



Навчання з підкріпленням

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	<i>Другий (магістерський)</i>
Галузь знань	<i>12 Інформаційні технології</i>
Спеціальність	<i>121 Інженерія програмного забезпечення 123 Комп'ютерна інженерія 126 Інформаційні системи та технології</i>
Освітня програма	<i>Інженерія програмного забезпечення комп'ютерних систем Інженерія програмного забезпечення інформаційних систем Комп'ютерні системи та мережі Інтегровані інформаційні системи Інформаційні управляючі системи та технології Інформаційне забезпечення робототехнічних систем Інформаційні системи та технології</i>
Статус дисципліни	<i>Вибіркова</i>
Форма навчання	<i>очна(денна)</i>
Рік підготовки, семестр	<i>2 курс, осінній семестр</i>
Обсяг дисципліни	<i>4 кредити/120 год. денна форма: лекцій 36 годин, лабораторних занять 18 год., СРС 66 год.</i>
Семестровий контроль/ контрольні заходи	<i>Залік</i>
Розклад занять	<i>schedule.kpi.ua</i>
Мова викладання	<i>Українська</i>
Інформація про керівника курсу / викладачів	Лектор: ст. викладач Кочура Юрій Петрович, iuriy.kochura@gmail.com Лабораторні роботи: ст. викладач Кочура Юрій Петрович, iuriy.kochura@gmail.com
Розміщення курсу	https://classroom.google.com/c/ODIzNjQzNTgyOTM2?cjc=tdzrpjsk

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчення та результати навчання

Навчання з підкріпленням (англ. reinforcement learning, RL) — це тип машинного навчання для автоматизованого прийняття рішень на основі взаємодій: агент навчається приймати оптимальні рішення, беручи до уваги винагороди, які отримує від середовища у ході взаємодії. Починаючи з 2014 року глибоке навчання з підкріпленням (deep RL) зазнало значних успіхів та стало одним з найінтенсивніших напрямків досліджень у галузі штучного інтелекту. Сьогодні навчання нейронних мереж з підкріпленням активно використовується у робототехніці, фінансовій сфері та іграх. У ряді завдань досягнуто надлюдської продуктивності: відео ігри, покер, а також у настільних іграх, включаючи го та шахи.

Цей курс познайомить Вас з сімейством статистичних алгоритмів, які вивчають оптимальну стратегію, метою якої є максимізація загальної винагороди, отриманої агентом при взаємодії з навколишнім середовищем.

Предмет навчальної дисципліни – методи та засоби навчання з підкріпленням, які спрямовані на моделювання процесів прийняття рішень агентами з метою максимізації сумарної винагороди у різних середовищах.

Навчальна мета

- Познайомитись з фундаментальною теорією та особливостями RL;
- Навчитись вирішувати прості задачі за допомогою методів RL;
- Познайомитись з дослідженнями в галузі RL ($\leq$ останні 10 років).

Вивчення дисципліни формує наступні **компетенції**:

- Здатність аналізувати та моделювати процеси прийняття рішень агентами у середовищах з невизначеністю, використовуючи методи навчання з підкріпленням.
- Здатність застосовувати алгоритми навчання з підкріпленням та глибокого навчання з підкріпленням для побудови інтелектуальних агентів і систем автоматизованого прийняття рішень.

Програмні результати навчання, на формування яких спрямована дисципліна:

- Уміти обирати та реалізовувати алгоритми навчання з підкріпленням для розв'язання прикладних задач.
- Уміти оцінювати ефективність стратегій навчання агентів, аналізувати результати експериментів та робити висновки щодо оптимальності поведінки агентів.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Пререквізити

Для опанування дисципліни необхідні базові знання з вищої математики, аналітичної геометрії та лінійної алгебри, дискретної математики, а також уміння програмувати та розуміння основ нейронних мереж і систем штучного інтелекту.

Постреквізити

Розглянута теорія та отриманий практичний досвід в рамках даної навчальної дисципліни буде корисним для розв'язання прикладних задач у робототехніці, автономному керуванні, фінансовому прогнозуванні, оптимізації бізнес-процесів та створенні інтелектуальних ігрових агентів.

Знання, одержані студентами при вивченні дисципліни можуть бути використані у дисциплінах «Науково-дослідна практика» та «Виконання магістерської дисертації».

3. Зміст навчальної дисципліни

Розділ 1. Вступ до навчання з підкріпленням

Тема 1.1. Основні поняття навчання з підкріпленням.

Тема 1.2. Концепція взаємодії «агент – середовище».

Тема 1.3. Ключові елементи задачі навчання з підкріпленням: стан, дія, винагорода, стратегія, функція цінності.

Тема 1.4. Підходи до навчання з підкріпленням: навчання з моделлю середовища та без моделі.

Тема 1.5. Приклади практичних застосувань навчання з підкріпленням.

Розділ 2. Марковські процеси

Тема 2.1. Марковські процеси станів.

Тема 2.2. Марковські процеси винагороди.

Тема 2.3. Марковські процеси прийняття рішень.

Тема 2.4. Властивість Маркова, матриця переходів, функції винагород.

Тема 2.5. Математичний апарат опису взаємодії агента із середовищем.

Розділ 3. Планування за допомогою динамічного програмування

Тема 3.1. Основні ідеї динамічного програмування.

Тема 3.2. Принцип оптимальності Беллмана.

Тема 3.3. Ітерація цінності (Value Iteration).

Тема 3.4. Ітерація стратегії (Policy Iteration).

Тема 3.5. Обчислювальна складність методів динамічного програмування та їх практичні обмеження.

Розділ 4. Методи навчання Монте-Карло та методи з часовими різницями (TD-learning)

Тема 4.1. Стохастичне моделювання та оцінка функцій очікування.

Тема 4.2. Оцінка функції цінності методом Монте-Карло.

Тема 4.3. Методи з часовими різницями TD(0) та TD(λ).

Тема 4.4. Порівняння Монте-Карло та TD-методів.

Тема 4.5. Поєднання підходів і практичні приклади застосування.

Розділ 5. Методи апроксимації функції цінності

Тема 5.1. Необхідність апроксимації у великих або неперервних просторах станів.

Тема 5.2. Лінійні методи апроксимації.

Тема 5.3. Нелінійні методи апроксимації.

Тема 5.4. Використання нейронних мереж.

Розділ 6. Градієнтні методи стратегій (Policy Gradient Methods)

Тема 6.1. Пряме оптимізаційне навчання стратегій.

Тема 6.2. Стохастичні градієнтні методи.

Тема 6.3. Переваги та обмеження підходів на основі стратегії.

4. Навчальні ресурси та матеріали

4.1. Базова література

1. Кочура, Ю. П. Нейронні мережі [Електронний ресурс] : підручник для здобувачів ступеня бакалавра/магістра за спец. 123 «Комп'ютерна інженерія» та 121 «Інженерія програмного забезпечення» / Ю. П. Кочура, Ю. Г. Гордієнко ; КПІ ім. Ігоря Сікорського. – Електрон. текст. дані (1 файл: 3,98 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2024. – 114 с. – (Серія «Штучний інтелект для індустрії»). – Назва з екрана. URL <https://ela.kpi.ua/handle/123456789/73660>

4.2. Додаткова література

2. Sutton, R. S., & Barto, A. G. (2018). [*Reinforcement learning: An introduction*](#). MIT press.

3. Zai, A., & Brown, B. (2020). [*Deep reinforcement learning in action*](#). Manning Publications.

5. Методика опанування навчальної дисципліни (освітнього компоненту)

5.1. Лекційні заняття

№ розділу	Назва лекції та перелік основних питань	Кількість ауд. годин
1	Розділ 1. Вступ до навчання з підкріпленням Лекція 1. Базові поняття та структура задачі навчання з підкріпленням Тема 1.1. Основні поняття навчання з підкріпленням. Тема 1.2. Концепція взаємодії «агент – середовище». Тема 1.3. Ключові елементи задачі навчання з підкріпленням: стан, дія, винагорода, стратегія, функція цінності. Лекція 2. Підходи та практичні застосування навчання з підкріпленням Тема 1.4. Підходи до навчання з підкріпленням: навчання з моделлю середовища та без моделі. Тема 1.5. Приклади практичних застосувань навчання з підкріпленням.	4
2	Розділ 2. Марковські процеси Лекція 3. Основи марковських процесів Тема 2.1. Марковські процеси станів. Тема 2.2. Марковські процеси винагорода. Лекція 4. Марковські процеси прийняття рішень та математичний апарат опису взаємодії Тема 2.3. Марковські процеси прийняття рішень. Тема 2.4. Властивість Маркова, матриця переходів, функції винагород. Тема 2.5. Математичний апарат опису взаємодії агента із середовищем.	4
3	Розділ 3. Планування за допомогою динамічного програмування Лекція 5. Вступ до динамічного програмування Тема 3.1. Основні ідеї динамічного програмування. Тема 3.2. Принцип оптимальності Беллмана. Лекція 6. Ітерація цінності Тема 3.3. Ітерація цінності (Value Iteration). Виведення рівняння Беллмана для цінності. Приклади ітераційного обчислення функції	8

	цінності. Лекція 7. Ітерація стратегії Тема 3.4. Ітерація стратегії (Policy Iteration). Оцінювання стратегії та її покращення. Порівняння з ітерацією ціності. Лекція 8. Обмеження методів динамічного програмування Тема 3.5. Обчислювальна складність методів динамічного програмування та їх практичні обмеження. Проблеми масштабованості у великих просторах станів. Використання наближених методів.	
4	Розділ 4. Методи навчання Монте-Карло та методи з часовими різницями (TD-learning) Лекція 9. Стохастичне моделювання у навчанні з підкріпленням Тема 4.1. Стохастичне моделювання та оцінка функцій очікування. Роль випадковості у процесі навчання. Очікуване значення винагороди та його оцінка на основі вибірки. Лекція 10. Методи Монте-Карло для оцінки функцій ціності Тема 4.2. Оцінка функції ціності методом Монте-Карло. Епізодичне навчання. Оцінка стратегій та оновлення функцій ціності. Лекція 11. Методи з часовими різницями Тема 4.3. Методи з часовими різницями TD(0) та TD(λ). Зв'язок між Монте-Карло і динамічним програмуванням. Накопичення слідів та роль параметра λ. Лекція 12. Порівняння та поєднання підходів Тема 4.4. Порівняння Монте-Карло та TD-методів. Тема 4.5. Поєднання підходів і практичні приклади застосування.	8
5	Розділ 5. Методи апроксимації функції ціності Лекція 13. Масштабування Тема 5.1. Необхідність апроксимації у великих або неперервних просторах станів. Обмеження табличних методів. Узагальнення досвіду агента через функціональні наближення. Лекція 14. Лінійні методи апроксимації функції ціності Тема 5.2. Лінійні методи апроксимації. Основні принципи лінійної апроксимації. Вибір базисних функцій. Застосування до TD-методів. Лекція 15. Нелінійні методи апроксимації Тема 5.3. Нелінійні методи апроксимації. Поліноміальні, сплайн- та радіально-базисні функції. Порівняння з лінійними підходами. Лекція 16. Нейронні мережі для апроксимації функцій	8

	Тема 5.4. Використання нейронних мереж. Архітектури нейронних мереж для навчання з підкріпленням. Проблеми стабільності та збіжності. Основи глибокого навчання з підкріпленням.	
6	Розділ 6. Градієнтні методи стратегій (Policy Gradient Methods) Лекція 17. Градієнтні методи стратегій Тема 6.1. Пряме оптимізаційне навчання стратегій. Тема 6.2. Стохастичні градієнтні методи. Лекція 18. Переваги та обмеження Тема 6.3. Переваги та обмеження підходів на основі стратегії. Проблеми стабільності та масштабованості. Сучасні варіанти актор-критик, A2C, PPO. Рекомендації щодо застосування у практичних задачах. МКР.	4
	Разом:	36

5.2. Лабораторні заняття

№ з/п	Назва лабораторної роботи	Кількість ауд. годин
1	Навчання агента у середовищах різної складності	6
2	Компроміс між отриманням нових знань та використанням раніше набутого досвіду	6
3	Передбачення найкращих станів та дій агента	6
	Разом:	18

5.3 Модульна контрольна робота

Модульна контрольна робота виконується за усіма темами, що розглядаються в дисципліні.

6. Самостійна робота здобувача вищої освіти денної форми навчання

До СРС відносяться такі види робіт:

- вивчення теоретичного матеріалу лекційних занять;
- виконання завдань лабораторних робіт та оформлення протоколів;
- підготовка до модульної контрольної роботи та семестрового контролю.

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Академічна доброчесність

Під час проходження цього курсу студенти зобов'язані дотримуватись політики та принципів

академічної доброчесності визначених Кодексом честі КПІ ім. Ігоря Сікорського (Детальніше Розділ 3: <https://kpi.ua/code>) та усі наступні правила:

1. Кожен із здобувачів повинен відправляти на перевірку власно виконану роботу. Використання чужих розв'язків або програмного коду і представлення їх за свої напрацювання є плагіатом та серйозним порушенням основних академічних стандартів.
2. Здобувачі не повинні ділитися своїми розв'язками з іншими студентами, а також просити інших ділитися своїми розв'язками з Вами.
3. Якщо студенти отримували допомогу у вирішенні певного завдання — потрібно зазначити це у звіті, а саме: від кого та яку допомогу отримували.

Норми етичної поведінки

Норми етичної поведінки студентів і працівників університету визначені у розділі 2 Кодексу честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського». Детальніше: <https://kpi.ua/code>.

Виконання лабораторних робіт

Викладачем встановлюється граничний термін для виконання кожного практичного завдання з метою послідовного та збалансованого засвоєння студентами навчального матеріалу, визначеного даним силабусом. Здача на перевірку виконаних завдань після встановлених термінів призводить до зниження балів. Кінцевим терміном для здачі на перевірку усіх завдань лабораторних робіт є останнє лабораторне заняття. Студент(-ка) має можливість переробити або доопрацювати завдання лабораторних робіт на перездачі з метою підвищення свого поточного рейтингу.

Оцінювання контрольних заходів у разі пропуску

У разі пропуску студентом(-кою) модульної контрольної роботи, студент(-ка) отримує нуль балів за цей контрольний захід. Повторне проходження модульної контрольної роботи для покращення свого результату непередбачено.

Процедура оскарження оцінок

Якщо студент(-ка) вважає, що його(її) роботу недооцінено або переоцінено — потрібно звернутися до викладача, який здійснював оцінювання та повідомити про це із зазначенням короткого обґрунтування (оцінка буде переглянута).

8. Види контролю та рейтингова система оцінювання результатів навчання (PCO)

8.1 Поточний контроль

Поточний контроль успішності засвоєння знань виконується шляхом перевірки завдань лабораторних робіт та МКР. Таким чином, семестровий рейтинг студента з дисципліни складається з балів, що він отримує за:

- виконання та захист завдань 3-х завдань лабораторних робіт;
- модульну контрольну роботу.

8.2 Система рейтингових (вагових) балів та критерії оцінювання

Завдання лабораторних робіт

Ваговий бал кожного завдання – 20 балів. Максимальна кількість балів за всі завдання лабораторних робіт дорівнює $3 \times 20 = 60$ балів.

Складові оцінки та відповідний максимальний бал:

- програмна реалізація алгоритмів – 5 балів;
- експериментальне дослідження/порівняння алгоритмів – 5 балів;
- звіт (відповідність структури, презентація результатів роботи, якість та обґрунтування рішень, висновки) – 5 балів;
- захист результатів комп'ютерного практикуму (теоретичні знання за темою лабораторної роботи; обґрунтування прийнятих рішень; якість, точність та відповіді) – 5 балів.

Бал може бути знижений у разі неповного виконання завдань, недостатньої якості звіту, непереконливого захисту результатів та порушення строку здачі. Отриманий бал знижується за задачу після встановленого строку: -1 бал за кожен наступний тиждень після дедлайну, але в сумі не більше 10 балів.

Модульна контрольна робота (МКР)

МКР містить 4 питання, ваговий бал кожного становить 10 балів.

Критерії оцінювання кожного завдання модульної контрольної роботи:

- повна відповідь (не менше 95% потрібної інформації) – 10 балів;
- достатньо повна відповідь (не менше 75% потрібної інформації), або повна відповідь з незначними помилками – 7-9 балів;
- неповна відповідь (не менше 60% потрібної інформації) та незначні помилки – 6 балів;
- незадовільна відповідь – 0-5 балів;

8.3 Календарний контроль

Календарний контроль проводиться двічі на семестр з метою моніторингу поточного стану виконання студентами вимог силабусу.

Умови позитивного календарного контролю:

- за результатами навчальної роботи на першому календарному контролі (7-й тиждень семестру) студент отримує статус «атестований», якщо його поточний рейтинг становить не менше ніж 50% від максимально можливої кількості балів, які можна було набрати протягом перших 6 тижнів;
- за результатами навчальної роботи на другому календарному контролі (13-й тиждень семестру) студент отримує статус «атестований», якщо його поточний рейтинг становить не менше ніж 50% від максимально можливої кількості балів, які можна було набрати протягом перших 12 тижнів.

Максимальна сума вагових балів контрольних заходів протягом семестру складає:

$$3 \cdot 20 + 40 = 100.$$

Сума балів, які отримує студент за роботу протягом семестру, обчислюється за формулою:

$$R_D = \sum_{i=1}^3 r_{ЛРi} + r_{МКР},$$

де $r_{ЛРi}$ – бали, отримані за i -ту лабораторну роботу (0...20);

$r_{МКР}$ – бали за МКР (0...40);

Залік

Умовою допуску до заліку є зарахування всіх лабораторних робіт та модульної контрольної роботи.

Студенти, які виконали умову допуску до заліку та набрали $R_D \geq 60$, отримують підсумкову рейтингову оцінку без потреби проходження заходу семестрового контролю.

Студенти, які виконали умови допуску до заліку, але набрали $R_D \in 60$, а також ті, хто бажають підвищити свою рейтингову оцінку, виконують залікову контрольну роботу.

На заліку студенти виконують письмову контрольну роботу. Кожен білет містить два завдання. Кожне завдання оцінюється у 20 балів.

Критерії оцінювання кожного завдання залікової контрольної роботи:

- повна відповідь (не менше 95% потрібної інформації) – 20 балів;
- достатньо повна відповідь (не менше 75% потрібної інформації, або незначні неточності) – 15-19 балів;
- неповна відповідь (не менше 60% потрібної інформації та деякі помилки) – 12-15 балів;
- незадовільна відповідь – 0-11 балів.

У випадку виконання залікової контрольної роботи бали за неї додаються до R_D , а рейтингова оцінка трансформується до університетської системи оцінювання згідно з таблицею 1.

Таблиця 1. Переведення рейтингових балів до оцінок за університетською шкалою

Кількість балів	Оцінка
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску/ Порушення Кодексу честі	Не допущено

9. Нормативні положення політики навчальної дисципліни (освітнього компонента)

Правила відвідування занять. Відвідування занять є вільним, бали за присутність на лекціях та лабораторних заняттях не нараховуються, відповідно штрафні бали у разі відсутності на занятті не передбачаються.

Дистанційний режим навчання. У разі запровадження обмежень на відвідування університету, пов'язаних з введенням карантину або режиму воєнного стану в державі, освітній процес здійснюється у дистанційному режимі відповідно до Положення про дистанційне навчання в КПІ ім. Ігоря Сікорського (<https://osvita.kpi.ua/index.php/node/188>), Регламенту організації освітнього процесу в дистанційному режимі (<https://profkom.kpi.ua/reglament-organizatsiyi-osvitnogo-protsesu-v-distantsiynomu-rezhimi>) та Регламенту проведення семестрового контролю в дистанційному режимі (<https://osvita.kpi.ua/node/148>). У режимі дистанційного навчання заняття відбуваються у вигляді онлайн-конференції на платформах BigBlueButton, Google Meet, Zoom. Посилання на конференцію видається на початку семестру і розміщується в АС «Електронний кампус». Результати оцінювання висвітлюють у АС «Електронний кампус» на особистій сторінці здобувача (<https://ecampus.kpi.ua>).

Правила поведінки на заняттях. На заняттях слід дотримуватись норм етичної поведінки визначених у Кодексі честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» (<https://kpi.ua/code>). На території університету здобувачі мають поводити себе відповідно до Правил внутрішнього розпорядку (<https://kpi.ua/admin-rule>).

Визнання результатів навчання, набутих у неформальній/інформальній освіті. Положення про визнання результатів навчання, набутих у неформальній / інформальній освіті (<https://osvita.kpi.ua/index.php/node/179>) регламентує визнання результатів навчання, набутих у неформальній/інформальній освіті:

- у разі навчання на неформальній освіті, яку студент обрав самостійно і хоче зарахувати результати навчання в рамках дисципліни, проходиться процедура валідації, що передбачає подання здобувачем заяви на ім'я декана, декларації підтверджувальних документів. Рішення про визнання чи не визнання приймається комісією у складі завідувача кафедри, викладача, гаранта освітньо-професійної програми.

Політика використання штучного інтелекту. Використання штучного інтелекту регламентується «Політикою використання штучного інтелекту для академічної діяльності в КПІ ім. Ігоря Сікорського» (<https://osvita.kpi.ua/node/1225>). Усі завдання, як під час виконання навчальних завдань з дисципліни, так і індивідуальні завдання, мають бути результатом власної оригінальної роботи здобувача. Використання ШІ для автоматичної генерації відповідей без подальшого їх аналізу та доопрацювання заборонено.

Робочу програму навчальної дисципліни (силабус)

Складено: ст. викладач кафедри обчислювальної техніки Кочура Ю. П.

Ухвалено кафедрою обчислювальної техніки (протокол № 12 від 23.06.2025р.).

Погоджено методичною комісією ФІОТ (протокол № 11 від 27.06.2025 р.).