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

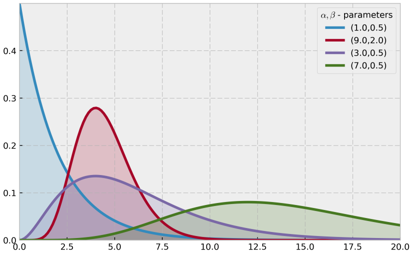
A Gamma random variable, denoted $X \sim \text{Gamma}(\alpha, \beta)$, is also

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

The additional parameter gives flexibility which helps us better express subjective priors. The density function for a $\text{Gamma}(\alpha, \beta)$ 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

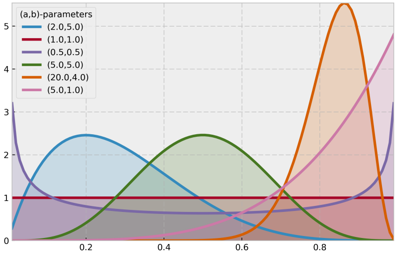


The Beta distribution

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 $X \sim \text{Binomial}(N, p)$, with p still unknown.
- Then our posterior *is again a Beta distribution*, i.e. $p|X \sim \text{Beta}(\alpha + X, \beta + N - X)$.
- If we start with a Beta$(1, 1)$ prior on p (which is a Uniform), observe data $X \sim \text{Binomial}(N, p)$, 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.

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?

```
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)
```

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

References I