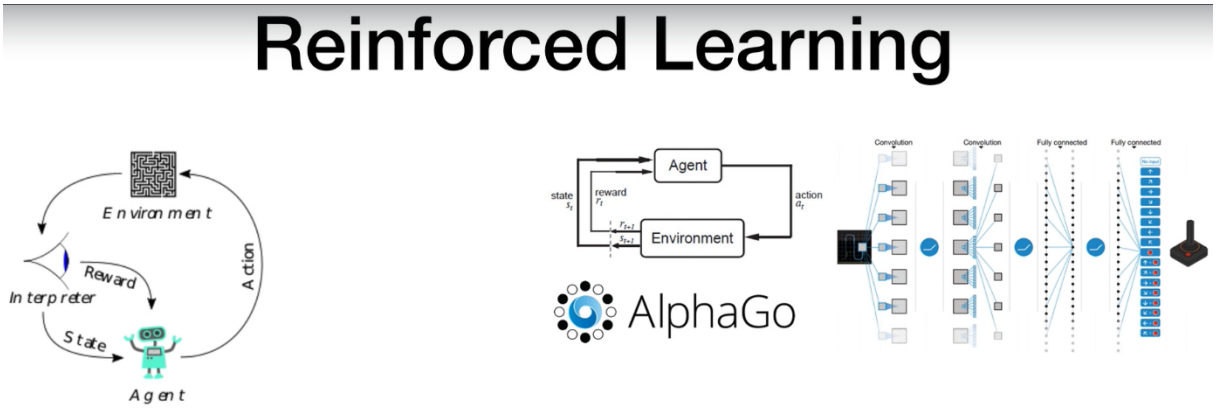
REINFORCED LEARNING (PEKİŞTİRMELİ/TAKVİYELİ ÖĞRENME)

Algoritma labeled training set'e bakarak kendini eğitmek yerine, rasgele eylemler deniyor ve ödül ve ceza kavramları ile kendi kendini eğitiyor.

Bir robotun yürümeyi öğrenmesi gibi. Makine kendi hatalarından da öğreniyor.

Temelde şöyle çalışıyor bir Agent yazıyoruz ve bu agent ortam içinde bir aksiyon yapıyor. Bu aksiyon sonunda gözlem yapılıyor ve bu aksiyonun ne kadar iyi veya kötü olduğu analiz ediliyor ve iyise aksiyon ödüllendiriliyor kötüyse cezalandırılıyor. Buna göre agent kendi kendine öğreniyor.



Alpha Go oyunu için de önce dünyada bilinen Go yazılımlarını yüklediler yani Go kurallarını bilerek agent aksiyonlara başladı. Daha sonra Go yazılımı kendi kendine maçlar yapmaya başladı. Ve sonuçta artık tüm insanlardan daha iyi bir hale geldi.

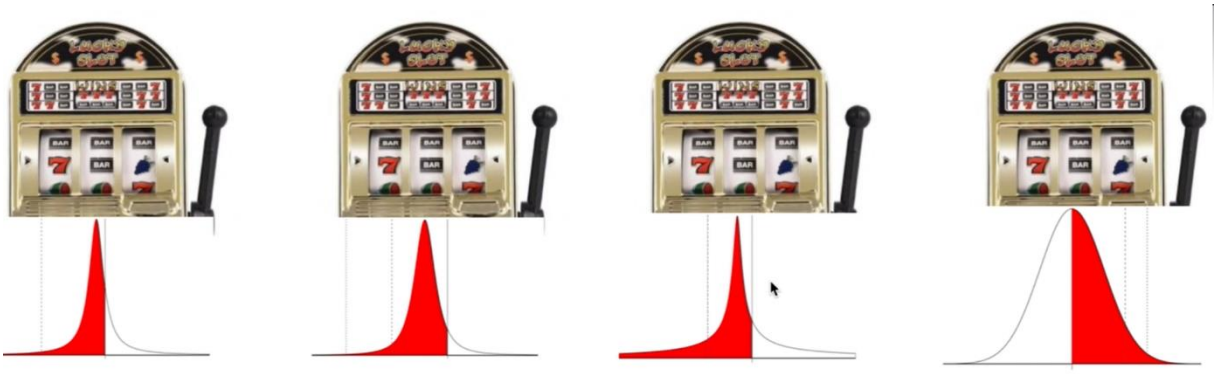
ONE ARMED BANDIT

Bu aslında bildiğimiz slot makinesi.

Mesela bir kumarhaneye gittik ve bir sürü slot makinesi var. Nasıl oynamalıyız ki para kazanma ihtimalimiz artsın?

Biliyoruz ki bu makine aldığı paranın sadece bir kısmını geri veriyor mesela %80 ini geri verirken kalan %20 yi cebine atıyor.

Bu makinelerin her birinin farklı dağılımları var:



Biz de gözlem yapıyoruz ve dağılımlarına bakıyoruz. Sonuçta da bize hangi kombinasyon en fazla kazandırır ona bakıyoruz. Gözlem için her makine de oyunlar oynuyoruz. Sonuçta hangisiyle oynamamız gerektiğini seçiyoruz. One armed bandir teorisi buna dayanıyor.

A/B TEST

Benzer şekilde A/B test kavramından bahsedebiliriz. İnternette bir çok reklam veriliyor hangisi daha başarılıdır? Ya da bir sitenin kullanıcı arayüzü değişti hangisi daha başarılıdır?

A/B test şuna dayanıyor birden fazla reklamı kullanıcalara gösteriyoruz ve kullanıcıların tepkisine (tıklamasına) göre hangi reklamın daha iyi olduğunu seçer.

Yani hangi reklamı göstermesi gerektiğine karar vermeye çalışan Agent bir aksiyon yapıyor ve reklamları gösteriyor ortamdan gelen tepkiye yani tıklamalar göre agent bir sonraki adımda hangi reklamı göstereceğine karar vermeye çalışıyor. Aynı sayıda kişiye A B C D gibi değişik reklamları gösterip geri dönüşlere göre hangi reklamın daha iyi olduğuna karar veririz.

Sonuçta reinforced learning çok farklı yerlerde kullanılabilir.

UCB (UPPER CONFIDENCE BOUND)

Reinforced learning algoritmalarından ilki olan upper confidence bound algoritmasına bakacağız.

Algoritmanın kabulü her oyunun arkasında istatiksel bir dağılım olduğu. Slot makinelerinin de bir dağılımı var. Reklam tıklamalarında da aynı şekilde.

UCB'nin dayandığı bantık şu:

- ➔ Kullanıcı her seferinde bir eylem yapıyor. (event e)
- ➔ Bu eylem karşılığında bir skor döner (örneğin web tıklaması 1 tıklamaması 0)
- ➔ Amaç tıklamaları maximuma çıkarmak.

Agent reklamı yayınlıyor, reklam tıklandıysa ödül 1 geliyor tıklanmadıysa ceza 0 geliyor veya agent ona göre davranışlarını şekillendiriyor.

Diyelim ki elimizde 4 reklam var:

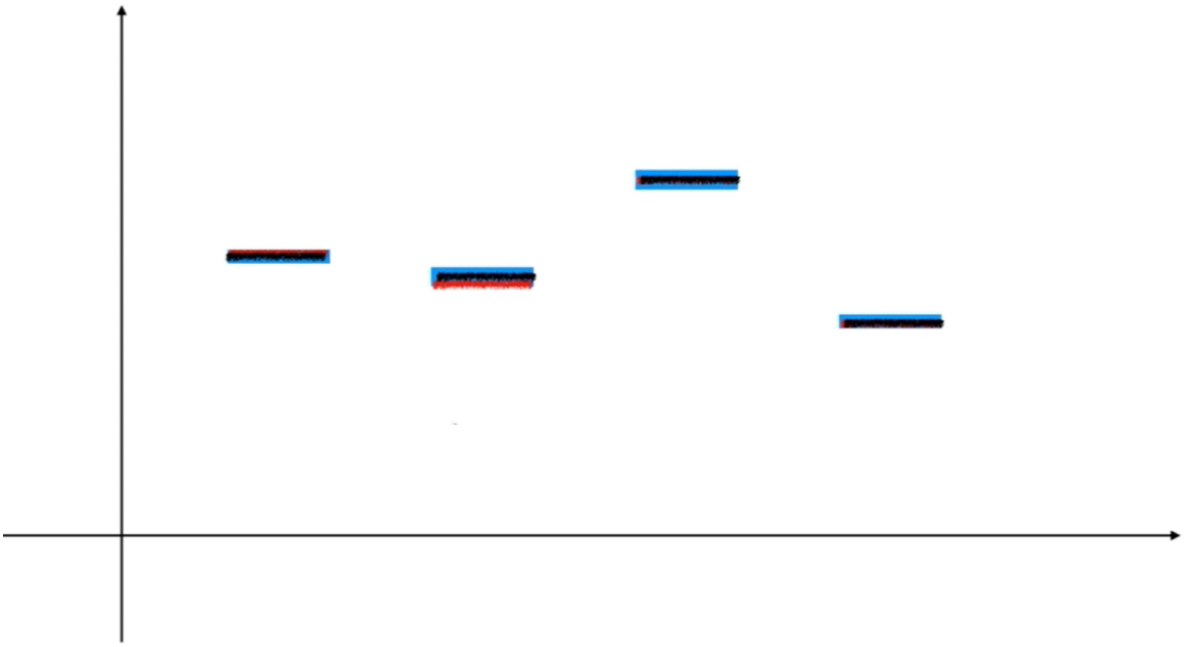
Kırmızı ile gösterilen yerler 4 reklamın da tıklanma potansiyellerini gösteriyor diyelim. Mavi ile gösterilen de confidence interval yani min ve max tıklanma ihtimalleri.

Slot makinesi olarak düşünürsek kazanacağımız en yüksek ve en düşük parayı gösteriyor. Yani en üstü jackpot'u temsil ederken en alt kaybetme olayını temsil ediyor.

Şimdi bizim agent çeşitli action'lar gerçekleştiriyor mesela reklam A yı gösteriyor ve bir tıklanma oluyor ve bu tıklanma rakamı beklentinin üzerinde siyah ile gösteriliyor.

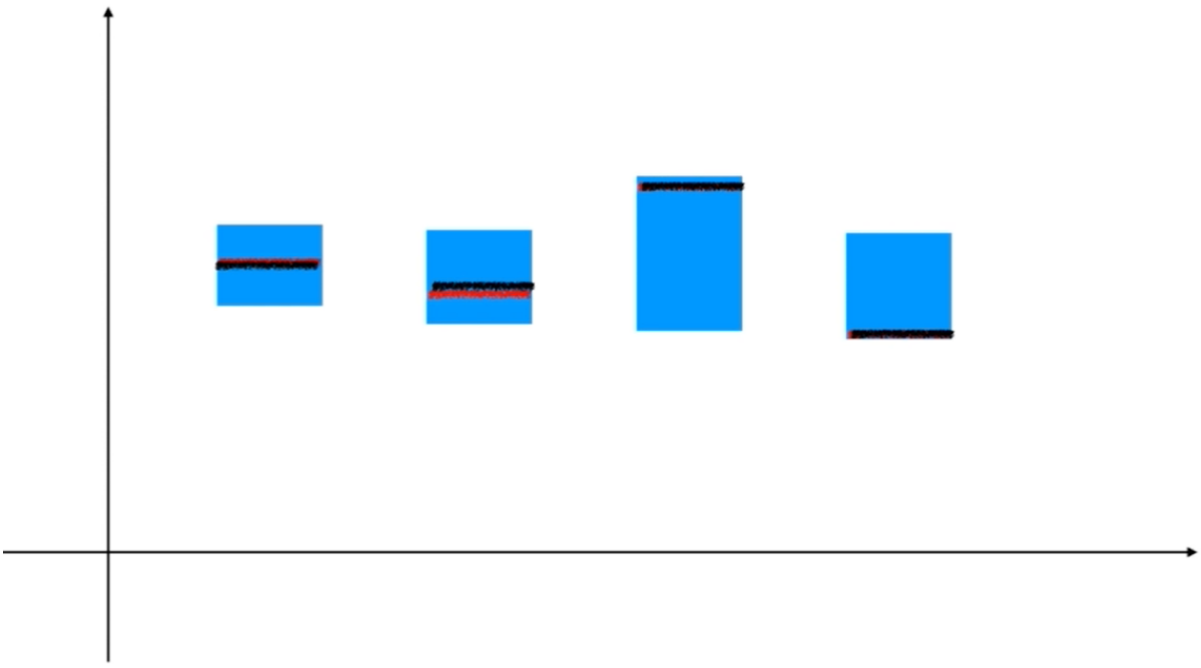
Veya robot ayakta durmaya çalışıyor 1. Joint'e bir tork uyguladı ve bu tork sonunda aldığı ödül siyahın seviyesinde, maximum ödülü değil de daha düşük bir ödül aldı.

Sonuçta oynadıkça bu intervallar küçülüyor ve sonunda bu kırmızı ile gösterilen gizli dağılım durumunu keşfetmiş oluyoruz.



Mesela 4 reklamdan hangisi gösterilsin 3. Olanı gösterelim çünkü bunun geri dönüşü en fazla. Çok fazla oynayınca buna karar verebiliriz dağılımı keşfetmiş oluyoruz.

Ama marifet çok fazla oynamadan buna karar verebilmek.



Mesela her slotla 4-5 kez oynadık ve böyle bir confidence graph elde ettik. En yüksek upper confidence olanı seçip bununla oynamamız daha mantıklı olur.

TAM OLARAK ANLAMADIM AMA KODLAMA KISMINDA DAHA ANLAŞILIR OLACAK DİYOR BAKALIM ORADA AÇIKLAMALARA DEVAM EDEBİLİRİM.