Lecture 10: Discrete Distributions

4F13: Machine Learning

Carl Edward Rasmussen and Zoubin Ghahramani

Department of Engineering
University of Cambridge

http://mlg.eng.cam.ac.uk/teaching/4f13/

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 that tails?

  Wait... can you throw 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 throws:

  HHTH

  How much are you *now* willing to bet $p(\text{head}) > 0.5$?

The Bernoulli discrete distribution

The *Bernoulli* discrete probability distribution over binary random variables:

- Binary random variable X: outcome x of a single coin throw.
- The two values x can take are
  - $X = 0$ for tail,
  - $X = 1$ for heads.
- Let the probability of heads be $\pi = p(X = 1)$.
  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! Maximise $p(H|\pi)$ with respect to π:

$$p(H|\pi) \; = \; p(x = 1|\pi) \; = \; \pi, \qquad \text{argmax}_{\pi \in [0,1]} \, \pi = 1$$

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

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

The Binomial distribution: counts of binary outcomes

We observe a sequence of throws rather than a single throw:

$$\text{HHTH}$$

- 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 don't really care about the order of the outcomes, only about the *counts*. In our example the probability of 3 heads out of 4 throws is: $4\pi^3(1-\pi)$.

The *Binomial* distribution gives the probability of observing k heads out of n throws

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

- This assumes independent throws 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 under a Binomial distribution

If we observe k heads out of n throws, what do we think π is?

We can maximise the likelihood of parameter π 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 π

$$\begin{aligned} \log p(k|\pi, n) &= k \log \pi + (n-k) \log(1-\pi) + \text{Constant} \\ \frac{\partial \log p(k|\pi, n)}{\partial \pi} &= \frac{k}{\pi} - \frac{n-k}{1-\pi} = 0 \iff \boxed{\pi = \frac{k}{n}} \end{aligned}$$

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 throwing the coin

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

(That for example out of 10,000,000 throws 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 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.

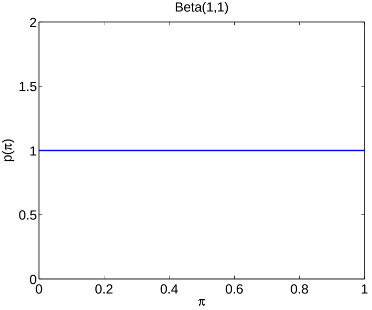
Depending on the strength of your prior assumptions, it takes a different number of actual observations to change your mind.

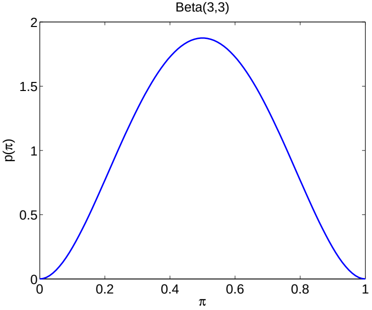
The Beta distribution: distributions on *probabilities*

Continuous probability distribution defined on the interval $(0, 1)$

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

- $\alpha > 0$ and $\beta > 0$ are the shape *parameters*.
- the parameters correspond to 'one plus the pseudo-counts'.
- $\Gamma(\alpha)$ is an extension of the factorial function. $\Gamma(n) = (n-1)!$ for integer n.
- $B(\alpha, \beta)$ is the beta function, it normalises the Beta distribution.
- The mean is given by $E(\pi) = \frac{\alpha}{\alpha+\beta}$. [Left: $\alpha = \beta = 1$, Right: $\alpha = \beta = 3$]





Posterior for coin tossing

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

$$\mathcal{D} = \{k = 1\}, \quad \text{where} \quad n = 1.$$

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

$$p(\mathcal{D}|\pi) = \pi$$

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

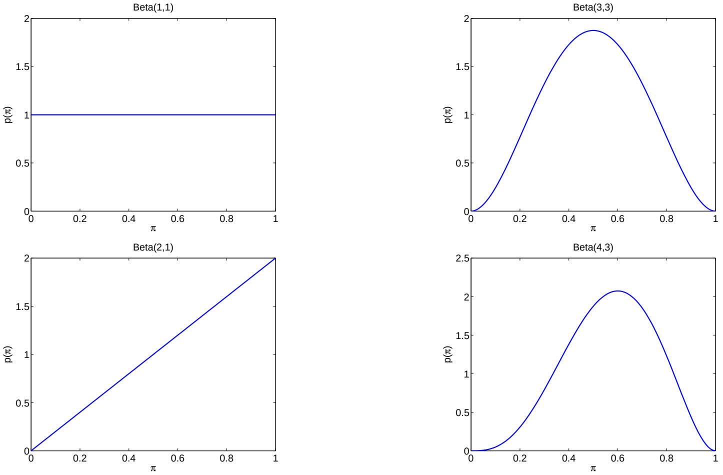
$$\begin{aligned} p(\pi|\mathcal{D}) &= \frac{p(\pi|\alpha,\beta)p(\mathcal{D}|\pi)}{p(\mathcal{D})} \propto \pi\,\text{Beta}(\pi|\alpha,\beta) \\ &\propto \pi\,\pi^{(\alpha-1)}(1-\pi)^{(\beta-1)} \propto \text{Beta}(\pi|\alpha+1,\beta) \end{aligned}$$

The Beta distribution is a *conjugate* prior to the Binomial distribution:

- The resulting posterior is also a Beta distribution.
- The posterior parameters are given by: $\begin{aligned} \alpha_{\text{posterior}} &= \alpha_{\text{prior}} + k \\ \beta_{\text{posterior}} &= \beta_{\text{prior}} + (n-k) \end{aligned}$

Before and after observing one head

Prior



Posterior

Making predictions - posterior mean

Under the Maximum Likelihood approach we report the value of π that maximises the likelihood of π given the observed data.

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

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

This corresponds to reporting the mean of the *posterior* distribution.

- Learner A with $\text{Beta}(1, 1)$ predicts $p(x = 1|\mathcal{D}) = \frac{2}{3}$
- Learner B with $\text{Beta}(3, 3)$ predicts $p(x = 1|\mathcal{D}) = \frac{4}{7}$

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?"

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

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

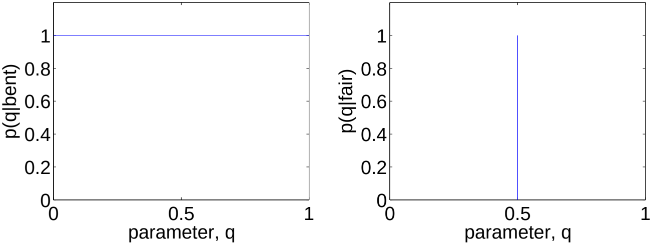
Note that for any $l > 1$ and fixed α and β, the two posteriors $\text{Beta}(\pi|\alpha, \beta)$ and $\text{Beta}(\pi|l\alpha, l\beta)$ have the same *average* π, but give different values for $p(\pi > 0.5)$.

Learning about a coin, multiple models (1)

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

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

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



We make 10 tosses, and get: T H T H T T T T T T

Learning about a coin, multiple models (2)

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

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

The posterior for the models, by Bayes rule:

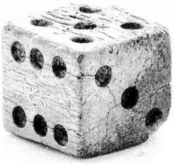
$$p(\text{fair}|\mathcal{D}) \propto 0.0008, \qquad p(\text{bent}|\mathcal{D}) \propto 0.0004,$$

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:

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

The Multinomial distribution (1)



Generalisation of the Binomial distribution from 2 outcomes to m outcomes. Useful for random variables that take one of a finite set of possible outcomes.

Throw a die $n = 60$ times, and count the of observed (6 possible) outcomes.

Outcome	Count
$X = x_1 = 1$	$k_1 = 12$
$X = x_2 = 2$	$k_2 = 7$
$X = x_3 = 3$	$k_3 = 11$
$X = x_4 = 4$	$k_4 = 8$
$X = x_5 = 5$	$k_5 = 9$
$X = x_6 = 6$	$k_6 = 13$

Note that we have one parameter too many. We don't need to know all the k_i and n, because $\sum_{i=1}^{6} k_i = n$.

The Multinomial distribution (2)

Consider a discrete random variable X that can take one of m values $x_1, \ldots, x_m$.

Out of n independent trials, let k_i be the number of times $X = x_i$ was observed. It follows that $\sum_{i=1}^{m} k_i = n$.

Denote by π_i the probability that $X = x_i$, with $\sum_{i=1}^{m} \pi_i = 1$.

The probability of observing a vector of occurrences $\mathbf{k} = [k_1, \ldots, k_m]^\top$ is given by the *Multinomial* distribution parametrised by $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_m]^\top$:

$$p(\mathbf{k}|\boldsymbol{\pi}, n) \;=\; p(k_1, \ldots, k_m|\pi_1, \ldots, \pi_m, n) \;=\; \frac{n!}{k_1!k_2!\ldots k_m!} \prod_{i=1} \pi_i^{k_i}$$

- Note that we can write $p(\mathbf{k}|\boldsymbol{\pi})$ since n is redundant.
- The multinomial coefficient $\frac{n!}{k_1!k_2!\ldots k_m!}$ is a generalisation of $\binom{n}{k}$.

Example: word counts in text

Consider describing a text document by the frequency of occurrence of every distinct word.

The UCI *Bag of Words* dataset from the University of California, Irvine. [1]

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/bag-of-words/