Beta-Bernoulli distributions

David Mareček

October 12, 2022

EUROPEAN UNION
European Structural and Investment Fund
Operational Programme Research,
Development and Education

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Many of the slides in this presentation were taken from the presentations of Carl Edward Rasmussen (University of Cambridge)

Coin tossing

You are presented with a coin.

What is the probability of heads?
How is the probability defined?
We need data!

We toss once and it's head (H).
How much are you willing to bet $p(\mathsf{H}) > 0.5$?

Bernoulli distribution

The Bernoulli probability distribution over binary random variables:

- Binary random variable X: outcome x of a single coin toss.
- The two values x can take are
  - $X = 0$ for tails,
  - $X = 1$ for heads.
- Let the probability of heads be $\pi = p(X = 1)$.
- π is the parameter of the Bernoulli distribution.
- The probability of tail is $p(X = 0) = 1 - \pi$. We can compactly write

$$p(X = x|\pi) = p(x|\pi) = \pi^x(1 - \pi)^{1-x}$$

What do we think π is after observing a single heads outcome?

- Maximum likelihood! We maximise the probability of data $p(H|\pi)$ with respect to π:

$$p(H|\pi) = p(x = 1|\pi) = \pi, \qquad argmax_{\pi\in[0,1]}\pi = 1$$

- Ok, so the answer is $\pi = 1$. This coin only generates heads.

Is this reasonable? How much are you willing to bet $p(H) > 0.5$?

Binomial distribution

We observe a sequence of tosses rather than a single toss: HHTH

- The probability of this particular sequence is: $p(\text{HHTH}) = \pi^3(1-\pi)$.
- But so is the probability of THHH, of HTHH and of HHHT.
- We often don't care about the order of the outcomes, only about the counts. In our example the probability of 3 heads out of 4 tosses is: $4\pi^3(1-\pi)$.

The binomial distribution gives the probability of observing k heads out of n tosses

$$p(k|\pi, n) = \binom{n}{k}\pi^k(1-\pi)^{n-k}$$

- This assumes n independent tosses from a Bernoulli distribution $p(x|\pi)$.
- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient, also known as "n choose k"

Maximum likelihood estimation

If we observe k heads out of n tosses, what do we think π is?
We can maximise the likelihood of the observed data given the parameter π.

$$p(k|\pi, n) \propto \pi^k (1-\pi)^{n-k}$$

It is convenient to take the logarithm and derivatives with respect to π:

$$\log p(k|\pi, n) = k \log \pi + (n-k) \log(1-\pi) + \text{Constant}$$

$$\frac{\partial \log p(k|\pi, n)}{\partial \pi} = \frac{k}{\pi} - \frac{n-k}{1-\pi} = 0 \iff \pi = \frac{k}{n}$$

Is this reasonable?

- For HHTH we get $\pi = 3/4$.

How much would you bet now that $p(\text{H}) > 0.5$?
We would need a probability over a probability...

Prior beliefs about coins

So we have observed 3 heads out of 4 tosses but are unwilling to bet much money that $p(\mathsf{H}) > 0.5$? (That for example out of 10,000,000 tosses at least 5,000,001 will be heads.) Why?

- You might believe that coins tend to be fair. $\pi \simeq 1/2$
- A finite set of observations updates your opinion about π.
- But how to express your opinion about π before you see any data?

Pseudo-counts: You think the coin is fair and... you are...

- Not very sure. You act as if you had seen 2 heads and 2 tails before.
- Pretty sure. It is as if you had observed 20 heads and 20 tails before.
- Totally sure. As if you had seen 1000 heads and 1000 tails before.

Depending on the strength of your prior assumptions, it takes a different number of actual observations to change your mind.

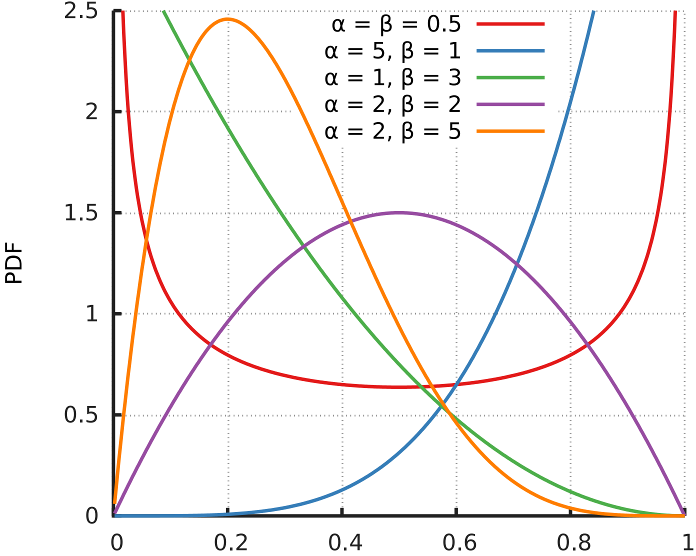
Beta distribution: distributions on probabilities

Continuous probability distribution defined on the interval [0, 1]

$$Beta(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1} = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

- $\alpha > 0$ and $\beta > 0$ are the shape parameters.
- these parameters correspond to 'one plus the pseudo-counts'.
- $\Gamma(\alpha)$ is an extension of the factorial function.
  (`https://en.wikipedia.org/wiki/Gamma_function`)
- $\Gamma(n) = (n-1)!$ for integer n.
- $B(\alpha,\beta) = \int_0^1 \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi$ is the normalization function so that it sums up to one.
- The mean is given by $E(\pi) = \frac{\alpha}{\alpha+\beta}$

Beta distribution: examples



Beta distribution: online demo

https://mathlets.org/mathlets/beta-distribution

Beta distribution: exercise

Could you imagine the shape of coin that would fit the following parameters for the prior Beta distributions?

- $\alpha = 100, \quad \beta = 100$
- $\alpha = 2, \quad \beta = 3$
- $\alpha = 2, \quad \beta = 10$
- $\alpha = 0.1, \quad \beta = 0.1$
- $\alpha = 1, \quad \beta = 1$

Posterior $\propto$ Prior $\times$ Likelihood

Imagine we observe k heads out of n tosses.

The probability of the observed data given π is the likelihood:

$$p(D|\pi) \propto \pi^k (1-\pi)^{n-k}$$

We use our prior $p(\pi|\alpha,\beta) = Beta(\pi|\alpha,\beta)$ to get the posterior probability (Bayes' theorem):

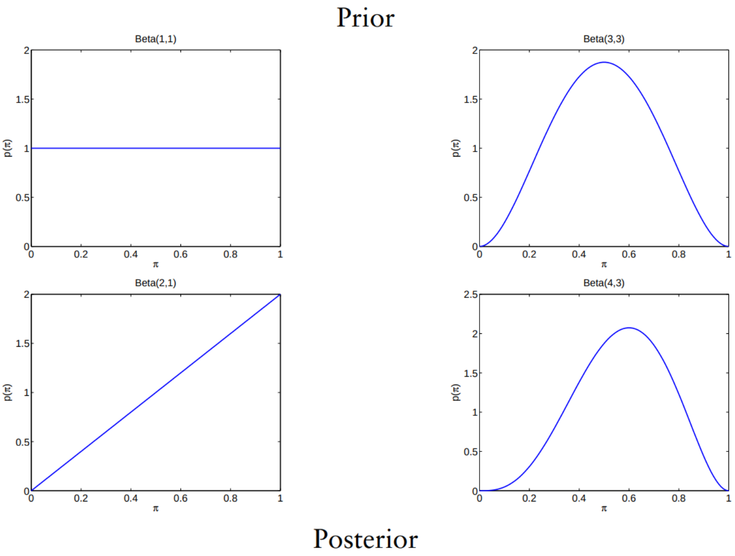
$$p(\pi|D) = \frac{p(\pi|\alpha,\beta)p(D|\pi)}{p(D)} \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}\pi^k(1-\pi)^{n-k} = \pi^{\alpha+k-1}(1-\pi)^{\beta+n-k-1}$$

$$p(\pi|D) = Beta(\pi|\alpha+k, \beta+n-k)$$

The Beta distribution is a *conjugate* prior to the Bernoulli/binomial distribution:

- The resulting posterior is also a Beta distribution with parameters.
- The posterior parameters are: $\alpha_{posterior} = \alpha_{prior} + k, \quad \beta_{posterior} = \beta_{prior} + (n-k)$

Before and after observing one head



Posterior Beta distribution: online demo

http://www.randomservices.org/random/apps/BetaCoin.html

Making predictions

Given some data D, what is the probability of the next toss being heads, $x_{next} = 1$?

Under the **Maximum Likelihood approach** we predict using the value of π_{ML} that maximises the likelihood of π given the observed data D:

$$p(x_{next} = 1|\pi_{ML}) = \pi_{ML}$$

With the **Bayesian approach**, we average over all possible parameter settings:

$$p(x_{next} = 1|D) = \int_0^1 p(x_{next} = 1|\pi)p(\pi|D)d\pi = \int_0^1 \pi Beta(\pi|\alpha{+}k, \beta{+}n{-}k)d\pi = \frac{\alpha + k}{\alpha + \beta + n}$$

The prediction for heads happens to correspond to the mean of the posterior distribution. E.g. for $D = (x = 1)$:

- Learner A with $Beta(1, 1)$ predicts $p(x_{next} = 1|D) = 2/3$
- Learner B with $Beta(3, 3)$ predicts $p(x_{next} = 1|D) = 4/7$

Expected value of the Beta distribution

Predictive probability of heads is equal to the expected value of the posterior distribution. Here, we derive the expected value for $X \sim Beta(\alpha, \beta)$:

$$
\begin{aligned}
E[X] &= \int_0^1 \pi \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} d\pi \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{\alpha}(1-\pi)^{\beta-1} d\pi \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \pi^{\alpha+1-1}(1-\pi)^{\beta-1} d\pi \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{1}{B(\alpha+1, \beta)} \\
&= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+\beta+1)} \\
&= \frac{\alpha\Gamma(\alpha)\Gamma(\alpha+\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)\Gamma(\alpha)} = \frac{\alpha}{\alpha+\beta}
\end{aligned}
$$

Making predictions - other statistics

Given the posterior distribution, we can also answer other questions such as "what is the probability that $\pi > 0.5$ given the observed data?"

$$p(\pi > 0.5|D) = \int_{0.5}^{1} p(\pi'|D)d\pi' = \int_{0.5}^{1} Beta(\pi'|\alpha', \beta')d\pi'$$

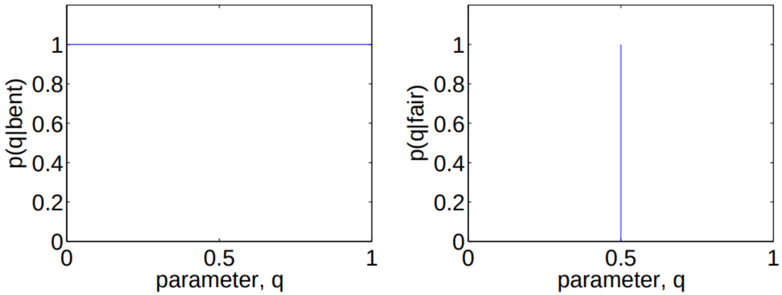
- Learner A with prior $Beta(1, 1)$ predicts $p(\pi > 0.5|D) = 0.75$
- Learner B with prior $Beta(3, 3)$ predicts $p(\pi > 0.5|D) = 0.66$

Learning about a coin, multiple models

Consider two alternative models of a coin, “fair” and “bent”. A priori, we may think that “fair” is more probable, e.g.:

$$p(fair) = 0.8, \qquad p(bent) = 0.2$$

For the bent coin, (a little unrealistically) all parameter values could be equally likely, where the fair coin has a fixed probability:



Learning about a coin, multiple models

We make 10 tosses, and get data D: T H T H T T T T T T
The evidence for the fair model is: $p(D|fair) = (1/2)^10 \simeq 0.001$ and for the bent model:

$$p(D|bent) = \int p(D|\pi, bent)p(\pi|bent)d\pi = \int \pi^2(1-\pi)^8 d\pi = B(3,9) \simeq 0.002$$

Using priors $p(fair) = 0.8$, $p(bent) = 0.2$, the posterior by Bayes rule:

$$p(fair|D) \propto 0.0008, \quad p(bent|D) \propto 0.0004,$$

i.e., two thirds probability that the coin is fair.
How do we make predictions? By weighting the predictions from each model by their probability. Probability of Head at next toss is:

$$\frac{2}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{3}{12} = \frac{5}{12}$$

Summary

Frequentist statistics

- Predictions on the underlying truths of the experiment use only data from the current experiment.
- Maximum likelihood estimation - we maximize the probability of data.

Bayesian statistics

- Predictions take past knowledge of similar experiments into account. These are known as *prior*. This prior is combined with current experiment data to get a *posterior*.
- Maximum a posteriori estimation

Definition and Posterior probability

Beta distribution:

$$Beta(\pi|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}$$

Dirichlet distribution (generalization of Beta distribution to m outcomes):

$$Dir(\vec{\pi}|\alpha_1,\dots,\alpha_m) = \frac{\Gamma(\sum_{i=1}^{m}\alpha_i)}{\prod_{i=1}^{m}\Gamma(\alpha_i)}\prod_{i=1}^{m}\pi_i^{\alpha_i-1}$$

Posterior probability for Beta-Bernoulli distribution:

$$p(\pi|D) = \frac{p(\pi|\alpha,\beta)p(D|\pi)}{p(D)} \propto \pi^{\alpha-1}(1-\pi)^{\beta-1}\pi^{k}(1-\pi)^{n-k} = \pi^{\alpha+k-1}(1-\pi)^{\beta+n-k-1}$$

$$p(\pi|D) = Beta(\pi|\alpha+k,\beta+n-k)$$

Posterior probability for Dirichlet-Categorical distribution:

$$p(\vec{\pi}|D) = \frac{p(\vec{\pi}|\vec{\alpha})p(D|\vec{\pi})}{p(D)} \propto \prod_{i=1}^{m}\pi_i^{\alpha_i-1}\pi_i^{k_i} = \prod_{i=1}^{m}\pi_i^{\alpha_i+k_i-1} \propto Dir(\vec{\pi}|\vec{\alpha}+\vec{k})$$

Predictive distributions

Beta-Bernoulli predictions:

$$p(x_{next} = 1|D) = \int_0^1 p(x_{next} = 1|\pi)p(\pi|D)d\pi = \int_0^1 \pi Beta(\pi|\alpha{+}k, \beta{+}n{-}k)d\pi = \frac{\alpha + k}{\alpha + \beta + n}$$

Dirichlet-Categorical predictions:

$$p(x_{next} = j|D) = \int_{\triangle} p(x_{next} = j|\vec{\pi})p(\vec{\pi}|D)d\vec{\pi} = \int_{\triangle} \pi_j Dir(\vec{\pi}|\vec{\alpha} + \vec{k})d\vec{\pi} = \frac{\alpha_j + k_j}{\sum_{i=1}^{m}(\alpha_i + k_i)}$$

The sign $\triangle$ indicates a simplex: the integral is taken across all vectors $\vec{\pi}$ that are valid probability distributions, i.e. $\sum_{i=1}^{m} \pi_i = 1$

The integrals are equal to expected values of given Beta/Dirichlet posterior distributions.