SPR Day 2

Readings from Bishop PRML

1.2.1 Probability Density functions

1.2.2 - 1.2.3 Expectations + Bayesian probabilities

2.1 Binary variables

2.1.1 Beta distribution

Gamma functions

Goals

Review ML estimation for Bernoulli data

Review continuous r.v. basic definition

Introduce Beta random variables

PDF formula

What is a Gamma function?

Introduce Beta-Bernoulli model & MAP estimation

interpret α, β parameters as pseudo counts

Next time Big Idea: Bayesian predictive distribution

Binary Random Variables & the Bernoulli distribution.

Consider observing outcome of a random coin toss
We are not sure if coin is fair or not.
Let X be a random variable indicating outcome
Sample space?

1 "heads"

0 "tails"

Probability Mass function?

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

where $0 \leq \mu \leq 1$

The variable μ here is a "parameter", not a "random variable" (yet)

We will often write the PMF as a conditional

$$p(X=1 \mid \mu),$$

just to remind ourselves that μ is required for that calculation

The name for this kind of distribution is Bernoulli

$$p(X=x \mid \mu) = \text{BernPMF}(X=x \mid \mu) = \mu^x (1-\mu)^{1-x}$$

$$= \begin{cases} \mu & \text{if } x=1 \\ 1-\mu & \text{if } x=0 \end{cases}$$

Expectation

Given a discrete random variable X,
we might have a function of X in mind for a task
(e.g. the cost of the fruit selected,
the points I get for each dice role,
etc)

Let $f(x)$ map from sample space of X
to a real value or
vector of real values.

We want to know the "average" value of $f(x)$
weighting each outcome by its probability.

$$\mathbb{E}[f(x)] = \sum_{x \in \mathcal{X}} P(X=x) f(x)$$

When each value x is represented as a real value
or vector,
we can compute the mean of random variable X

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} P(X=x)\, x$$

Expectations of Bernoulli random vars.

$X \sim \text{Bern}(\mu)$

$$\mathbb{E}[X] = \sum_{x \in \{0,1\}} x \, p(X=x)$$

$$= \underline{1} \cdot \mu + 0 \cdot (1-\mu)$$

$$= \mu$$

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

$$= \underline{1}^2 \cdot \mu + 0^2 \cdot (1-\mu)$$

$$= \mu$$

Recall variance is measure of how much, on average, a random variable's value will differ from its mean.

X is R.V.

$$\text{var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad \text{by definition}$$

$$= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$$

$$= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Thus, variance of a Bernoulli:

$$\underset{X \sim \text{Bern}(\mu)}{\text{var}}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mu - \mu^2 = \mu(1-\mu)$$

Independence & Conditional Independence

Two random variables X and Y are "independent" if the distribution of Y given knowledge of X is the same as the marginal distribution of Y

$$Pr(Y=y \mid X=x) = Pr(Y=y) \quad \text{for all } y, x$$

That is, the distribution of Y never depends on the value of X

The joint distribution of two indep. r.v.s looks like a product of marginals

$$Pr(X=x, Y=y) = Pr(Y=y \mid X=x) Pr(X=x) \quad \text{product rule}$$

$$= Pr(Y=y) Pr(X=x) \quad \text{defn of independence}$$

Conditional Independence

We say two rand vars X and Y are "conditionally independent" given a third random variable Z iff

$$Pr(Y=y \mid X=x, Z=z) = Pr(Y=y \mid Z=z) \quad \text{for all } x, y, z \text{ values}$$

again, this means the joint distr. looks like

$$Pr(X=x, Y=y, Z=z) = Pr(Y=y \mid Z=z) Pr(X=x \mid Z=z) Pr(Z=z)$$

Models for (independent) coin-flips

Suppose we coin toss N times, the outcome can be represented as

$$X_1 = x_1,\ X_2 = x_2 \cdots \quad X_N = x_N$$

In general, how many parameters are needed to describe this joint distribution?

ans: $2^N - 1$

Joint proba mass function requires 2^N table entries

but sum must be one so last entry is determined

Now, let us assume each toss is independent

$$p(X_1 = x_1, \cdots X_N = x_n) = \prod_{n=1}^{N} p(X_n = x_n)$$

How many parameters? N

Further assume identically distributed $\longrightarrow$

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

How many params? 1

Insight: We always observe N values

Assumptions like independence & identical distribution help us reduce number of parameters we need to estimate/learn

Estimation of μ from data

Given N observations, $X_1 = x_1 \cdots X_N = x_n$, of binary outcomes, and assuming model $\prod_{n=1}^{N} \text{Bern}(X_n = x_n | \mu)$, what is value of μ?

What principle should we use?

A good one: maximize likelihood

$$\hat{\mu} = \underset{\mu \in [0,1]}{\arg\max} \prod_{n=1}^{N} \text{Bern}(X_n = x_n | \mu)$$

"find value of μ that makes our observed data most likely / most probable under the model"

Issue: $\prod_{n=1}^{N} \text{Bern}(x_n | \mu)$ is a product of N numbers between 0 and 1

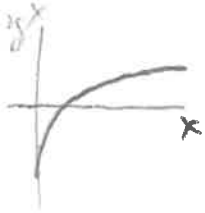
What can happen in a computer?

underflow!

$0.9 \cdot 0.9 \cdot 0.9 \cdot \cdots = 0.0$ using finite precision

but should be > 0.0

Correction: for any objective function $f(x)$, $\mathbb{R} \to \mathbb{R}$, consider $\log f(x)$ as objective instead



$\log(a)$ is a monotonic function, so

$$\underset{x}{\text{argmax}}\ f(x) = \underset{x}{\text{argmax}}\ \log f(x)$$

Maximizing log likelihood for ML estimation

$$\hat{\mu} = \underset{\mu \in [0,1]}{\operatorname{argmax}} \sum_{n=1}^{N} \log \text{BernPMF}(X_n = x_n \mid \mu)$$

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

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

where $s(x) = \sum_{n} x_n \geq 0$

$r(x) = N - s(x) \geq 0$

This is a constrained optimization problem, b/c valid solutions μ must be within $[0, 1]$

In general, 2 ways to solve

(1) act like there is no constraint, hope that answer works

(2) method of Lagrange multipliers (next class)

$\underset{\mu}{\operatorname{argmax}} \; f(\mu)$, $\quad f(\mu) = s \log \mu + r \log(1-\mu)$

set $\frac{\partial f}{\partial \mu} = 0$ and solve for μ

then, verify second derivative is negative so μ^* is a maximum.

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

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

$$s - s\mu = r\mu$$

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

$$\boxed{\mu^* = \frac{s}{s+r} = \frac{\sum_{n} x_n}{N}}$$

does this meet our constraint? yes!

Problems w/ ML estimation

Suppose 3 coins all heads

$$\hat{\mu}^{ML} = 1.0$$

do we believe that?

ML is vulnerable to overfitting on small training sets

Advantages of ML estimation

Can prove several general properties (we won't prove, just good to know)

(1) ML estimates are consistent

Suppose I generate data from coin flip model with μ^{true}

then, you estimate μ^{ML} from data

can prove that as $N \to \infty$, $\mu^{ML} = \mu^{true}$

(2) ML estimates are equivariant to different parameterizations

model A: $\prod_n Bern(X_n = x_n \mid \mu) \quad \mu \in [0, 1]$

B: $\prod_n Bern(X_n = x_n \mid s^2) \quad s \in [0, 1] \quad \mu(s) = s^2$

C: $\prod_n Bern(X_n = x_n \mid \sigma(r)) \quad r \in \mathbb{R}, \ \sigma(r) = \frac{e^r}{1+e^r}$

As long as you can map parameters to one-another,

$$\mu^{ML} = map(s^{ML}) = map(r^{ML})$$

Continuous Random Variables

Examples

Probability coin ends up heads

μ sample space $\mu \in [0,1]$

Height of avg. student

h $h \in [0, \infty)$

Distance of random location from Medford

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

Sample spaces are continuous intervals of real line

Can talk coherently about the probability of interval events

what's probability that height is above 75 cm

$$p(H \geq 75)$$

Need different language for specific events

should not say what is probability height equals 75

instead, think about limits

$$\text{pdf}(H=75) = p(H \in [75, 75+\delta]) \text{ as } \delta \to 0$$

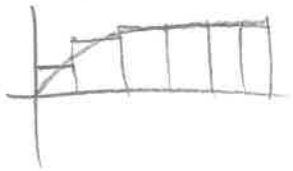
↑ probability density function of random variable H

Two rules for a PDF function for continuous r.v. X

$$pdf(x) \geq 0 \quad \text{for all outcomes } x \in \mathcal{X}$$

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

Think of the Riemannian sum of this integral



$$\lim_{\Delta x \to 0} \sum_{n=1}^{N} pdf(x_n)\,\Delta x_n = 1$$

density of each bin | volume of each bin | mass

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

Can density be larger than one? Yes, if mass of bin is small enough

Joint probabilities also obey these rules

$X \in \mathbb{R}$, $Y \in \mathbb{R}$

$$pdf(x, y) \geq 0 \quad \text{and} \quad \iint pdf(x,y)\,dx\,dy = 1$$

Sum Rule

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

Product Rule

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

Beta distribution

Random variable μ

Sample space $\mu \in [0, 1]$

PDF function parameters $a > 0$, $b > 0$

$$\text{BetaPdf}(\mu|a,b) = \underbrace{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}_{\text{const. wrt } \mu} \underbrace{\mu^{a-1}(1-\mu)^{b-1}}_{\text{"interesting part" is a function of } f(\mu|a,b)}$$

we'll call this "normalizing constant"

Gamma function: see notebook online.

$$\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$$

$\Gamma : \mathbb{R} \to \mathbb{R}$

scalar → scalar real

Remember:

$$\int_0^1 \text{BetaPDF}(\mu|a,b)\, d\mu = 1$$

$$\int_0^1 c(a,b)\, f(\mu|a,b)\, d\mu = 1$$

implies

$$\int_0^1 f(\mu|a,b)\, d\mu = \frac{1}{c(a,b)}$$

thus,

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}\, d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Now, consider a joint model for
μ and the coin observations $x_1 \cdots x_N$

$$p(x_1, x_2 \cdots x_N, \mu) = \text{BetaPDF}(\mu | a, b) \prod_{n=1}^{N} \text{Bern}(x_n | \mu)$$

discrete continuous

We observe $x_1 \cdots x_N$, how to estimate μ?

several ways:

(1) find μ that maximizes $p(\mu | x_1 \cdots x_N)$
called maximum a-posteriori (MAP)

(2) take a more Bayesian approach (next time)

MAP estimator

$$\mu^* = \underset{\mu \in [0,1]}{\arg\max} \; p(\mu | x_1 \cdots x_N)$$

$$= \quad " \quad \frac{p(\mu) p(x_1 \cdots x_N | \mu)}{\text{const wrt } \mu}$$

$$p(x_1 \cdots x_N) = \int_\mu p(\mu, x) d\mu$$

$$= \underset{\mu \in [0,1]}{\text{argmax}} \; \log \text{BetaPDF}(\mu | a, b) + \sum_n \log \text{BernPMF}(x_n | \mu)$$

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

MAP estimator (cont'd)

$$\mu^* = \underset{\mu \in [0,1]}{\arg\max} \; \underbrace{(s(x)+a-1)}_{s'} \log \mu + \underbrace{(r(x)+b-1)}_{r'} \log(1-\mu)$$

We've solved this before!

$$\mu^* = \frac{s'}{s'+r'} \quad \text{when} \quad \begin{matrix} s' \geq 0 \\ r' \geq 0 \end{matrix}$$

$$\mu^* = \frac{s(x)+a-1}{\underbrace{N}_{=s(x)+r(x)} + a + b - 2}$$

requires
$a \geq 1$
$b \geq 1$

otherwise MAP does not exist

Interpretation

a acts like a "pseudocount" for heads

b acts like a "pseudocount" for tails