MS&E 235: Lecture 5
Prof. Ashish Goel
Apr 14, 2009
Notes by Rio Goodman

Lecture 5: Uncertainty

Uncertainty and risk are important to understand in any industry, but due to the vast amounts of data and automated decision making, it is particularly important in Internet Commerce applications. This lecture will introduce uncertainty through the mutli-armed bandit problem, and provide a decision technique using the Gittins Index.

Two-Armed Bandit

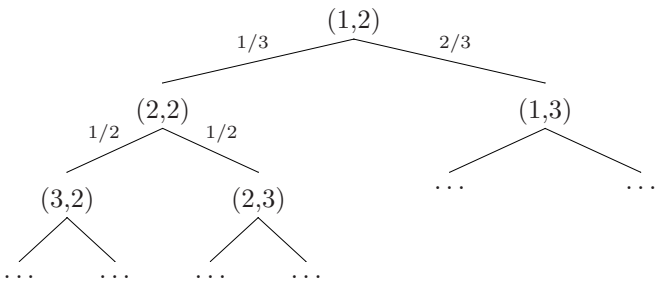
Imagine that you live in Las Vegas and every day, for n days, you go to a casino and put \$1 in a slot machine. There are two arms on this slot machine, and you have some historical data on both. This slot machine has only 2 possible rewards: \$0 (failure) and \$1 (success).
The historical data is represented: (number of successes, number of failures)

Arm 1: $(1, 2)$
Arm 2: $(10, 10)$

We will assume that a machine with no historical data starts with $(1, 1)$, and refer to the number of successes experienced with arm i as α_i, and the number of failures β_i. So for the the above two arms, we have:

$\alpha_1 = 1, \beta_1 = 2$
$\alpha_2 = 10, \beta_2 = 10$

Arm 2 has a lot of historical data, and we can be fairly certain that it has nearly a 50% success probability. However, Arm 1's success probability is less certain. If we go by the historical data alone, it has an expected success probability of 1/3. But if we play Arm 1, the evolution of historical data will look like the tree below, with the associated probabilities of success and failure shown along the edges:



This tree represents not just our initial belief on the success probability, but also encapsulates our belief in how this success probability will evolve. Such a tree is called a prior. If the first play is a success (the left branch), then the expected success probability immediately jumps to 50%, and we have only lost one day to experimentation. It seems reasonable if we are going to play for a year or 1000 days, to try Arm 1 for a few days to see whether our expected success probability is accurate. If Arm 1 experiences a lot of failures, then we can always switch to Arm 2, where we are nearly guaranteed a 50% success probability. If the first arm evolves to a point where its success probability is more than half, then we have lucked out. In other words, we believe that our success probability for Arm 1 is 1/3, but we do not have much confidence in our belief.

By creating this tree, we are implicitly planning out all the future scenarios.

Multi-Armed Bandit

In order to define this mathematically, we must introduce some notation. An (α, β) prior corresponds to a success probability of $\alpha/(\alpha+\beta)$ in the current step, and the arm becomes an $(\alpha+1, \beta)$ in the event of a success, and an $(\alpha, \beta+1)$ arm otherwise.

We also need a discount factor, θ, which can be considered to be the "present value of tomorrow's reward." If you are going to earn \$1 tomorrow, it is worth $\$\theta$ to you today. If you are going to earn \$1 the day after tomorrow, it is worth $\$\theta^2$ to you today. Another way to think of θ is as $\theta = 1 - r$, where r is the interest rate. As a rule of thumb, $\theta \approx 1 - \frac{1}{T}$, where T is the number of periods for which we are planning. For example, if we are considering 10 periods, $\theta \approx 0.9$. If we are considering 10000 periods, $\theta \approx .9999$.

Interesting fact: A typical investor is around 45 years old, and the life expectancy is around 75. So a typical investor plans for 30 years. It is believed by many economists that this is why the interest rate is usually around 3%.

Now, given N arms:
(α_1, β_1)
(α_2, β_2)
. . .
(α_N, β_N)

Discount factor $= \theta$

Which arm should we play?

The total expected discounted reward is the expected value of the rewards from all periods, multiplied by their respective discount factors, to keep them as present values. We want to maximize the total expected discounted reward.

A greedy algorithm will always pull the arm with the highest Beta Prior. But sometimes this may not be optimal. Consider the following example.

Example: $(\alpha_1 = 10^{16}, \beta_1 = 10^{16})$, $(\alpha_2 = 1, \beta_2 = 2)$, $(\alpha_3 = 1, \beta_3 = 2)$, ..., $(\alpha_N = 1, \beta_N = 2)$
$\theta = 0.999$
We have Arm 1 generating a steady success rate of 50%, and all the other arms relatively unknown, but with lower beta priors. The greedy algorithm would pull Arm 1. But as we saw in the previous section, we are not very confident in the beta priors for the other arms, and it is a virtual guarantee, for large N, that at least one of them will be much more profitable than 50%. So the optimum strategy will spend some time exploring these other arms.

Gittins Index Theorem

There exists a function g of three variables, $g(\alpha, \beta, \theta)$, such that an optimum strategy for maximizing total expected discounted reward in the multi-armed bandit problem with Beta priors is to play the arm i with the largest value of $g(\alpha_i, \beta_i, \theta)$. This function, g, is known as the Gittins Index.

With this theorem, we can compute g for each arm individually. Then, at each period, we can answer the question: "which arm has the highest Gittins index?" and then play that arm.

Now we will see that just having the knowledge that there is such an index will help us solve the problem.

First, an example to constrain the Gittins index: $(\alpha_1 = 10^{16}, \beta_1 = 10^{16})$, $(\alpha_2 = 10^{16}, \beta_2 = 2 \times 10^{16})$
$\theta = 0.9$
Clearly, we should play Arm 1 in this situation, so we want the Gittins index to be higher for Arm 1 than for Arm 2. To capture this constraint in a general problem, we will introduce a special "reference" machine.

$$R_p = (pk, (1-p)k) \to (\alpha, \beta) \text{ as } k \to \infty$$

Here, R_p is an ideal reference machine, whose Beta Prior will be p. Note that for $0 < p < 1$, $\frac{\alpha}{\alpha+\beta} = \frac{pk}{k} = p$ as $k \to \infty$. We know that the Gittins index of R_p must lie somewhere above that of R_0 and below that of

R_1, and if we can find the point, p, where we are indifferent between playing R_p and machine i, then we can assign p to be the Gittins index, $g(\alpha_i, \beta_i, \theta)$.

Suppose we play R_p at every period. Then, $E[\text{total discounted reward}] = p + p\theta + p\theta^2 + ... = \frac{p}{1-\theta}$. We take this value to be the Gittins index of R_p.

Let us define the value function, $V(p; \alpha, \beta, \theta)$, to be the maximum expected discounted reward of any strategy that starts with two arms, R_p and (α, β). If R_p is played in the first period, then R_p will always be played because R_p does not change, and the other arm only changes if it is played.

Now we find the indifference point between the two arms. We know that we can get an expected discounted reward of at least $\frac{p}{1-\theta}$, because we are guaranteed $\frac{p}{1-\theta}$ by playing R_p. Hence, $V(p; \alpha, \beta, \theta)$ is at least $\frac{p}{1-\theta}$. The smallest p for which the value function is exactly equal to $\frac{p}{1-\theta}$ will correspond to the Gittins index, since it must be better to play the (alpha, beta) machine for smaller p, but better to play the reference machine for higher p, making this the indifference point.

Thinking back to our original problem, we had: $(\alpha_1 = 1, \beta_1 = 2), (\alpha_2 = 10, \beta_2 = 10)$, $\theta = 0.99$
Forgetting about Arm 2, we can calculate the indifference point for Arm 1. Suppose, for argument's sake that it is 0.60: $p_1 = 0.60 \Rightarrow g(\alpha_1, \beta_1, \theta) = 60$
Forgetting about Arm 1, we can calculate the indifference point for Arm 2. Suppose, for argument's sake, that it is 0.55: $p_2 = 0.55 \Rightarrow g(\alpha_1, \beta_1, \theta) = 55$
$\Rightarrow$ play Arm 1.

We will now see a formula for computing the value function, which in turn will be useful for calculating the Gittins index. To compute the value function, we must find the best strategy to play R_p and the (α, β) machine. In the first step, this strategy has just two choices.

1. Play R_p. In this case, R_p will be played in all future periods as well, yielding a reward (as explained earlier) of $\frac{p}{1-\theta}$

2. Play (α, β). In this case, we must add the value this period and the *expected value* of two possibilities next period: success and failure. We have a $\frac{\alpha}{\alpha+\beta}$ probability of success and a $\frac{\beta}{\alpha+\beta}$ probability of failure. We also need to multiply next period's reward by θ. Thus, the expected discounted reward for playing (α, β) will be:
$$\frac{\alpha}{\alpha+\beta} + \theta\left(\frac{\alpha}{\alpha+\beta}V(p; \alpha+1, \beta, \theta) + \frac{\beta}{\alpha+\beta}V(p; \alpha, \beta+1, \theta\right)$$

So the maximization becomes:

$$V(p; \alpha, \beta, \theta) = max\left\{\frac{p}{1-\theta}, \frac{\alpha}{\alpha+\beta} + \theta\left(\frac{\alpha}{\alpha+\beta}V(p; \alpha+1, \beta, \theta) + \frac{\beta}{\alpha+\beta}V(p; \alpha, \beta+1, \theta\right)\right\}$$