COMP2610/COMP6261 - Information Theory

Lecture 5: Useful Discrete Probability Distributions

Mark Reid and **Aditya Menon**

Research School of Computer Science
The Australian National University

THE AUSTRALIAN NATIONAL UNIVERSITY

August 5th, 2014

Last time

- Examples of application of Bayes' rule
  - Formalizing problems in language of probability
  - Eating hamburgers, detecting terrorists, document classification
- Frequentist vs Bayesian probabilities

The Bayesian Inference Framework

Bayesian Inference

Bayesian inference provides us with a a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\underbrace{p(Z|X)}_{\text{posterior}} = \frac{\overbrace{p(X|Z)}^{\text{likelihood}}\overbrace{p(Z)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

$$= \frac{p(X|Z)p(Z)}{\sum_{Z'} p(X|Z')p(Z')}$$

Prior: Belief that someone is sick

Likelihood: Probability of testing positive given you are sick

Posterior: Probability of being sick given you test positive

This time

- The Bernoulli and binomial distribution
- Estimating probabilities from data
- Bayesian inference for parameter estimation

Outline

The Bernoulli Distribution:

Introduction

Consider a binary variable $X \in \{0, 1\}$. It could represent many things:

- Whether a coin lands heads or tails
- The presence/absence of a word in a document
- A transmitted bit in a message
- The success of a medical trial

Often, these outcomes (0 or 1) are not equally likely

What is a general way to model such an X?

The Bernoulli Distribution

Definition

The variable X takes on the outcomes

$$X = \begin{cases} 1 & \text{probability } \theta \\ 0 & \text{probability } 1 - \theta \end{cases}$$

Here, $0 \leq \theta \leq 1$ is a parameter representing the probability of success

For higher values of θ, it is more likely to see 1 than 0

- e.g. a biased coin

The Bernoulli Distribution

Definition

By definition,

$$p(X = 1|\theta) = \theta$$
$$p(X = 0|\theta) = 1 - \theta$$

More succinctly,

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

The Bernoulli Distribution

Definition

By definition,

$$p(X = 1|\theta) = \theta$$
$$p(X = 0|\theta) = 1 - \theta$$

More succinctly,

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

This is known as a Bernoulli distribution over binary outcomes:

$$p(X = x|\theta) = \mathrm{Bern}(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

Note the use of the conditioning symbol for θ; will revisit later

The expected value (or mean) is given by:

$$\begin{aligned}\mathbb{E}[X|\theta] &= \sum_{x\in\{0,1\}} x \cdot p(x|\theta) \\ &= 1 \cdot p(X=1|\theta) + 0 \cdot p(X=0|\theta) \\ &= \theta.\end{aligned}$$

The variance (or squared standard deviation) is given by:

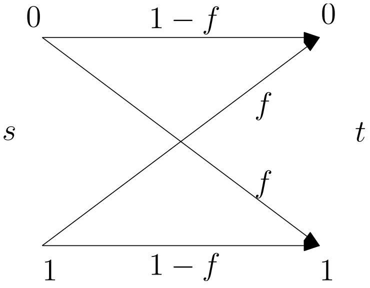
$$\mathbb{V}[X|\theta] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \theta(1-\theta).$$

Example: Binary Symmetric Channel

Suppose a sender transmits messages s that are sequences of bits

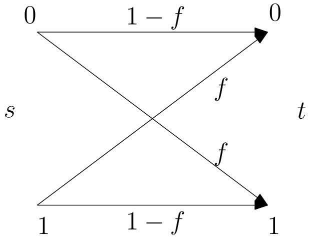
The receiver sees the bit sequence (message) t

Due to noise in the channel, the message is flipped with probability $0 \leq f \leq 1$



Example: Binary Symmetric Channel

We can think of r as the outcome of a random variable, with conditional distribution given by:



$$p(t=0|s=0)=1-f \quad p(t=0|s=1)=f$$
$$p(t=1|s=0)=f \quad p(t=1|s=1)=1-f$$

If E denotes whether an error occurred, clearly

$$p(E=e)=\text{Bern}(e|f).$$

The Binomial Distribution

Introduction

Suppose we perform N independent Bernoulli trials

- e.g. we toss a coin N times
- e.g. we transmit a sequence of N bits across a noisy channel

Each trial has probability θ of success

What is the distribution of the number of times (m) that $X = 1$?

- e.g. the number of times we obtained m heads
- e.g. the number of errors in the transmitted sequence

The Binomial Distribution
Definition

Let

$$Y = \sum_{n=1}^{N} X_n$$

where $X_n \sim \text{Bern}(\theta)$

Then Y has a binomial distribution with parameters N, θ:

$$p(Y = m) = \text{Bin}(m|N, \theta) = \binom{N}{m} \theta^m (1-\theta)^{N-m}$$

where

$$\binom{N}{m} = \frac{N!}{(N-m)!m!}$$

is the $\#$ of ways we can we obtain m heads out of N coin flips

The Binomial Distribution:

Mean and Variance

It is easy to show that:

$$\mathbb{E}[Y] = \sum_{m=0}^{N} m \cdot \mathrm{Bin}(m|N, \theta) = N\theta$$

$$\mathbb{V}[Y] = \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \cdot \mathrm{Bin}(m|N, \theta) = N\theta(1 - \theta)$$

- Follows from linearity of mean and variance

The Binomial Distribution:
Example

Ashton is an excellent off spinner. The probability of him getting a wicket during a cricket match is $\frac{1}{4}$.

His coach, Darren, commands him to get 10 wickets in a particular game.

1. What is the probability that he will get exactly three wickets?
2. What is the expected number of wickets he will get?
3. What is the probability that he will get at least one wicket?

The Binomial Distribution:

Example: Distribution of the Number of Wickets

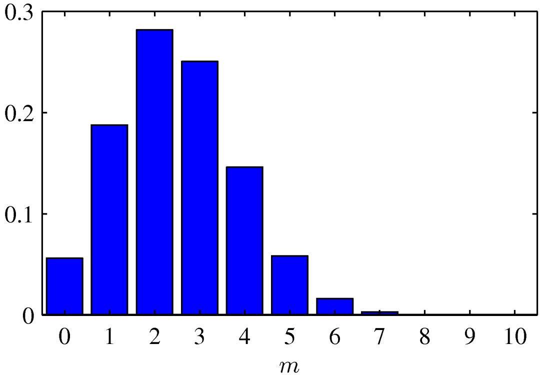


Figure: Histogram of the binomial distribution with $N = 10$ and $\theta = 0.25$. From Bishop (PRML, 2006)

The Bernoulli Distribution: Parameter Estimation

Consider the set of observations $\mathcal{D} = \{x_1, \ldots, x_N\}$ with $x_i \in \{0, 1\}$:

- The outcomes of a sequence of coin flips
- Whether or not there are errors in a transmitted bit string

Each observation is the outcome of a random variable X, with distribution

$$p(X = x) = \text{Bern}(x|\theta) = \theta^x (1-\theta)^{1-x}$$

for some parameter θ

The Bernoulli Distribution: Parameter Estimation

We know that

$$X \sim \mathrm{Bern}(x|\theta) = \theta^x (1-\theta)^{1-x}$$

But often, we don't know what the value of θ is

- The probability of a coin toss resulting in heads
- The probability of the word *defence* appearing in a document about sports

What would be a reasonable estimate for θ from $\mathcal{D}$?

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Say that we observe

$$\mathcal{D} = \{0, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

Intuitively, which seems more plausible: $\theta = \frac{1}{2}$? $\theta = \frac{1}{5}$?

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Say that we observe

$$\mathcal{D} = \{0, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

If it were true that $\theta = \frac{1}{2}$, **then** the probability of this sequence would be

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{10} p(x_i|\theta) \\ &= \prod_{i=1}^{10} \frac{1}{2} \\ &= \frac{1}{2^{10}} \\ &\approx 0.001. \end{aligned}$$

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Say that we observe

$$\mathcal{D} = \{0, 0, 0, 1, 0, 0, 1, 0, 0, 0\}$$

If it were true that $\theta = \frac{1}{5}$, **then** the probability of this sequence would be

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{10} p(x_i|\theta) \\ &= \left(\frac{1}{5}\right)^2 \cdot \left(\frac{4}{5}\right)^8 \\ &\approx 0.007. \end{aligned}$$

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

We can write down how likely $\mathcal{D}$ is under the Bernoulli model. Assuming independent observations:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

We call $L(\theta) = p(\mathcal{D}|\theta)$ the likelihood function

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

We can write down how likely $\mathcal{D}$ is under the Bernoulli model. Assuming independent observations:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(x_i|\theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

We call $L(\theta) = p(\mathcal{D}|\theta)$ the likelihood function

Maximum likelihood principle: We want to maximize this function wrt θ

The parameter for which the observed sequence has the highest probability

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Maximising $p(\mathcal{D}|\theta)$ is equivalent to maximising $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = \sum_{i=1}^{N} [x_i \log \theta + (1 - x_i) \log(1 - \theta)]$$

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Maximising $p(\mathcal{D}|\theta)$ is equivalent to maximising $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = \sum_{i=1}^{N} [x_i \log \theta + (1 - x_i) \log(1 - \theta)]$$

Setting $\frac{d\mathcal{L}}{d\theta} = 0$ we obtain:

$$\theta_{\text{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Maximising $p(\mathcal{D}|\theta)$ is equivalent to maximising $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$

$$\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta) = \sum_{i=1}^{N} [x_i \log \theta + (1 - x_i) \log(1 - \theta)]$$

Setting $\frac{d\mathcal{L}}{d\theta} = 0$ we obtain:

$$\theta_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

The proportion of times $x = 1$ in the dataset $\mathcal{D}$!

The Bernoulli Distribution:
Parameter Estimation — Issues with Maximum Likelihood

Consider the following scenarios:

- After $N = 3$ coin flips we obtained 3 'tails'
  - What is the estimate of the probability of a coin flip resulting in 'heads'?
- In a small set of documents about sports, the words *defence* never appeared.
  - What are the consequences when predicting whether a document is about sports (using Bayes' rule)?

Consider the following scenarios:

- After $N = 3$ coin flips we obtained 3 'tails'
  - What is the estimate of the probability of a coin flip resulting in 'heads'?
- In a small set of documents about sports, the words *defence* never appeared.
  - What are the consequences when predicting whether a document is about sports (using Bayes' rule)?

These issues are usually referred to as overfitting

- Need to "smooth" out our parameter estimates
- Alternatively, we can do Bayesian inference by considering priors over the parameters

The Bernoulli Distribution:

Parameter Estimation: Bayesian Inference

Recall:

$$\underbrace{p(Z|X)}_{\text{posterior}} = \frac{\overbrace{p(X|Z)}^{\text{likelihood}}\overbrace{p(Z)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

If we treat θ as a random variable, we may have some prior belief $p(\theta)$ about its value

- e.g. we believe θ is probably close to 0.5

Our prior on θ quantifies what we believe θ is likely to be, before looking at the data

Our posterior on θ quantifies what we believe θ is likely to be, after looking at the data

The likelihood of X given θ is

$$\mathrm{Bern}(x|\theta) = \theta^x (1-\theta)^{1-x}$$

For the prior, it is mathematically convenient to express it as a Beta distribution:

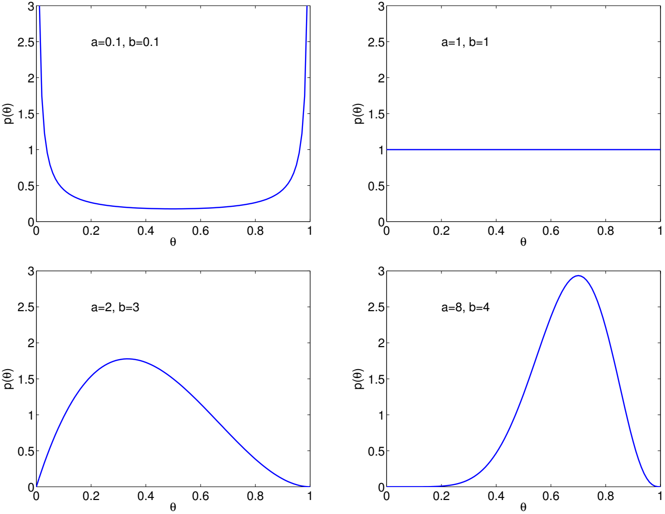
$$\mathrm{Beta}(\theta|a,b) = \frac{1}{Z(a,b)} \theta^{a-1}(1-\theta)^{b-1},$$

where $Z(a,b)$ is a suitable normaliser

We can tune a, b to reflect our belief in the range of likely values of θ

Beta Prior

Examples



Beta Prior and Binomial Likelihood:
Beta Posterior Distribution

Recall that for $\mathcal{D} = \{x_1, \ldots, x_N\}$, the likelihood under a Bernoulli model is:

$$p(\mathcal{D}|\theta) = \theta^m (1-\theta)^\ell,$$

where $m = \sharp(x = 1)$ and $\ell \stackrel{\text{def}}{=} N - m = \sharp(x = 0)$.

Beta Prior and Binomial Likelihood:

Beta Posterior Distribution

Recall that for $\mathcal{D} = \{x_1, \ldots, x_N\}$, the likelihood under a Bernoulli model is:

$$p(\mathcal{D}|\theta) = \theta^m (1-\theta)^\ell,$$

where $m = \sharp(x = 1)$ and $\ell \stackrel{\text{def}}{=} N - m = \sharp(x = 0)$.

For the prior $p(\theta|a, b) = \text{Beta}(\theta|a, b)$ we can obtain the posterior:

$$\begin{aligned} p(\theta|\mathcal{D}, a, b) &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{p(\mathcal{D}|a, b)} \\ &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{\int_0^1 p(\mathcal{D}|\theta)p(\theta|a, b)d\theta} \\ &= \text{Beta}(\theta|m + a, \ell + b). \end{aligned}$$

Beta Prior and Binomial Likelihood:

Beta Posterior Distribution

Recall that for $\mathcal{D} = \{x_1, \ldots, x_N\}$, the likelihood under a Bernoulli model is:

$$p(\mathcal{D}|\theta) = \theta^m (1-\theta)^\ell,$$

where $m = \sharp(x = 1)$ and $\ell \stackrel{\text{def}}{=} N - m = \sharp(x = 0)$.

For the prior $p(\theta|a, b) = \text{Beta}(\theta|a, b)$ we can obtain the posterior:

$$\begin{aligned} p(\theta|\mathcal{D}, a, b) &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{p(\mathcal{D}|a, b)} \\ &= \frac{p(\mathcal{D}|\theta)p(\theta|a, b)}{\int_0^1 p(\mathcal{D}|\theta)p(\theta|a, b)d\theta} \\ &= \text{Beta}(\theta|m + a, \ell + b). \end{aligned}$$

Can use this as our new prior if we see more data!

Now suppose we choose θ_{MAP} to maximise $p(\theta|\mathcal{D})$

One can show that

$$\theta_{\text{MAP}} = \frac{m + a - 1}{N + a + b - 2}$$

c.f. the estimate that did not use any prior,

$$\theta_{\text{ML}} = \frac{m}{N}$$

The prior parameters a and b can be seen as adding some "fake" trials!

Summary

- Distributions involving binary random variables
  - Bernoulli distribution
  - Binomial distribution
- Bayesian inference: Full posterior on the parameters
  - Beta prior and binomial likelihood $\rightarrow$ Beta posterior
- Reading: Mackay §23.1 and §23.5; Bishop §2.1 and §2.2

Next time

- The entropy and its properties