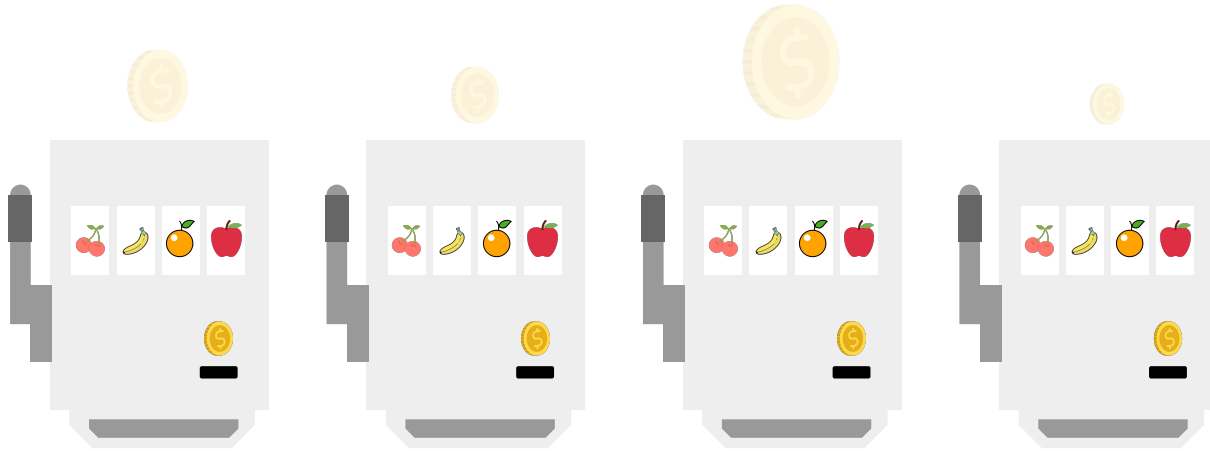


Multi-armed bandits



O que é e como pode ser útil?

Formulação do problema





Como escolher a sequência de caça-níqueis que
maximiza o lucro?

Exploration vs. Exploitation



Buscar **novas possibilidades**
em busca de algo
possivelmente melhor



Explorar as alternativas que
deram **maior ganho** no
passado

Cenários reais

- Propagandas (internet)
- Testes clínicos
- Sistemas de recomendação
- Escolhendo comida

Cenários reais

- **Propagandas (internet)**
- Testes clínicos
- Sistemas de recomendação
- Escolhendo comida



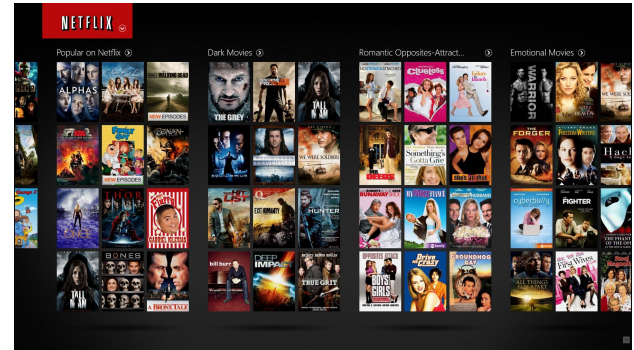
Cenários reais

- Propagandas (internet)
- **Testes clínicos**
- Sistemas de recomendação
- Escolhendo comida



Cenários reais

- Propagandas (internet)
- Testes clínicos
- **Sistemas de recomendação**
- Escolhendo comida



Cenários reais

- Propagandas (internet)
- Testes clínicos
- Sistemas de recomendação
- **Escolhendo comida**

Clássico
Nunca falha
7/10



ou



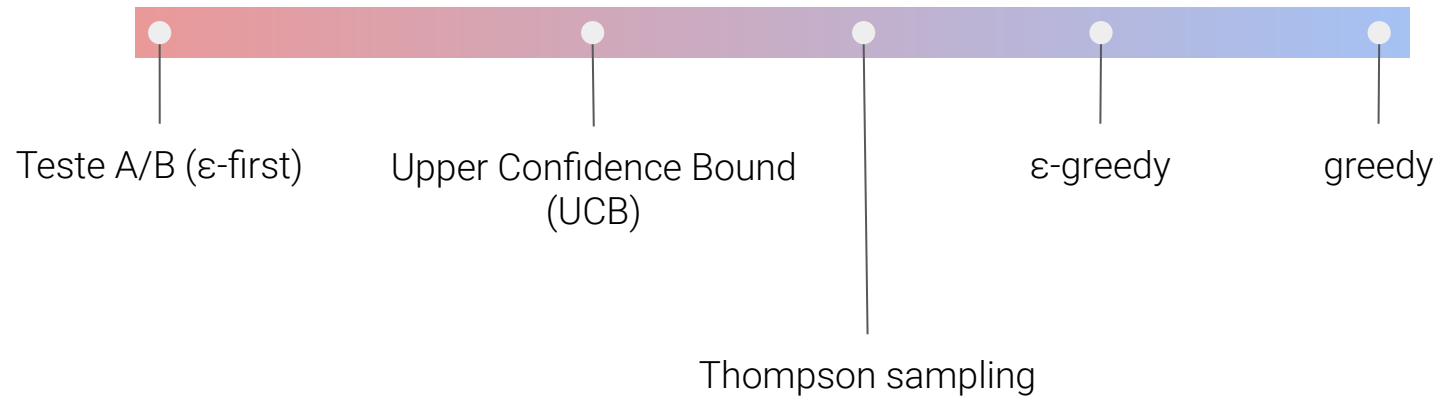
↑
Novo
Muito bom ou muito ruim
0~10/10

Como resolver esse problema?

Estratégias

Exploration

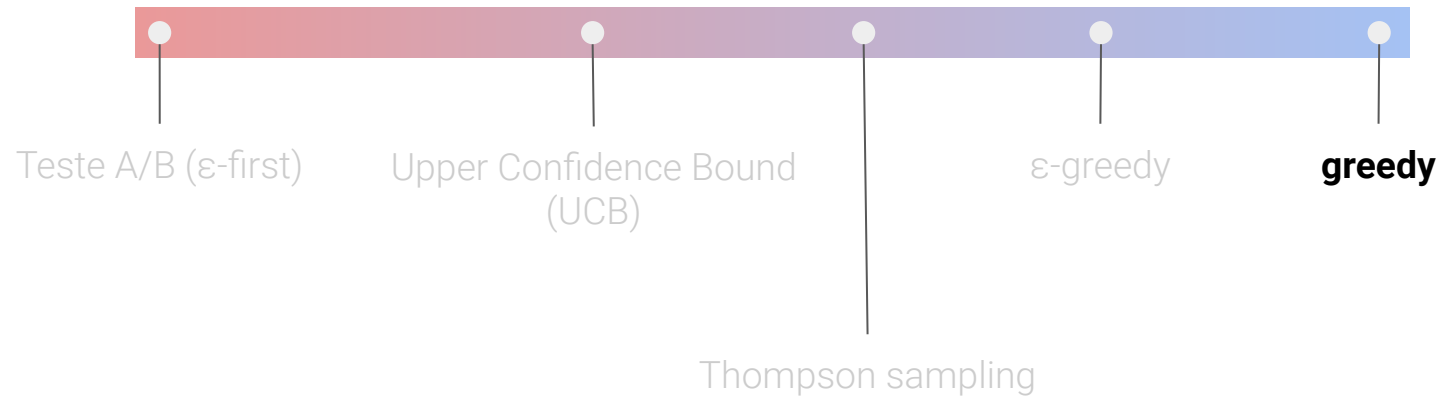
Exploitation



Estratégias

Exploration

Exploitation

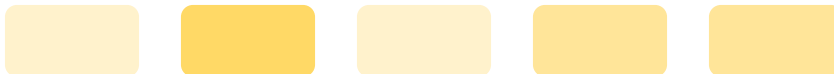


Greedy

Real (desconhecido)



Observações iniciais (2 rodadas)

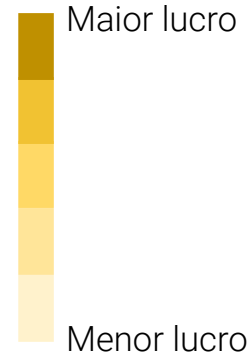


Algoritmo

A cada nova rodada, sempre escolhe o que deu mais lucro nas observações iniciais.

Problema

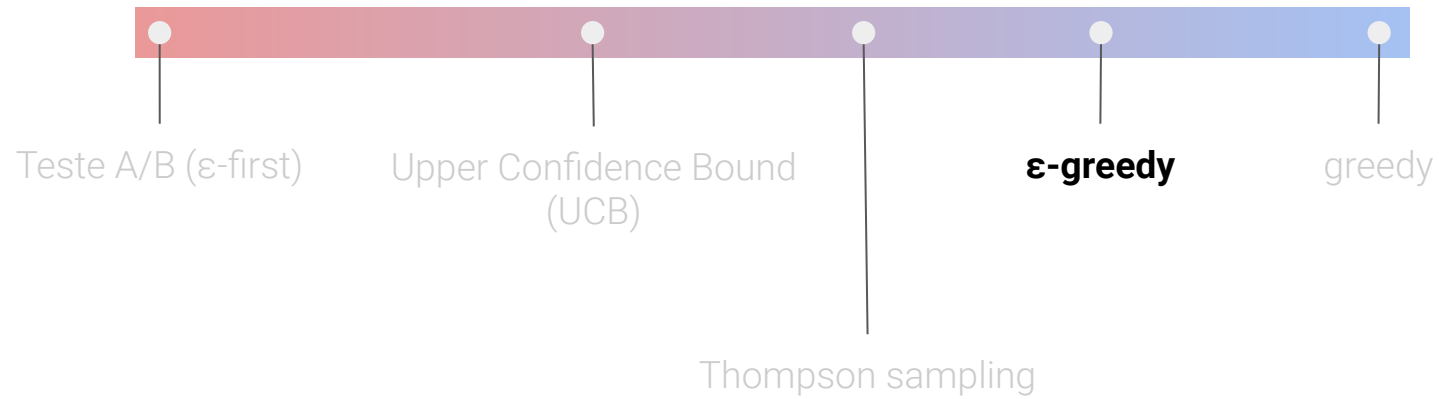
Se as observações iniciais estão enviesadas, a alternativa selecionada pode não ser a melhor.



Estratégias

Exploration

Exploitation



ϵ -Greedy

Real (desconhecido)

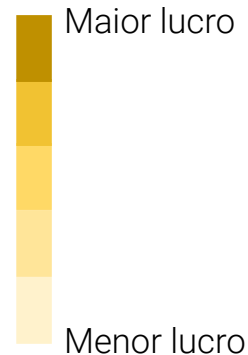
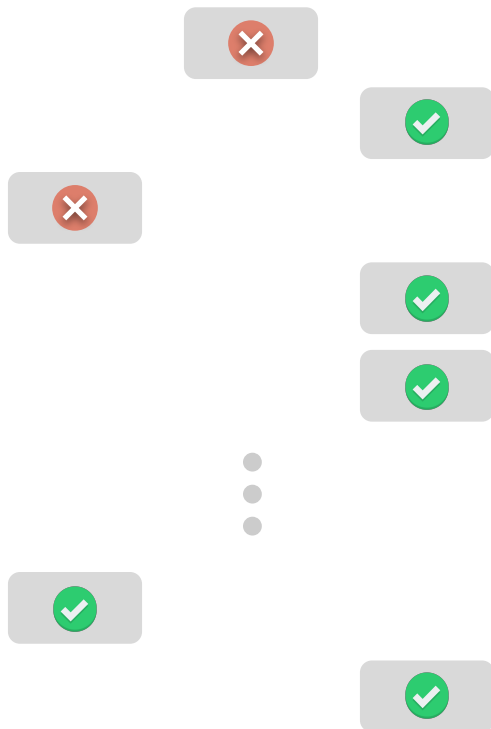


Algoritmo

A cada nova rodada, com probabilidade ϵ , escolhe uma alternativa aleatória. Do contrário, $(1 - \epsilon)$ escolhe a alternativa com maior lucro até o momento.

Problema

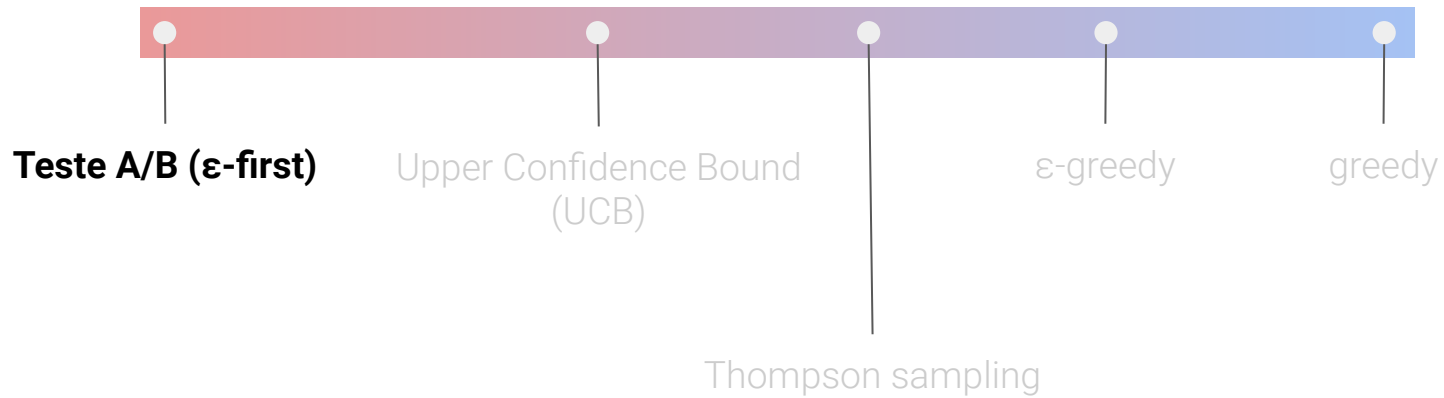
O nível de exploração depende do ϵ . Se for pequeno, a exploração será subótima. Se for grande, acabamos explorando mesmo quando há uma alternativa claramente superior



Estratégias

Exploration

Exploitation



Teste A/B (ϵ -first)

Real (desconhecido)



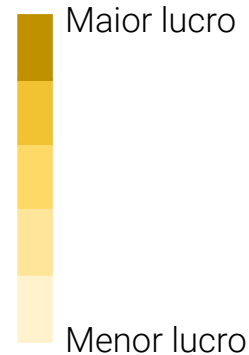
Algoritmo

Seleciona alternativas aleatoriamente durante todo o processo. Avalia o resultado final e escolhe o melhor entre todas as possibilidades.

Problema

O foco em 100% de exploração pode ser arriscado em cenários em que o custo de apostar em alternativas ruins é alto.

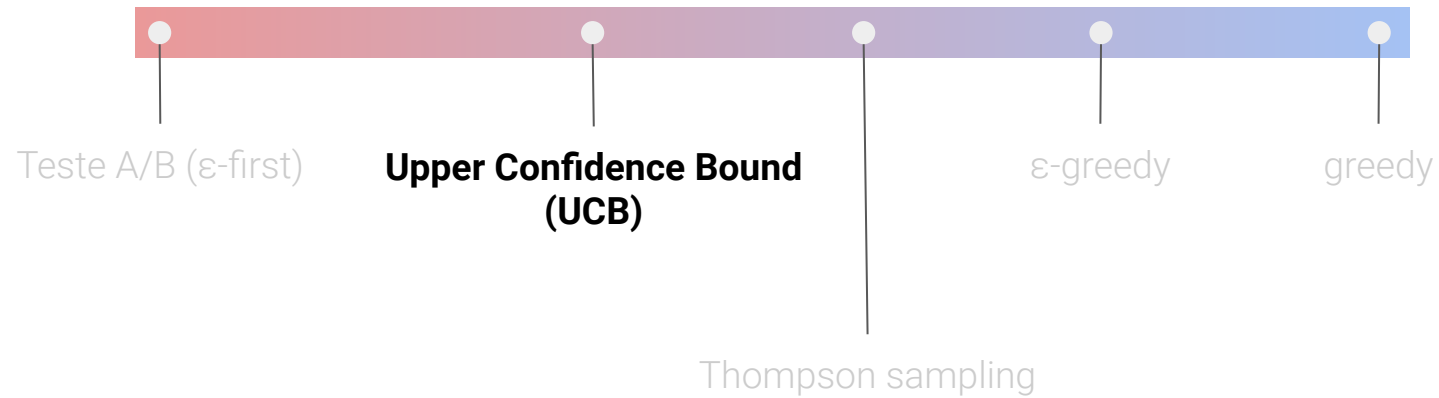
Observado final



Estratégias

Exploration

Exploitation



Upper Confidence Bound (UCB)

"Otimismo em cenário de incertezas"

Real (desconhecido)



Algoritmo

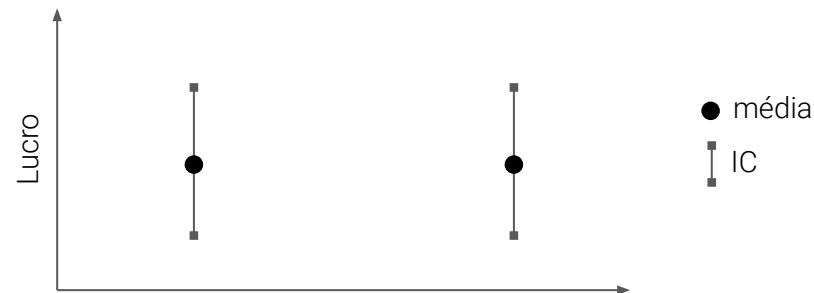
A cada rodada, seleciona a alternativa que tem o maior limite de confiança superior (UCB) dado por:

$$\bar{r}_i + \sqrt{\frac{3 \log n}{2N_i(n)}}$$

Onde:

\bar{r}_i É o lucro médio da alternativa i até o momento

$N_i(n)$ É a quantidade de vezes que a alternativa i foi selecionada até então



Upper Confidence Bound (UCB)

"Otimismo em cenário de incertezas"

Real (desconhecido)



Algoritmo

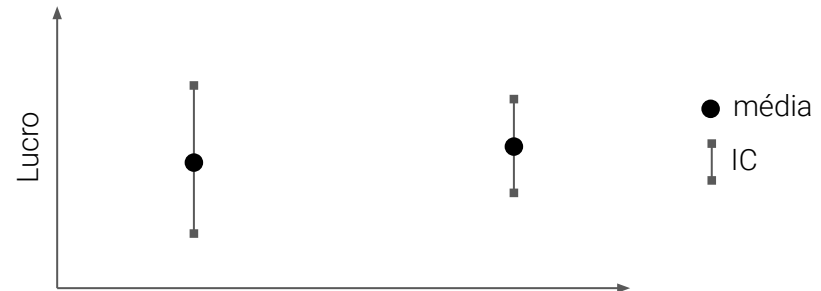
A cada rodada, seleciona a alternativa que tem o maior limite de confiança superior (UCB) dado por:

$$\bar{r}_i + \sqrt{\frac{3 \log n}{2N_i(n)}}$$

Onde:

\bar{r}_i É o lucro médio da alternativa i até o momento

$N_i(n)$ É a quantidade de vezes que a alternativa i foi selecionada até então



Upper Confidence Bound (UCB)

"Otimismo em cenário de incertezas"

Algoritmo

A cada rodada, seleciona a alternativa que tem o maior limite de confiança superior (UCB) dado por:

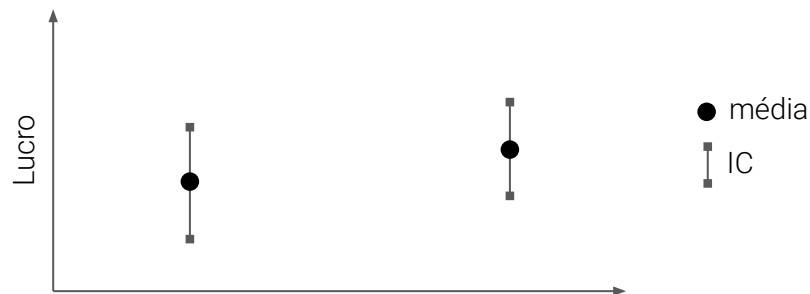
$$\bar{r}_i + \sqrt{\frac{3 \log n}{2N_i(n)}}$$

Onde:

\bar{r}_i É o lucro médio da alternativa i até o momento

$N_i(n)$ É a quantidade de vezes que a alternativa i foi selecionada até então

Real (desconhecido)



Upper Confidence Bound (UCB)

"Otimismo em cenário de incertezas"

Algoritmo

A cada rodada, seleciona a alternativa que tem o maior limite de confiança superior (UCB) dado por:

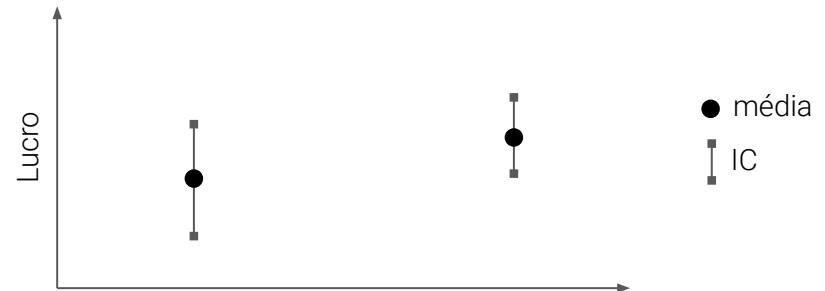
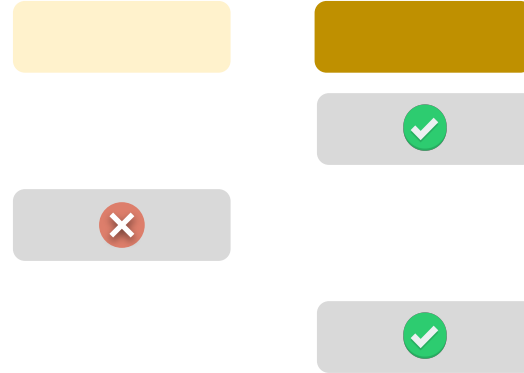
$$\bar{r}_i + \sqrt{\frac{3 \log n}{2N_i(n)}}$$

Onde:

\bar{r}_i É o lucro médio da alternativa i até o momento

$N_i(n)$ É a quantidade de vezes que a alternativa i foi selecionada até então

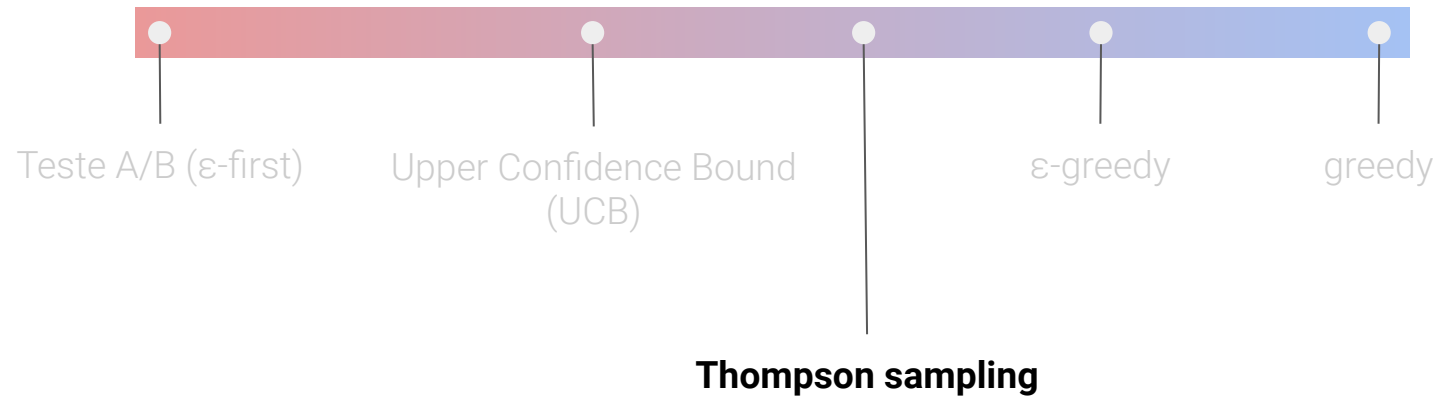
Real (desconhecido)



Estratégias

Exploration

Exploitation



Thompson Sampling

Real (desconhecido)



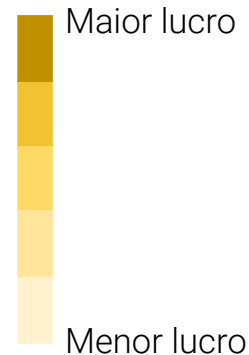
Distribuições a priori

$Beta(\alpha = 1, \beta = 1)$

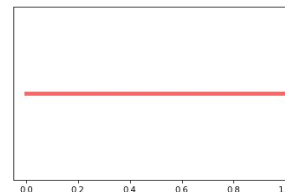
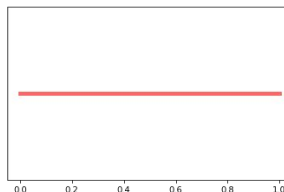
$Beta(\alpha = 1, \beta = 1)$

Algoritmo

A cada rodada, amostra cada distribuição a priori e seleciona a alternativa com o maior valor. Utiliza o ganho/perda da rodada para atualizar a distribuição a posteriori.



Distribuições a posteriori



Thompson Sampling

Real (desconhecido)



Distribuições a priori

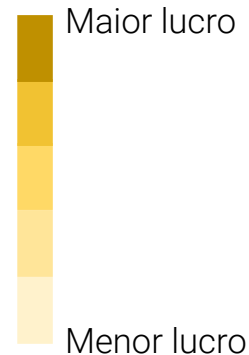
$Beta(\alpha = 1, \beta = 1)$

$Beta(\alpha = 1, \beta = 1)$

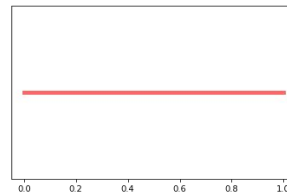
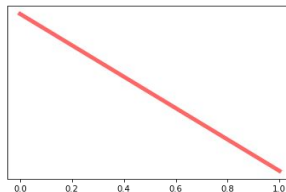


Algoritmo

A cada rodada, amostra cada distribuição a priori e seleciona a alternativa com o maior valor. Utiliza o ganho/perda da rodada para atualizar a distribuição a posteriori.



Distribuições a posteriori



Thompson Sampling

Real (desconhecido)



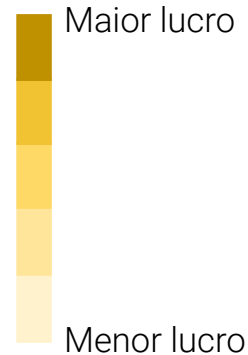
Distribuições a priori

$Beta(\alpha = 1, \beta = 1)$

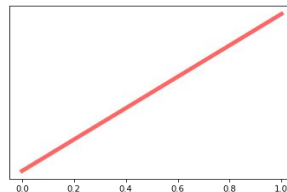
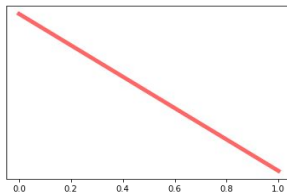
$Beta(\alpha = 1, \beta = 1)$

Algoritmo

A cada rodada, amostra cada distribuição a priori e seleciona a alternativa com o maior valor. Utiliza o ganho/perda da rodada para atualizar a distribuição a posteriori.

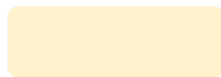


Distribuições a posteriori



Thompson Sampling

Real (desconhecido)



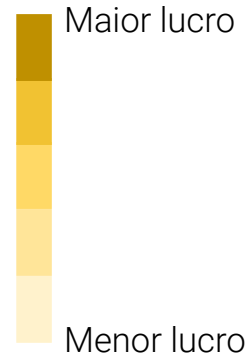
Distribuições a priori

$Beta(\alpha = 1, \beta = 1)$

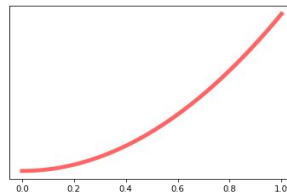
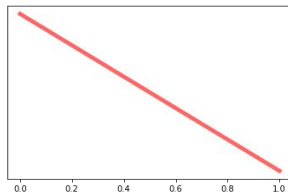
$Beta(\alpha = 1, \beta = 1)$

Algoritmo

A cada rodada, amostra cada distribuição a priori e seleciona a alternativa com o maior valor. Utiliza o ganho/perda da rodada para atualizar a distribuição a posteriori.



Distribuições a posteriori



Thompson Sampling

Real (desconhecido)



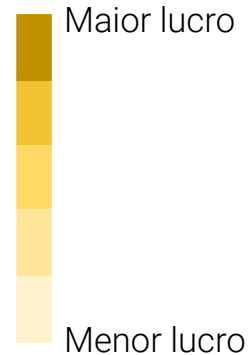
Distribuições a priori

$Beta(\alpha = 1, \beta = 1)$

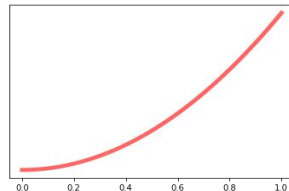
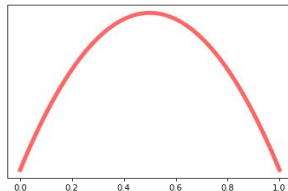
$Beta(\alpha = 1, \beta = 1)$

Algoritmo

A cada rodada, amostra cada distribuição a priori e seleciona a alternativa com o maior valor. Utiliza o ganho/perda da rodada para atualizar a distribuição a posteriori.



Distribuições a posteriori



Thompson Sampling

Real (desconhecido)



Distribuições a priori

$Beta(\alpha = 1, \beta = 1)$

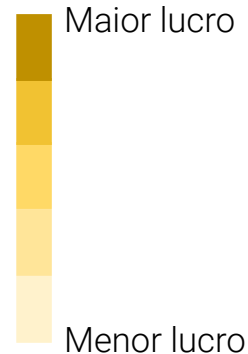
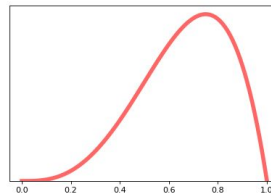
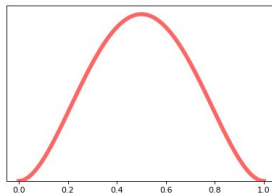
$Beta(\alpha = 1, \beta = 1)$

Algoritmo

A cada rodada, amostra cada distribuição a priori e seleciona a alternativa com o maior valor. Utiliza o ganho/perda da rodada para atualizar a distribuição a posteriori.



Distribuições a posteriori

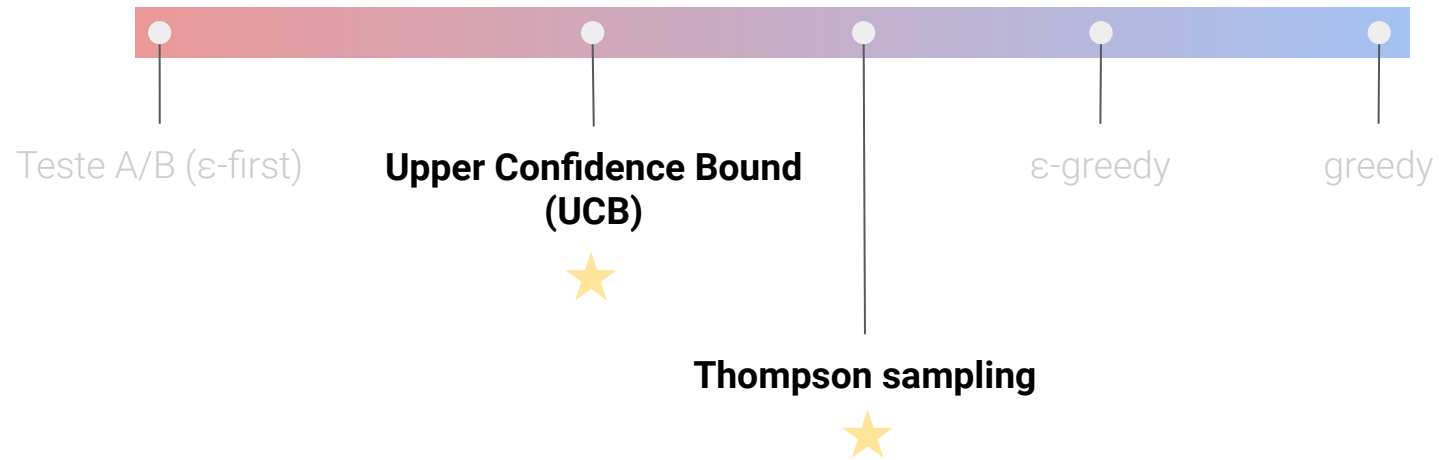


Estratégias

Conclusão

Exploration

Exploitation





WE NEED TO GO DEEPER

Variantes

- Contextual bandits
- Non-stochastic bandits
- Interleaving
- Provavelmente muitas outras possibilidades

