Pekiştirmeli Öğrenme

Yapay Zekâ ve Veri Dersleri Rehberi · Kısım II: Yapay Zekânın Dalları

Etmen ve çevre | Markov karar süreci | Değer fonksiyonu | Keşif ve sömürü | Derin PÖ | Komşu dersler

Sercan KÜLCÜ · Giresun Üniversitesi · Bilgisayar Mühendisliği

© Sercan Külcü. Tüm hakları saklıdır. · Sunumun etkileşimli HTML sürümü: https://sercankulcu.github.io/files/ai_courses/sunumlar/Bolum_07_Pekistirmeli_Ogrenme.html

GİRİŞ

Ders Hangi Soruyu Sorar?

> **Yalnız ödül sinyaliyle, deneme yanılma yoluyla iyi kararlar nasıl öğrenilir?**

Pekiştirmeli öğrenme, bir etmenin çevreyle etkileşerek, yalnız aldığı **ödül sinyaline** bakıp uzun vadede toplam ödülü en büyük yapan davranışı deneme yanılmayla öğrenmesini inceler. Doğru eylem önceden söylenmez; eylemlerin sonucu çoğu zaman gecikmeli gelir.

Bir bakışta

YAKLAŞIM	GİRDİ	ÇIKTI
ödülle deneme yanılma	ortamla etkileşim: durum ve ödül	politika: durumdan eyleme kural

BAŞARI ÖLÇÜTÜ	MATEMATİK YÜKÜ	VERİ İHTİYACI
toplam beklenen ödül	yüksek: olasılık, dinamik programlama	etkileşimle üretilir

Kısa Tarihçe

1957
Bellman, Dynamic Programming kitabıyla Bellman denklemini ve değer fonksiyonu fikrini ortaya koyar.

1988
Sutton zamansal fark (TD) öğrenmesini yayımlar: tahmini bir sonraki tahminle düzeltmek.

1989
Watkins doktora tezinde Q-öğrenmeyi önerir.

1992
Tesauro'nun TD-Gammon programı kendine karşı oynayarak tavlada usta düzeyine yaklaşır.

2015
DeepMind'ın derin Q-ağı (DQN) Atari oyunlarını ham piksellerden öğrenir (Nature).

2016
AlphaGo, Go'da Lee Sedol'ü yener: ağaç araması, derin ağlar ve pekiştirmeli öğrenme birlikte.

DERS

Temel Kavramlar

Etmen ve çevre
Etmen eylem yapar; çevre yeni durum ve ödül döndürür.

Markov karar süreci
Durumlar, eylemler, geçiş olasılıkları ve ödüller.

Getiri
İndirimli (discounted) gelecekteki ödüllerin toplamı.

Politika
Her durumda hangi eylemin seçileceğini söyleyen kural.

Değer fonksiyonu
Bir durumdan ya da eylemden beklenen getiri.

Keşif ve sömürü
Yeni eylem denemek ile bilineni kullanmak arasındaki denge.

Zamansal fark
TD, Q-öğrenme, SARSA: modelsiz, adım adım öğrenme.

Politika gradyanı
Politikayı doğrudan, getirinin gradyanıyla iyileştirmek.

DERS

Tipik İş Akışı



Örnek: Bir depo robotu ızgara üzerinde konumunu gözler, çoğu zaman en iyi bildiği yönü bazen de rastgele bir yönü seçer, rafa ulaşınca +10, duvara çarpınca −1 ödül alır. Q-öğrenme her adımda eylem değerlerini günceller; binlerce denemeden sonra kısa ve güvenli yol ortaya çıkar.

DERS

Derste Neler İşlenir?

1. Çok kollu haydut problemi ve keşif
2. Markov karar süreçleri ve Bellman denklemleri
3. Dinamik programlama: politika ve değer yinelemesi
4. Monte Carlo yöntemleri
5. Zamansal fark öğrenmesi: TD(0) ve SARSA
6. Q-öğrenme ve politika dışı öğrenme
7. n-adımlı yöntemler ve uygunluk izleri
8. Fonksiyon yaklaşıklamayla değer kestirimi
9. Derin Q-ağları: deneyim tekrarı, hedef ağ
10. Politika gradyanı ve REINFORCE
11. Aktör-eleştirmen yöntemleri ve PPO
12. Model tabanlı PÖ ve Monte Carlo ağaç araması

Kazandırdığı beceriler

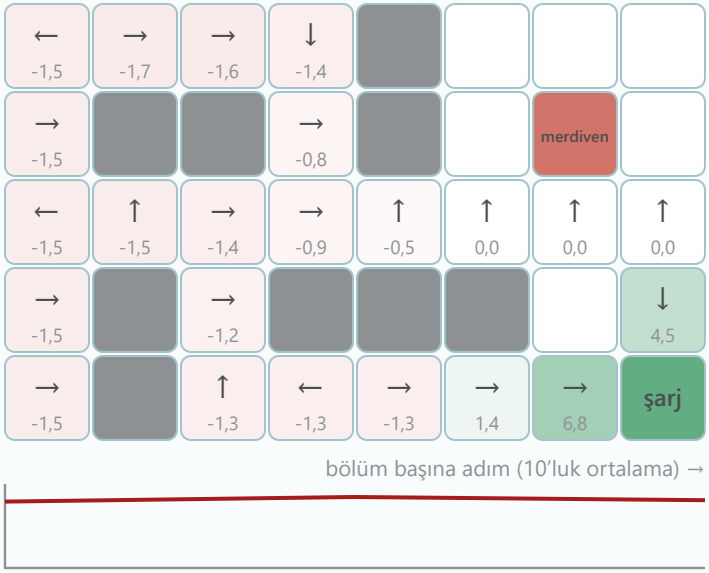
Gymnasium ile ortam kurmak | Q-öğrenmeyi sıfırdan yazmak | Stable-Baselines3 ile etmen eğitmek | ödül fonksiyonu tasarlamak | çok tohumlu deney yürütüp raporlamak

Tipik bir dönemin akışıdır; sıra ve kapsam üniversiteden üniversiteye değişir.

DENE

Süpürge Şarjı Kendi Buluyor

Robot süpürgeye harita verilmedi; her adım −1, şarj istasyonu +10, merdivenden düşmek −10 puan. Bölümler boyunca deneyerek öğrensin: eğitin, sonra robotu sol üst köşeden gönderin.



10 bölüm eğit | 100 bölüm eğit

Robotu gönder | Sıfırla

Keşif oranı ε = **0,20**

Toplam **3** bölüm. Son 10 bölümde ortalama **80,7** adım, 0 kez merdivenden düştü.

Okların gösterdiği yol sol üstten henüz şarja varmıyor: daha çok eğitin.

Kimse robota yol göstermedi: Q-öğrenme her adımdan sonra o karedeki eylemin değerini **ödül + gelecekteki en iyi değer** ile günceller. Keşif yoksa yeni yollar bulunmaz, çoksa robot sık düşer.

KONUM

Ön Koşullar

Bu dersten önce bilinmesi yararlı olan dersler ve neden gerektikleri.

Ders	Neden gerekir
Makine Öğrenmesi	etiket yerine ödülle öğrenme
Algoritmalar	dinamik programlama ve çizgede en kısa yol
Programlama	Python ile simülasyon döngüleri ve diziler

KONUM

Bu Ders Sana Göre mi?

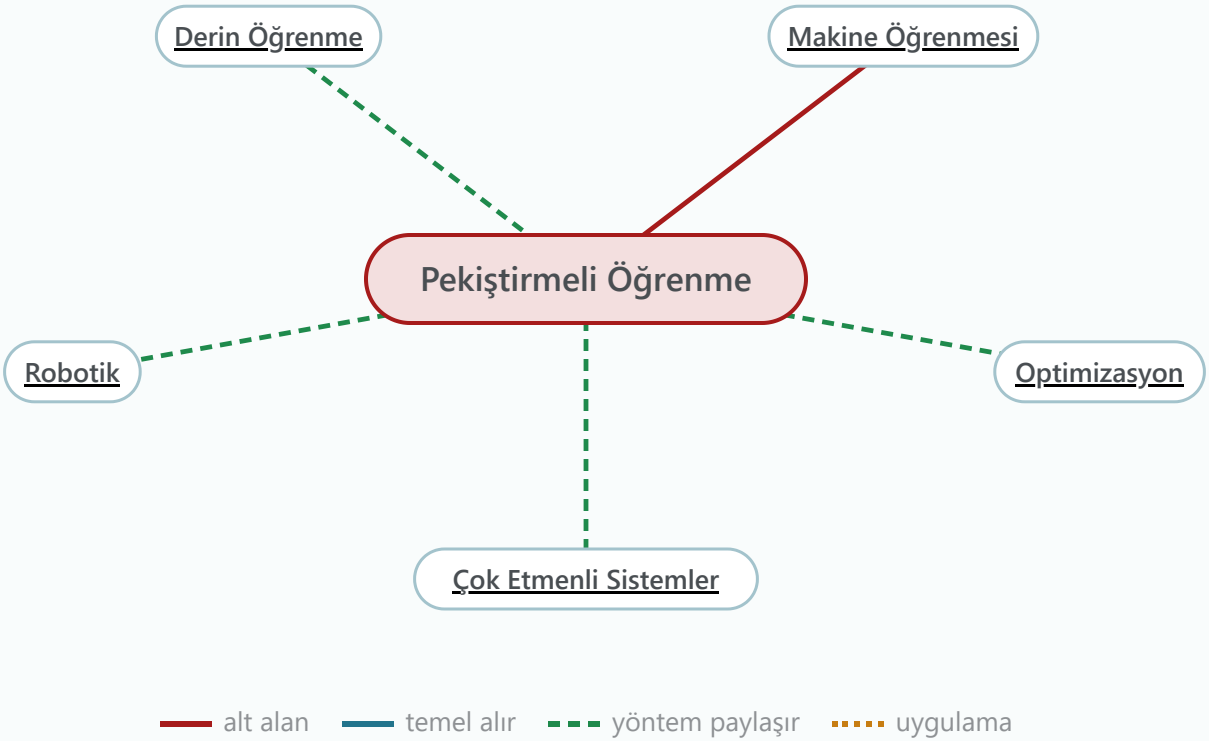
Kendine sor; "hayır" dediklerin için sağdaki sütun nereden başlayacağını gösterir.

Kendine sor	Hayırsa önce
Beklenen değeri ve koşullu olasılığı hesaplayabiliyor musun?	Olasılık ve İstatistik
Dinamik programlamayla bir en kısa yol problemini çözebiliyor musun?	Algoritmalar dersindeki dinamik programlama konusu
Kayıp fonksiyonunu ve gradyan inişini açıklayabiliyor musun?	Makine Öğrenmesi
Python'da bir simülasyon döngüsü yazıp sonucu grafikle çizebiliyor musun?	Programlamaya giriş, ardından Matplotlib'e kısa bir giriş

Çoğuna "evet" diyebiliyorsan derse hazırsın; bir ya da iki "hayır" dönem içinde kapanabilir.

KONUM

Dersin Komşuları



ÖRTÜŞME

Örtüştüğü Dersler

İlişki	Ders	Ortak konu
alt alan	Bu ders, Makine Öğrenmesi dersinin alt alanıdır	etiket yerine ödülle öğrenme
yöntem paylaşır	Derin Öğrenme ile yöntem paylaşır	derin Q-ağları, derin politika ağları
yöntem paylaşır	Çok Etmenli Sistemler ile yöntem paylaşır	çok etmenli pekiştirmeli öğrenme, oyun kuramı
yöntem paylaşır	Robotik ve Robot Görüsü ile yöntem paylaşır	denetim, hareket öğrenme
yöntem paylaşır	Optimizasyon ile yöntem paylaşır	dinamik programlama, politika gradyanı

Ayrıştığı Noktalar (1/2)

Ölçüt	Pekiştirmeli Öğrenme	Makine Öğrenmesi	Optimizasyon
Temel soru	Yalnız ödül sinyaliyle, deneme yanılma yoluyla iyi kararlar nasıl öğrenilir?	Bir program örneklerden nasıl genelleme yapar?	Bir amaç fonksiyonunu en iyi yapan değerler nasıl bulunur?
Yaklaşım	ödülle deneme yanılma	istatistiksel öğrenme	amaç ve kısıtların matematiksel modeli
Girdi	ortamla etkileşim: durum ve ödül	etiketli ya da etiketsiz örnekler	amaç fonksiyonu ve kısıtlar
Çıktı	politika: durumdan eyleme kural	tahmin yapan model	en iyi ya da ona yakın çözüm
Başarı ölçütü	toplam beklenen ödül	görülmemiş veride hata	amaç değeri, yakınsama hızı
Matematik yükü	yüksek: olasılık, dinamik programlama	yüksek: doğrusal cebir, olasılık, optimizasyon	kuramın kendisi
Veri ihtiyacı	etkileşimle üretilir	orta ve çok	gerekmez

- **Makine Öğrenmesi:** Denetimli öğrenmede her örneğin doğru yanıtı verilir; PÖ'de yalnız gecikmeli bir ödül vardır ve veri etmenin kendi eylemleriyle toplanır.
- **Optimizasyon:** Optimizasyon modeli bilinen bir problemi çözer; PÖ çevrenin modelini bilmeden etkileşerek öğrenir.

Ayrıştığı Noktalar (2/2)

Ölçüt	Pekiştirmeli Öğrenme	Çok Etmenli Sistemler	Robotik
Temel soru	Yalnız ödül sinyaliyle, deneme yanılma yoluyla iyi kararlar nasıl öğrenilir?	Birden çok özerk etmen birlikte ya da birbirine karşı nasıl karar verir?	Bir makine fiziksel dünyayı nasıl algılar, konumunu nasıl bilir ve nasıl hareket eder?
Yaklaşım	ödülle deneme yanılma	etkileşen özerk etmenler	algıla, planla, eyle döngüsü
Girdi	ortamla etkileşim: durum ve ödül	etmenler, hedefleri, iletişim kuralları	algılayıcı ölçümleri
Çıktı	politika: durumdan eyleme kural	ortak davranış, denge, anlaşma	motor komutları, yol
Başarı ölçütü	toplam beklenen ödül	toplam fayda, kararlılık	görev başarısı, güvenlik
Matematik yükü	yüksek: olasılık, dinamik programlama	orta: oyun kuramı	yüksek: dönüşümler, denetim
Veri ihtiyacı	etkileşimle üretilir	az; benzetim	fiziksel deney ve benzetim

- **Çok Etmenli Sistemler:** Çok etmenli sistemler etmenlerin eşgüdümünü ve çatışmasını inceler; bu ders çoğunlukla tek etmenin öğrenmesine odaklanır.
- **Robotik:** Robotik algılama, kinematik ve denetimi kapsar; PÖ bunlardan yalnız davranışın deneyimle öğrenilmesine katkı yapar.

UYGULAMA

Uygulama Alanları

Oyunlar
Go, satranç, Atari ve strateji oyunları.

Robot denetimi
Yürüme, kavrama ve hareket öğrenme.

Kaynak yönetimi
Veri merkezi soğutma, ağ trafiği yönlendirme.

Öneri sistemleri
Uzun vadeli kullanıcı memnuniyetini gözetmek.

Trafik sinyalleri
Kavşaklarda bekleme süresini azaltmak.

Dil modeli ince ayarı
İnsan geri bildirimiyle pekiştirme (RLHF).

UYGULAMA

Riskler ve Etik

Ödül sömürüsü

OpenAI'ın tekne yarışı etmeni yarışı bitirmek yerine aynı puan hedeflerini vurmak için durmadan daire çizdi.

Güvensiz keşif

Deneme yanılma gerçek dünyada pahalıdır: keşif yapan bir robot ya da tedavi politikası denerken zarar verebilir.

Tekrarlanamayan sonuçlar

Aynı algoritma farklı rastgele tohumlarla çok farklı sonuç verir; az tohumla raporlanan üstünlük yanıltıcı olabilir.

Kullanıcıyı şekillendirmek

Tıklamayı en büyük yapan öneri politikası kullanıcıyı daha uç içeriğe ve bağımlılığa itebilir.

Ayrıntılı tartışma: Açıklanabilir Yapay Zekâ ve Yapay Zekâ Etiği.

UYGULAMA

Araçlar ve Örnek Veriler

Araçlar ve kütüphaneler

Gymnasium | Stable-Baselines3 | PyTorch | RLlib | MuJoCo | PettingZoo

Klasik ortamlar

Izgara dünyası (gridworld) | Çok kollu haydut | CartPole | MountainCar | Atari oyunları (ALE) | MuJoCo denetim görevleri

PÖ'de hazır veri seti yerine **ortam** verilir: etmen veriyi kendi eylemleriyle üretir, bu yüzden simülasyon dersin laboratuvarıdır.

UYGULAMA

Sık Karıştırılanlar

Yaygın düşünce	Doğrusu
PÖ denetimli öğrenmenin başka bir adıdır.	Doğru eylem hiç söylenmez; etmen yalnız ödülden ve kendi topladığı deneyimden öğrenir.
Ödülü yazmak kolaydır.	Kötü tasarlanmış ödülü etmen beklenmedik kestirmelerle sömürür (reward hacking).
Simülasyonda öğrenen robot gerçekte de çalışır.	Simülasyon ile gerçek dünya arasındaki fark (sim-to-real) ayrıca ele alınmalıdır.
PÖ yalnız oyunlar içindir.	Denetim, kaynak yönetimi ve dil modellerinin insan geri bildirimiyle eğitimi de PÖ kullanır.

SONRASI

Bu Dersten Sonra

Bu dersi temel alan ya da genişleten dersler

Bu ders çoğunlukla uç noktadır; öteki dersleri temel alır.

İlgili meslekler

Pekiştirmeli öğrenme mühendisi

Robotik yazılım mühendisi

Oyun YZ geliştiricisi

Araştırmacı

Pekiştirmeli öğrenmeden sonra birden çok etmenin birlikte öğrenmesi için çok etmenli sistemlere, fiziksel dünyada öğrenen makineler için robotiğe geçilebilir. Derin öğrenme bilgisi derin PÖ'nün, optimizasyon bilgisi ise politika gradyanı yöntemlerinin anlaşılmasını kolaylaştırır.

SONRASI

Kaynaklar

- Sutton ve Barto, *Reinforcement Learning: An Introduction*, 2. baskı, MIT Press, 2018.
- Szepesvári, *Algorithms for Reinforcement Learning*, Morgan & Claypool, 2010.
- Bertsekas, *Reinforcement Learning and Optimal Control*, Athena Scientific, 2019.
- Mnih ve ark., "Human-level control through deep reinforcement learning", *Nature* 518, 2015.

Son

Bölüm 7: Pekiştirmeli Öğrenme

← Önceki bölüm · Alanın haritası · Sonraki bölüm →