Discrete Binary Distributions

Carl Edward Rasmussen

November 11th, 2016
Key concepts

- Bernoulli: probabilities over binary variables
- Binomial: probabilities over counts and binary sequences
- Inference, priors and pseudo-counts, the Beta distribution
- model comparison: an example
Coin tossing

• You are presented with a coin: what is the probability of heads?  
  *What does this question even mean?*

• How much are you willing to bet $p(\text{head}) > 0.5$?
  *Do you expect this coin to come up heads more often than tails?  
  Wait... can you toss the coin a few times, I need data!*

• Ok, you observe the following sequence of outcomes (T: tail, H: head):
  
  H

  *This is not enough data!*

• Now you observe the outcome of three additional tosses:
  
  HHTH

  How much are you *now* willing to bet $p(\text{head}) > 0.5$?
The Bernoulli discrete binary 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 tail,
  - $X = 1$ for heads.
- Let the probability of heads be $\pi = p(X = 1)$. $\pi$ 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 $\pi$ is after observing a single heads outcome?
- Maximum likelihood! Maximise $p(H | \pi)$ with respect to $\pi$:
  \[
p(H | \pi) = p(x = 1 | \pi) = \pi, \quad \text{argmax}_{\pi \in [0,1]} \pi = 1
  \]
- Ok, so the answer is $\pi = 1$. This coin only generates heads.
  
  *Is this reasonable? How much are you willing to bet $p(\text{heads}) > 0.5$?*
The binomial distribution: counts of binary outcomes

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

\[ \text{HHTH} \]

- The probability of this particular sequence is: \( p(\text{HHTH}) = \pi^3(1 - \pi) \).
- But so is the probability of \( \text{THHH} \), of \( \text{HTHH} \) and of \( \text{HHHT} \).
- 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”.
## Naming of discrete distributions

<table>
<thead>
<tr>
<th></th>
<th>binary</th>
<th>multi-valued</th>
</tr>
</thead>
<tbody>
<tr>
<td>sequence</td>
<td>binary categorical $\pi^k(1 - \pi)^{n-k}$</td>
<td>categorical $\prod_{i=1}^{k} \pi_{i}^{k_i}$</td>
</tr>
<tr>
<td>counts</td>
<td>binomial $\binom{n}{k} \pi^k(1 - \pi)^{n-k}$</td>
<td>multinomial</td>
</tr>
</tbody>
</table>

For binary outcomes with $k$ successes in $n$ trials. For multi-dimensional distributions we have $k$ possible outcomes, a sequence $x_1, \ldots, x_N$ and counts $c_i = \sum_{n=1}^{N} \delta(x_n, i)$. For all distributions the parameter of the distribution is the vector $\pi$, which has either one element $\pi \in [0; 1]$ or multiple entries such that $\pi_i > 0$ and $\sum \pi_i = 1$. 
If we observe \( k \) heads out of \( n \) tosses, what do we think \( \pi \) is? We can maximise the likelihood of parameter \( \pi \) given the observed data.

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

It is convenient to take the logarithm and derivatives with respect to \( \pi \)

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

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

Is this reasonable?

- For HHTH we get \( \pi = 3/4 \).
- How much would you bet now that \( p(\text{heads}) > 0.5? \)

What do you think \( p(\pi > 0.5) \) is?

Wait! This is a probability over ... a probability?
Prior beliefs about coins – before tossing the coin

So you have observed 3 heads out of 4 tosses but are unwilling to bet £100 that $p(\text{heads}) > 0.5$?

(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 ($\pi \simeq \frac{1}{2}$).
• A finite set of observations *updates your opinion* about $\pi$.
• But how to express your opinion about $\pi$ *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.
The Beta distribution: distributions on *probabilities*

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

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

- \(\alpha > 0\) 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\(^1\). \(\Gamma(n) = (n - 1)!\) for integer \(n\).
- \(\text{B}(\alpha, \beta)\) is the beta function, it normalises the Beta distribution.
- The mean is given by \(\mathbb{E}(\pi) = \frac{\alpha}{\alpha + \beta}\). \([\text{Left: } \alpha = \beta = 1, \text{ Right: } \alpha = \beta = 3]\)

\[1\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx\]
Posterior for coin tossing

Imagine we observe a single coin toss and it comes out heads. Our observed data is:

\[ D = \{k = 1\}, \quad \text{where} \quad n = 1. \]

The probability of the observed data given \( \pi \) is the likelihood:

\[ p(D|\pi) = \pi \]

We use our prior \( p(\pi|\alpha, \beta) = \text{Beta}(\pi|\alpha, \beta) \) to get the posterior probability:

\[
p(\pi|D) = \frac{p(\pi|\alpha, \beta)p(D|\pi)}{p(D)} \propto \pi \text{Beta}(\pi|\alpha, \beta) \\
\propto \pi \pi^{(\alpha-1)} (1 - \pi)^{(\beta-1)} \propto \text{Beta}(\pi|\alpha + 1, \beta)\]

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

- The resulting posterior is also a Beta distribution.
- The posterior parameters are given by:
  \[
  \alpha_{\text{posterior}} = \alpha_{\text{prior}} + k \\
  \beta_{\text{posterior}} = \beta_{\text{prior}} + (n - k)
  \]
Before and after observing one head

Prior

Posterior
Making predictions

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

Under the Maximum Likelihood approach we predict using the value of $\pi_{\text{ML}}$ that maximises the likelihood of $\pi$ given the observed data, $\mathcal{D}$:

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

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

$$p(x_{\text{next}} = 1 | \mathcal{D}) = \int p(x = 1 | \pi) p(\pi | \mathcal{D}) d\pi$$

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

- Learner A with Beta$(1, 1)$ predicts $p(x_{\text{next}} = 1 | \mathcal{D}) = \frac{2}{3}$
- Learner B with Beta$(3, 3)$ predicts $p(x_{\text{next}} = 1 | \mathcal{D}) = \frac{4}{7}$
Given the posterior distribution, we can also answer other questions such as “what is the probability that $\pi > 0.5$ given the observed data?”

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

- **Learner A with prior Beta(1, 1)** predicts $p(\pi > 0.5 | D) = 0.75$
- **Learner B with prior Beta(3, 3)** predicts $p(\pi > 0.5 | D) = 0.66$
Consider two alternative models of a coin, “fair” and “bent”. A priori, we may think that “fair” is more probable, eg:

\[ p(\text{fair}) = 0.8, \quad p(\text{bent}) = 0.2 \]

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 (2)

We make 10 tosses, and get data $\mathcal{D}$: T H T H T T T T T

The **evidence** for the fair model is: $p(\mathcal{D}|\text{fair}) = (1/2)^{10} \approx 0.001$

and for the bent model:

$$p(\mathcal{D}|\text{bent}) = \int p(\mathcal{D}|\pi, \text{bent}) p(\pi|\text{bent}) \, d\pi = \int \pi^2 (1-\pi)^8 \, d\pi = B(3,9) \approx 0.002$$

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

$$p(\text{fair}|\mathcal{D}) \propto 0.0008, \quad p(\text{bent}|\mathcal{D}) \propto 0.0004,$$

ie, 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:

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