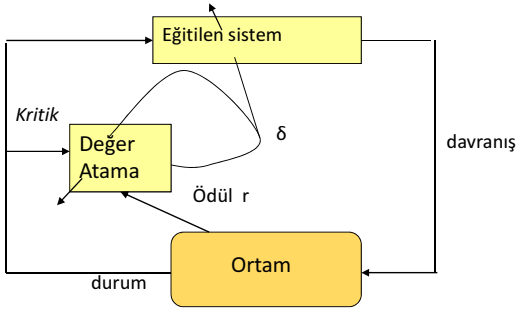
Hatırlatma

Eğiticisiz Öğrenme

- Pekiştirmeli Öğrenme (reinforcement learning)



Öğrenme işleminin her adımında istenilen yanıtı sağlayan bir eğitici yok

Eğitilen sistem, sonuçta elde edilecek yanıta erişmek için gerekli davranışı eleştiriyi gözönünde tutarak bulmak bulmak zorunda

Psikoloji açısından Pekiştirmeli öğrenme

- Biz kararlarımızı nasıl veriyoruz?
- Verdiğimiz kararlar daha sonraki davranışlarımızı nasıl etkiliyor?
- Verdiğimiz kararların sonuçları öğrenmemizi sağlıyor mu?

Şartlanma-Pekiştirmeli öğrenme

İlişkilendirme (association): $O_1 \longrightarrow T_1$, $O_2 \longrightarrow T_2$, $O_1 \longrightarrow T_2$

Klasik Şartlanma

Throndike'nin Yasası: $U_1 \longrightarrow Te_1$  $U_1 \Longrightarrow Te_1$

$U_2 \longrightarrow Te_2$  $U_2 \longrightarrow Te_2$

Etkin Şartlanma

δ

Psikolojide pekiştirmeli öğrenme

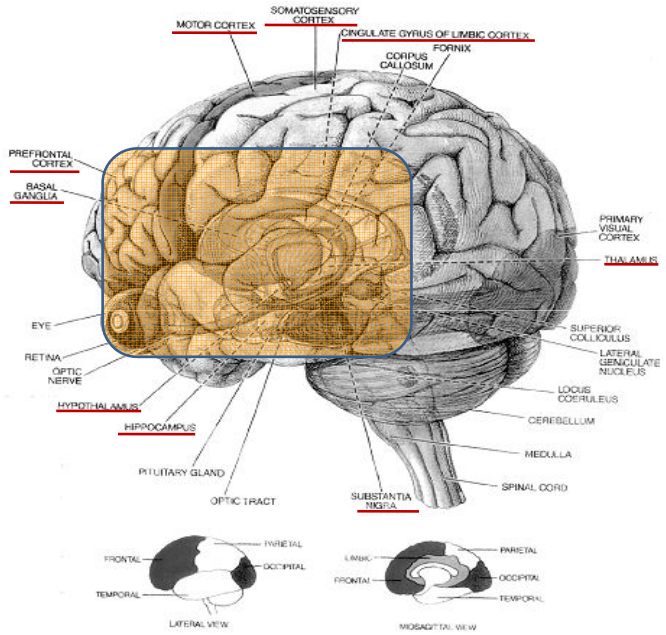
- Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond. (Thorndike, 1911, p. 244)

Psikolojide pekiştirmeli öğrenme

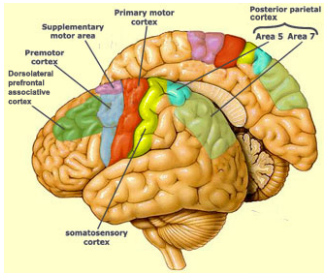
- Throndike (1898): uyaran-yanıt ilişkilendirmesi
  (stimulus-response association)
- Skinner (1938): davranışsal düzenleme
  (behavioral regulation)

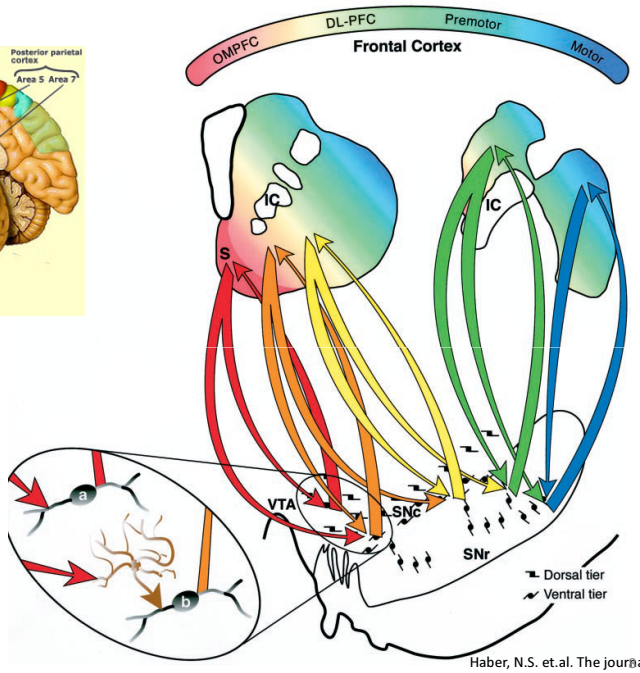
Nörobilim açısından Pekiştirmeli öğrenme

- Beyindeki hangi bölgeler yer alıyor?
- Bu bölgelerin birbirleriyle bağlantıları neler?
- Bağlantıları etkileyen mekanizmalar neler?



http://thebrain.mcgill.ca/





Haber, N.S. et.al. The journal of neuroscience, 2000.

Makina öğrenmesinde pekiştirmeli öğrenme
(Machine learning)

- Ortamdaki belirsizliğe rağmen bir amaca erişmek için aktif karar veren bir aracının ortamla ilişkisi inceleniyor.
- Aracı davranışlarını seçerken yararlanma-arama ikilemi ile yüzleşir. (exploit-explore)
- Pekiştirmeli öğrenme sistemi:

  π yaklaşım (policy)

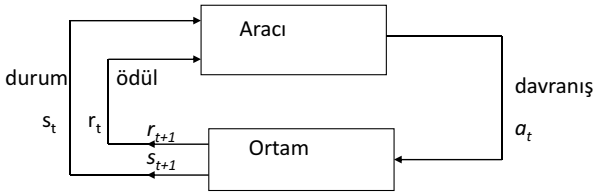
  r ödül fonksiyonu (reward function)

  Q^π , V^π değer fonksiyonu (value function)

  s *ortam modeli*

Makina öğrenmesinde pekiştirmeli öğrenme

- öğrenen, karar veren aracı
- etkileşim içinde olduğu ortam



- amaç: π^* optimal yaklaşımı bulmak

$$V^*(s) = \max_\pi V^\pi(s) \qquad Q^*(s,a) = \max_\pi Q^\pi(s,a)$$

Makina öğrenmesinde pekiştirmeli öğrenme

$$V^\pi(s) = E_\pi\{R_t | s_t = s\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\}$$

$$= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}\left[R^a_{ss'} + \gamma V^\pi(s')\right]$$

Makina öğrenmesinde pekiştirmeli öğrenme

$$Q^\pi(s,a) = E_\pi\{R_t | s_t = s,\ a_t = a\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s,\ a_t = a\right\}$$

$$= \sum_{s'} P^a_{ss'}\left[R^a_{ss'} + \gamma V^\pi(s')\right]$$

Bir Pekiştirmeli öğrenme metodu: Dinamik programlama

- ortam modeli : Markov karar işlevi
  (Markov Decision Process (MDP))

- yaklaşım belirleme : ardışıl

$$V_{k+1}(s) = E_\pi\left\{r_{t+1} + \gamma V_k(s_{t+1}) \middle| s_t = s\right\}$$
$$= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'}\left[R^a_{ss'} + \gamma V_k(s')\right]$$

Bir Pekiştirmeli öğrenme metodu: Dinamik programlama

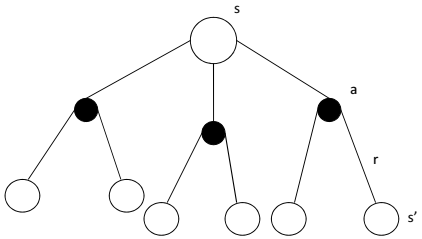
- yaklaşım iyileştirme :

$$Q^\pi(s,a) = E_\pi\left\{r_{t+1} + \gamma V^\pi(s_{t+1}) \middle| s_t = s,\ a_t = a\right\}$$
$$= \sum_{s'} P^a_{ss'}\left[R^a_{ss'} + \gamma V^\pi(s')\right]$$

- yaklaşım:

$$\pi'(s) = \arg\max_a Q^\pi(s,a) = \arg\max_a E\left\{r_{t+1} + \gamma V^\pi(s_{t+1}) \middle| s_t = s,\ a_t = a\right\}$$

Bir Pekiştirmeli öğrenme metodu: Dinamik programlama



Bir pekiştirmeli öğrenme metodu: Monte Carlo

- ortam modeli: deneyim
  gerçek deneyim
  (on-line)
  benzeşim deneyim
  (simulated)

- yaklaşımla ve yaklaşım ötesinde
  (on-policy) (off-policy)

Bir pekiştirmeli öğrenme metodu: Monte Carlo

$$V^{\pi}(s) = (1-\varepsilon)\max_{a}\sum_{s'} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma V^{\pi}(s')\right]$$
$$+\frac{\varepsilon}{|\mathrm{A(s)}|}\sum_{a}\sum_{s'} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma V^{\pi}(s')\right]$$

Bir pekiştirmeli öğrenme metodu: Monte Carlo

$$Q^{\pi}(s,\pi'(s)) = \sum_{a}\pi'(s,a)Q^{\pi}(s,a)$$
$$= \frac{\varepsilon}{|A(s)|}\sum_{a} Q^{\pi}(s,a) + (1-\varepsilon)\max_{a} Q^{\pi}(s,a)$$

Bir pekiştirmeli öğrenme metodu: Monte Carlo

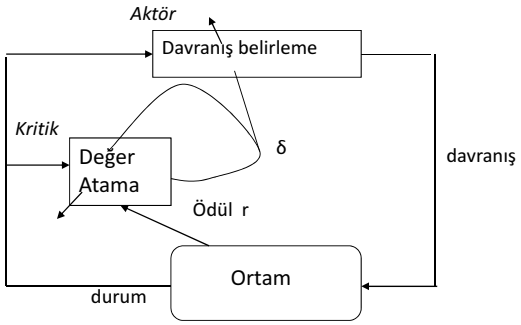
Bir pekiştirmeli öğrenme metodu: Zamansal fark

- Monte Carlo metoduna benziyor: ortamın tam modeline gereksinimi yok
- Dinamik programlamaya benziyor: en son çıktıyı beklemeden güncelleme yapabiliyor

- Yaklaşımla: Sarsa
  yaklaşım ötesinde: Q-öğrenme (Q-learning)

- Aktör-Kritik

Bir zamansal fark metodu: Aktör-kritik



Bir Zamansal fark metodu: Aktör-kritik

$$V(t) = w^T(t)s(t)$$
$$\delta(t) = r(t+1) + \gamma V(t+1) - V(t)$$
$$w(k+1) = w(k) + \eta\delta(k)x(k)$$

$$\hat{w}(t+1) = \hat{w}(t) + \alpha\delta(t)e(t)$$
$$e(t+1) = \lambda e(t) + (1-\lambda)a(t)s(t)$$
$$a(t) = f\left[\hat{w}^T(t)s(t) + n(t)\right]$$

Pekiştirmeli öğrenmeye ilişkin biliş bilimde bir uygulama

Biliş bilim ne ile ilgileniyor?

- Davranışsal: girişe karşılık gelen çıkış ne?
- Fonksiyonel: çıkış nasıl oluşuyor?
- Fiziksel: çıkışı ne üretiyor?

Pekiştirmeli öğrenme için geliştirilecek bir hesaplamalı modelde nelere dikkat edilmeli?

Davranışsal: uyaran ⟷ yanıt
yanıt ⟷ ödül/ceza
ödül ⟶ yararlan (exploit)
ceza ⟶ ara (explore)

Fonksiyonel: geçmişi değerlendir ⟶ beklenti oluştur

Fiziksel: nöral yapıların/bağlantıların özelikleri

Pekiştirmeli öğrenme için önerilen bazı hesaplamalı modeller

- Barto & Sutton & Anderson (1983)
  makina öğrenmesi
  TD (temporal difference)

- Schultz & Dayan & Montague (1997)
  Kritik,TD
  Kritik: VTA

- Suri & Scultz (1998)
  Aktör-Kritik, TD
  Kritik: nigrostriatal dopamin nöronları
  Aktör: Striatum

Bir pekiştirmeli öğrenme metodu: Zamansal fark (Temporal Difference(TD))

Barto , A.G.
IEEE,Syst.
Man&Cyber.1983

Gelecekteki ödülü öngörme

t anındaki öngörü

t+1 anındaki öngörü

$$P(t) = E\{r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + ...\}$$

$$P(t+1) = E\{r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + ...\}$$

$$P(t) = E\{r(t) + \gamma(\underbrace{r(t+1) + \gamma r(t+2) + ...}_{P(t+1)})\}$$

$$P(t) = E\{r(t)\} + \gamma P(t+1)$$

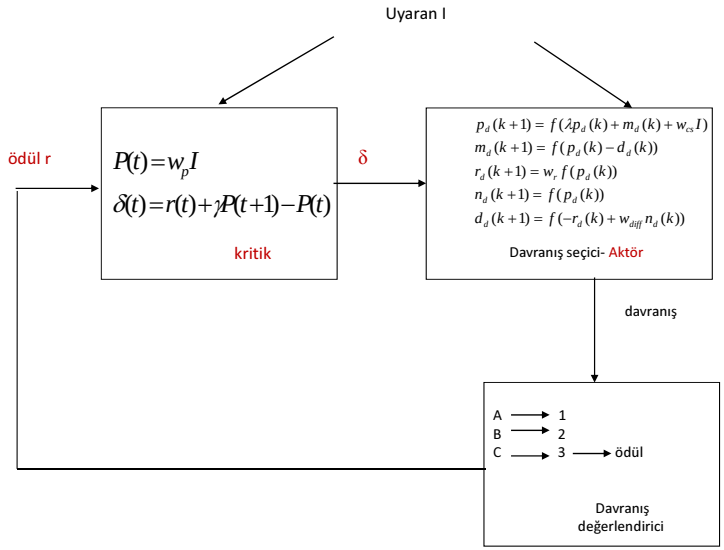
$$\delta(t) \hat{=} r(t) + \gamma P(t+1) - P(t)$$ ← Hata

Ardışıl eşleştirme ödevi

- Amaç: Bir dizi öğrenmek A → 1, B → 2, C → 3

- Yöntem: 1) U_1 = C Te_1 = 3
  2) U_2 = B Te_2 = 2
  3) U_3 = A Te_3 = 1

  $U_1 \rightarrow Te_1$ ödül
  $U_2 \rightarrow Te_2$
  $U_1 \rightarrow Te_1$ ödül
  $U_3 \rightarrow Te_3$
  $U_2 \rightarrow Te_2$
  $U_1 \rightarrow Te_1$ ödül



Davranış seçici sistem

$$p_d(k+1) = f(\lambda p_d(k) + m_d(k) + w_{cs} I)$$
$$m_d(k+1) = f(p_d(k) - d_d(k))$$
$$r_d(k+1) = w_r f(p_d(k))$$
$$n_d(k+1) = f(p_d(k))$$
$$d_d(k+1) = f(-r_d(k) + w_{diff} n_d(k))$$

$$f(x) = 0.5\tanh(ax - \beta)$$

w_r ve w_{cs} öğrenme ile değiştirilecek

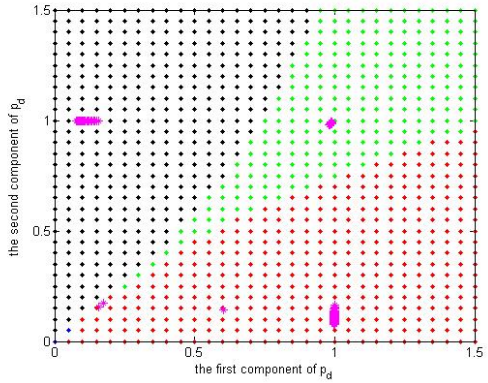
Güncelleme terimleri

$$w_r(n+1) = w_r(n) + \eta_{w_r} . \delta(n) . w_r . f(p_d(k))$$

$$w_v(n+1) = w_v(n) + \eta_{w_v} . \delta(n) . I^T(n-1)$$

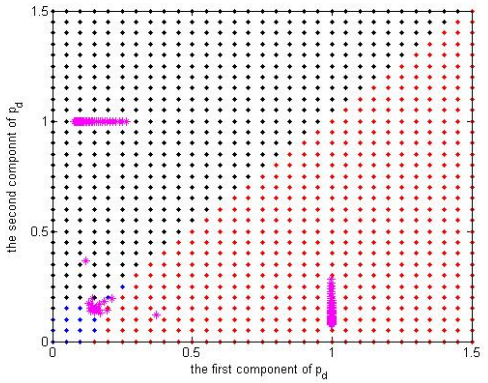
$$w_{cs}(n+1) = w_{cs}(n) + \eta_{w_{cs}} . \delta(n) . a(n-1) . I^T(n-1)$$

Dinamik sistemin davranışı



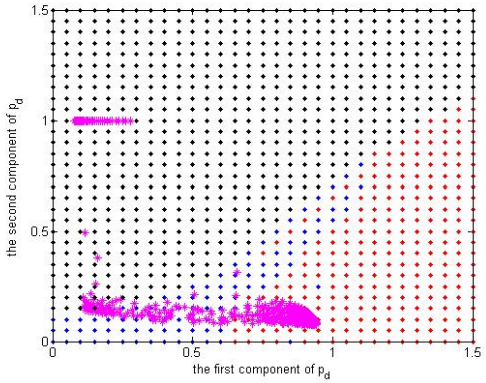
$$w_r = \begin{bmatrix} 1.25 & 0 \\ 0 & 1.25 \end{bmatrix}$$

Dinamik sistemin davranışı



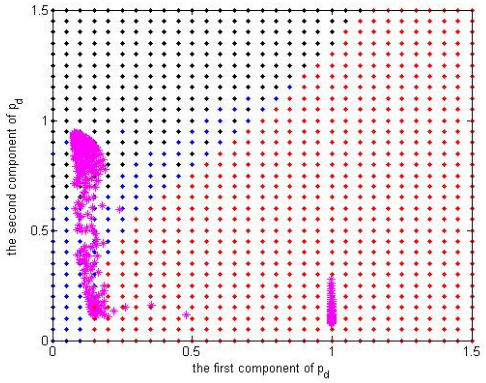
$$w_r = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Dinamik sistemin davranışı



$$w_r = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}$$

Dinamik sistemin davranışı



$$w_r = \begin{bmatrix} 1 & 0 \\ 0 & 0.5 \end{bmatrix}$$