Studend science session 20.05.2021

Topic: Analysis of the change in the policy of action of an agent in case of a change in the values of a reward for an action performed in the field of machine learning without supervision

Тема: Анализ промяната на политиката на действие на агент при промяна стойностите на награда при извършено действие в областа на машино обучение без надзор

Изготвил: Маг.инж. Мартин Калоев

Abstract

- One of the great advances in the field of artificial intelligence is the creation of general artificial intelligence (AGI). Or artificial intelligence allowing an agent to perform tasks of human complexity. This is due to the development of neural networks that implement in practice the concepts described by adaptive programming.
- Adaptive programming is a set of algorithms and methods describing the relationships between actions and states.
- Artificial intelligence of this type plays an important role in the modern world, in the areas like targeted advertising; digital assisted trade; robotics; transport and others.
- What all these agents have in common is the ability to learn from observed processes and interact with users.
- This article discusses the work of such an agent.The focus of the studie is the relationship between the change in the policy used by the agent and the change in the value of the rewards received from an action.
- The agent's task is to solve a two-dimensional spatial problem using a greedy epsilon policy. The algorithms used by the agent are an algorithm for determine the quality of actions, a sarsa algorithm and algorithms for evaluation quality of the policy.

- Един от големите напредаци в областа на изкуствения интелект е създаването на общ изкуствен интелект(AGI). Или изкуствен интелект позволяващ на агент да изпълява задачи с човешка сложност. Това се дължи на развитието на невроните мрежи реализиращи на практика концепциите описани от адаптивното програмиране.
- Адаптивното програмиране е сбор от алгоритми и методи описващи връзките между действия и състояния.
- Изкуствения интелект от този тип има изключително голямо роля в съвремения свят. Областите на "целева реклама", "дигитални подпомогната търговия", "роботика" и транспорт и други.
- Общото между всички тези агенти е способноста да се учат от наблюдавани процеси и взаимодейтвие с потребители.
- Изучава се връзката между промята в използвана политика от агента и промяната на стойноста на наградите получени при действие.
- Тази статия разглежда работата на такъв агент. Изследването се средоточава върху промяната на поведението и използваната политика при промяна стойностите на наградата получена при направено действие.
- Задачата на агента е да реши двуизмерен пространствен проблем като използва алчна епсилон политика. Алгоритмите използвани от агента са алгоритм за да определяне качеството на действия, sarsa алгоритъм и алгоритми за оценка качествата на политиката.

Tea cups

Related work

- Articles on a similar topic can be found with the following titles:
- An Analysis of Q-Learning Algorithms with Strategies of Reward Function
- https://www.researchgate.net/publication/50247491_An_Analysis_of_Q-Learning_Algorithms_with_Strategies_of_Reward_Function
- Q-Learning Algorithms: A Comprehensive Classification and Applications
- https://www.researchgate.net/publication/335805245_Q-Learning_Algorithms_A_Comprehensive_Classification_and_Applications
- Playing Atari with Deep Reinforcement Learning
- https://arxiv.org/pdf/1312.5602v1.pdf
- Low-rank State-action Value-function Approximation
- https://arxiv.org/abs/2104.08805v1

- Статии на подобна тема могат да бъдат намерени със следните заглавия:
- An Analysis of Q-Learning Algorithms with Strategies of Reward Function
- https://www.researchgate.net/publication/50247491_An_Analysis_of_Q-Learning_Algorithms_with_Strategies_of_Reward_Function
- Q-Learning Algorithms: A Comprehensive Classification and Applications
- https://www.researchgate.net/publication/335805245_Q-Learning_Algorithms_A_Comprehensive_Classification_and_Applications
- Playing Atari with Deep Reinforcement Learning
- https://arxiv.org/pdf/1312.5602v1.pdf
- Low-rank State-action Value-function Approximation
- https://arxiv.org/abs/2104.08805v1

Tea cups

Algorithms list

- The list of used algorithms:
- *-Sarsa:*
- *-Quality of action equation by R.E.Bellman*
- *-Value of action equation by R.E.Bellman*
- *-Policy based on action equation.*
- Those equations belongs to the family of discrete mathematics.
- Discrete mathematics is the study of mathematical structures that are countable or otherwise distinct and separable. Examples of structures that are discrete are combinations, graphs, and logical statements.
- Please note the slide named: **Model and equation**
- Manuals for application of algorithms:
- **Reinforcement Learning:**
- **An Introduction**
- by:
- *Richard S. Sutton and Andrew G. Barto*
- A Bradford Book
- The MIT Press
- Cambridge, Massachusetts
- London, England
- ***at: http://incompleteideas.net/book/first/ebook/node41.html***

Sarsa:

$$Q(s_t, a_t) = Q(s_t, a_t) + a^*(r_t + y^*Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

Q_l belman curent quality for action

$$Q(s_t, a_t) = Q(s_t, a_t) + a^*\left(r_t + y^*\max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right)$$

which can also be written as

$$Q(s_t, a_t) = (1 - \alpha)\,^*Q(s_t, a_t) + \alpha^*\left(r_t + y^*\max_a Q(s_{t+1}, a)\right)$$

Value:

$$v(s) = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s]$$

Quality

$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$$

Policy:

$$\text{TDerror}\ (s) = V^\pi(s) - \sum_{s'} T\left(s, \pi(s), s'\right)\left[r\left(s, \pi(s), s'\right) + \gamma V^\pi\left(s'\right)\right]$$

Tea cups

Algorithms list

- Списъка от алгоритми използвани при тестове е:
- *-Sarsa:*
- *-Quality of action equation by R.E.Bellman*
- *-Value of action equation by R.E.Bellman*
- *-Policy based on action equation.*
- Общото между тези уравниея е че, те са от раздела на приложна математика и дискретна математика описваща процес чрез промяна на състоянията. Решаването на тези уравнения изисква дефиниране на модел описващ процес.
- Моля обърнете внимание на слайд с име: **Model and equation**
- Наръчници за прилагане на алгоритмите:
- **Reinforcement Learning:**
- **An Introduction**
- by:
- *Richard S. Sutton and Andrew G. Barto*
- A Bradford Book
- The MIT Press
- Cambridge, Massachusetts
- London, England
- ***at : http://incompleteideas.net/book/first/ebook/node41.html***

Sarsa:

$$Q(s_t, a_t) = Q(s_t, a_t) + a^*(r_t + y^*Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t))$$

Q_l belman curent quality for action

$$Q(s_t, a_t) = Q(s_t, a_t) + a^*\left(r_t + y^*\max_a Q(s_{t+1}, a) - Q(s_t, a_t)\right)$$

which can also be written as

$$Q(s_t, a_t) = (1 - a)\,^*Q(s_t, a_t) + a^*\left(r_t + y^*\max_a Q(s_{t+1}, a)\right)$$

Value:

$$v(s) = \mathbb{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s]$$

Quality

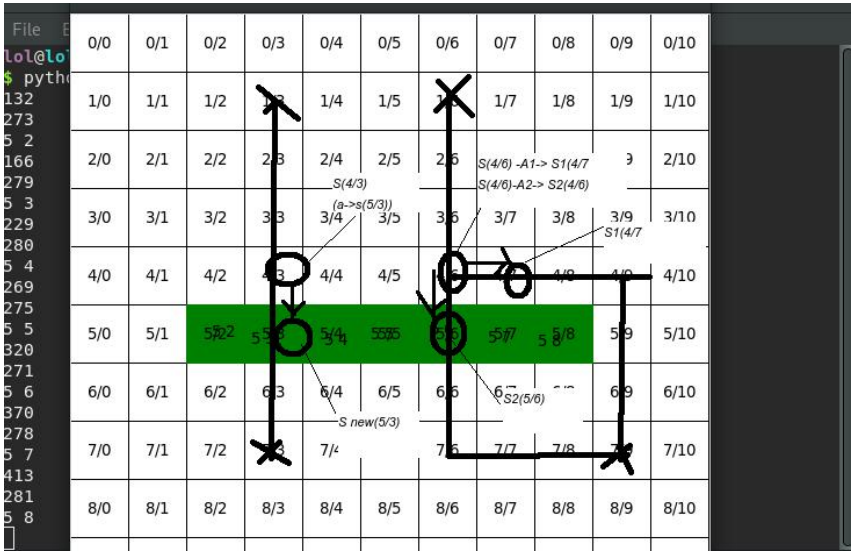
$$q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$$

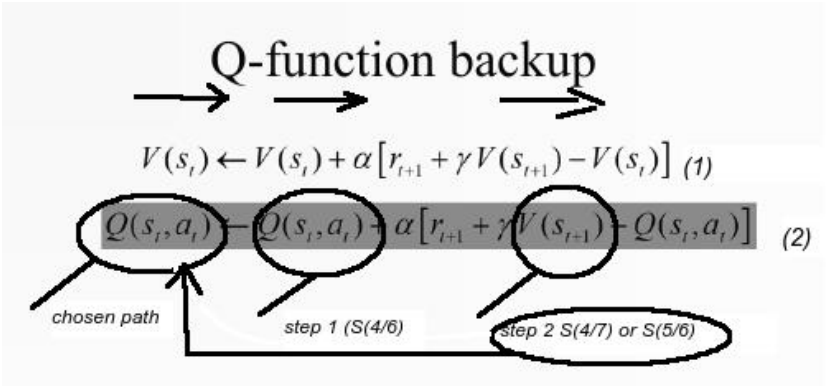
Policy:

$$\text{TDerror}\ (s) = V^\pi(s) - \sum_{s'} T(s, \pi(s), s')[r(s, \pi(s), s') + \gamma V^\pi(s')]$$

Tea cups

Model and equation





Tea cups

Model and equation explained

- In the slide: Model and equation you can see an equation that can be interpreted as:
- The quality of a certain action (Q) is determined by the change from initial state (S) to another state (St).
- Q has two attributes (S-state) and (A or V - value of action). Actions of higher quality have priority in choosing a policy or strategy.

- In this particular model, S (4/6) can choose to go into several new states. The state S (4/7) has a higher quality compared to the state S (5/6). From the agent's point of view, it's a better policy to go around a wall instead of crashing into it.

- В слайда: Model and equation се вижда уравнение което може да се тълкува като:
- Качеството на определено действие(Q) се определя промяната начално състояние(S) във друго състояние(St).
- Q има два атрибута(S-състояние) и (A или V - стойност на действие). Действията с по-високо качество имат приоритет при избор на политика или стратегия.

- В този определен модел S(4/6) може да избере да премине в няколко нови състояния. Състоянието S(4/7) има по-голямо качество сравнено със състоянието S(5/6). От гледна точка на агента е по-добра политика да заобиколи стена вместо да се блъсне в нея.

Tea cups

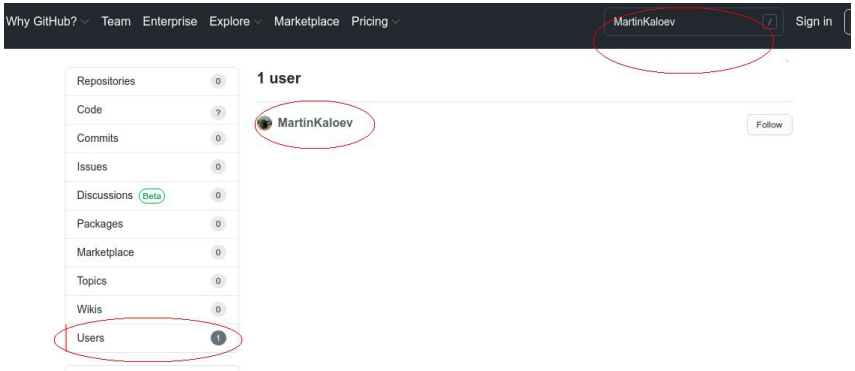
Test recreation

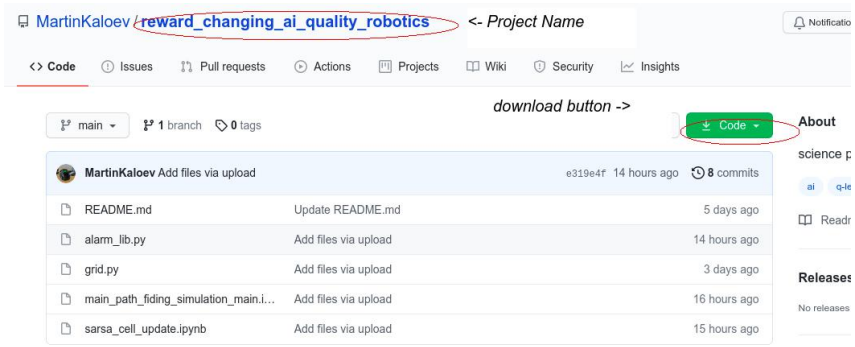
- To repeat the tests performed, please follow the steps below:
- 1) Use a virtual tool of your choice to navigate the World Wide Web to reach current github address
- 2) Download a copy of the software solution used in the tests. (Note **the github navigation** slides)
- 3) Use a virtual tool of your choice to navigate the World Wide Web to reach current jypiterlab address
- 4) Use the menus to access the online version of the software lab (Note **the jypiterlab navigation** slides)
- 5) Load a file named main _ ** _ main in the online version of jypiterlab and run the simulation. (Note the slides **jypiterlab: how to use it**)

- За да повторите извършените тестове, моля да последвате следните стъпки:
- 1) Използвайте виртуален инструмент по ваш избор за навигация в интернет за да, достигнете актуалния адрес на github
- 2) Изтеглете копие от програмното решение използвано в тестовете. (Обърнете внимание на слайдове github navigation)
- 3) Използвайте виртуален инструмент по ваш избор за навигация в интернет за да, достигнете актуалния адрес на jypiterlab
- 4) Използвайте менюта за да достигнете до онлайн версията на програмна лаборатория (Обърнете внимание на слайдове jypiterlab navigation)
- 5) Заредете файл с име main_**_main в онлайн версията на jypiterlab и стартирайте симулацията. (Обърнете внимание на слайдове jypiterlab: how to use it)

Tea cups

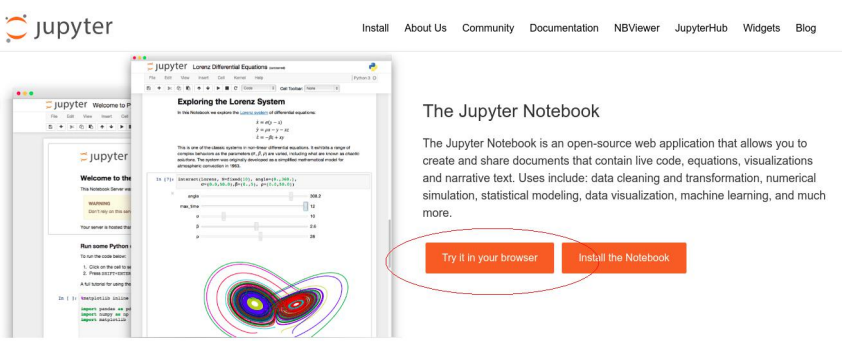
the github navigation

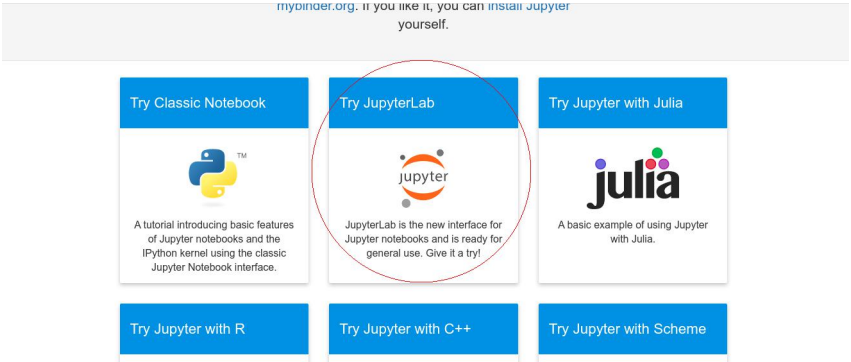




Tea cups

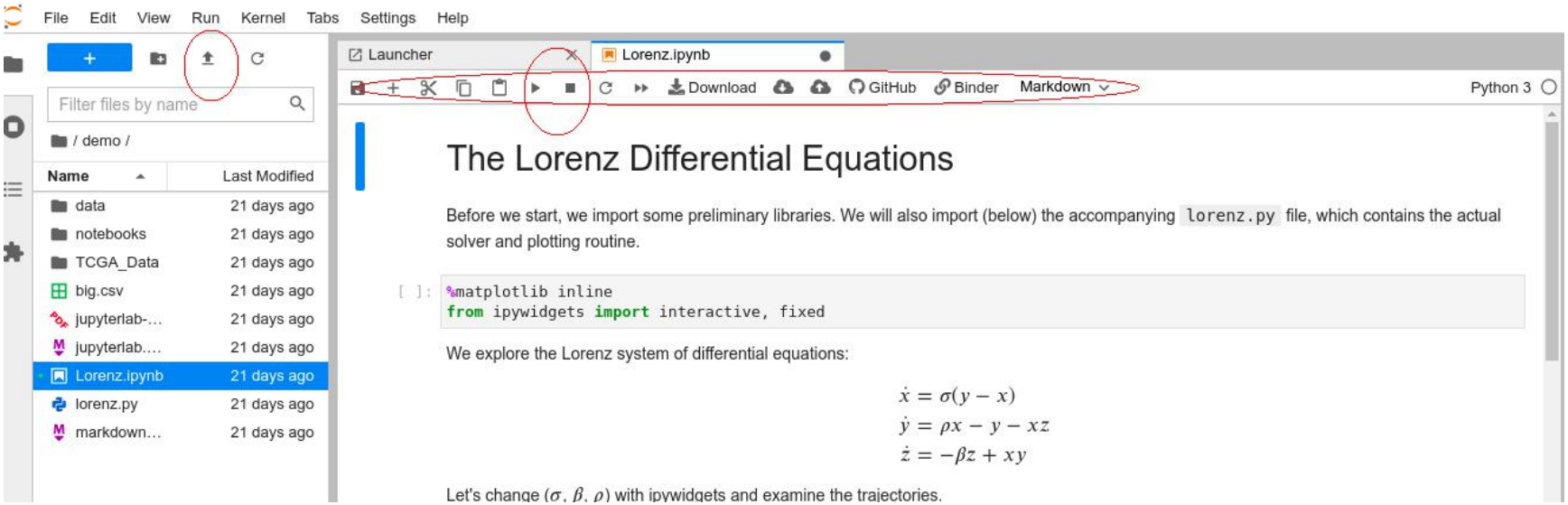
the jypiterlab navigation





Tea cups

jypiterlab: how to use it



Tea cups

test explained

- The virtual tool main _ ** _ main creates a simulation of an agent solving a randomly generated maze in an optimal way. The user can change the value (1) of the actions for each step and the penalties on the assessment of action in violation of the rules of the maze.
- This allows different agent behavior to be observed at different values set by the user.
- In addition, the functionality in cell 7 can be replaced by a Bellman algorithm with a Sarsa algorithm (2).

- 1*see Slide Model and equation explained again
- 2*see sarsa_cell_update sniped

- Виртуалният инструмента main_**_main създава симулация на агент решаващ случайно генериран лабиринт по оптимален начин. Потрибителя може да променя стойноста (1) на действията за всяка стъпка и наказанията върху оценката на действие при нарушаване на правилата на лабиринта.
- Това позволява да се наблюдава различно поведение на агент при различните стойности зададени от потребителя.
- В допълние функционалноста в cell 7 може да бъде заменена от алгоритъм на Bellman с алгоритъм Sarsa (2).

Results

- Tests are performed exhausting all possible combinations. After the tests are completed, a statistical analysis is performed as to which factors play the biggest role in the violation of the rules of the labyrinth by the agent. The results of the tests show that the difference between the value of the detour compared to the value of the penalty in the event of a breach plays the biggest role. The detour path means the path connecting two points that are closest to the breakthrough point.

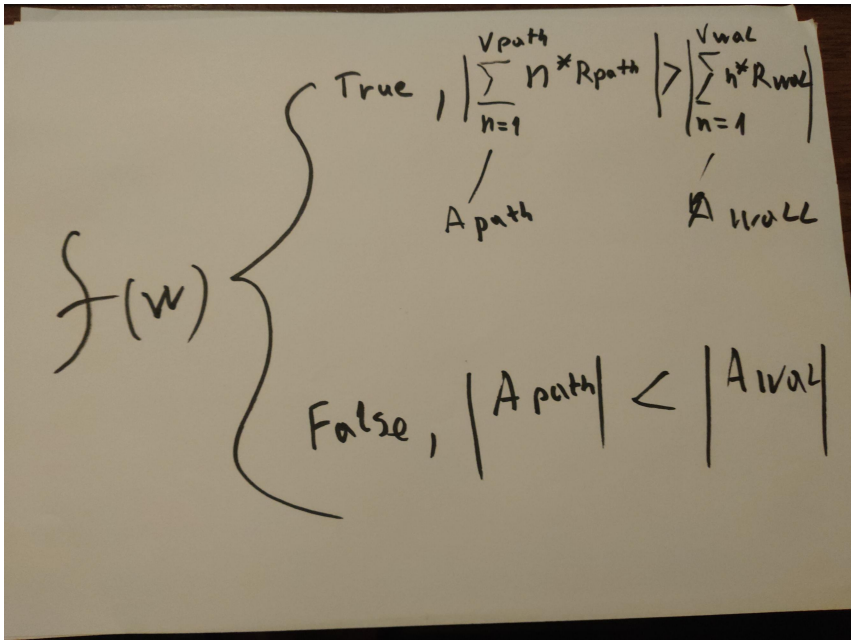
- Провежда се тестове изтощаващи всички възможно комбинации. След като тестовете се приключат се извършва статистически анализ кои фактори имат най-голяма роля при нарушаване на правилата на лабиринта от агента. Резултатите от тестовете показват че, най-голяма роля има взаимотношенията от стойноста на пътя на заобикаляне сравнено със стойноста от наказанието в случай на пробив. Път на заобикаляне означава пътя свързващ две точки които са най-близки до точката на пробив.

Tea cups

discussions and optimizations

- A possibility to optimize the virtual agent is the addition of a library signaling a possible breakthrough. An example of such a library is shown in alarm_lib.py provided in the project repository.

- Възможност за оптимизация на виртуалния агент е добавка на библиотека сигнализираща възможен пробив. Пример за такава библиотека е показан в alarm_lib.py предоставен в хранилището към проекта.



Where are we?

- Let's take a few minutes to discuss the role of Q-learning as the basis of many of the technologies used in the modern world.

- Да отделим няколко минути за дискусия ролята на Q-learning като основа на много от технологиите използвани в съвремният свят.

conclusion

- The article analyzes the change in the behavior of an agent when changing the values of actions.
- Answers to questions about why rewards often have a negative value instead of a positive one. The moments in which the agent violates the rules of the given task are considered and analyzed. A solution for the specific problems of a problem derived in an equation and a software library is proposed.

- В статията се анализира промяната на поведение на агент при промяна на стойностите на действията.
- Отговаря се на въпроси защо наградите често има отрицателна вместо положителна стойност. Разглежда и се анализират моментите в които агента нарушава правилата на дадената му задача. Предлага се решение за специфичните проблеми на задача изведени в уравние и софтуерна библиотека.

Thank you for your attention