**Комп’ютерний практикум №3**

**MDP і навчання з підкріпленням**

**Мета роботи:** ознайомитись з методами пошуку в умовах невідомості та навчання з підкріпленням в моделях на основі станів; дослідити їх використання для інтелектуального агента в типовому середовищі.

***З*авдання:**

1. Обрати середовище моделювання та задачу, що містить агента, який може бути навчений методом «з підкріпленнями».
2. В обраному середовищі вирішити задачу знаходження найкращої стратегії поведінки, реалізувавши метод Q-learning (Q-навчання) або Value iteration, залежно від завдання.

Дозволяється в якості альтернативи використати інший алгоритм навчання з підкріпленням: метод оцінювання стратегій (SARSA, TD-learning), глибокого навчання (DQN та ін.)., розібравшись з принципом роботи і реалізації такого алгоритму.

1. Виконати дослідження реалізованого методу: задача дослідження визначається варіантом
2. Описати середовище, використаний метод та результати його застосування і дослідження у звіті.

**Середовище.**

В якості середовища рекомендується використане двомірне середовище(Grid World) з чітко вираженими просторово станами, наприклад:

* середовище Volcano Crossing, розглянуте на лекціях
* просте середовище з бомбою та золотом з практики
* середовище Pacman з наступної роботи;
* середовища з класичних задач «кіт-миша»[3]:
* світ Вампуса[2];
* середовище для робота-пилососа у двомірному середовищі

Розмір такого середовища – мінімум 10х10, краще більше. Середовище буде містити «перешкоди» та «стіни», що відповідають областям (станам), в які не можна потрапити і стани з «винагородами» (велика позитивна нагорода, епізод завершується при її досягненні), виконання кожної дії(переміщення) зменшує загальну винагороду на невелику постійну величину(вартість руху). Опціонально можуть бути також «штрафні» стани, з великою негативною винагородою і завершенням епізоду при потраплянні в них.

Для класичних методів MDP cередовище або дії агента мають мати елементи стохастичності, наприклад, має бути імовірність помилкового виконання дії (агент потрапляє в сусідній від запланованого стан), а для навчання з підкріпленням має бути реалізовано дослідження середовища. Для навчання з підкріпленням агент не знає конфігурацію приміщення, положення і значення винагород та імовірності відмови при виконанні дії.

Дозволяється для виконання роботи обрати також деяке інше типове середовище невисокої складності, наприклад: **OpenAI Gym** (<https://gym.openai.com/>)- бібліотека Python, яка надає API для розробки і порівняння RL алгоритмів з величезною кількістю віртуальних середовищ і загальним API. Gym вже має безліч готових оточень, в тому числі для ігор Atari. При цьому завдання(алгоритм і задача дослідження не змінюється)

**Звіт має містити:**

1. Постановку задачі, короткий опис середовища, його станів, дій та ін.
2. Короткий опис алгоритму
3. Результати роботи алгоритму
4. Опис експерименту та його результати
5. Висновки

**Контрольні запитання:**

1. Порівняйте навчання з підкріпленням з іншими парадигмами навчання.
2. Охарактеризуйте область застосування навчання з підкріпленням.
3. Назвіть основні компоненти системи навчання з підкріпленням та охарактеризуйте їх значення
4. Розкрийте суть методу Q-навчання та Value iteration. В чому їх відмінність від SARSA, Police iteration.
5. Які переваги та обмеження мають методи навчання з підкріпленням? Окресліть сферу застосування даних методів.

**Рекомендовані матеріали:**

1. Електронний конспект (презентація). Теми: марківські процеси прийняття рішень, методи навчання з підкріпленням.
2. Рассел С., Норвіг П. «Штучний інтелект. Сучасний підхід». Розділи 10, 17 та 21.
3. Деякі електронні ресурси:

<https://habr.com/ru/company/newprolab/blog/343834/>

<https://habr.com/ru/post/345656/>

<https://proglib.io/p/reinforcement-learning/>

<http://incompleteideas.net/book/RLbook2018.pdf>

**Варіанти завдань**

Варіант індивідуального завдання визначається **за номером в списку групи**, для бригади – **номером першого студента бригади в списку групи** (якщо інше не визначено викладачем):

|  |  |  |  |
| --- | --- | --- | --- |
| **Номер студента/бригади** | **Форма карти** | **Задача дослідження** | **Алгоритм** |
| 1 | кімнати з вузькими проходами-дверми | вплив імовірності невірного результату дії | Value iteration |
| 2 | кімнати з вузькими проходами-дверми | вплив імовірності невірного результату дії | Q-learning |
| 3 | кімнати з вузькими проходами-дверми | вплив коефіцієнта знецінення | Value iteration |
| 4 | кімнати з вузькими проходами-дверми | вплив коефіцієнта знецінення | Q-learning |
| 5 | кімнати з вузькими проходами-дверми | вплив коефіцієнта швидкості навчання α | Value iteration |
| 6 | кімнати з вузькими проходами-дверми | вплив коефіцієнта швидкості навчання α | Q-learning |
| 7 | кімнати з вузькими проходами-дверми | вплив коефіцієнта дослідження ε | Q-learning |
| 8 | кімнати з вузькими проходами-дверми | Влив значення винагороди та вартості руху | Q-learning |
| 9 | кімнати з вузькими проходами-дверми | Влив значення винагороди та вартості руху | Value iteration |
| 10 | приміщення з окремими перешкодами | вплив імовірності невірного результату дії | Value iteration |
| 11 | приміщення з окремими перешкодами | вплив імовірності невірного результату дії | Q-learning |
| 12 | приміщення з окремими перешкодами | вплив коефіцієнта знецінення | Value iteration |
| 13 | приміщення з окремими перешкодами | вплив коефіцієнта знецінення | Q-learning |
| 14 | приміщення з окремими перешкодами | вплив коефіцієнта швидкості навчання α | Value iteration |
| 15 | приміщення з окремими перешкодами | вплив коефіцієнта швидкості навчання α | Q-learning |
| 16 | приміщення з окремими перешкодами | вплив коефіцієнта дослідження ε | Q-learning |
| 17 | приміщення з окремими перешкодами | Влив значення винагороди та вартості руху | Q-learning |
| 18 | приміщення з окремими перешкодами | Влив значення винагороди та вартості руху | Value iteration |
| 19 | лабіринт | вплив імовірності невірного результату дії | Value iteration |
| 20 | лабіринт | вплив імовірності невірного результату дії | Q-learning |
| 21 | лабіринт | вплив коефіцієнта знецінення | Value iteration |
| 22 | лабіринт | вплив коефіцієнта знецінення | Q-learning |
| 23 | лабіринт | вплив коефіцієнта швидкості навчання α | Value iteration |
| 24 | лабіринт | вплив коефіцієнта швидкості навчання α | Q-learning |
| 25 | лабіринт | вплив коефіцієнта дослідження ε | Q-learning |
| 26 | лабіринт | Влив значення винагороди та вартості руху | Q-learning |
| 27 | лабіринт | Влив значення винагороди та вартості руху | Value iteration |
| 28 | приміщення без перешкод | вплив імовірності невірного результату дії | Value iteration |
| 29 | приміщення без перешкод | вплив імовірності невірного результату дії | Q-learning |
| 30 | приміщення без перешкод | вплив коефіцієнта знецінення | Value iteration |
| 31 | приміщення без перешкод | вплив коефіцієнта знецінення | Q-learning |
| 32 | приміщення без перешкод | вплив коефіцієнта швидкості навчання α | Q-learning |
| 33 | приміщення без перешкод | вплив коефіцієнта швидкості навчання α | Q-learning |
| 34 | приміщення без перешкод | вплив коефіцієнта дослідження ε | Q-learning |
| 35 | приміщення без перешкод | Влив значення винагороди та вартості руху | Q-learning |
| 36 | приміщення без перешкод | Влив значення винагороди та вартості руху | Value iteration |
| 37 | кімнати+опенспейс | вплив карти | Value iteration |
| 38 | перешкоди+опенспейс | вплив карти | Value iteration |
| 39 | лабіринт+опенспейс | вплив карти | Value iteration |
| 40 | кімнати+лабіринт | вплив карти | Value iteration |
| 41 | перешкоди+ лабіринт | вплив карти | Value iteration |
| 42 | кімнати+перешкоди | вплив карти | Value iteration |
| 43 | кімнати+опенспейс | вплив карти | Q-learning |
| 44 | перешкоди+опенспейс | вплив карти | Q-learning |
| 45 | лабіринт+опенспейс | вплив карти | Q-learning |
| 46 | кімнати+лабіринт | вплив карти | Q-learning |
| 47 | перешкоди+ лабіринт | вплив карти | Q-learning |
| 48 | кімнати+перешкоди | вплив карти | Q-learning |