Multi-armed bandits

Part 2

Probability to be the best

With Epsilon-Greedy, the probability to be selected for a certain advertisement depends on its rank:

- $(1 - \varepsilon) + \varepsilon / K$ if rank is first
- ε / K otherwise

where is the number of competing ads.

Relative qualities are not relevant: the winner takes it all (almost all), be it winner with a minuscule or huge edge, it does not matter.

Clearly, it is too raw to be an optimal criterion.

With Softmax, we do not use ranking, but scoring.

Each ad receives a certain score, the observed CTR or something else usable as quality estimation.

The score is transformed in probability to be selected,

Ranks are preserved: a better ad gets bigger probability to be selected than a worse one.

Moreover, "distances" in quality are preserved and intensified, depending on the Temperature parameter.

More sophisticated than Epsilon-Greedy.

Though, artificial: the temperature is not easy to interpret and tune.

A different approach is Thompson Sampling.

The idea is very appealing for intuition:

The probability to be selected

is equal to

the probability to be the best one.

If ad A is twice more likely to be better than B (depending on their histories) then A get twice greater probability to be selected than B.

It is strongly convincing.

Indeed, theorems exist that assert Thompson Sampling is really a very good method.

It *converges* to optimal solutions, in the long run.

I.e. it progressively focus on the best ads and eventually on the best one only.

It is *self-adaptive*: new observations progressively change ads' probabilities to be selected in the right way.

Indeed, it is learning inside a *Bayesian* framework.

It does *not* require parameters like exploration rate in Epsilon-Greedy or temperature in Softmax.

It is very *easy to use and interpret*, provided that you have *enough computational resources*.

Thompson Sampling

Let us explain the algorithm with an example.

We have two ads with this history:

Imps	Clicks		Imps	Clicks
100	3		200	3

Clearly, A is more credible as better than B.

The point is: how much more credible?

More precisely: which is the probability that A has greater CTR than B?

We can compute

Prob(A is better) and Prob(B is better)

We can build two Beta distributions:

Beta(4, 98) for A

Beta(4, 198) for B

At the next round, we sample a random number x from the Beta associated to A and another random number y from the Beta associated to B.

If $x > y$ then we select A, otherwise we select B.

Then we update the history for the selected ad, recording the impression and the click, if it happens.

Then we update the history for the selected ad, recording the impression and the click, if it happens.

E.g., if we select A and it does not get a click, a t the next step the Beta associated to A is Beta(4, 99). If it gets a click, the Beta becomes Beta(5, 98).

[Remember: the parameters are #hits + 1 and #failures + 1]

We repeat this procedure at each round.

The key point is that the probabilities of selecting A or B are the probabilities that A or B is the best ad.