فرادرس
فراتر از یک کلاس درس
www.faradars.org

آموزش یادگیری تقویتی عمیق در پایتون
پیاده‌سازی بازی مار

فصل سوم: پیاده‌سازی عامل

مدرس:

سیّد علی کلامی هریس

دانشجوی دکترای حرفه‌ای داروسازی، فعال در حوزه یادگیری عمیق

faradars.org/fvdrlp101

اپیزود و گام

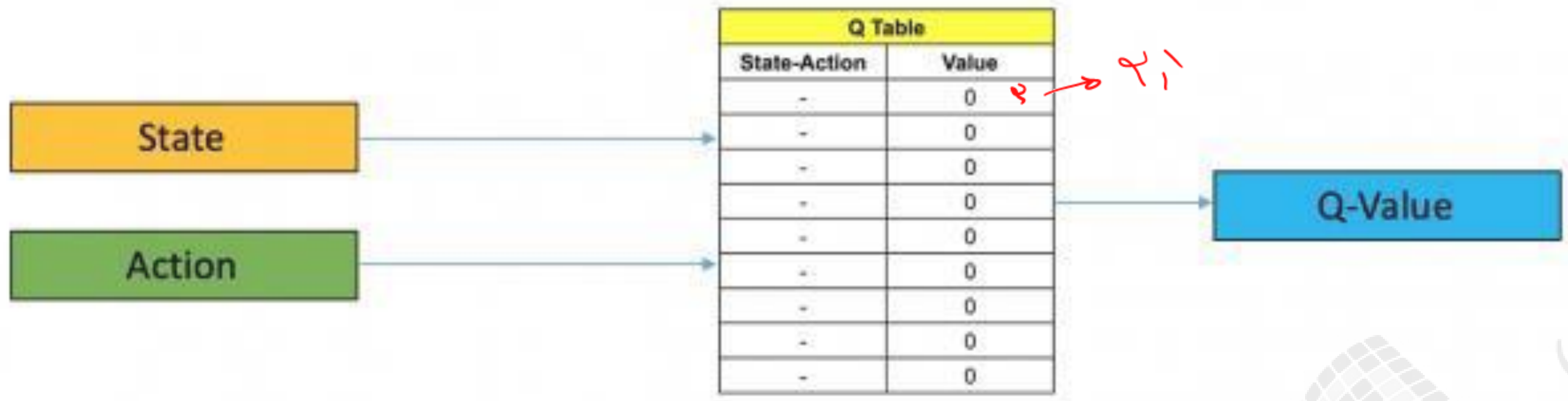
اپیزود (Episode): از شرایط ابتدایی (Initial State) شروع و به شرایط انتهایی (Final State) ختم می‌شود.

گام (Step): هر گام معادل یک بار تصمیم‌گیری و انجام عمل (Action) توسط عامل (Agent) می‌باشد.

یادگیری Q

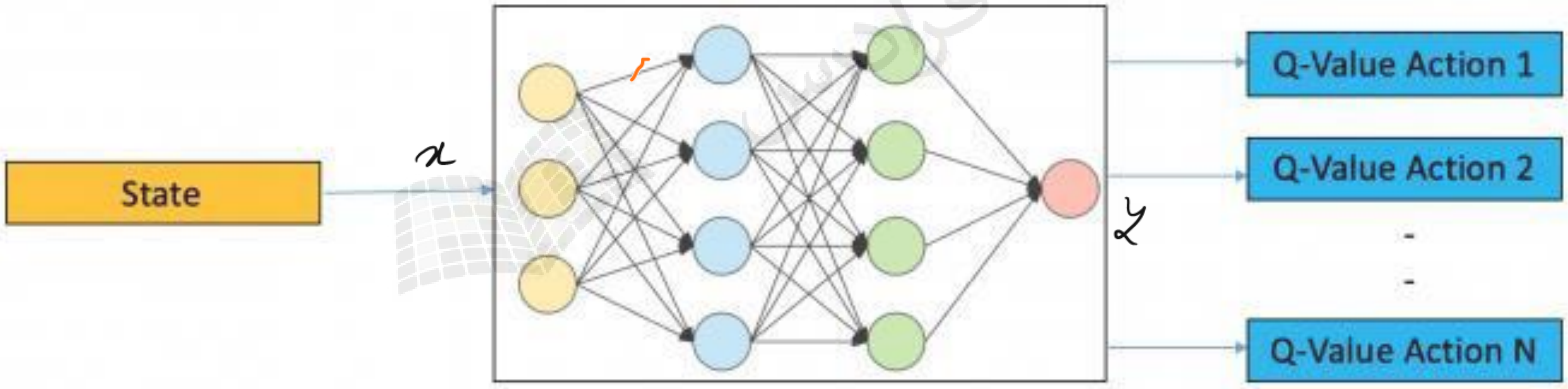
در مبحث Q-Learning در انتهای محاسبات به معادله بلمن (Bellman Equation) می‌رسیدیم که اساس آموزش مدل بود:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \overset{۰٫۱}{\underset{۱}{\alpha}} \cdot \overbrace{\left(r_t + \underset{۰٫۹}{\gamma} \cdot \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right)}^{\text{TD}}$$

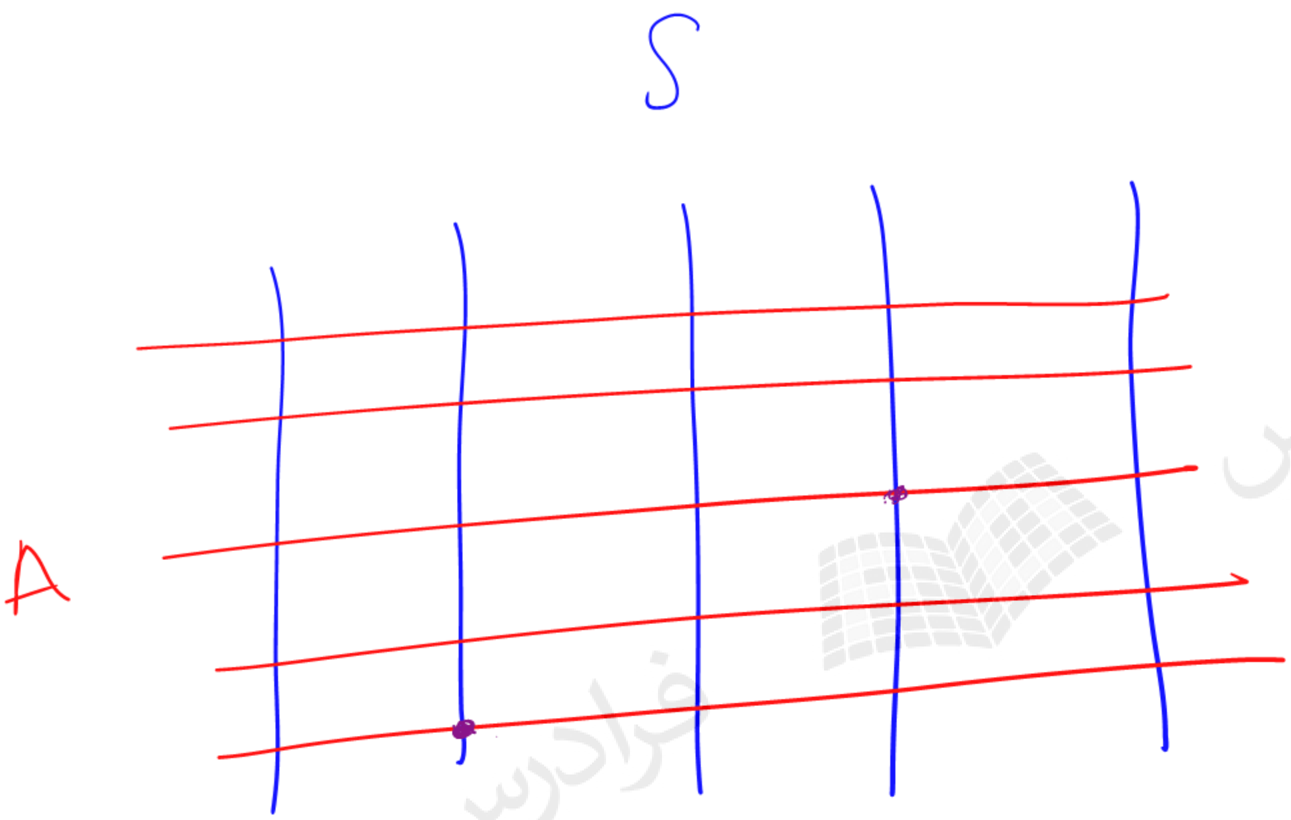


Q Learning

شبکه عصبی



Deep Q Learning


S
A

شبکه عمیق Q

برای این مدل اگر یک رشته (Trajectory) به شکل زیر داشته باشیم:

$$\underbrace{s_1, a_1, r_1, s_2}, a_2, r_2, \ldots, s_n, a_n, r_n$$

پیش‌بینی مدل به شکل زیر خواهد بود:

$$y = r + \gamma \cdot \max_{a'} Q(s', a'; \theta_{t-1})$$

شبکه عمیق Q

تابع زیان (Loss Function) نیز به شکل زیر قابل محاسبه خواهد بود:

$$L(\theta_t) = \mathbb{E}_{s,a,r,s' \sim \rho(\cdot)} \left[\left(y - Q(s, a; \theta_t) \right)^2 \right]$$

سیاست‌های مورد استفاده

سیاست تصادفی (Random Policy):

$$P(a_t = i) = \frac{1}{n}$$

$$a_t = u\{1, n\}$$

سیاست‌های مورد استفاده

سیاست حریصانه (Greedy Policy):

$$A_m = \operatorname*{argmax}_a Q(s_t, a; \theta_t)$$

$$P(a_t = i) = \begin{cases} 1 & i = A_m \\ 0 & \text{otherwise} \end{cases}$$

$$a_t = A_m$$

۱ $\boxed{\text{۲}}$ -۱ -۴

۰ ۱ ۰ ۰

سیاست‌های مورد استفاده

سیاست حریصانه اپسیلون (Epsilon Greedy Policy):

$$A_m = \underset{a}{\operatorname{argmax}}\, Q(s_t, a; \theta_t)$$

$$P(a_t = i) = \begin{cases} (1-\epsilon) + \epsilon/n & i = A_m \\ \epsilon/n & \text{otherwise} \end{cases}$$

$$a_t = \begin{cases} u\{1, n\} & U(0,1) < \epsilon \\ A_m & \text{otherwise} \end{cases}$$

$\epsilon = 1$ | $\epsilon = 0$

$\frac{1}{n}$ | 1

$\frac{1}{n}$ | 0

سیاست‌های مورد استفاده

سیاست بولتزمن (Boltzmann Policy):

$$P(a_t = i) = \frac{e^{\frac{Q(s_t,i;\theta_t)}{T}}}{\sum_{j=1}^{n} e^{\frac{Q(s_t,j;\theta_t)}{T}}}$$

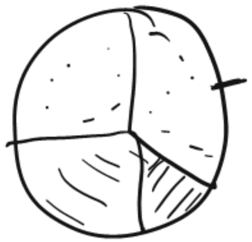
۱   -۲   ۳

T=۱

$e^{\frac{۱}{۱}}$ $e^{\frac{-۲}{۱}}$ $e^{\frac{۳}{۱}}$

۰٫۰۹   ۰٫۰۱   ۰٫۹

$\underbrace{۰ \quad ۰٫۰۹}_{۱} \underbrace{\quad ۰٫۱}_{۲} \underbrace{\quad ۱}_{۳}$



$$a_t = \min\left\{ i \,|\, U(0,1) < \sum_{j=1}^{i} P(a = j)\,,\, i \in \{1,2,\ldots,n\} \right\}$$

تعادل بین جستجو و بهره‌برداری

جستجو (Exploration)	بهره‌برداری (Exploitation)
در ابتدا باید زیاد باشد.	در انتها باید زیاد باشد.
بدون توجه به Qها کار می‌کنند.	با توجه به Qها کار می‌کنند.

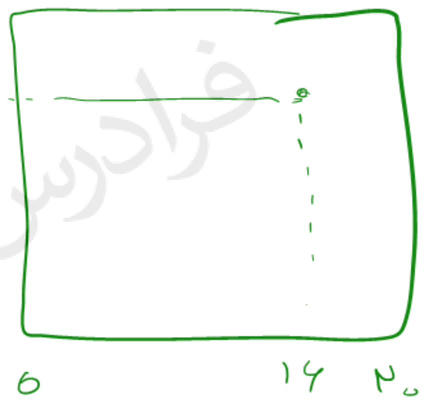
میانگین متحرک چیست؟

میانگین متحرک (Moving Average) یک ابزار بسیار کارآمد می‌باشد که برای نمایش روندها مفید می‌باشد.

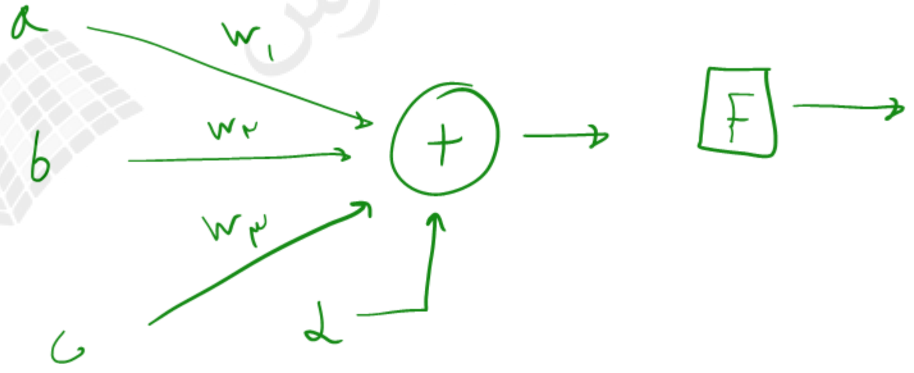
این ابزار یک میانگین با طول پنجره مشخص در طول زمان یا در طول یک آرایه محاسبه می‌کند.

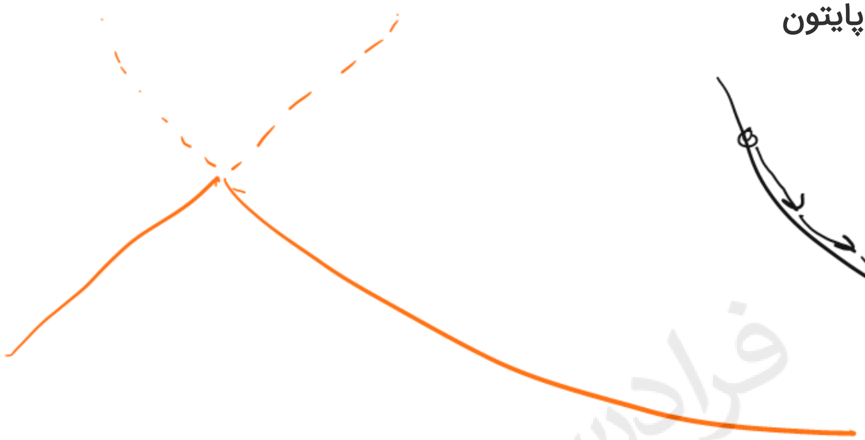
میانگین متحرک، انواع مختلفی دارد که میانگین متحرک ساده (Simple Moving Average یا SMA) ساده‌ترین آن‌ها می‌باشد.

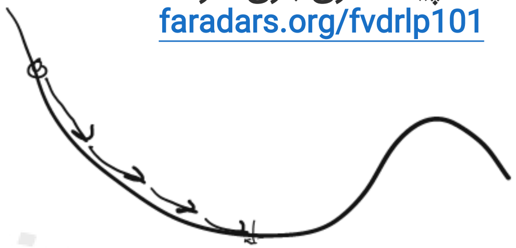

Time Series
SMA(20)
Value
Time
0
1
2
3
4
5
0
200
400
600
800
1000

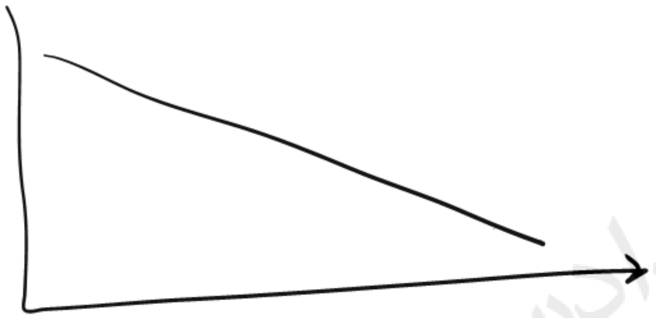


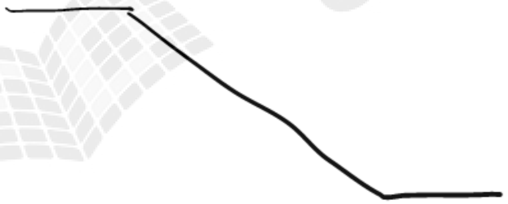
$$\frac{۱۶-۰}{۲۰-۰} = \frac{۱۶}{۲۰} = ۰٫۸$$

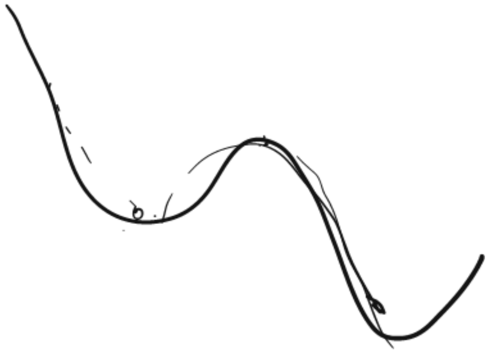


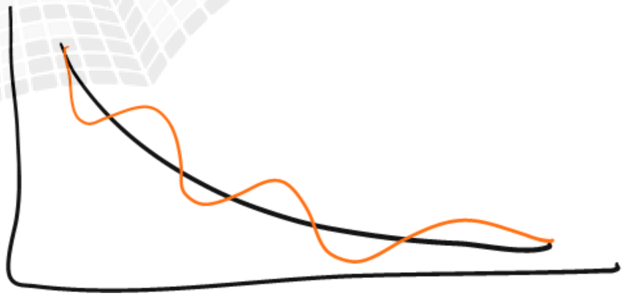


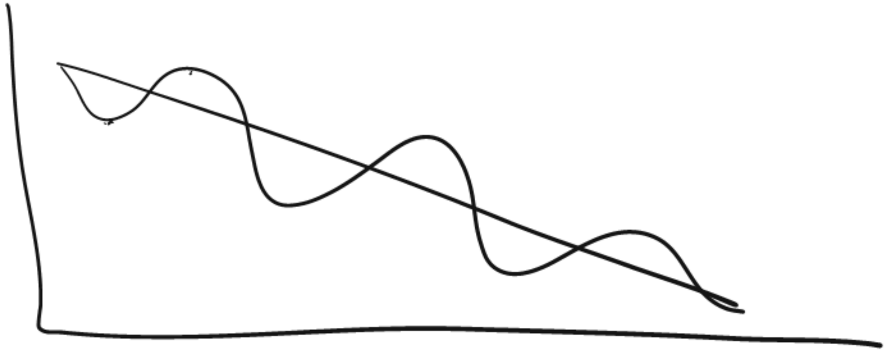


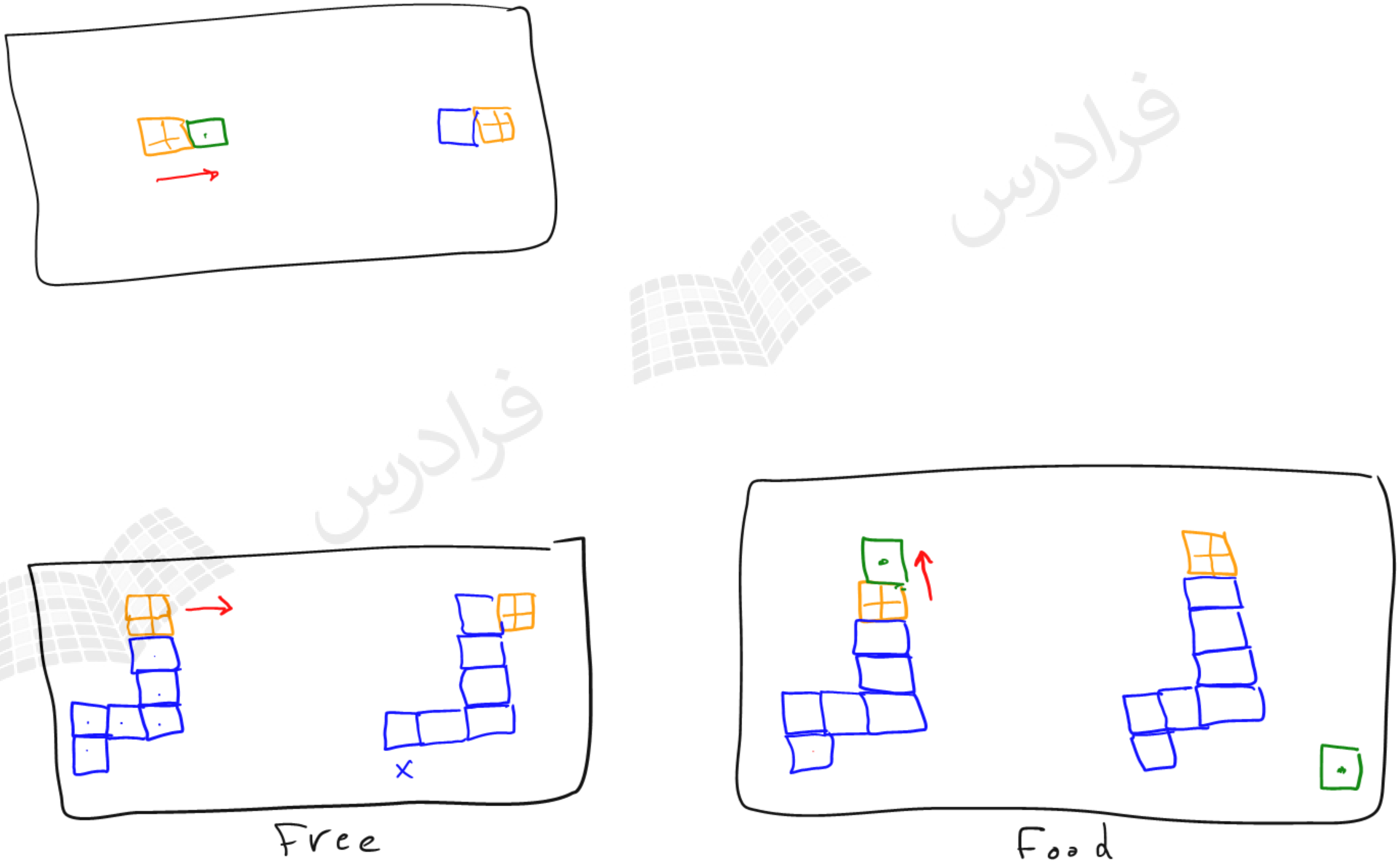


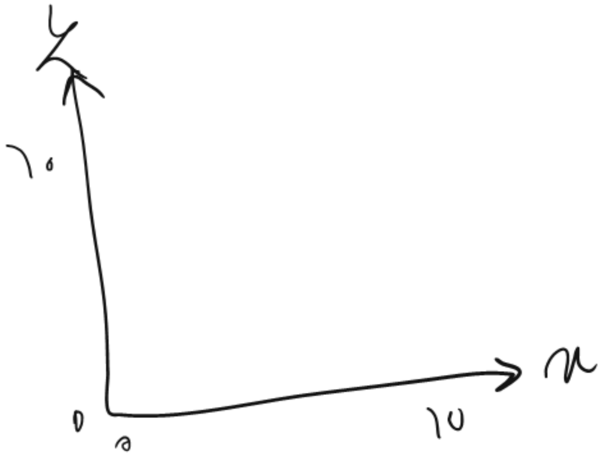




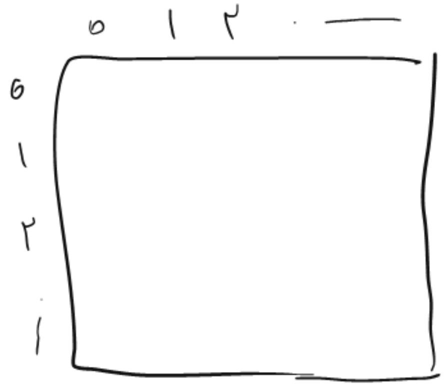


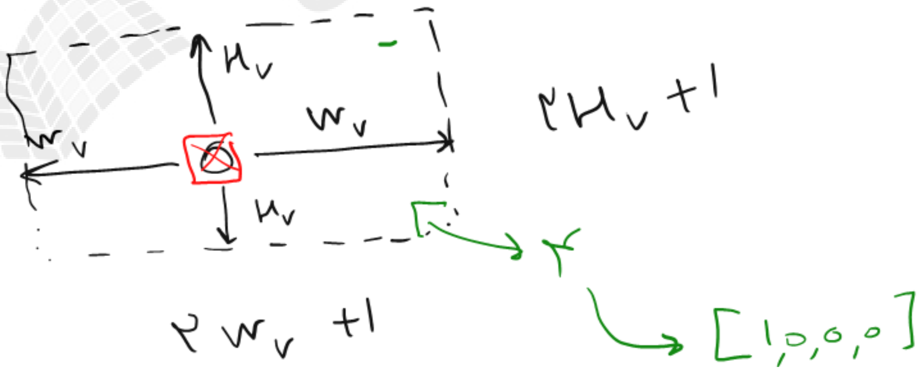



Free
Food



۰	0001
۱	0010
۲	0100
۳	1000





میانگین متحرک چیست؟

اگر یک سری زمانی (Time Series) به شکل زیر داشته باشیم:

$$S = \{S_1, S_2, \ldots, S_n\}$$

برای محاسبه میانگین متحرک ساده از رابطه زیر استفاده می‌کنیم:

$$SMA_t = \frac{1}{L} \cdot \sum_{i=1}^{L} S_{t-i+1}$$

این اسلایدها بر مبنای نکات مطرح شده در فرادرس

آموزش یادگیری تقویتی عمیق در پایتون

پیاده‌سازی بازی مار

تهیه شده است.

برای کسب اطلاعات بیشتر در مورد این آموزش به لینک زیر مراجعه نمایید.

faradars.org/fvdrlp101