Федеральное государственное автономное образовательное учреждение высшего образования "Национальный исследовательский университет "Высшая школа экономики"

Факультет Компьютерных Наук

LAMBDA lab

Рабочая программа дисциплины/общеуниверситетского факультатива Обучение с подкреплением

для уровня подготовки – бакалавриат/специалитет/магистратура/аспирантура

Разработчик(и) программы
Устюжанин Андрей Евгениевич — зав. лаборатории
Ратников Фёдор Дмитриевич — старший научный сотрудник
Васильев Олег Юрьевич — исследователь
Фрицлер Александр Александрович — исследователь
Швечиков Павел Дмитриевич — исследователь

Одобрена к реализации на заседании комиссии
«___»_____ 201_ г.

Утверждена «___»_____ 201_ г.

Руководитель Методического центра ДООП
«___»_____ 201_ г.
[Введите И.О. Фамилия] _____ [подпись]

Москва, 201_

Настоящая программа может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы.

1. Область применения

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину [введите название общеуниверситетского факультатива], студентов и слушателей, желающих принять участие в работе факультатива.

2. Цели освоения дисциплины

В результате освоения учебной дисциплины студенты и слушатели должны овладеть следующими образовательными результатами:

- Уметь строить алгоритмы нахождения оптимальной стратегии в процессах принятия решений.
- Уметь выявлять и решать прикладные задачи обучения с подкреплением в прикладных областях (робототехика, машинный перевод, диалоговые системы, информационный поиск, баннерная реклама и т. п.)
- Понимать теоретическую базу обучения с подкреплением.

3. Тематический план учебной дисциплины

№	Название раздела	Подразделение НИУ ВШЭ, за которой закреплен раздел[1]	Всего часов	Аудиторные часы			Самостоя-тельная работа
				Лекции	Семинары	Практические занятия	
1	Процессы принятия решений, black-box методы оптимизации в обучении с подкреплением.	LAMBDA	16	4		4	8
2	Value-based подход в обучении с подкреплением.	LAMBDA	16	4		4	8
3	Аппроксиммационный подход к обучению с подкреплением. Deep reinforcement learning.	LAMBDA	16	4		4	8
4	Policy-based методы в обучении с подкреплением	LAMBDA	16	4		4	8
5	Прикладные задачи обучения с подкреплением	LAMBDA	16	4		4	8

[1] Если ОУФ реализуется одним подразделением, столбец можно убрать.

6	Вспомогательные занятия (CNN-ликбез, RNN-ликбез)	LAMBDA	8	4		4	0

4. Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год				Кафедра/подразделение [2]	Параметры **
		1	2	3	4		
Вялотекущий	Практическое задание	6	6			LAMBDA	12 заданий «на закодить», из которых на отлично нужно выполнить примерно 8 (в зависимости от качества выполнения). Jupyter;numpy/scipy;gym; tf/theano. Заданий может быть больше, но требования «на отлично» более строгими не станут.

5. Содержание дисциплины

The syllabus is approximate: the lectures may occur in a slightly different order and some topics may end up taking two weeks. However, everything announced here is guaranteed to occur at least once.

Section I: Decision processes

- Welcome to Reinforcement Learning
    - Lecture: RL problems around us. Decision processes. Basic genetic algorithms
    - Seminar: Welcome into openai gym, basic genetic algorithms
    - Homework description - see week0/README.md
        - HSE Homework deadline: *23.59 15.09.17*
- RL as blackbox optimization
    - Lecture: Recap on genetic algorithms; Evolutionary strategies. Stochastic optimization, Crossentropy method. Parameter space search vs action space search.
    - Seminar: Tabular CEM for Taxi-v0, deep CEM for box2d environments.
    - Homework description - see week1/README.md
        - HSE Homework deadline: *23.59 22.09.17*

Section II: Value-based methods

[2] Если ОУФ реализуется одним подразделением, столбец можно убрать

- Value-based methods
    - Lecture: Discounted reward MDP. Value-based approach. Value iteration. Policy iteration. Discounted reward fails.
    - Seminar: Value iteration.
        - HSE Homework deadline: *23.59 29.09.17*
- Model-free reinforcement learning
    - Lecture: Q-learning. SARSA. Off-policy Vs on-policy algorithms. N-step algorithms. TD(Lambda).
    - Seminar: Qlearning Vs SARSA Vs Expected Value SARSA
    - HSE Homework deadline: *23.59 6.10.17*

Section III: Approximate reinforcement learning

- deep learning recap
    - Lecture: Deep learning 101
    - Seminar: Simple image classification with convnets
    - HSE Homework deadline: _23.59 6.10.17
- Approximate reinforcement learning
    - Lecture: Infinite/continuous state space. Value function approximation. Convergence conditions. Multiple agents trick; experience replay, target networks, double/dueling/bootstrap DQN, etc.
    - Seminar: Approximate Q-learning with experience replay. (CartPole, Atari)

Section IV: Policy-based methods

- Policy gradient methods I
    - Lecture: Motivation for policy-based, policy gradient, logderivative trick, REINFORCE/crossentropy method, variance reduction(baseline), advantage actor-critic (incl. GAE)
    - Seminar: REINFORCE, advantage actor-critic
- Policy gradient methods II
    - Lecture: Trust region policy optimization. NPO/PPO. Deterministic policy gradient. DDPG. Bonus: DPG for discrete action spaces.
    - Seminar: Approximate TRPO for simple robotic tasks.

[Maybe a bonus lecture here]

Section V: Practical applications

- Exploration in reinforcement learning
    - Lecture: Contextual bandits. Thompson Sampling, UCB, bayesian UCB. Exploration in model-based RL, MCTS. "Deep" heuristics for exploration.
      Seminar: bayesian exploration for contextual bandits. UCB for MCTS.
      [Recurrent neural nets recap goes here]

- Partially observable MDPs
  - Lecture: POMDP intro. POMDP learning (agents with memory). POMDP planning (POMCP, etc)
  - Seminar: Deep kung-fu & doom with recurrent A3C and DRQN
- Applications II
  - Lecture: Reinforcement Learning as a general way to optimize non-differentiable loss. G2P, machine translation, conversation models, image captioning, discrete GANs. Self-critical sequence training.
  - Seminar: Simple neural machine translation with self-critical sequence training

6. Критерии оценки знаний, навыков

Оценка за курс выставляется по сумме баллов, баллы набираются за практические задания. Для уже знакомых с дисциплиной слушателей есть возможность получить дополнительные баллы за проектную (исследовательскую или прикладную) работу прямо использующию методы обучения с подкреплением.

7. Образовательные технологии

Все задания сдаются в opentask - http://anytask.org/course/228

7.1 Методические указания студентам

В осеннем семестре 2017 курс живёт в репозитории https://github.com/yandexdataschool/practical_rl/tree/fall17 (вся информация в README.md). Если вы слушаете курс в другом семестре, вам по тому же адресу в /tree/master, а ещё лучше — с вопросом к преподавателям и коллегам-слушателям.

8. Оценочные средства для текущего контроля и аттестации студента

Все задания размещены в репозитории (см. пункт 7.1).

Оценка курса строится на основании практических заданий (пример - http://bit.ly/2fFFZgq). Каждое задание состоит из основной части, за которую можно получить 10 баллов, и нескольких дополнительных задач опционального характера за дополнительные баллы.

Слушатель может самостоятельно выбрать какие задания он хочет выполнить, оценка выставляется по сумме баллов.

9. Порядок формирования оценок по дисциплине

Критерии могли устареть. Актуальная информация на текущий семестр - http://bit.ly/2khP3NF

Сумма баллов	Оценка

0-9	0
10-19	1
20-29	2
30-39	3
40-49	4
50-59	5
60-69	6
70-79	7
80-89	8
90-99	9
100+	10

10. Учебно-методическое и информационное обеспечение дисциплины

10.1 Основная литература

Обязательной литературы нет.

10.2 Дополнительная литература

Дополнительная литература по каждой неделе есть в README.md каждой неделе (под заголовком «Materials»). Пример такого списка - http://bit.ly/2hIkoIA . Рекомендуемые материалы помечены как [main] или [recommended].

10.3 Справочники, словари, энциклопедии

http://yandex.ru/
http://google.com/

10.4 Программные средства

Курс проводится на стандартном data science стэке: python + jupyter + numpy + scipy + matplotlib + sklearn. Слушателю не знакомому хотя бы с 3 из них настоятельно рекомендуется пройти ликбез по ним (в составе курса по Машинному Обучению).

Задания совместимы как с python 2.7.*, так и python 3.5+. В меру собственной извращённости и стремления к приключениям слушатель может выбрать другую среду выполнения заданий с согласия семинариста.

Около половины заданий курса используют модели глубинного обучения(deep learning). Строить таковые модели слушателю предлагается в theano или tensorflow (поддерживаются на осенний семестр 2017, если вы живёте в другом году — уточняйте у преподавателей).

10.5 Дистанционная поддержка дисциплины

Курс можно сдавать дистанционно, для чего слушателю нужно осваивать материалы и сдавать задания в том же порядке, что и очному слушателю. Видео и текстовые материалы по каждой неделе выглядят примерно так - http://bit.ly/2hIkoIA .

11. Материально-техническое обеспечение дисциплины

Минимальные системные требования — 2+ Gb ram, x86-совместимый процессор, 64-битная ОС.

Комфортный уровень — 4gb ram, видеокарта с CUDA или побольше ядер CPU.

Если очень хочется, курс можно пройти на чём угодно, если на нём работает браузер с javascript (и даже если не работает).