Bayesian Bandits

CS 486 / 686

University of Waterloo

Lecture 22 Supplementary: July 17, 2017

Multi-Armed Bandits

- Problem:
  - N bandits with unknown average reward $R(a)$
  - Which arm a should we play at each time step?
  - Exploitation/exploration tradeoff
- Common frequentist approaches:
  - ϵ-greedy
  - Upper confidence bound (UCB)
- Alternative Bayesian approaches
  - Thompson sampling
  - Gittins indices

Bayesian Learning

- Notation:
  - r^a: random variable for a's rewards
  - $\Pr(r^a; \theta)$: unknown distribution (parameterized by θ)
  - $R(a) = E[r^a]$: unknown average reward
- Idea:
  - Express uncertainty about θ by a prior $\Pr(\theta)$
  - Compute posterior $\Pr(\theta | r_1^a, r_2^a, \dots, r_n^a)$ based on samples $r_1^a, r_2^a, \dots, r_n^a$ observed for a so far.

- Bayes theorem:

$$\Pr(\theta | r_1^a, r_2^a, \dots, r_n^a) \propto \Pr(\theta) \Pr(r_1^a, r_2^a, \dots, r_n^a | \theta)$$

Distributional Information

- Posterior over θ allows us to estimate
  - Distribution over next reward r^a

$$\Pr(r^a|r_1^a, r_2^a, \dots, r_n^a) = \int_\theta \Pr(r^a; \theta) \Pr(\theta|r_1^a, r_2^a, \dots, r_n^a)\, d\theta$$

  - Distribution over $R(a)$ when θ includes the mean

$\Pr(R(a)|r_1^a, r_2^a, \dots, r_n^a) = \Pr(\theta|r_1^a, r_2^a, \dots, r_n^a)$ if $\theta = R(a)$

- To guide exploration:
  - UCB: $\Pr\big(R(a) \leq bound(r_1^a, r_2^a, \dots, r_n^a)\big) \geq 1 - \delta$
  - Bayesian techniques: $\Pr\big(R(a)|r_1^a, r_2^a, \dots, r_n^a\big)$

Coin Example

- Consider two biased coins C_1 and C_2

$$R(C_1) = \Pr(C_1 = head)$$
$$R(C_2) = \Pr(C_2 = head)$$

- Problem:
  - Maximize # of heads in k flips
  - Which coin should we choose for each flip?

Bernoulli Variables

- r^{C_1}, r^{C_2} are Bernoulli variables with domain $\{0,1\}$
- Bernoulli dist. are parameterized by their mean
  - i.e. $\Pr(r^{C_1}; \theta_1) = \theta_1 = R(C_1)$

    $\Pr(r^{C_2}; \theta_2) = \theta_2 = R(C_2)$

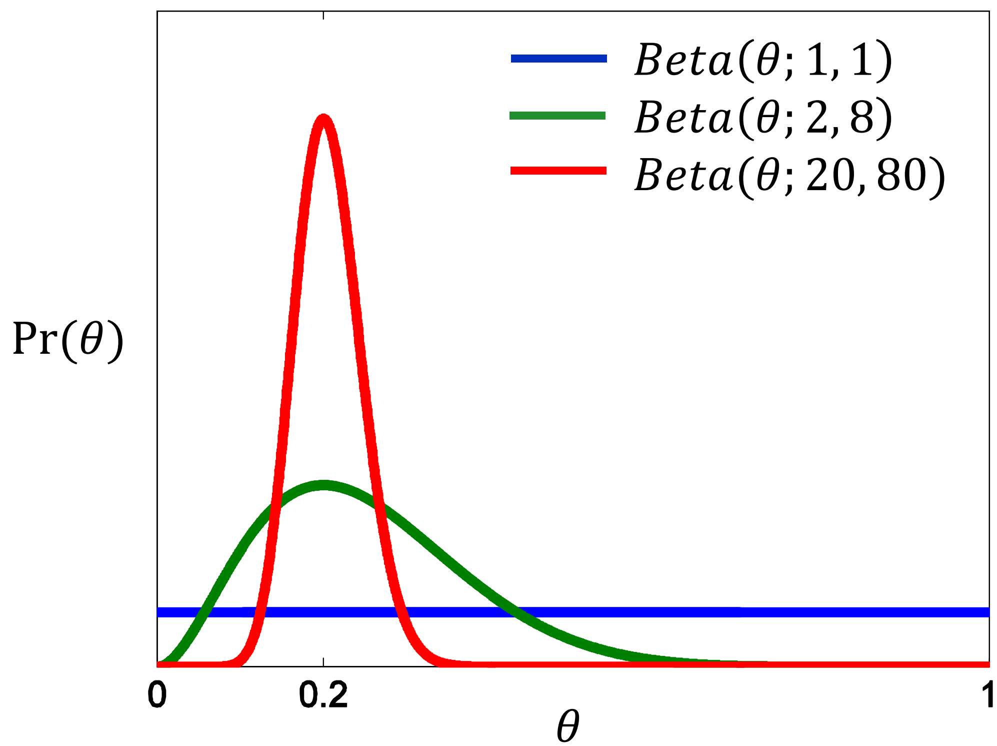
Beta distribution

- Let the prior $\Pr(\theta)$ be a Beta distribution

$$Beta(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- $\alpha - 1$: # of heads
- $\beta - 1$: # of tails

- $E[\theta] = \alpha/(\alpha + \beta)$



Belief Update

- Prior: $\Pr(\theta) = Beta(\theta; \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$
- Posterior after coin flip:

$$\begin{aligned}
\Pr(\theta|head) &\propto \quad \Pr(\theta) \qquad\qquad \Pr(head|\theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \qquad \theta \\
&= \theta^{(\alpha+1)-1}(1-\theta)^{\beta-1} \\
&\propto {\color{darkred} Beta(\theta; \alpha+1, \beta)}
\end{aligned}$$

$$\begin{aligned}
\Pr(\theta|tail) &\propto \quad \Pr(\theta) \qquad\qquad \Pr(tail|\theta) \\
&\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \quad (1-\theta) \\
&= \theta^{\alpha-1}(1-\theta)^{(\beta+1)-1} \\
&\propto {\color{darkred} Beta(\theta; \alpha, \beta+1)}
\end{aligned}$$

Thompson Sampling

- Idea:
  - Sample several potential average rewards:
    $R_1(a), \ldots R_k(a) \sim \Pr(R(a)|r_1^a, \ldots, r_n^a)$ for each a
  - Estimate empirical average
    $\hat{R}(a) = \frac{1}{k}\sum_{i=1}^{k} R_i(a)$
  - Execute $argmax_a \hat{R}(a)$

- Coin example
  - $\Pr(R(a)|r_1^a, \ldots, r_n^a) = \mathrm{Beta}(\theta_a; \alpha_a, \beta_a)$
    where $\alpha_a - 1 = \#heads$ and $\beta_a - 1 = \#tails$

Thompson Sampling Algorithm Bernoulli Rewards

ThompsonSampling(h)

$V \leftarrow 0$

For $n = 1$ to h

Sample $R_1(a), \ldots, R_k(a) \sim \Pr(R(a)) \;\; \forall a$

$\hat{R}(a) \leftarrow \frac{1}{k} \sum_{i=1}^{k} R_i(a) \quad \forall a$

$a^* \leftarrow \mathrm{argmax}_a \hat{R}(a)$

Execute a^* and receive r

$V \leftarrow V + r$

Update $\Pr(R(a^*))$ based on r

Return V

Comparison

Thompson Sampling

- Action Selection

  $a^* = \text{argmax}_\text{a}\, \hat{R}(a)$

- Empirical mean

  $\hat{R}(a) = \frac{1}{k}\sum_{i=1}^{k} R_i(a)$

- Samples

  $R_i(a) \sim \Pr(R_i(a) | r_1^a \dots r_n^a)$

  $r_i^a \sim \Pr(r^a; \theta)$

- Some exploration

Greedy Strategy

- Action Selection

  $a^* = \text{argmax}_\text{a}\, \tilde{R}(a)$

- Empirical mean

  $\tilde{R}(a) = \frac{1}{n}\sum_{i=1}^{n} r_i^a$

- Samples

  $r_i^a \sim \Pr(r^a; \theta)$

- No exploration

Sample Size

- In Thompson sampling, amount of data n and sample size k regulate amount of exploration

- As n and k increase, $\hat{R}(a)$ becomes less stochastic, which reduces exploration
  - As $n \uparrow$, $\Pr(R(a)|r_1^a \dots r_n^a)$ becomes more peaked
  - As $k \uparrow$, $\hat{R}(a)$ approaches $E[R(a)|r_1^a \dots r_n^a]$

- The stochasticity of $\hat{R}(a)$ ensures that all actions are chosen with some probability

Continuous Rewards

- So far we assumed that $r \in \{0,1\}$

- What about continuous rewards, i.e. $r \in [0,1]$?
  - NB: rewards in $[a, b]$ can be remapped to $[0,1]$ by an affine transformation without changing the problem

- Idea:
  - When we receive a reward r
  - Sample $b \sim Bernoulli(r)$ s.t. $b \in \{0,1\}$

Thompson Sampling Algorithm Continuous rewards

ThompsonSampling(h)
- $V \leftarrow 0$
- For $n = 1$ to h
  - Sample $R_1(a), \dots, R_k(a) \sim \Pr(R(a))\ \ \forall a$
  - $\hat{R}(a) \leftarrow \frac{1}{k}\sum_{i=1}^{k} R_i(a)\quad \forall a$
  - $a^* \leftarrow \mathrm{argmax_a}\ \hat{R}(a)$
  - Execute a^* and receive r
  - $V \leftarrow V + r$
  - Sample $b \sim Bernoulli(r)$
  - Update $\Pr(R(a^*))$ based on b

Return V

Analysis

- Thompson sampling converges to best arm
- Theory:
  - Expected cumulative regret: $O(\log n)$
  - On par with UCB and ϵ-greedy
- Practice:
  - Sample size k often set to 1
  - Used by Bing for ad placement
    - Graepel, Candela, Borchert, Herbrich (2010) Web-scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine, ICML.