a) + \alpha[R + \gamma \max Q(s',a') - Q(s,a)]$

$\alpha = 0.1$ (tốc độ học),

$\gamma = 0.9$ (hệ số chiết khấu).

Áp dụng chiến lược ϵ-greedy ($\epsilon = 0.1$) bằng cách gieo xác suất np.random.rand(). Nếu xác suất gieo nhỏ hơn ϵ-greedy thì chọn hành động ngẫu nhiên. Ngược lại thì chọn hành động tối ưu từ Q-table

c) Huấn luyện:

Chạy 200 episodes, mỗi episode bắt đầu từ (0,0) và kết thúc khi đến (4,4) hoặc sau 100 bước.

Lưu đường đi tối ưu sau khi huấn luyện.

Lưu danh sách tích luỹ reward mỗi lần chạy để vẽ biểu đồ

d) Kiểm tra:

Chạy tác nhân với Q-table đã huấn luyện (không khám phá) và in ra đường đi.

2. Bài tập:

Bạn hãy huấn luyện mô hình học tăng cường để tìm đường đi tối ưu với mê cung sau:

```
S . . . .
. W W . W
. . . W .
```

. . . .
W . W . G

Sau đó trình bày kết quả trong file pdf với tên B5_RL_MSSV bao gồm:

a) đường đi tối ưu của bạn, đồ thị phần thưởng quá trình học.
b) đường đi tối ưu, đồ thị phần thưởng với tốc độ học (alpha) 0.05, hệ số chiết khấu (gamma) 0.8, xác suất khám phá (epsilon) 0.1
c) đường đi tối ưu, đồ thị phần thưởng với tốc độ học (alpha) 0.1, hệ số chiết khấu (gamma) 0.7, xác suất khám phá (epsilon) 0.15
d) đường đi tối ưu, đồ thị phần thưởng với tốc độ học (alpha) 0.15, hệ số chiết khấu (gamma) 0.9, xác suất khám phá (epsilon) 0.15
e) Giải thích sự khác nhau ở đồ thị tích luỹ tổng phần thưởng b,c,d
f) hãy tạo 1 ô bẫy trên đường đi với phần thưởng là -30 và huấn luyện lại. In mê cung đã tạo bẫy, và in đường đi sau khi huấn luyện
g) chép và dán toàn bộ code vào cuối file pdf và đặt tên theo định dạng B5_RL_MSSV