Multi-armed Bandit

A. Richards

02.08.2017

Objectives

Morning

- Overview of Frequentist A/B testing
- Overview of Bayesian A/B testing
- Review of Bayes' Theorem
- Conjugate Priors
- Is CTR_A is better than CTR_B through Code

Frequentist A/B testing

1. define a metric
2. determine parameters of interest for study (number of observations, power, significance threshold, and so on)
3. run test, without checking results, until number of observations has been achieved
4. calculate p-value associated with hypothesis test
5. report *p*-value and suggestion for action

Bayesian A/B testing

1. Define a metric
2. Run test, continually monitor results
3. At any time calculate probability that $A \geq B$ or vice versa
4. Suggest course of action based on probabilities calculated

Discussion

Obtaining significance depends on power

- What are the factors that influence power?
- Which A/B testing framework relies on significance testing?
- What about the other framework what does it use to guide decisions?

With the Bayesian framework you:

- have a degree of Belief?
- can say *it is 95% likely that site A is better than site B*?
- can stop a test early based on surprising data

That formula again

$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)} \tag{1}$$

- prior - $P(\theta)$ - one's beliefs about a quantity before presented with evidence
- posterior - $P(\theta|y)$ - probability of the parameters given the evidence
- likelihood - $P(y|\theta)$ - probability of the evidence given the parameters
- normalizing constant - $P(y)$

Lets talk about priors

Subjective vs Objective priors

Bayesian priors can be classified into two classes: objective priors, which aim to allow the data to influence the posterior the most, and subjective priors, which allow the practitioner to express his or her views into the prior.

- If we added more probability mass to certain areas of the prior, and less elsewhere, we are biasing our inference towards the unknowns existing in the former area.
- The prior's influence changes as our dataset increases

Empirical Bayes

It is not a true Bayesian method. Empirical Bayes combines frequentist and Bayesian inference. The prior distribution, instead of being selected beforehand is estimated directly from the data generally with frequentist methods.

https://en.wikipedia.org/wiki/Empirical_Bayes_method

CAUTION

Many people feel that empirical bayes is *double counting* or *double dipping* from the data

The Gamma distribution

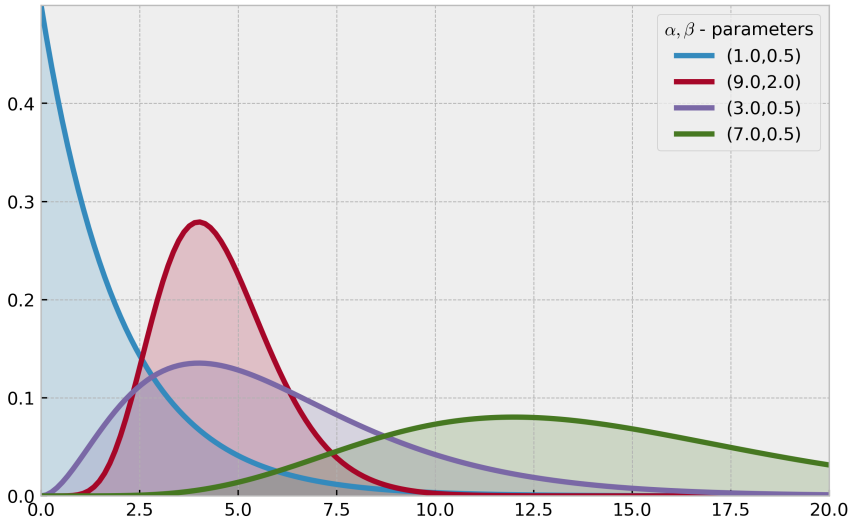
A Gamma random variable, denoted $X \sim \text{Gamma}(\alpha, \beta)$, is also

$$\text{Exp}(\beta) \sim \text{Gamma}(1, \beta) \tag{2}$$

The additional parameter gives flexibility which helps us better express subjective priors. The density function for a Gamma(α, β) random variable is:

$$f(x \mid \alpha, \beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{3}$$

where $\Gamma(\alpha)$ is the Gamma function

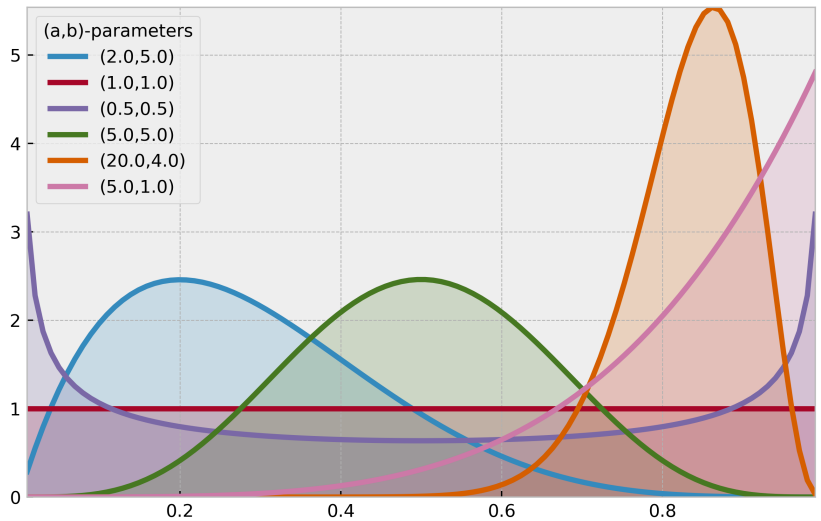


The Beta distribution

The Beta distribution is very useful in Bayesian statistics. A random variable X has a Beta distribution, with parameters (α, β), if its density function is:

$$f_X(x|\ \alpha, \beta) = \frac{x^{(\alpha-1)}(1-x)^{(\beta-1)}}{B(\alpha, \beta)} \tag{4}$$

where B is the Beta function (hence the name). The random variable X is only allowed in [0,1], making the Beta distribution a popular distribution for decimal values, probabilities and proportions. The values of α and β, both positive values, provide great flexibility in the shape of the distribution.



So what?

There is an interesting connection between the Beta distribution and the Binomial distribution. Suppose we are interested in some unknown proportion or probability p. (think of coin example)

- We assign a Beta(α, β) prior to p. We observe some data generated by a Binomial process, say $X \sim$ Binomial(N, p), with p still unknown.
- Then our posterior *is again a Beta distribution*, i.e. $p|X \sim$ Beta$(\alpha + X, \beta + N - X)$.
- If we start with a Beta$(1, 1)$ prior on p (which is a Uniform), observe data $X \sim$ Binomial(N, p), then our posterior is Beta$(1 + X, 1 + N - X)$.

A Beta prior with Binomial observations creates a Beta posterior. This is a very useful property, both computationally and heuristically.

Beta conjugate to the binomial proof

If you are curious how we can identify a conjugate prior you can follow the proof (5 minutes).

https://www.youtube.com/watch?v=hKYvZF9wXkk

Conjugate priors

Recall that

$$P(y|\theta) \propto p(\theta|y)p(\theta) \tag{5}$$

- Conjugate families of priors arise when the likelihood times the prior produces a recognizable posterior kernel
- For mathematical convenience, we construct a family of prior densities that lead to simple posterior densities.
- Conjugate prior distributions have the practical advantage, in addition to computational convenience, of being interpretable as additional data
- Probability distributions that belong to an exponential family have natural conjugate prior distributions

Is CTR_A better than CTR_B?

```
import numpy as np

num_samples = 10000
A = np.random.beta(1 + num_clicks_A,
1 + num_views_A - num_clicks_A,
size=num_samples)
B = np.random.beta(1 + num_clicks_B,
1 + num_views_B - num_clicks_B,
size=num_samples)
# Probability that A wins:
print(np.sum(A > B) / float(num_samples))
```

Check-in questions

Core questions

1. Why might we consider a Bayesian approach over a frequentist in an A/B testing scenario?
2. What is a conjugate prior? (Can you provide an example)

Bonus questions

1. What is meant by empirical Bayes?
2. How is the Gamma related to the Exponential distribution?
3. What are the limits on α and β in the Beta distribution?

Objectives

Morning

- ✓ Overview of Frequentist A/B testing
- ✓ Overview of Bayesian A/B testing
- ✓ Review of Bayes' Theorem
- ✓ Conjugate Priors
- ✓ Determining whether CTR_A is better than CTR_B through Code

Afternoon

- Introduction and use cases of Multi-Arm Bandits
- Zen and the Art of Minimizing Regret
- Overview of Common Strategies
  - Epsilon-Greedy
  - Softmax
  - UCB1
  - Bayesian Bandit
- Other forms of Multi-Arm Bandits

What's up with the name?

Multi-armed bandit

Suppose you are faced with N slot machines. Each bandit has an unknown probability of distributing a prize (assume for now the prizes are the same for each bandit, only the probabilities differ). Some bandits are very generous, others not so much. Of course, you don't know what these probabilities are. By only choosing one bandit per round, our task is devise a strategy to maximize our winnings.

The bandits problem

- If we knew the bandit with the largest probability, then always picking this bandit would yield the maximum winnings. So our task can be phrased as Find the best bandit, and as quickly as possible.
- The task is complicated by the stochastic nature of the bandits. A suboptimal bandit can return many winnings, purely by chance, which would make us believe that it is a very profitable bandit. Similarly, the best bandit can return many duds. Should we keep trying losers then, or give up?
- A more troublesome problem is, if we have found a bandit that returns *pretty good* results, do we keep drawing from it to maintain our *pretty good score*, or do we try other bandits in hopes of finding an *even-better* bandit? This is the exploration vs. exploitation dilemma.

Exploration vs Exploitation

- Exploration: Trying out different options to try and determine the reward associated with the given approach (i.e. acquiring more knowledge)
- Exploitation: Going with the approach that you believe to have the highest expected payoff (i.e. optimizing decisions based on existing knowledge)

Multi-armed bandit applications go beyond A/B

- Internet display advertising: What ad strategy will maximize sales? Naturally minimizing strategies that do not work (generalizes to A/B/C/D strategies)
- Ecology: How do the animals maximize its fitness w.r.t energy?
- Finance: which stock option gives the highest return, under time-varying return profiles.
- Clinical trials: a researcher would like to find the best treatment, out of many possible treatment, while minimizing losses.
- Psychology: How does punishment and reward affect our behaviour? How do humans learn?

Traditional A/B testing

The process

- Start with pure exploration in which groups A and B are assigned equal number of users
- Once you think you have determined the better option, switch to pure exploitation in which you stop the experiment and send all users to the better performer

Potential issues

- Equal number of observations are routed to A and B for a preset amount of time or iterations
- Only after that preset amount of time or iterations do we stop and use the better performer
- Waste time (and money!) showing users the site that is not performing as well

Multi-armed bandit solutions

- Shows a user the site that you think is best most of the time (exactly how is dictated by the strategy chosen)
- As the experiment runs, we update the belief about the true CTR (Click Through Rate)
- Run for however long until we are satisfied the experiment has determined the better site
- Balances exploration and exploitation rather than doing only one or the other

Formalization

- The model is given by a set of real distributions $B = R_1, ..., R_K$,
- where each distribution is associated with the a reward delivered by one of the $K \in N +$ levers.
- We will let $\mu_1, ..., \mu_K$ be the mean values associated with these reward distributions.
- The gambler plays one lever per round and observes the associated reward.
- The goal is to maximize the sum of the collective rewards, or alternatively minimize the agent's regret

Regret

The regret p that an agent experiences after T rounds is the difference between the reward sum associated with an optimal strategy and the sum of collected rewards

$$p = T\mu^* - \sum_{t=1}^{T} \hat{r}_t \tag{6}$$

- μ^* - is the maximal reward mean, $\mu^* = max_k\{\mu_k\}$
- $\hat{r}_t$ - is the reward at time t
- Regret is simply a measure of how often you choose a suboptimal bandit. We can think of this as the cost function we are trying to minimize

A zero-regret strategy

- A zero-regret strategy is a strategy whose average regret per round p/T tends to zero when the number of rounds played tends toward infinity
- Interestingly enough, a zero-regret strategy does not guarantee you will never choose a sub-optimal outcome, but rather guarantees that, over time, you will tend to choose the optimal outcome

Epsilon-Greedy strategy

- Explore with some probability *epsilon* (often 10%)
- All other times we will exploit (i.e. choose the bandit with the best performance so far)
- After we choose a given bandit we update the performance based on the result.

Other strategies

UCB1 - Upper confidence bound

For the UCB1 algorithm we will choose whichever bandit that has the largest value.

softmax

For the softmax algorithm we will choose the bandit randomly in proportion to its estimated value.

Bayesian Bandits

The Bayesian bandit algorithm involves modeling each of our bandits with a beta distribution with the following shape parameters:

- $\alpha = 1 +$ number of times bandit has won
- $\beta = 1 +$ number of times bandit has lost

We will then take a random sample from each bandit's distribution and choose the bandit with the highest value.

Bayesian Bandits

There are also many approximately-optimal solutions which are quite good. One of the few solutions that can scale incredibly well. The solution is known as Bayesian Bandits.

These strategies are an example of online algorithms because they are continuously-being-updated, aka reinforcement learning algorithm. The algorithm starts in an ignorant state, where it knows nothing, and begins to acquire data by testing the system. As it acquires data and results, it learns what the best and worst behaviors are (in this case, it learns which bandit is the best).

- Like in a normal multi-arm problem, an agent must choose between arms during each iteration
- Before making the choice, the agent sees a d-dimensional feature vector (context vector), associated with the current iterations state
- The agent uses the context vector as well as the history of past rewards to choose the arm to play in the current iteration
- Over time, the aim is for the agent to learn how the context vectors relate to the associated rewards so as to pick the optimal arm

Summary

- The problem originated with a gambler standing in front of a row of slot machines (referred to as *one-armed bandits*)
- The agent has to decide which machines to play, how many times to play each machine, and it which order
- Each bandit will provide a reward from a unknown distribution
- Objective is to maximize the sum of rewards earned through a series of lever pulls
- There are several classes of fairly optimal solutions

Check-in questions

Core questions

1. Can you contrast exploration vs exploitation
2. Can you use plain English to explain what regret means in this context?
3. Cay you explain any of the following strategies: Epsilon-Greedy, Softmax, or UCB1 in your own words.
4. What is the Bayesian Bandit strategy and how is it an example of online learning?

References I