Thompson Sampling for Bernoulli Bandits

David S. Rosenberg

1 Basic Idea

Suppose we have K bandits (i.e. slot machines). Our game proceeds in rounds, and in each round we can select one of the K bandits to play. In the Bernoulli bandit setting, every time bandit k is played, it pays off a reward of 1 with probability θ_k and 0 with probability $1 - \theta_k$.

If we knew $\theta_1, \ldots, \theta_K \in [0, 1]$, the optimal action would be to always play bandit $k_{\text{best}} = \arg\max_k \theta_k$. However, we start with no information, so we must trade off between trying various bandits to estimate their probability of payoff (exploring) with committing to the bandit that seems best so far (exploiting). There are various approaches to this tradeoff.

2 Thompson Sampling

Thompson sampling is a Bayesian approach to this problem. We start with a prior distribution on each of the K bandits: $\pi(\theta_1), \ldots, \pi(\theta_K)$. In this setting, it's easiest to use a prior from the Beta family of probability distributions, since that's a conjugate prior.

Let $\mathcal{D}_t$ be all the data we have collected after t rounds of play, which we can represent as a set of pairs $((n_{10}, n_{11}), (n_{20}, n_{21}), \ldots, n_{K0}, n_{K1}))$, where n_{k1} is the number of times bandit k was played with a payoff of 1 and n_{k0} is the number of times bandit k was played with a payoff of 0. So if we look at the data at round t, we'll have $\sum_{k=1}^{K} (n_{k1} + n_{k0}) = t$.

Suppose we've just completed round t and we have data $\mathcal{D}_t$. We can then compute the posterior distributions on the unknown parameters: $\pi(\theta_1 \mid \mathcal{D}_t), \ldots, \pi(\theta_K \mid \mathcal{D}_t)$.

Which bandit should we play at the $t+1$'st round? The greedy approach would be to play the bandit that has the highest probability of being the best. This is all exploitation, and no exploration. The Thompson sampling approach is an interesting compromise: For each k, sample $\hat{\theta}_k \sim \pi(\theta_k \mid \mathcal{D}_t)$. Then play bandit $k = \arg\max_k \hat{\theta}_k$. Then k is a sample from the posterior distribution over which bandit is the best. As we become more confident about which is the best bandit, we'll choose that bandit more frequently. All that's left is to work out the details for computing the posterior distributions.

2.1 Details for using the Beta prior

We'll use the following parameterization of the Beta family of distributions:

$$\begin{aligned} \theta &\sim \text{Beta}(\alpha, \beta) \\ \pi(\theta) &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \end{aligned}$$

for $\alpha, \beta > 0$, and where the support of the distribution is $(0, 1)$. Some probability densities are shown in the Figure[1] below:

2.1.1 Posterior Prior / Posterior

For the case of a single bandit, let's let θ be the probability of a payoff of 1, and let n_1 be the number of 1's we've observed so far, and n_0 be the number of 0's

[1] Figure by Horas based on the work of Krishnavedala (Own work) [Public domain], via Wikimedia Commons.

observed so far. The likelihood function for θ for this data is then $p(n_0, n_1 \mid \theta) = \theta^{n_1}(1-\theta)^{n_0}$. If our prior on θ is Beta(α_0, α_1), then the posterior after observing data $\mathcal{D} = (n_0, n_1)$ is given by

$$\begin{aligned} p(\theta \mid \mathcal{D}) &\propto p(\theta)p(\mathcal{D} \mid \theta) \\ &\propto \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_0 - 1} \times \theta^{n_1}(1-\theta)^{n_0} \\ &= \theta^{\alpha_1 - 1 + n_1}(1-\theta)^{\alpha_0 - 1 + n_0}. \end{aligned}$$

Thus the posterior distribution on θ is $\theta \mid \mathcal{D} \sim$ Beta$(\alpha_0 + n_0, \alpha_1 + n_1)$.

For K bandits, we just compute the posteriors separately for each θ_k in the obvious way. So if $((n_{11}, n_{10}), (n_{21}, n_{20}), \ldots, n_{K1}, n_{K0}))$ is our data $\mathcal{D}_t$ after the t'th round, and if we use the priors $\theta_k \sim$ Beta $(\alpha_{k0}, \alpha_{k1})$ then the posterior distribution for bandit k is

$$\theta_k \mid \mathcal{D}_t \sim \text{Beta}(\alpha_{k0} + n_{k0}, \alpha_{k1} + n_{k1}).$$