مفاهیم یادگیری ماشین:

یادگیری ماشین به نظارت شده، نظارت نشده و تقویتی تقسیم میشن. نظارت شده و نظارت نشده که مفاهیم آشنایی دارن. توی یادگیری تقویتی یه عامل یادگیری داریم. همون عامل هوشمند که توی هوش مصنوعی هم داشتیم. توی تقویتی عامل یاد میگیره که چطور توی محیط با انجام یه سری اقدامات و دیدن نتیجه اونا، رفتار کنه. دقیقا مثل یه انسان که در طول زندگی و طی تجربیاتی که بدست میاره در آینده عمل میکنه. خب حالا این یادگیری چطور انجام میشه؟ با سیستم جزا و پاداش، ینی هربار که عامل کاری انجام میده در ازای کاری که انجام میده یا پاداش دریافت میکنه که ینی نتیجه کارش مثبت بوده، یا تنبیه میشه ینی نتیجه کارش منفی بوده. خب یه سری از کلماتی که میبنی اینان:

Environment = محیط

Actuator = عامل

Action(s) = عمل (اعمال)

Step(s) = گام(ها)

Result = نتیجه

State = حالت (حالت محیط)

Task = وظیفه

فرایند یادگیری:

فرض کن یه عامل در حال یادگیری داری. عامل حالت0 رو از محیط دریافت میکنه و با توجه به اون عمل0 رو انجام میده، محیط به حالت1 میره، محیط پاداش نتیجه1 رو به عامل میده (فرض کن +1 امتیاز بهش میده) خب اینجا یه دنباله از حالت، عمل، نتیجه(پاداش) رو داریم (که مدام تکرار میشه). عامل میخواد که پاداششو ماکسیمم کنه.

وظایف یا اپیزودیک یا مستمر(غیر اپیزودیک):

یه وظیفه، یه نمونه از یه مساله یادگیری هه.

توی وظایف اپیزودیک(مرحله ای) یه نقطه شروع داریم یه نقطه پایان(حالت ترمینال) مثل اینکه زمانی که بازی رو شروع میکنی تا زمانی یا تو اون مرحله به نقطه مقصدت برسی یا جونت تموم بشه و ببازی. عامل میخواد پاداشش رو توی هر اپیزود ماکسیمم کنه.

وظایف مستمر همیشه ادامه دارن و نقطه شروع و پایان نداره و تا زمانی که عامل انسانی متوقفش نکنه ادامه میده. مثل عامل هوشمندی که وظیفه اش خرید و فروش سهام بورسه، این تا زمانی که دستی متوقف نشه به خرید و فروش ادامه میده.

رویکرد های یادگیری تقویتی:

سه رویکر وجود داره: ارزش محور، سیاست محور و مدل محور. که چیزی که ما داریم بررسی میکنیم بحث مدل محور هست. در رویکرد مدل محور، محیط مدل میشه. ینی یه مدل از رفتار محیط ساخته میشه. در نتیجه هر محیط نیاز به مدل متفاوت داره.

من این مفاهیم رو خیلی خلاصه گفتم ولی کامل ترش رو میتونی از توی این لینک بخونی:

<https://blog.faradars.org/an-introduction-to-reinforcement-learning/>

Gym toolkit

خب حالا میرسیم به این کتابخونه. این کتاب خونه در واقع شامل یه سری نمونه مسائل یا به عبارت دیگه همون محیط ها هست که میتونی از اون برای الگوریتم های یادگیری تقویتی استفاده کنی. خب الان محیطی که تو انتخاب کردی محیط پکمن هه.

تابع step:

این محیط یه تابع استپ داره که چهار تا مقدار رو برمیگردونه

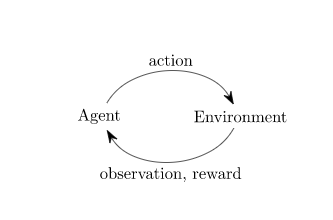
(تابع استپ در واقع همون عملی(اکشن) هست که عامل در تعامل با محیط انجام میده)

observation (object): an environment-specific object representing your observation of the environment. For example, pixel data from a camera, joint angles and joint velocities of a robot, or the board state in a board game. حالت محیط یا همون استیت رو برمیگردونه

reward (float): amount of reward achieved by the previous action. The scale varies between environments, but the goal is always to increase your total reward.مقدار پاداش رو برمیگردونه که میتونه یه عدد منفی یا مثبت باشه و هدف بیشینه کردن این عدد هست

done (boolean): whether it’s time to reset the environment again. Most (but not all) tasks are divided up into well-defined episodes, and done being True indicates the episode has terminated. (For example, perhaps the pole tipped too far, or you lost your last life.)نشون دهنده اینه که به اخر بازی یا حالت ترمینال رسیدی یا نه (مثلا مرحله رو تموم کردی یا اینکه سوختی)

info (dict): diagnostic information useful for debugging. It can sometimes be useful for learning (for example, it might contain the raw probabilities behind the environment’s last state change). However, official evaluations of your agent are not allowed to use this for learning.یه سری اطلاعات که برا دیباگ کردن خوبه ولی عامل برای یادگیری اجازه نداره از این اطلاعات استفاده کنه)



در واقع ما داریم یه حلقه میزنیم. عامل یه عمل انجام میده محیط تاثیر میگیره و یه حالت و پاداش رو به عامل برمیگردونه.

Reset()

با صدا زدن تابع ریست مرحله شروع میشه(به حالت اولیه محیط برمیگردیم) که در واقع مقدار فلگ دان رو ست میکنه. در واقع کار درست تر تغییر همین فلگ هست.

)done flag(

Spaces:

Every environment comes with an action\_space (این که چه اعمالی مجاز هستند. فضای اعمال)

and an observation\_space. (فضای مشاهدات. چه مشاهداتی امکان پذیر هستند)

These attributes are of type Space, and they describe the format of valid actions and observations

The Discrete space allows a fixed range of non-negative numbers, so in this case valid actions are either 0 or 1. The Box space represents an n-dimensional box, so valid observations will be an array of 4 numbers.

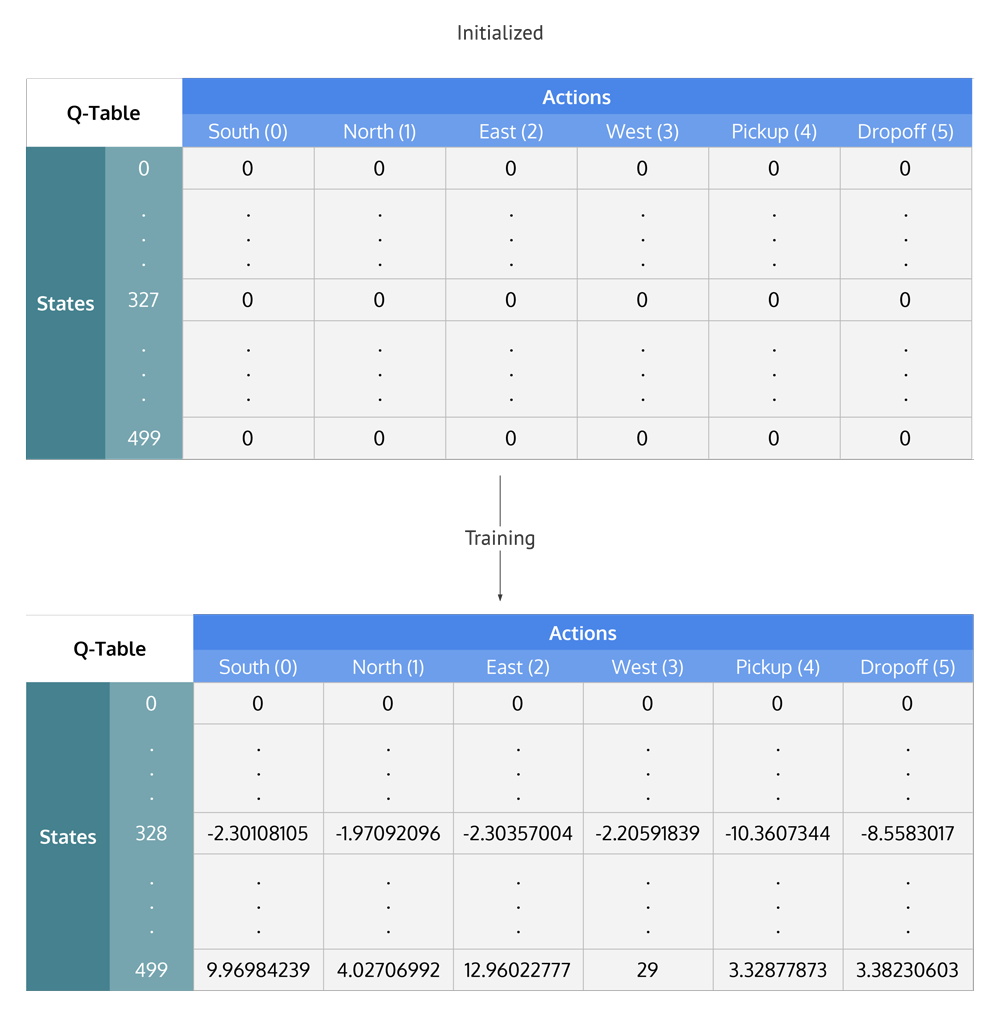
خب تا این جا چیزایی بود که خود داکیومنتش گفته بود.

اینی که این الان استفاده کرده اسمش

1. network هه

این یادگیری میاد یه جدول کیو درست میکنه که که توش دو تایی های حالت و عمل هستن. ینی مث یه ارایه که ستونش عمل هه سطرش حالت ها. ینی به ازای یه حالت عامل فلان عمل رو انجام میده و وقتی پاداش این گام رو گرفت مقدارشو میزاره تو خونه اش.

یه چیزی شبیه این عکس پایین



مقادیر اولیه این ماتریس خالیه. صفره. در حین ترین پر میشه. توی خود یادگیری-کیو به این حالته که یه اپیزود تموم میشه، بعد ارزیابی انجام میشه، ولی توی این حالت که با دیپ لرنینگ کار کرده توی خود اپیزود هم چند بار ارزیابی انجام میشه و این خونه ها اپدیتمیشن تا عملکرد بهتر بشه.