Reinforcement learning - In detail

Mustafa Muhammad

27th October 2021

Two opposing forces

1. Exploration

2. Exploitation

Explore - Exploit delemma

Applications of explore exploit dilemma

The nature of probability

We would need an infinite number of samples to get an absolutely precise estimate.

As we collect more samples, the confidence interval shrinks.

Therefore, the more samples we collect the better.

If we talk about click through rate for advertisements

If one advertisement is better, that means the other is worse.

If I show the sub optimal add 1 million times, I've wasted 1 million impressions to get a sub optimal CTR.

My desire to show only the best advertisement (and hence make more money is fundamentally at odds with my desire to have an accurate click through rate)

The CTR is needed to show which advertisement is best in the first place.

Epsilon greedy theory.

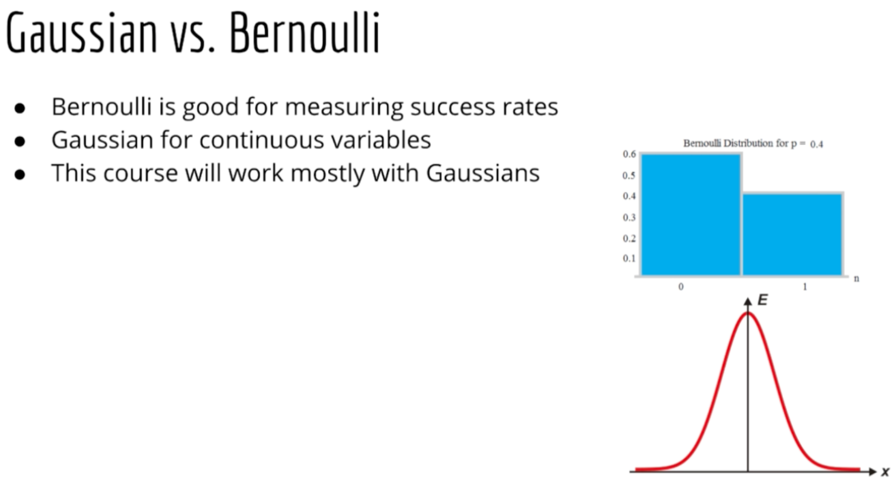
Need to balance explore/exploit

Observe: Choosing best maximum likelyhood winrate doesn't work

We call this greedy

For us: greedy means picking the bandit with the highest MLE win rate, with no regard to confidence in prediction or amount of data collected.

The name "epsilon-greedy" makes it obvious that we're going to modify the greedy strategy in some way.



Gaussian vs. Bernoulli

- Bernoulli is good for measuring success rates
- Gaussian for continuous variables
- This course will work mostly with Gaussians

Idea: have a small probability of doing something random (non-greedy).

That small probability is given by eplision.

Typical values: 5%, 10%.

```
#Greedy
while True:
    j = argmax(predicted bandit means)
    x = play bandit j and get reward
    bandits[j].update_mean(x)

#Greedy-Epsilon
while True:
    p = random number in [0, 1]
    if p < epsilon:
        j = choose a random bandit
    else:
        j = argmax(predicted bandit means)
    x = play bandit j and get reward
    banjits[j].update_mean(x)
```

Additional details

The purpose of exploration(non zero epsilon) is so we can collect data about each bandit.

$E(R) = (1-\epsilon)0.9 + \epsilon\frac{(0.8+0.9)}{2}$