COMP2610 / COMP6261 - Information Theory

Lecture 5: Bernoulli, Binomial, Maximum Likelihood and MAP

Australian
National
University

6 August 2018

Last time

- Exa
  - https://eduassistpro.github.i
- 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

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

likelihood prior

evidence

Prior: Belief that someone is sick

Likelihood: Probability of testing positive given someone is sick

Posterior: Probability of being sick given someone tests positive

This time

- The Bernoulli and binomial distribution (we will make much use of this henceforth in studying binary channels)

- Esti

- Bayesian inference for parameter estimati

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

- A tra

- 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

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

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,

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

More suc

The Bernoulli Distribution

Definition

By definition,

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

More suc

This is known as a Bernoulli distribution over binary o

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

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

The Bernoulli Distribution

Mean and Variance

The expected value (or mean) is given by:

$$\mathbb{E}[X|\theta] = \sum_{x \quad 0,1} x \cdot p(x|\theta)$$

The variance (or squared standard deviation) is giv

$$\begin{aligned}\mathbb{V}[X|\theta] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[(X - \theta)^2] \\ &= (0 - \theta)^2 \cdot p(X = 0|\theta) + (1 - \theta)^2 \cdot p(X = 1|\theta) \\ &= \theta(1 - \theta).\end{aligned}$$

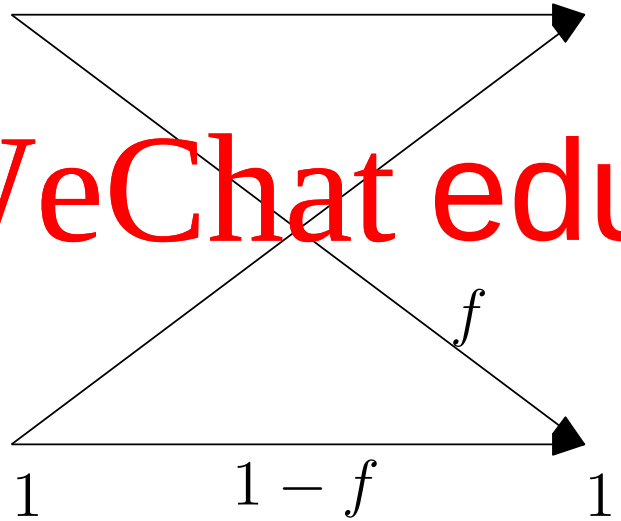
Example: Binary Symmetric Channel

Suppose a sender transmits messages s that are sequences of bits

The receiver sees the bit sequence (message) r

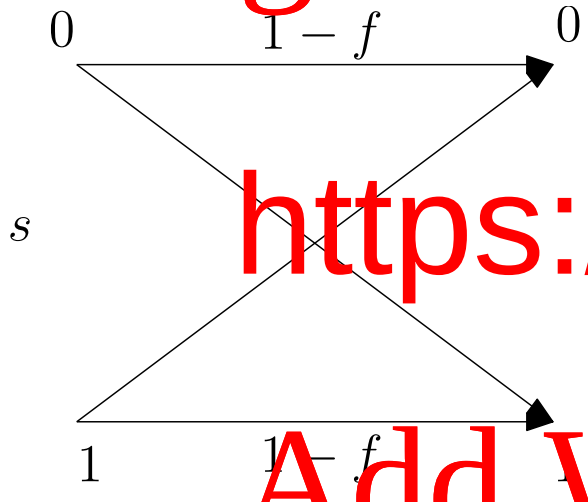
Due to noi

$0 \leq f \leq$



Example: Binary Symmetric Channel

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



$= 1) = f$

$= 1) = 1 - f$

If E denotes whether an error occurred, clearly

$$p(E = e) = \text{Bern}(e|f), \qquad e \in \{0, 1\}.$$

The Binomial Distribution

Introduction

Suppose we perform N independent Bernoulli trials

- e.g. we toss a coin N times

- e.g.

Each trial h θ

What is the distribution of the number of times (

- e.g. the number of times we obtained

- e.g. the number of errors in the transmitted sequence

The Binomial Distribution

Definition

Let

$$Y = \sum_{i=1}^{N} X_i$$

where X_i

Then Y

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

for $m \in \{0, 1, \ldots, N\}$. Here

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

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:

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

- Follows from linearity of mean and variance

$$\mathbb{E}[Y] = \mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbb{E}[X_i]$$

$$\mathbb{V}[Y] = \mathbb{V}\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \mathbb{V}[X_i] = N\theta(1-\theta)$$

The Binomial Distribution:

Example

Ashton is an excellent off spinner. The probability of him getting a wicket during a cricket match is $\frac{1}{4}$. (That is, on each attempt, there is a 1/4 chance he will get a wicket.)

His coach
game.

1. Wh
   $\text{Bin}(3|10, 0.25)$
2. What is the expected number of wickets he will g
   $\mathbb{E}[Y]$, where $Y \sim \text{Bin}(\cdot|10, 0.25)$.
3. What is the probability that he will get at least one wicket?
   $\sum_{m=1}^{10} \text{Bin}(m|N = 10, \theta = 0.25) = 1 - \text{Bin}(m = 0|N = 10, \theta = 0.25)$

The Binomial Distribution:

Example: Distribution of the Number of Wickets

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

- Wh

Each observation is the outcome of a random variab tion

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

for some parameter θ

The Bernoulli Distribution: Parameter Estimation

We know that

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

But often,

- The
- The probability of the word *defence* a sports

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

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Say that w

Intuitively, which seems more plausible: θ

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

Say that we observe

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

If it were tr e

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

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

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{10} p(x_i$$

$$= \left(\frac{1}{5}\right)^2 \cdot \frac{}{5}$$

$$\approx 0.007.$$

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

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

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:

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

Maximum likelihood principle: We want to maximize 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 $\mathcal{L}(\theta) = \log p(\mathcal{D}|\theta)$.

$$\mathcal{L}(\theta) = \text{lo} \qquad \text{og}(1-\theta)]$$

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

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

$$\mathcal{L}(\theta) = \text{lo} \qquad \text{og}(1-\theta)]$$

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

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

The Bernoulli Distribution: Parameter Estimation:

Maximum Likelihood

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

$$\mathcal{L}(\theta) = \text{lo} \qquad \text{og}(1-\theta)]$$

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

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

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 s *ce* never app
  - https://eduassistpro.github.i

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 s *ce* never app
  - https://eduassistpro.github.i

These issues are usually referred to as overfitting

- Need to "smooth" out our parameter estimat

- Alternatively, we can do Bayesian inference by considering priors over the parameters

The Bernoulli Distribution:

Parameter Estimation: Bayesian Inference

Recall:

$$p(\theta|X) = \frac{\overbrace{p(X|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{p(X)}$$

If we treat $p(\theta)$ about its value

- e.g. we believe θ is probably close to 0

Our prior on θ quantifies what we believe the data

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

The Bernoulli Distribution:

Parameter Estimation: Bayesian Inference

The likelihood of X given θ is

$$\text{Bern}(x\ \theta) = \theta^x(1 \quad \theta)^{1-x}$$

For the pri
distribution:

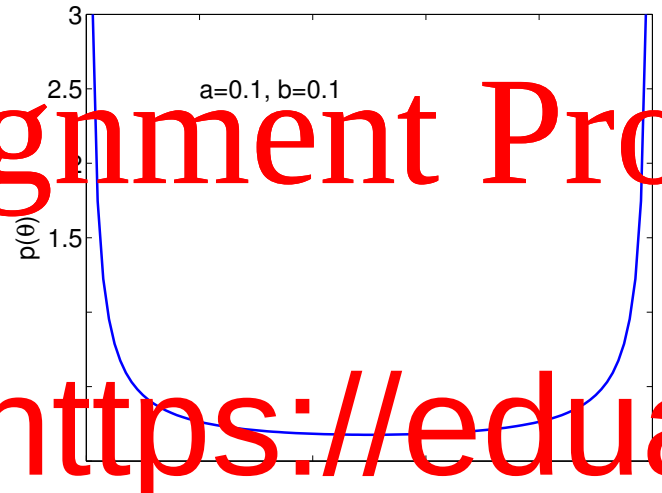
$$\text{Beta}(\theta|a, b) = \frac{1}{Z(a, b)}\theta$$

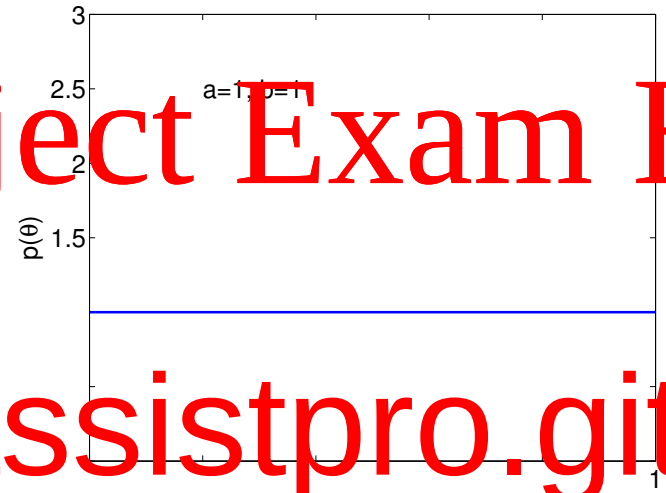
where $Z(a, b)$ is a suitable normaliser

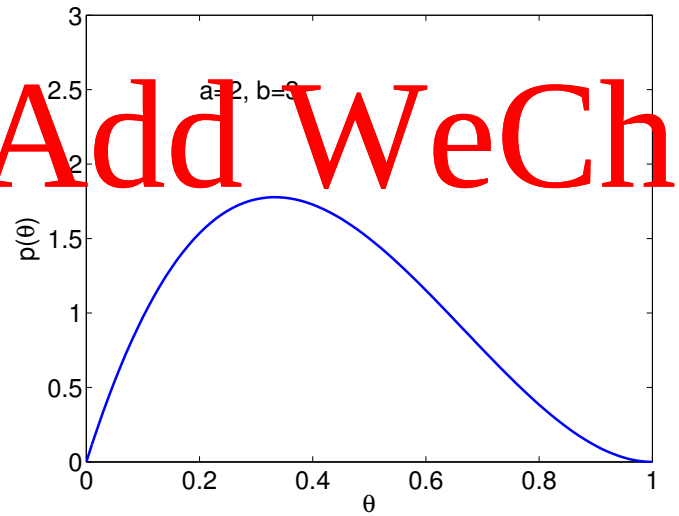
We can tune a, b to reflect our belief in the range of likely values of θ

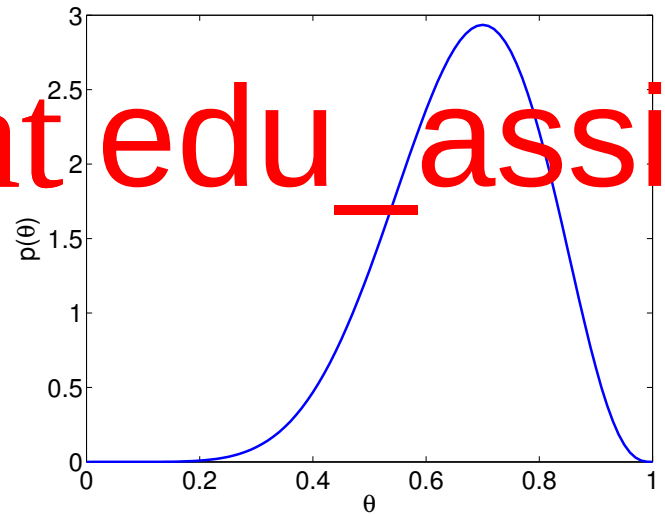
Beta Prior

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 $m \overset{\text{def}}{=}$

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 $m \overset{\text{def}}{=}$

For the pri

$$p(\theta|\mathcal{D}, a, b) = \frac{\mathcal{D}|}{p(\mathcal{D}|}$$

$$= \frac{p(\mathcal{D}|}{\int_0^1 p(\mathcal{D}|}$$

$$= \text{Beta}(\theta|m+a, \ell+b).$$

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 $m \overset{\text{def}}{=}$

For the pri

$$p(\theta|\mathcal{D}, a, b) = \frac{\mathcal{D}|}{p(\mathcal{D}|}$$

$$= \frac{p(\mathcal{D}|}{\int_0^1 p(\mathcal{D}|}$$

$$= \text{Beta}(\theta|m+a, \ell+b).$$

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

Beta Prior and Binomial Likelihood:

Beta Posterior Distribution

Now suppose we choose θ_{MAP} to maximise $p(\theta|\mathcal{D})$
(MAP= Maximum *A Posteriori*)

One can show that

cf. the esti

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

The prior parameters *a* and *b* can be seen a

What values of *a* and *b* ensure $\theta_{\text{MAP}} = \theta_{\text{ML}}$? $a = b = 1$. Make sense?
(Note that the choice of the beta distribution was not accidental here — it is the “conjugate prior” for the binomial distribution.)

Summary

- Distributions involving binary random variables
  - Bernoulli distribution

  -
- Bayesian inference: Full posterior on the par
  - Beta prior and binomial likelihood

- Reading: Mackay §23.1 and §23.5; Bis

Next time

- Ent