SPR day 2

Topic: Maximum Likelihood Estimation for Bernoulli Likelihoods

Continuous Random Variables

Reading:

Bishop Sec. 1.2.3 Contrast ML & Bayesian approach

Bishop Sec. 2.1 ML estimation

Outline:

(1) Bernoulli Distributions
(2) Independence Assumptions to simplify models
(3) ML estimation for Bernoulli coin flips
(4) Pros & Cons of ML estimation
(5) Continuous R.V.s, PDFs and CDFs

Binary Random Variables & Bernoulli Distribution

Let X be random variable indicating outcome of a coin toss. We are not sure if coin is fair.

Sample space?

Let $X=1$ indicate "heads" so $\Omega = \{0,1\}$

$X=0$ indicate "tails"

Probability Mass Function?

$p(X=1) = \mu$ μ is a parameter

$p(X=0) = 1-\mu$ $0 \leq \mu \leq 1$

This definition ensures the PMF is valid (sums to 1, all entries ≥ 0)

We can write the PMF in two ways:

$$p(X=x|\mu) = \begin{cases} \mu & \text{if } x=1 \\ 1-\mu & \text{if } x=0 \end{cases}$$ as 2 cases

$$= \mu^x (1-\mu)^{1-x}$$ as 1 function

The name for this distribution is Bernoulli.

Bernoulli Distribution Facts

Let $X \sim \text{Bernoulli}(\mu)$. $p(x|\mu) = \text{BernPMF}(x|\mu)$
$= \mu^x (1-\mu)^{1-x}$

What is mean? $\mathbb{E}[x] = \sum_{x \in \{0,1\}} p(x|\mu)\, x$

$= \mu^0 (1-\mu)^1 0 + \mu^1 (1-\mu)^0 1$

$= \mu$

What is $\mathbb{E}[x^2] = \sum_{x \in \{0,1\}} p(x|\mu)\, x^2$

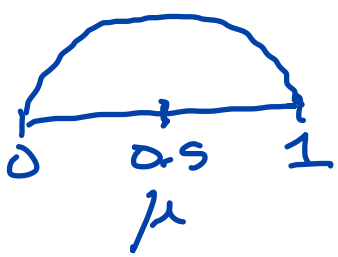
$= \mu^0 (1-\mu)^1 0^2 + \mu^1 (1-\mu)^0 1^2$

$= \mu$

What is Variance? $= \mathbb{E}[x^2] - \mathbb{E}[x]^2$ recall identity $\text{Var} = \mathbb{E}[x^2] - \mathbb{E}[x]^2$

$= \mu - \mu^2$

$= \mu(1-\mu)$



maximum variance when $\mu = 0.5$ (fair coin)

Models for Many Coin Flips

We perform N coin tosses.

Model each w/ Bernoulli random variable

$$X_1, X_2, \cdots X_N$$

We can model the joint of all tosses in several ways.

assumptions	joint PMF	number of free parameters
most general	$p(x_1, x_2, \cdots x_N)$	$2^N - 1$
each toss is independent	$p(x_1)\,p(x_2)\cdots p(x_N)$ Product of N separate Bernoulli PMFs	N one param $0 \le \mu_n \le 1$ for each toss
each toss independent and identically distributed	$p(x_1\|\mu)\,p(x_2\|\mu)\cdots p(x_N\|\mu)$ N Bernoulli PMFs that all share one parameter	1 one parameter $0 \le \mu \le 1$

total number of scalar values we need to define to make a unique valid PMF

each N-digit binary config has own probability

2^N total configs, but need to sum to 1 so last number is not free but determined

Assumptions can simplify models, but need to carefully decide when appropriate.

Estimating parameters from data

Given N observations of coin flip outcomes

$x_1, x_2, \dots x_N$ where $x_n \in \{0, 1\}$.

Assume an i.i.d model

$$p(X_1 = x_1, X_2 = x_2, \dots X_N = x_N) = \prod_{n=1}^{N} \text{BernPMF}(x_n | \mu)$$

$$= \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{1-x_n}$$

Goal: What value for μ plausibly generated this data?

To answer this question, let's pick a strategy

Maximum likelihood → Find a value of μ that makes our observations most likely (highest probability) under assumed model

Formally, as an optimization problem:

$$\hat{\mu} = \underset{\mu \in [0,1]}{\text{argmax}} \prod_{n=1}^{N} \text{BernPMF}(x_n | \mu)$$

- $\hat{\mu}$: value of μ that achieves best objective
- μ: variable to be optimized
- $[0, 1]$: space to search for allowed solutions
- $\prod_{n=1}^{N} \text{BernPMF}(x_n | \mu)$: objective function $J(\mu)$

Numerical Issues w/ Maximum Likelihood

Consider evaluating the ML objective in general

$$J(\mu) = \prod_{n=1}^{N} \text{BernPMF}(x_n|\mu)$$

product of N numbers between 0 and 1

What can happen in a practical computer with large N?

Even if each PMF evaluated to 0.9,

we'd have $J(\mu) = 0.9 \cdot 0.9 \cdots 0.9$

$$= 0.9^N$$

$$= \begin{cases} 2.65 \times 10^{-5} & \text{when } N = 10^2 \\ 6 \times 10^{-45} & \text{when } N = 10^3 \\ 6 \times 10^{-45} & N = 10^4 \end{cases}$$

Real results in NumPy using float32 precision

The problem is that the product 0.9^N is way too small

problem! inaccuracies arise in real implementations

Will either underflow to 0.0 or give wrong results.

Solution: Work in log space instead.

if all PMFs = 0.9

Likelihood $J(\mu) = \prod_n \text{BernPMF}(x_n|\mu)$ $\approx 0.9^N$

log likelihood $\mathcal{L}(\mu) = \log J(\mu) = \sum_{n=1}^{N} \log \text{BernPMF}(x_n|\mu)$ $\approx N \log 0.9$

much easier to represent in computer for large N!

Maximizing log likelihood finds the same optimal parameter $\hat{\mu}$ as maximizing likelihood because log is a monotonic increasing function

$$\hat{\mu} = \underset{\mu}{\text{argmax}}\ \mathcal{L}(\mu) = \underset{\mu}{\text{argmax}}\ J(\mu)$$

Solving ML Optimization for Coin Flips

Given N observations, we want to solve

$$\hat{\mu} = \underset{\mu \in [0,1]}{\operatorname{argmax}} \underbrace{\sum_{n=1}^{N} \log \text{BernPMF}(x_n | \mu)}_{\mathcal{L}(\mu)}$$

$$= \underset{\mu \in [0,1]}{\operatorname{argmax}} \sum_{n=1}^{N} \log \left[\mu^{x_n} (1-\mu)^{1-x_n} \right]$$

$$= \underset{\mu \in [0,1]}{\operatorname{argmax}} \sum_{n=1}^{N} x_n \log \mu + \sum_{n=1}^{N} (1-x_n) \log (1-\mu)$$

$$= \underset{\mu \in [0,1]}{\operatorname{argmax}} \; s(x) \log \mu + r(x) \log (1-\mu)$$

$$s(x) = \sum_n x_n \geq 0$$

$$r(x) = \sum_n 1 - x_n \geq 0$$

We have reduced to a constrained optimization problem

Solutions $\hat{\mu}$ must lie in interval $[0,1]$

Generally, two ways to satisfy constraint

(1) solve as if no constraint, see if answer satisfies → Our approach today

(2) method of Lagrange multipliers (later class)

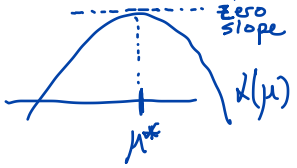
Solving ML optimization for coin flips

Ignoring constraints:

$$\mu^* = \underset{\mu}{\operatorname{argmax}}\; s(x)\log\mu + r(x)\log(1-\mu)$$

Calculus tells us that at an optima μ^*, first derivative evaluated at μ^* equals zero

$$\ell'(\mu^*) = 0$$



Technically should do 2nd derivative test to verify if maxima or minima. You can trust us, it is global max.

So, we set $\ell'(\mu) = 0$ and solve for μ

$$\ell'(\mu) = \frac{\partial}{\partial\mu}\ell(\mu) = \frac{s}{\mu} - \frac{r}{1-\mu}$$

(set to zero)

$$\frac{s}{\mu} - \frac{r}{1-\mu} = 0$$

$$\frac{s}{\mu} = \frac{r}{1-\mu}$$

$$s(1-\mu) = r\mu$$

$$s = (s+r)\mu$$

(solve for μ)

$$\mu^* = \frac{s}{s+r}$$

Check. Does solution to unconstrained problem solve our constaints?

Yes!

$$\mu^* = \frac{s(x)}{s(x)+r(x)} = \frac{\sum_n x_n}{\sum_n x_n + \sum_n (1-x_n)} = \frac{\sum_n x_n}{N}$$

($\sum_n x_n$: #1s; $\sum_n (1-x_n)$: #0s)

$$= \frac{\text{number of 1s in data}}{\text{total size of data}}$$

μ^* cannot be larger than 1 ✓ smaller than 0 ✓

Advantages of ML estimation

Maximizing the likelihood as a strategy to learn parameter values has several advantages

(1) ML estimates are consistent

This means roughly that if
(1) our assumed likelihood distribution matches the true data-generating process
(2) we have enough data
that our ML estimate is guaranteed to recover the "true" data-generating parameter.

Suppose we pick some μ^{true},
and then draw N observations $x_n \sim p(x|\mu^{true})$
As $N \to \infty$, we can prove that $\hat{\mu}^{ML}(x_1 \cdots x_N) \to \mu^{true}$.

(2) ML estimates are equivariant to different parameterizations

3 possible parameterizations for coin toss example

A: $\mu \in [0,1]$ $\quad p(x|\mu) = \prod_{n=1}^{N} Bern(x_n|\mu)$

B: $s \in [0,1]$ $\quad p(x|s) = \prod_{n} Bern(x_n|s^2)$

C: $r \in \mathbb{R}$ $\quad p(x|r) = \prod_{n} Bern(x_n|\sigma(r))$ where $\sigma(r) = \frac{e^r}{1+e^r}$ sigmoid function

Can show that as long as there is an invertible mapping between parameterizations, then can estimate the ML parameter for any model (A, B, C) and apply map to get any other model's ML estimate

Problems with ML Estimation

ML estimation has several limitations.

(1) Overfitting

Suppose you observe $N=3$ coin flips. All are heads.

$$\hat{\mu}(x_1, x_2, x_3) = \hat{\mu}(1,1,1) = \frac{3}{3} = 1$$

Do we really believe all future flips will be heads?

ML is vulnerable to overfitting on small datasets

(2) Difficulty of obtaining a solution

Some models (like coin flip) allow closed-form formula. Easy to use!

More complex models do not enjoy closed-form (the set gradient to zero and solve strategy doesn't simplify)

Instead, we use more expensive numerical estimation techniques such as gradient descent.

(3) Is solution unique? Not always guaranteed.

Fundamentals: Continuous Random Variables

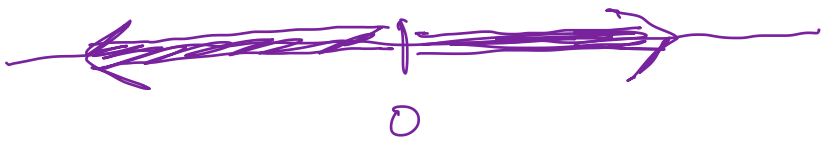
Thus far we've discussed discrete rand. vars.

But many random quantities take continuous values

Height of a random student

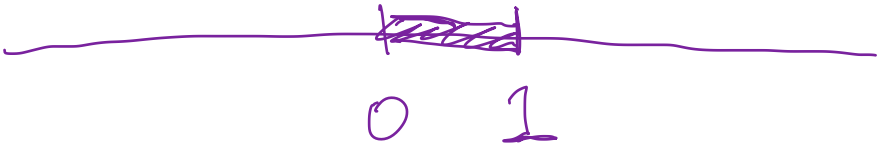
$h \in (0, \infty)$

Distance of random student from Medford (along latitude only)



$d \in (-\infty, +\infty)$

Probability that a coin ends up heads



$\mu \in [0, 1]$

Sample spaces are continuous intervals of real line.

Can coherently talk about probability of interval events

event: $h \in (a, b)$ — h is between a and b

probability: $p(H \in (a, b))$ or $p(a \leq H \leq b)$

Need different language for specific values.

Think about limits as interval size goes to zero

$$\lim_{\delta \to 0} \frac{p(H \in (h, h+\delta))}{\delta} = \frac{\text{proba of tiny interval}}{\text{width of tiny interval}}$$

Probability Density Functions

Let random variable X have continuous sample space $\mathbb{R}$.

Define the cumulative distribution function F for X as:

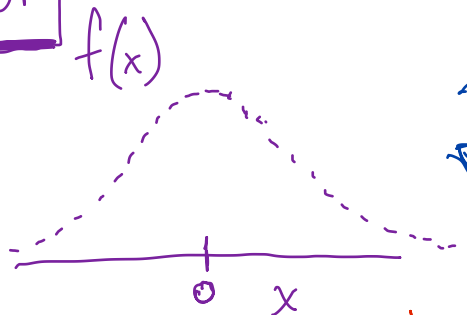
$$F(b) = p(X \leq b)$$

cumulative (aggregate) probability that x is less than value b.

We define the PDF f of X as the function satisfying

$$f(b) = \lim_{\delta \to 0} \frac{F(b+\delta) - F(b)}{\delta} = \frac{\partial}{\partial x} F(x) \Big|_{x=b}$$

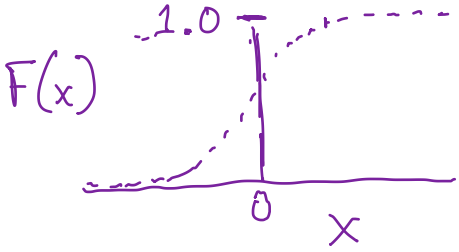
PDF measure density at value



integral

derivative

CDF measure mass over interval



no limit on maximum value
always non-negative

$$f(x) \geq 0$$

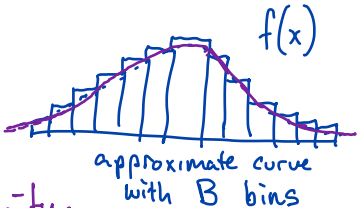
must integrate to one

$$\int_{x \in X} f(x)\,dx = 1$$

always non-negative
strictly increasing
minimum value 0.0
maximum value 1.0

Remember, PDFs indicate density.

$$\text{density} = \frac{\text{mass}}{\text{volume}}$$



Think of integral as a Riemannian sum as # bins goes to infinity.

$$\int_{x \in X} f(x)\,dx \approx \lim_{\Delta x \to 0} \sum_{b=1}^{B} f(x_b)\,\Delta x = 1$$

$f(x_b)$: density of bin b; Δx: volume of bin b; 1: mass

Density function can be larger than one. Implies very small volume. Example: For any $\varepsilon > 0$, Uniform over $(0, +\varepsilon)$ has pdf $\frac{1}{\varepsilon}$

Sum and Product Rules for continuous r.v.

Let $X \in \mathbb{R}$, $Y \in \mathbb{R}$ be random variables.

We can define joint PDF as $p(x,y)$, which satisfies

- non-negative: $p(x,y) \geq 0$
- integrates to one: $\int_{y \in \mathbb{R}} \int_{x \in \mathbb{R}} p(x,y)\, dx\, dy = 1$

Key Takeaway: just replace sums with integrals

Sum Rule:

$$p(x) = \int p(x,y)\, dy$$

$$p(y) = \int p(x,y)\, dx$$

Product Rule:

$$p(x,y) = p(x|y)\, p(y)$$

$$= p(y|x)\, p(x)$$