UCB Formulation: Establishes Some Confidence

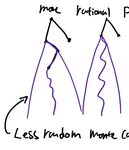
2024年5月20日 星期一 17:01

More expansion, more correct

Essentially Putting the tree to 2 parts
1. rational tree growing
2. Data that help to understand the quality better

↑

make you grow the tree, less random more Carlo Sample
make rational part on the top



→ Enough coverage, Correct values
(when growing, distribution is dynamic)

Less random monte carlo sampling

What is the optimal in 2 degree of stochastic

$\mu_1 = 0.8$, $\mu_2 = 0.2$] True
$\mathbb{E}[X_1]$
$\mathbb{E}[X_2]$

X_1 { 1 0.8 ; 0 0.2 } r.v.

X_2 { 1 0.2 ; 0 0.8 } r.v.

If you know this, $\mu_1(X,1)$ is better
but when don't know, essentially same with Q Learning

option → same Probabilistic Distribution
→ each time only see 1 trajectory of such distribution

both Stochastic over come
what to actually Choose, what is a mathematical strategy

In ε-greedy or 50%/50%/50% Choose : There would be certain % of operation that is wrong, in the long run
Essentially the same in expectation

"Regret"
Mathematical Construct

$\mathbb{E}[\sum_{i=1}^{T} X_{C_i}]$

Variable on coin also random 2
value of coin is random

Coin $C_i \in (C_{blue}, C_{red})$ random 1
Choosing Coin is random

2 Stochastic flavour baked into it

Optimalistically :
Always Play the better Coin μ^*
Best value : $\mu^* \cdot T$
Lowest value : $\mu^{low} \cdot T$

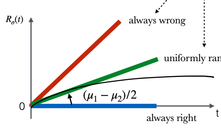
Regret is the gap between getting and optimal

$$R_\theta(T) = \mathbb{E}\left[\sum_{i=0}^{t} X_{C^*} - \sum_{i=0}^{t} X_{C_i}\right]$$

some one C^*, some one not
zero regret when $\sum X_{C_i} = \sum X_{C^*}$ → $(\mu^* - \mu) = \Delta$
Make one given Coin value not random

50% 50% Policy :
$\frac{1}{2}T(\mu^* - \mu)$

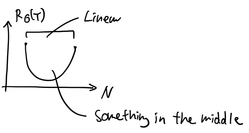
"Linear Regret" ← All naive Approach is linear regret (No learning, look at the scene with fresh eye)
Come up with something not linear?
—Non learning approach boils to the same idea



Monotonic increase, Count decrease
but Can regret less and less, which is the definition of
what learning actually is

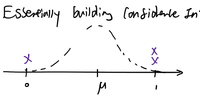
$\mathbb{E}[R(N)] \sim O(N^{2/3}(K \log N)^{1/3})$ better than linear a bit for ε greedy

$\mathbb{E}[R(N)] \sim O(\log N)$ for UCB (The best)



Give it a Confidence

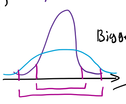
Essentially building Confidence Interval, Statistics is useful



$\bar{X} = \frac{1}{n}\sum_{i=1}^{k} X_i$, as $k \to \infty$, $\bar{x} \to \mu$ by CLT and LLN

Compare distributions with Concentration Inequality and Pin Point expectation!
(Shekychev inequality is a Concentration inequality)

Assuming the sample is somewhat
Consistent with statistical beliefs



Bigger Tail when not Concentrated

Give ε range and ask
how likely to have data point there

$$\mathbb{P}(|\bar{x} - \mu| > \varepsilon) \le 2e^{-2N\varepsilon^2}$$

Probability of deviation from ε neighborhood

Sample

Tolerance for deviation from the expectation

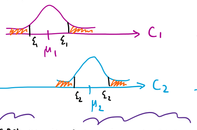
Big ε : Not Concentrated, less convergence
Small ε : Concentrated, more Convergence, more sure about actual value

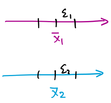
1. $\varepsilon \uparrow$, $P \downarrow$ (greater confidence, less accuracy to match μ)
2. $N \uparrow$, Tail decrease exponentially fast
→ More Concentrated
→ Less ε needed for same accuracy of knowing μ

Adaptive ε Choice : Banisit of Davent

Assume that we set a starting Point where set ε as $\sqrt{\frac{2\log T}{N_i}}$ → $2e^{-2N_i\frac{2\log T}{N_i}} = 2e^{-4\log T} \sim O(\frac{1}{T^4})$

Adaptive ε_i bound the tail in the same way





$\bar{x}_1 + \varepsilon_1$
$\bar{x}_2 + \varepsilon_2$

Optimistically :
Choose with Sample + bound
accordingly who have higher value

Quickly decrease when is outside of ε
As T (total sample) ↑, Probability of deviation shrinks really quickly

Optimism in face of uncertainty : Since μ is in the range, and we want to know the true μ to choose the better Coin, the best true can be do for μ_i is $\bar{x}_i \in \varepsilon_i$ (banisit of davent)

# of choosing less optimal coin (Accumulation of Regret)

Assume C_1 bad and C_2 good and the sample happen to say C_1 is good

$\bar{x}_1 + \varepsilon_1 \ge \bar{x}_2 + \varepsilon_2$

$\bar{x}_1 \in [\mu_1 - \varepsilon_1, \mu_1 + \varepsilon_1]$
$\bar{x}_2 \in [\mu_2 - \varepsilon_2, \mu_2 + \varepsilon_2]$

Assume true our sample is within true value + bounds
Assume Suboptimal highest and optimal lower

$\mu_1 + \varepsilon_1 \ge$ Relax to upper bound
$\ge \mu_2 - \varepsilon_2$ Constrict to lower bound

With particular ε, Prob of μ not in $\bar{x} \pm \varepsilon$ or
$\bar{x}$ not in $\mu \pm \varepsilon$ decrease exponentially

$$\mu_1 + \varepsilon_1 + \varepsilon_1 \ge \bar{x}_1 + \varepsilon_1 \ge \bar{x}_2 + \varepsilon_2 \ge \mu_2 + \varepsilon_2 - \varepsilon_2$$

← Connection between sample $\bar{x}$ and Expectation μ
Sample + bounds contain μ_1, μ_2 + bounds contain sample (assume)

$= \mu_1 + 2\varepsilon_1 \ge \mu_2$

$= 2\varepsilon_1 \ge \mu_2 - \mu_1 = 2\sqrt{\frac{2\log T}{N_1}} \ge \Delta$ (diff between μ_2 and μ_1)

$$N_1 \le \frac{8\log T}{\Delta^2}$$

No way for the number of choosing less optimal coin (N_1) to be greater than log growth for the total number of coin choosen.

Exponential less Play with less optimal Coin

→ only small enough sample misguide you

Intuition 水平名出 (ε是水, μ是名)

If banisit of davent choose the bad Coin,
well, Choose it, and then the bad Constrict really quickly, small ε
and the next round the true bad Coin's high value would be been shown

Only the the appears

More ties, less advantage of ε, more Constrict,
more Confident about the value, give less davent banisit when Comparison

UCB is everywhere in decision making and Algorithm

Practical : UCT

Assume the 2 option fixed distribution : UCB Brand to be drawing from i.i.d. samples
→ As you grow the tree, the distribution Changes, favour the better one
→ Not i.i.d. any more

$\bar{X} + C\sqrt{\frac{2\log T}{N}}$ Progress driven to reduce hyperparameters, when mean is not drawn right (clear), then are more hyperparameters attached

Hence the name UCB on tree with our former Proof → UCT