Сколковский институт науки и технологий (Сколтех)

Рабочая программа дисциплины	**Обучение с подкреплением**

Преподаватель	Осиненко Павел Валерьевич, старший преподаватель, PhD

Аннотация

Описание курса

В данном курсе изучается один из авангардных методов машинного обучения - обучение с подкреплением.
Этот метод основан на концепции агента, взаимодействующего со средой и оптимизирующего свои действия с целью достижения максимальной "выгоды" (в абстрактном смысле, который конкретизируется в приложениях).
Следующие темы адресованы в курсе:
- динамические среды;
- динамическое программирование;
- градиентные методы оптимизации политик;
- актор-критик;
- сходимость, безопасность и устойчивость обучения с подкреплением;
- глубокое и предиктивное обучение с подкреплением.

1. Основная информация

Академический уровень курса	Магистратура Аспирантура
Количество кредитов	6

Предварительные требования к курсу / рекомендации

Перед этим рекомендуется пройти курс численной линейной алгебры и методов оптимизации, а также курс по машинному обучению и приложениям. Кроме того, мы предполагаем, что слушатель свободно владеет линейной алгеброй, вероятностным и реальным анализом.

Тип оценки - дифференцированная

Отображение оценок в процентах

A:	86
B:	70
C:	56

D:	46
E:	36
F:	0

2. Содержание курса

Тема	Краткое содержание	Лекции (час)	Семинары (час)	Лабораторные занятия (час)	Самостоятельная работа (час)
1. Марковские процессы принятия решений	Основы динамических систем. Цепи Маркова. Управляемые цепи Маркова. Вознаграждение, ценность, дисконтирование. Примеры MDP	3	6		9
2. Динамическое Программирование и табличные методы	Принцип динамического программирования. Уравнения Гамильтона-Якоби-Беллмана. Итерация значений и политики. Q-обучение. Проклятие размерности	3	6		9
3. Базовый Политический уклон	Стохастическая оптимизация основана на обучении с подкреплением. Вывод понятия "ПОДКРЕПЛЕНИЕ".	3	3		12
4. Методы определения временных разниц	Буфер воспроизведения. Оценка значения. Глубокий Q-networks. Актер и критик. Преимущество. Конвергенция актора и критика. Асимптотическая конвергенция против единой конечной ограниченности. Абстрактная энергия, элементы теории устойчивости. Проблемы конвергенции. Постоянное возбуждение. Разведка против эксплуатации	6			12
5. Расширенный градиент политики	Естественный градиент политики. Базовый уровень градиента политики. Градиент политики "Субъект-критик". Оптимизация политики в регионе доверия (TRPO). Ближайшая оптимизация политики (PPO)	6	9		15
6. Продвинутые актеры-критики	Мягкий актер-критик. Исследование, основанное на энтропии. Анализ	3	6		18
7. Дополнительные темы	Конвергенция глубоких акторов и критиков. Обучение с подкреплением на основе моделей. Управление с предсказанием модели. Линейно -квадратичный регулятор.	6	9		18

3. Результаты обучения

Результаты обучения в Сколтехе указаны в соответствии со структурой результатов обучения в Сколтехе

1. ФУНДАМЕНТАЛЬНЫЕ ЗНАНИЯ
1.2. Знание прикладной науки и техники, науки, в том числе современные методы и инструменты
1.4. Междисциплинарное мышление, структура знаний и интеграция
2.1. ПОЗНАНИЕ И СПОСОБЫ РАССУЖДЕНИЯ
2.1.1. Аналитическое мышление и решение проблем
2.1.2. Системное мышление
2.1.3. Творческое мышление
2.1.4. Принятие решений (неоднозначных, срочных и т.д.)
2.1.5. Критическое мышление и метапознание
2.2. ОТНОШЕНИЕ И ПРОЦЕСС ОБУЧЕНИЯ
2.2.1. Инициатива и готовность идти на соответствующий риск
2.2.2. Готовность принимать решения в условиях неопределенности
2.2.3. Ответственность, интенсивность, настойчивость, безотлагательность и воля к достижению поставленных целей
2.2.4. Находчивость, гибкость и способность адаптироваться
2.2.5. Самосознание и стремление к самосовершенствованию, обучению на протяжении всей жизни и воспитанию
2.2.6. Развитие и поддержка преподавательского состава и обучающегося сообщества
2.3. ЭТИКА, СПРАВЕДЛИВОСТЬ И ДРУГИЕ ОБЯЗАННОСТИ
2.3.5. Проактивное видение и намерение в жизни
3.1. КОММУНИКАЦИЯ В МЕЖДУНАРОДНОЙ СРЕДЕ
3.1.1. Коммуникационная стратегия и структура
3.1.2. Письменная, электронная и графическая коммуникация
3.1.3. Устная презентация и обсуждение
3.1.4. Вопросы, слушание и диалог
3.1.5. Общение на английском языке в научной, деловой и общественной среде
3.1.6. Эффективное взаимодействие в различных культурных и международных условиях
3.2. КОМАНДНАЯ РАБОТА И ЛИДЕРСТВО
3.2.1. Формирование эффективных команд
3.2.2. Командная работа и управление проектами
3.2.3. Командная координация, принятие решений и лидерство
3.2.4. Рост и эволюция команды
3.2.5. Техническое и междисциплинарное сотрудничество
3.3. СОТРУДНИЧЕСТВО И ИЗМЕНЕНИЯ
3.3.1. Установление разнообразных связей и сетевого взаимодействия
3.3.2. Понимание различных ролей, перспектив и интересов
3.3.3. Переговоры и разрешение конфликтов
3.3.4. Вовлечение заинтересованных сторон
3.3.5. Осуществление намеренных изменений
4.2. ДАЛЬНОВИДНОСТЬ — ИЗОБРЕТЕНИЕ НОВЫХ ТЕХНОЛОГИЙ ПОСРЕДСТВОМ ИССЛЕДОВАНИЙ

4.2.1. Процесс исследования — гипотеза, доказательства и защита
4.2.2. Фундаментальные исследования, ведущие к новым научным открытиям
4.2.3. Исследования, направленные на разработку новых технологий
4.2.4. Представление о пользе новой науки и техники
4.2.5. Разработка концепций и внедрение их в практику

4. Задания и выставление оценок

Требование к физической посещаемости (% от числа занятий)	100

К требованиям о личном посещении:

студентам разрешается пропускать занятия только по уважительной причине, имеющей обоснование, например, медицинское заключение

Тип назначения	Краткое содержание задания	% от итоговой оценки за курс
Командный проект	Заданиями курса являются 5 мини-проектов, которые будут реализованы в группах. Темы мини -проектов: 1. Марковские процессы принятия решений. 2. УСИЛЕНИЕ. 3. Продвинутый градиент политики. 4. Мягкий актер-критик. 5. Обучение с подкреплением, основанное на модели. Конкретные задачи в рамках мини -проекта выбираются командами и утверждаются преподавателями курса. Например, команда может выбрать для решения марковский процесс принятия решений. Команды могут обратиться к преподавателям курса за помощью в выборе конкретной задачи.	70
Участие в занятиях	Посещаемость	30

5. Критерии оценки

Задание 1 Типа	Командный проект

Пример задания 1

Тема: расширенный градиент политики. Вы выбираете марковский процесс принятия решений и утверждаете его у преподавателей курса. В качестве альтернативы, попросите преподавателей курса помочь с выбором MDP. Выберите метод расширенного градиента политики, например, TRPO или PPO. Реализуйте агент MDP и policy gradient agent на Python. Предпочтительнее использовать PyTorch. Допускаются готовые движки для сред, таких как Gymnasium, однако агент должен быть реализован на основе

первых принципов с использованием PyTorch. То есть, никакие готовые настройки из фреймворков, таких как stablebaselines3, не допускаются. Кроме того, для первого командного проекта не допускается использование тренажерного зала, т.е. MDP должен быть реализован на основе первых принципов, например, с использованием только numpy в качестве ядра.
Предпочтительные формы реализации: хранилище кода или Google Colab. В обоих случаях конкретная проблема должна быть хорошо описана. В случае хранилища кода должен присутствовать понятный readme. Необходимо предоставить инструкции о том, как воспроизвести результаты.

Рекомендуемая структура хранилища:

1. src/ - основные компоненты программы (среды, агенты и т.д.)
2. run/ - запуск файлов и скриптов bash
3. artifacts/ - хранение статистики, например, csv-файлов, контрольных точек модели и т.д.
4. анализ/ - сценарии для оценки, например, интерактивные блокноты на Python для визуализации статистики

Команды должны тщательно и надлежащим образом соблюдать стандарты и соглашения по именованию на Python, правила чистого кода и документирования кода.
Письменные отчеты не требуются.

Критерии оценки для задания 1

Каждый мини-проект оценивается с помощью защиты, которая проводится в специально отведенные для этого часы на семинарах. Каждая защита должна включать описание конкретной проблемы, применяемых методов, основы реализации, презентацию результатов и краткое описание соответствующего хранилища кода или листа Colab. Преподаватели курса и другие команды могут задавать вопросы.

Задание 2 Типа	Участие в занятиях

Пример задания 2

Простая проверка посещаемости

Основные критерии оценки:
Посещаемость оценивается на основе принципа взвешивания, указанного в итоговом примечании. В случае явки менее В 50% случаев курс считается проваленным. 100%-ная посещаемость составляет 30% от итоговой оценки

6. Учебники и интернет-ресурсы

Необходимые учебники	ISBN-13 (or ISBN-10)

Sutton, R. and Barto, A. Reinforcement Learning: an Introduction, 2nd Edition. MIT Press, 2018	9780262039246
Рекомендуемые учебники	
Bertsekas, D. Reinforcement learning and optimal control, 2nd Edition. Link: https://www.mit.edu/~dimitrib/RLbook.html	опубликовано в Интернете
Puterman, M.. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2014	978-0-471-72782-8
Документы	DOI or URL
Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018, July). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning (pp. 1861-1870). PMLR.	http://proceedings.mlr.press/v80/haarnoja18b/haarnoja18b.pdf
Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv Preprint arXiv:1707.06347.	https://arxiv.org/pdf/1707.06347
Schulman, J. (2015). Trust Region Policy Optimization. arXiv Preprint arXiv:1502.05477.	https://people.engr.tamu.edu/guni/csce642/files/trpo.pdf
Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. Advances in neural information processing systems, 12.	https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf
Roderick, M., MacGlashan, J., & Tellex, S. (2017). Implementing the deep q-network. arXiv preprint arXiv:1711.07478.	https://arxiv.org/pdf/1711.07478

Веб-ресурсы (ссылки)	Описание
https://github.com/vwxyzjn/cleanrl	Высокое качество один файл реализации глубоких Алгоритмы обучения с подкреплением с научно-дружественных особенности (PPO, DQN, C51, DDPG, TD3, SAC, PPG)
https://stablebaselines3.readthedocs.io/en/master/	Stable-Baselines3 Docs - надежное обучение с подкреплением Реализации

7. Оборудование

Программное обеспечение
Python

Оборудование
Никакого дополнительного оборудования не требуется

8. Дополнительные примечания

Комментарии

Курс ориентирован на активное вовлечение студентов. Он способствует созданию качественной исследовательской и опытно-конструкторской работы в будущей карьере.