

Bayesian ML in Python

part 1 Bayes rule review

$$\text{rule} = \frac{P(A|B) = P(A, B)}{P(B)}$$

conditional

Joint

Marginal

example:

as $P(CA)$, $P(US)$, $P(MX)$?

		CA	US	MX	Σ	
		Buy	20	50	10	80
Buy	?	300	500	200	1000	
	Σ	320	550	210	$\Sigma 1080$	

$$P(CA) = \frac{320}{1080}$$

$$P(US) = \frac{550}{1080}$$

$$P(MX) = \frac{210}{1080}$$

note that all $P(x)$'s add to 1

$$P(Buy | MX) = \frac{P(Buy, MX)}{P(MX)} = \frac{\cancel{20}}{\frac{320}{1080}}$$

etc.

note that as a space \uparrow in size the $P \downarrow$

Sometimes P is so small that computer rounds to 0: log-probability used instead

Independence: $P(A, B) = P(A)P(B)$

- joint will be multiple of marginal unrelated events

$$P(\text{Buy} | \text{Country}) = \frac{P(\text{Buy} | \text{country})}{P(\text{country})}$$

$$= \frac{P(\text{Buy}) P(\text{country})}{P(\text{country})} = P(\text{Buy})$$

If $P(A|B) = \frac{P(A, B)}{P(B)}$ & $P(B|A) = \frac{P(B, A)}{P(A)}$

& $P(A, B) = P(B, A)$

then
$$P(A|B) = \frac{P(B|A) P(A)}{P(B)}$$

More
common
form.

VS : $P(B) = \sum_A P(A, B) = \sum_A P(B|A)P(A)$

(?)
Review
this

Becomes integral
w/ cont. dist.

Probability exercise

Fair coin $P(H) = P(T) = 0.5$

↳ 20 tosses, 15H & 5T

↳ what do we expect after 200 tosses?

↳ 150H & 50T?

X no.

$$200 - 20 = 180$$

↳ even outcomes = 90H, 90T

$$\therefore 90H + 15H = 105H \text{ & } 95T$$



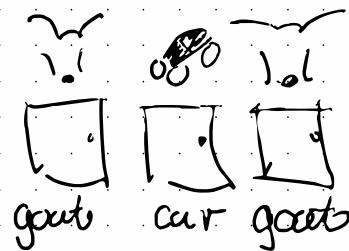
due to
independence.

Monty Hall problem

1. pick a door →

2. Monty reveals a goat

3. do you change?



When trying to calculate its chance to
analyze an ~~18~~ instance where an ~~car~~ fratty
door is chosen.

$C = \text{loc of car}$

$$P(H=2 | C=1) = 0.5$$

$H = \text{door M/H opens}$

$$P(H=2 | C=2) = 0$$

• we pick door 1

$$P(H=2 | C=3) = 1$$

Need to calculate

$$P(C=1 | H=2), P(C=3 | H=2)$$

prob of
occurrence

$$P(C=3 | H=2) = \frac{P(H=2 | C=3) \cdot P(C=3)}{\sum_{n=1}^3 p(H=2 | C=n) p(C=n)}$$

(2)

Normalized
over all poss.

$$= \frac{1 \times \frac{1}{3}}{(1 \times \frac{1}{3}) + (0 \times \frac{1}{3}) + (0.5 \times \frac{1}{3})}$$

$$= \frac{1}{3} - \left(\frac{1}{3} + \frac{0.5}{3} \right)$$

$$= \frac{1}{3} \times \frac{1}{1.5} = \frac{1}{1.5} = 0.6$$

you can show similarly that ie $0.5 \times \frac{1}{3}$

$$P(C=1 | H=2) = 0.3 \quad P(H=2 | C=1) P(C=1)$$
$$\frac{1}{3} + 0 + \frac{0.5}{3}$$

∴ you should switch.

$$= \frac{0.5}{3} \times \frac{3}{1.5} = 0.3$$

$$= \frac{0.5}{1.5} = 0.3$$

Imbalanced Classes

↳ classifier for disease

what is good accuracy? ↗ 80%
↗ 90% ↗ 95%

lets say $P(d) = 0.01$

if we say "no" always we get accuracy of 99%

we care about TP rate (sensitivity)

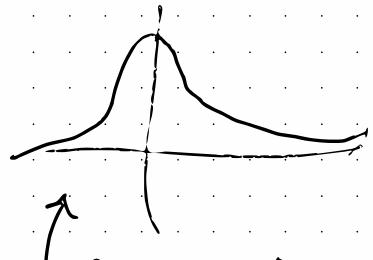
$$P(\text{prob} = 1 \mid d = 1) = \frac{P(\text{prob} = 1, d = 1)}{P(d = 1)}$$

<u>Metrics</u>	<u>equation</u>				
		Pres 1	Pres 0		
sensitivity	$\frac{TP}{TP + FN}$	d = 1	TP	FN	
specificity	$\frac{TN}{TN + FP}$	d = 0	FP	TN	
precision	$\frac{TP}{TP + FP}$	probability of disease given prediction is ?			
recall	$\frac{FN}{FN + TN}$	Check that this is right.			

Mean of gaussian distributed data

"maximum likelihood"

↳ what is this?



Joint prob.

$$p(x_1 \dots x_n) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$

$$\therefore p(x_1 \dots x_n | \mu) = \prod_{i=1}^N p(x_i | \mu)$$

↑ parameters depend on distribution

What are the parameters such that p is biggest?

$$L = \log p(x | \mu) = \sum_{i=1}^N \frac{-(x_i - \mu)^2}{2\sigma^2} + C \text{ w.r.t. } \mu$$

$$\frac{\partial L}{\partial \mu} = \frac{1}{N} \sum_{i=1}^N x_i \quad \begin{matrix} \leftarrow \text{Mean?} \\ \text{How did teacher} \\ \text{get from here to} \\ \text{there?} \end{matrix}$$

Drawbacks: we don't know precision.

Max Likelihood GTR

H = circle, T = -circle IID

↳ if we flip 2H & 3T what is total likelihood

$$S\left(\frac{1}{2}\right)? \quad L(N_H, N_T) = p^{N_H} (1-p)^{N_T}$$

$$= (0.5)^2 (1-0.5)^3$$

$$= \left(\frac{1}{2}\right)^5$$
 ← Multiply because they are all independent coin tosses

• What is max likelihood est. of p ?

$$L = \log(p^{N_H} (1-p)^{N_T})$$

$$\frac{\partial L}{\partial p} = \frac{N_H}{p} - \frac{N_T}{1-p} = 0 \quad \therefore \quad \frac{N_H}{N_H + N_T} \\ = \frac{2}{5}$$

↳ don't know how precise this is

Concept: log transformations make calculus easier

Confidence intervals

$$\bar{A} = \frac{1}{N} \sum_{i=1}^N x_i$$

↑ sum of random variables

$Z = X + Y$ where X & Y are also random

• Is Z random? yes. pdf, μ , var, etc.

$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{N}\right)$$

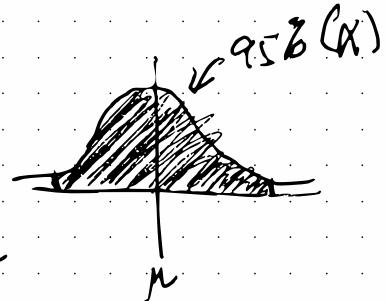
↑ ↑
mean var

CLT - iid random
vars will tend
towards this

$1-\alpha$ = confidence level

α = significance

(normally 0.05 but can be smaller)



$$0.95 = \int_{x_L}^{x_R} N(x; \mu, \frac{\sigma^2}{N}) dx$$

$$CI = \left[\mu + z_{\text{left}} \frac{\sigma}{\sqrt{n}}, \mu + z_{\text{right}} \frac{\sigma}{\sqrt{n}} \right]$$

↑ don't know sigma
can approx using T dist.

About bayesian paradigm

Frequentist

- Parameters are set & we don't know them
- data is generated according to these param.

$$\hat{\theta} = \arg \max_{\theta} P(X|\theta)$$

Bayesian

- Parameters are rand. w/ distributions
- Data is fixed
- model $p(\text{param.} | \text{data})$

$$P(\theta | X)$$

A/B Testing

web page \rightarrow conversion kpis (1/0)

$$\hookrightarrow \text{rate} = \frac{\# \text{conversion}}{\# \text{users}}$$

Idea: track rates across different versions of page.

"is difference in conversion rate between pages

stat. sig. @ sig.level α ?"

remember, α is usually 0.05 or 0.01.

Null hypothesis: Assumed outcome $H_0: \mu_1 = \mu_2$

Alternative: Measured outcome $H_1: \mu_1 \neq \mu_2$ two tailed

or

$\mu_1 > \mu_2$ one tailed

Simple A/B testing

→ difference in height of ♂ & ♀

↳ 2 lists of weights: \vec{m} & \vec{w}

$$\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_n \end{bmatrix} \quad \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

$l \times n \quad l \times n$

test stat:

$$t = \frac{\bar{m} - \bar{w}}{s_p \sqrt{\frac{2}{n}}} \quad \text{where } s_p = \sqrt{\frac{s_m^2 + s_w^2}{2}}$$

std dev

↑ unbiased, t instead of $\frac{1}{n-1}$
 not normally dist vs t-dist.

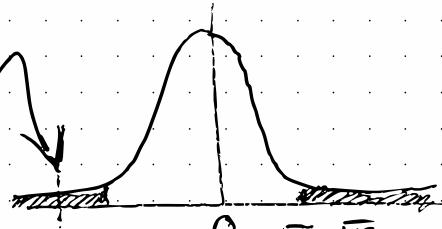
• PDF for t-dist is calculated using d.o.f (v)

where $D = 2N - 2$

thus...

if t falls in shaded area
then we can reject H_0

p-value: likelihood of wrongly
rejecting H_0 .



\sum of remaining
outcomes becomes
p-value

NOTE: You CANNOT ACCEPT H_0 SINCE
IT IS AN ASSUMPTION!

Testing Characteristics

- Larger $N = \text{Bigger } t$, $t = \frac{\bar{w} - \bar{v}}{S_p \sqrt{\frac{1}{n_m} + \frac{1}{n_w}}}$
- More confident of larger samples
- Larger $S_p = \text{smaller } t$, less confident w/ large var.

→ What if two groups have different sizes?

$$t = \frac{\bar{w} - \bar{v}}{S_p \sqrt{\frac{1}{n_m} + \frac{1}{n_w}}} \quad \text{where; } S_p = \sqrt{\frac{(n_m-1)s_m^2 + (n_w-1)s_w^2}{n_m + n_w - 2}}$$

Makes assumption \rightarrow re weighted
that var is the same
for both groups

↳ If they are then use Welch's t-test

Note that we assumed Normally dist data as well.

↳ What if you don't know dist or its different?

↳ non-parametric tests. (less stat "power")

t-test exercise

'advertisement_clicks.csv' - contains data

↳ Step① do we need to use Bonferroni Corr.?

No need, only comparing
2 things.

↳ used to solve problem
of multiple comparisons

↳ Step② use t-test to determine if
one advertiser is better than the other
↳ save us diff variance

Comments: my t & p values are different.

↳ two tablets vs one tablet?

↳ right direction, ~~one vs~~ both are
statistically significant.

↳ seems to be the sum of both vectors.
ten

0.01 vs 0.011

↳ Does this difference matter?

Maybe, depends on how many people 0.1% of people are...

χ^2 test statistic

↳ Also works w/ CTR & any categorical vars where things are counted

$$\chi^2 = \sum_i \frac{(obs_i - exp_i)^2}{exp_i}$$

	Click	7 Clicks
A	36	14
B	30	25
	(66)	39

ie for click, $A * exp_i = \frac{50 * 66}{105}$ & $obs_i = 36$

↳ thus $\chi^2 = 3.418$

↳ can also be calculated by

simplifying gives

$$\frac{n(ad - bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

→ What is p-value?

↳ $1 - \text{CDF}(\chi^2)$ if $p < \alpha$ then $A \neq B$

A/B... tests

↳ possible to get significant p-value by chance \Rightarrow p-hacking

• Benferroni correction $\Rightarrow \alpha_{\text{new}} = \frac{\alpha}{\text{No tests}}$.

↳ note that number of tests will depend upon how you choose to compare values.

Statistical Power \geq sensitivity

β = false neg. rate $P(\text{ rej. } H_0 \mid H_0 \text{ is true})$

$1 - \beta$ = power \leftarrow also gives how many samples we should collect.

A/B testing pitfalls

\rightarrow p-value can change as experiment progresses. DON'T CHECK IT.

\rightarrow Sample size? $N = 16 \frac{\sigma^2}{\delta^2}$ $\begin{matrix} \leftarrow \text{variance} \\ \leftarrow \text{smallest difference} \end{matrix}$
Lots of size calculators online

All this is quite confusing...

Bayesian A/B testing

Explore / Exploit

(i) Multi-armed bandit



Which slot machine?

(ii) What is the balance between best?

Exploring new possibilities & exploiting knowledge that is already known.

Intro to Bayesian A/B testing

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} \leftarrow \text{prior}$$

↑ ↑ ← "norm. const."
 posterior likelihood

prob of param = prob of data given param + what we knew about param

prob. of data

↳ integral of all

$$\int P(x|\theta)P(\theta)d\theta$$

↑ hard to solve

Conjugate priors

If we choose specific distributions for $P(X|\theta)$ & $P(\theta)$ we can make $P(\theta|X)$ the same dist as the prior!

e.g. CTR is Bernoulli distributed

$$P(X|\theta) = \prod_{i=1}^N \theta^{x_i} (1-\theta)^{1-x_i}$$

try &
combine
these...

$$0 < \theta < 1$$

↑ p of click ... beta distribution!

$$\theta \sim \text{Beta}(a, b) = \frac{\theta^{a-1} (1-\theta)^{b-1}}{B(a, b)}$$

where

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad \& \quad \Gamma(n) = (n-1)!$$

↑ Gamma
function

Here, we choose beta dist. because it ranges from 0 to 1 & can resemble many distributions.

→ What are a & b? Priors, if no exp was run or the same as what you got before...

$P(X|\theta)$ is Bernoulli because its what we can change depending on what you want to predict. the dist follows

Anyway... $P(X|\theta) * P(\theta)$

$$P(\theta|X) \propto \left[\prod_{i=1}^N \theta^{x_i} (1-\theta)^{1-x_i} \right] \theta^{a-1} (1-\theta)^{b-1}$$

↳ combine like terms using
by prob

$$P(\theta|X) \propto \theta^{a-1 + \sum_{i=1}^N x_i} (1-\theta)^{b-1 + \sum_{i=1}^N (1-x_i)}$$

we can ignore
normalization ↑

same shape as beta dist.

$$\therefore P(\theta|X) = \text{Beta}(a', b') \quad \begin{aligned} &= b + \sum_{i=1}^n x_i \\ &\quad \uparrow \quad = b + \# \text{ no-chicks} \\ &= a + \sum_{i=1}^n x_i \\ &= a + \# \text{ chicks} \end{aligned}$$

NB: $E(\theta) = \frac{a}{a+b}$ as $a, b \uparrow$
var ↓

$$\text{Var}(\theta) = \frac{ab}{(a+b)^2(a+b+1)} \quad \swarrow$$

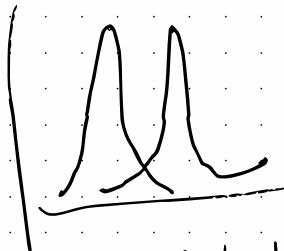
How do we choose a & b ?

$$\begin{aligned} a &= 1 ? \text{ uniform,} \\ b &= 1 \quad \{ \text{ all outcomes} \\ &\quad \text{ equally likely} \end{aligned}$$

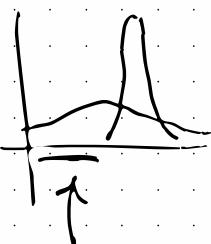
Bayesian A/B testing

How does sampling from $\text{Beta}(a', b')$ help us?

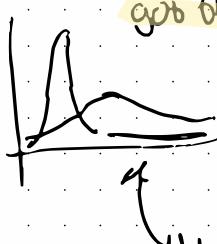
- ↳ organically solve multi armed bandit



equally likely to explore both



more likely to explore here



I probably
got this wrong

more likely to explore here

- ↳ Adjust accordingly. (Tompson sampling)
explore higher exp. low ver

Online Nature of Bayesian A/B Testing

Traditional A/B-T.

- ↳ Approximations
- ↳ Can't "peek" @ results
- ↳ Calculate sample sizes
- ↳ difficult to communicate
- ↳ lots of uninformative assumptions

Bayesian A/B-T.

- ↳ no prep necessary
- ↳ No threshold
- ↳ Naturally converges
- ↳ Updates for each obs.
- ↳ Close to how we think
- ↳ Multiple testing is easier
- ↳ can stop/review when you want.

Thresholds for Bayesian A/B T

ie $P(\mu_1 > \mu_2)$ where μ_i is Beta-distributed
 ↪ CTR

thus, $P(\mu_1 - \mu_2 > 0) = P(X > 0)$ given $\mu_1 - \mu_2 = X$

req PDF(X) = $\beta_1(x) * \beta_2(x)$ * conv. operator
 ↪ ↪ ↪

instead... Joint pdf = $P(\mu_1, \mu_2) = p(\mu_1) * p(\mu_2)$

↑ works due
to independence

thus, $P(\mu_1 > \mu_2) = \text{area under } p(\mu_1, \mu_2) \text{ where } \mu_1 > \mu_2$

$$\boxed{P(\mu_2 > \mu_1) = \sum_{i=0}^{\alpha_2-1} \frac{B(\alpha_1+i, \beta_1+b_2)}{(b_2+i) B(1+i, b_2) B(\alpha_2, b_2)}}$$

or...

↑ Quibe
the monster

$$L = \max(\mu_2 - \mu_1, 0) \rightarrow \text{Stop when } E_{\mu_1, \mu_2}(L) < \text{thresh}$$

Equation is too complicated, see lecture notes.

How does Bayesian Bandit work?

→ Bandit



$p \leftarrow \text{probability}$
 $a, b = (1, 1) \leftarrow \text{prior}$

3 actions:

- Pull
- Sample
- Update

Pull = Is random number ($0 < u < 1$) smaller than p ?

Sample = Sample from $B(a, b)$ $x = \text{pull from best bandit}$

Update = $a+ = x, b+ = 1-x$ ←

Exercise: Bayesian Learning

↳ Use Bayesian method on data, what is the problem?

↳ Bayesian bandit works in real time as opposed to batching data.

Solution: use Simulataneous

→ Notes of things I don't understand on next page

instead of maximizing a & b as 1

$$a = 1 + \text{clicks} \quad \& \quad b = 1 + \text{no-clicks}$$

as described in previous lectures

as vs chosen based on which sample gives

higher P : 'A' if Band A's sample > Band B's sample
else 'B'

(\hookrightarrow) relevant band A gets more

if click then increment click for that band A

\rightarrow over time, should converge to band A w/
better CTR (done live)

Question: how do we do this after the fact?

(\hookrightarrow) if 50:50 then both distributions will
be developed.

user-1's impression rec per row for each column

0	1	0	0
1	1	1	1
2	1	0	0
3	1	1	0
4	1	1	0

& case, make band A

\hookrightarrow rows & 1s

$$a = 1 + 1s$$

$$b = 1 + (rows - 1s)$$

\hookrightarrow test & control groups. (\hookrightarrow update priors for
each row.)

Very good!

Exercise: Compare strategies

① implement all 3 multi-armed bandit solution algorithms & compare their performance

- Bayesian Bandit
- E-greedy
- UCB1

- Which is easiest?
- Which converge faster?
- What is the loss relative to knowing best bandit & playing it all the time

↳ further study:

- How to calculate conj priors w/ any distribution for the likelihoods

↳ Think Bayes, Bayesian Methods for Hackers