**Обучение с подкреплением**

**О чём это:**

*Обучение с подкреплением - наука об управлении злобным агентом в выдуманном мире (с)*

Если приглядеться к нашей с вами жизни, можно заметить, что мы (люди) обычно занимаемся отнюдь не разметкой примеров и отображением X->y, минимизирующим эмпирический риск: мы существуем в мире, на который мы можем влиять, и который в свою очередь влияет на нас. А хотим мы в этом мире добиваться каких-то результатов: дойти до “из точки A точки B”, заработать побольше денег, привлечь и удержать пользователя - кому что ближе.

Задачи эти объединяет то, что в них вам придётся двигаться методом проб, ошибок, шишек и новых проб - у вас просто нет всеобъемлющей выборки, в которой есть правильная стратегия поведения во всех ситуациях. А ещё эти задачи объединяет то, что их можно решать автоматически. Да-да, творческий поиск решения задачи, выбор дизайна страницы или повышение “скилла” в любимой игрушке - машины это уже умеют.

Курс очень хочет дать своим слушателям понимание и практические навыки использования таких “машин”, именуемых “Алгоритмами обучения с подкреплением” или “Reinforcement learning”. В меню - теоретическая база, практические задания, инженерные “хаки” и неординарные предметные области. Даже биржевую торговлю завезли.

А ещё этот курс немного связан с методами глубокого обучения - их тут около трети. Так уж вышло, что многие задачи Reinforcement Learning сейчас лучше всего решаются с применением пресловутых “нейронок”. Если Вы их не знаете - не бойтесь, научим.

**Пререквизиты:**

* От студента:
  + Машинное обучение 1, Теория вероятностей
  + Полезно, но не обязательно — машинное обучение 2 и байесовские методы

**Отчётность:**

Модель отчётности: оценки ставятся по баллам, которые выдаются за выполнение форм контроля.

Основной источник баллов – домашние задания. Они выдаются после семинаров, делаются, отправляются, проверяются и зарабатывают баллы.

* Всего домашних заданий будет 10~12.
* Верно выполненная базовая часть ДЗ - 10 баллов.
* Крутое и изобретательное решение или любая значительная инициатива сверх минимума - бонусные баллы

Второй источник баллов — проекты. Их можно делать, а можно не делать, зато делать больше домашек. Проекты — решение любой нетривиальной задачи RL, воспроизведение статьи или что-то близкое по духу. Правила для проектов аналогичны нашему курсу deep learning - <https://github.com/yandexdataschool/HSE_deeplearning/wiki/Course-projects>

Разбалловка:

* “5” - 90+ баллов
* “4” - 70+ баллов
* “3” - 50+ баллов
* “2” - 0+ баллов

**Syllabus:**

1. Intro
   1. **Lecture:** Decision process. MDP. CDP. Full VS Partial observability; Deterministic Vs stochastic MDP. Game theory approach.
   2. **Seminar:** pacman via planning.
2. Classic reinforcement learning: Iterative solvers
   1. **Lecture:** Iterative MDP solving. Policy iteration, Value iteration, TD-algorithm.
   2. **Seminar:** Frozenlake (grid world) with policy and value iteration.
3. Classic reinforcement learning: Value-based
   1. **Lecture: “**online” reinforcement learning. Off-policy and on-policy algorithms. Q-learning, SARSA, K-step Q-learning. Exploration Vs Exploitation. Basic exploration strategies.
   2. **Seminar:** Pacman with Q-learning. E-greedy Vs softmax exploration.
4. Classic reinforcement learning: Alternative
   1. **Lecture:** Policy-based Vs Value-based solvers. Actor-critic.
   2. **Seminar:** Advantage Actor-critic Vs Q-learning on Atari given full RAM.
5. Approximate reinforcement learning: linear
   1. **Lecture:** Infinite/continuous state space. Q/Value function approximation. Convergence condition. Proof for linear models.
   2. **Seminar:** Approximate Q-learning with linear model for pacman.
6. Approximate Reinforcement Learning: deep learning
   1. **Lecture:** Nonlinear approximators. Deep Q-learning. Divergence problem. Stability tricks: experience replay, target network.
   2. **Seminar:** Playing atari with deep reinforcement learning (OpenAI gym, Agentnet)
7. Partially observable MDPs
   1. **Lecture:** POMDP. Model-based solvers Vs deep reinforcement learning with memory. DRQN. On-policy deep RL. Multiple agents trick. Deep A2c.
   2. **Seminar:** Deep kung-fu with recurrent DQN.
8. Hierarchical MDP
   1. **Lecture:** MDP Vs real world. Sparse and delayed rewards. Why Q-learning sucks. Hierarchical MDP. Hierarchy as temporal abstraction. MDP with symbolic reasoning.
   2. **Seminar:** Hierarchical RL for “Montezuma revenge” (with pre-trained DQN)
9. Continuous action space. Case study: robotics.
   1. **Lecture:** Continuous action space MDPs. Model-based approach (NAF). Actor-critic approach (dpg, svg). Trust Region Policy Optimization.
   2. **Seminar:** BipedalWalker with ddpg Vs qNAF. TRPO https://gym.openai.com/envs/BipedalWalker-v2

openAI RLLab for TRPO

1. Advanced exploration methods I: linear bandits. Case study: online advertising
   1. **Lecture:** Case study: Contextual bandits for RTB. Improved exploration methods. Bayesian approach.
   2. **Seminar:** RTB linear contextual MAB. Comparing exploration methods: greedy, e-greedy, quantile-based.

1. Advanced exploration methods II: deep RL
   1. **Bonus Lecture:** Case study: bayesian exploration methods for policy optimization. VIME and similar.
   2. **Seminar:** Vime Vs Bootstrap DQN
2. **Bonus Lecture:** Neural conversation models with deep reinforcement learning.
3. **Bonus Lecture:** Inverse Reinforcement Learning. Case study: Personalized medical treatment.
4. **Bonus Lecture:** Policy optimization without reinforcement learning. Case study: Portfolio management.