Bayesian ML in Python

- part 1 Bayes rule review

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

(conditional; Joint; Marginal)

example:

	CA	US	MX	Σ
Buy	20	50	10	80
¬Buy	300	500	200	1000
Σ	320	550	210	Σ 1080

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

$$P(CA) = \frac{320}{1080}$$
$$P(US) = \frac{550}{1080}$$
$$P(MX) = \frac{210}{1080}$$

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

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

⋮

etc.

note that as a space ↑ in size the P ↓

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

& $P(A,B) = P(B,A)$ ← more common form.

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

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

↑ Becomes integral w/ cont. dist.

(?) review this

Probability exercise

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

↳ 20 tosses, 15 H & 5 T

↳ what do we expect after 200 tosses?

↳ 150 H & 50 T?

✗ no.

$200 - 20 = 180$

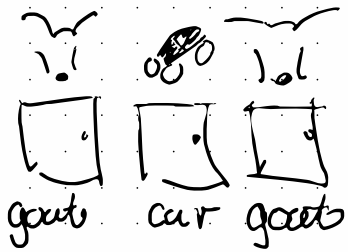
↳ even outcomes = 90 H, 90 T

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

due to independence.

Monty Hall problem

1. pick a door →
2. Monty reveals a goat
3. do you change?



When trying to calculate it's okay to analyze an instance where an arbitrary door is chosen.

C = loc of car

H = door MH opens

• we pick door 1

$P(H=2 \mid C=1) = 0.5$

$P(H=2 \mid C=2) = 0$

$P(H=2 \mid C=3) = 1$

need to calculate

$P(C=1 \mid H=2)$ (1), $P(C=3 \mid H=2)$ (2)

prob of occurance. →

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

normalized over all poss. →

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

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

$$= \frac{1}{3} \times \frac{3}{1.5} = \frac{1}{1.5} = 0.\dot{6}$$

you can show similarly that $P(C=1 \mid H=2)$ (1) $= 0.\dot{3}$ →

ie

$$\frac{\overset{0.5}{P(H=2 \mid C=1)}\,\overset{\frac{1}{3}}{P(C=1)}}{\frac{1}{3} + 0 + \frac{0.5}{3}}$$

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

$$= \frac{0.5}{1.5} = 0.\dot{3}$$

∴ you should switch.

Imbalanced Classes

↳ classifier for disease

what is good accuracy? ↳ 80?
↳ 90?
↳ 95?

Let's say $P(d) = 0.01$ 11

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

we care about TP rate (sensitivity)

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

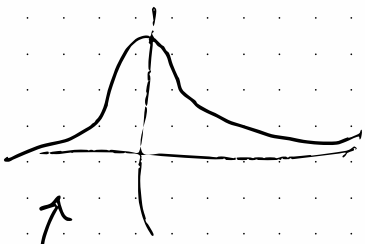
Metrics	equation
sensitivity	$\frac{TP}{TP + FN}$
specificity	$\frac{TN}{TN + FP}$
precision	$\frac{TP}{TP + FP}$ ← probability of disease given prediction is 1
recall	$\frac{FN}{FN + TN}$ ← Check that this is right.

	Pred 1	Pred 0
d = 1	TP	FN
d = 0	FP	TN

Mean of gaussian distributed data

"maximum likelihood"

↳ what is this?



$f(x) = p(x)$

$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$

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?

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

$$\frac{\partial L}{\partial \mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

← Mean? ↙ How did teacher get from here to there?

Drawbacks: we don't know precision.

Max Likelihood GR

H = circle, T = not circle IID

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

$5\left(\frac{1}{2}\right)$? ~~⊗~~ $L(N_H, N_T) = p^{N_H}(1-p)^{N_T}$

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

$= \left(\frac{1}{2}\right)^5$ ← 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

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

↑ sum of random variables

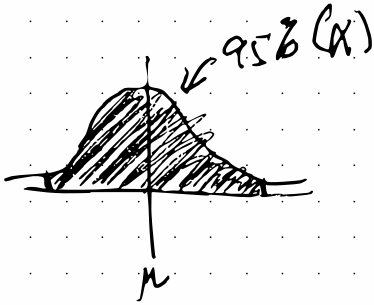
$Z = X + Y$ where X & Y are also random

- 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.

$$0.95 = \int_{x_L}^{x_R} N\left(x; 0, \frac{\sigma^2}{N}\right) dx$$

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

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

About bayesian paradigm

Frequentist	Bayesian
- Parameters are set & we don't know them	- Parameters are rvs. w/ distributions
- data is generated according to these param	- data is fixed - model $p(\text{param.} \mid \text{data})$

$$\hat{\Theta} = \underset{\Theta}{\arg\max}\, P(X \mid \Theta)$$

$P(\Theta \mid X)$

A/B Testing

web page → conversion kpis (1/0)

↳ rate = $\frac{\Sigma \text{ conversion}}{\Sigma \text{ users}}$

- 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 heights $\vec{m}$ & $\vec{w}$

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

$1 \times n$ $\quad$ $1 \times n$

test stat:

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

pooled std. dev

↑ not normally dist but t-dist.

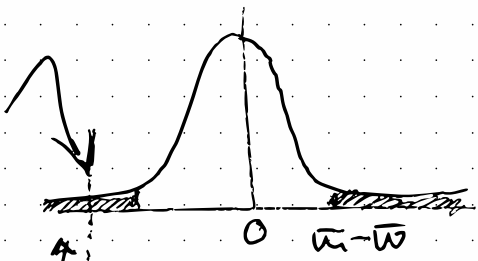
↖ unbiased, $\frac{1}{n}$ instead of $\frac{1}{n-1}$

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

where $\boxed{\nu = 2n - 2}$

thus...

if t lands in shaded area then we can reject H_0



p-value: Likelihood of wrongly rejecting H_0.

↳ Σ of remaining outcomes becomes p-value

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

Testing Characteristics

- larger N = bigger t, more confident of larger samples

$$t = \frac{\vec{m} - \vec{w}}{sp\sqrt{\frac{2}{n}}}$$

- larger sp = smaller t, less confident w/ large var.

→ what if two groups have different sizes?

$$t = \frac{\vec{m} - \vec{w}}{sp \cdot \sqrt{\frac{1}{n_m} + \frac{1}{n_w}}} \quad \text{where; } sp = \sqrt{\frac{(n-1)s_m^2 + (n-1)s_w^2}{n_m + n_w - 2}}$$

↑ makes assumption that var is the same for both groups

↑ is weighted change.

↳ if they are then use Welche's t-test

Note that we assumed normally dist data as well.

↳ what if you don't know dist or it's different?

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

t-test exercise

'advertisement_clicks.csv' - contains data.

↳ Step (1) do we need to use Bonferroni Corr.?

No need, only comparing 2 things.

↳ used to solve problem of multiple comparisons

↳ Step (2) use (t-test) to determine if one advertiser is ~~better~~ than the other.

↳ same vs diff variance

---

comments: my t & p values are different.

↳ two tailed vs one tailed?

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

N seems to be the sum of ~~both~~ ten both vectors.

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 counts

	Click	¬Click	
A	36	14	50
B	30	25	55
	66	39	105

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

re for click, $A_{exp} = \frac{50 * 66}{105}$ & $obs_i = 36$

↳ thus $\chi^2 = 3.418$ → can also be calculated by simplifying gives $\frac{N(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)}$

→ What is p-value?

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

A/B/... tests

↳ possible to get significant p-value by chance ⇒ p-hacking

- Bonferroni correction ⇒ $\alpha_{new} = \frac{\alpha}{\text{no tests}}$.
  ↳ note that number of tests will depend upon how you choose to compare values.

Statistical Power ↗ = sensitivity $P(\text{rej. } H_0 \mid H_0 \text{ is true})$

β = false neg. rate

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

A/B testing pitfalls

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

→ Sample size? $N = 16 \frac{\sigma^2}{\delta^2}$ ← variance, ← smallest difference

Lots of size calculators online

All this is quite confusing...

Bayesian A/B testing

Explore / Exploit

↳ Multi-armed bandit

↳ What is the balance between exploring new possibilities & exploiting knowledge that is already known.

Intro to bayesian A/B testing

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

- $P(\theta|X)$ ↑ posterior
- $P(X|\theta)$ ↑ likelihood
- $P(\theta)$ ← prior
- $P(X)$ ← Σ "norm. const."

prob of param given data = (prob of data given param * what we know about param) / prob. of data

↳ integral of all $P(X|\theta)P(\theta)d\theta$

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

↑ 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 it's what the data follows

can change depending on what you want to predict.

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 log 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.

$\therefore P(\theta|X) = \text{Beta}(a', b')$

$b' = b + \sum_{i=1}^{n} x_i$

$= b + \#$ no-chirps

$a' = a + \sum_{i=1}^{N} x_i$

$= a + \#$ chirps

NB: $E(\theta) = \frac{a}{a+b}$

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

as $a, b \uparrow$ var $\downarrow$

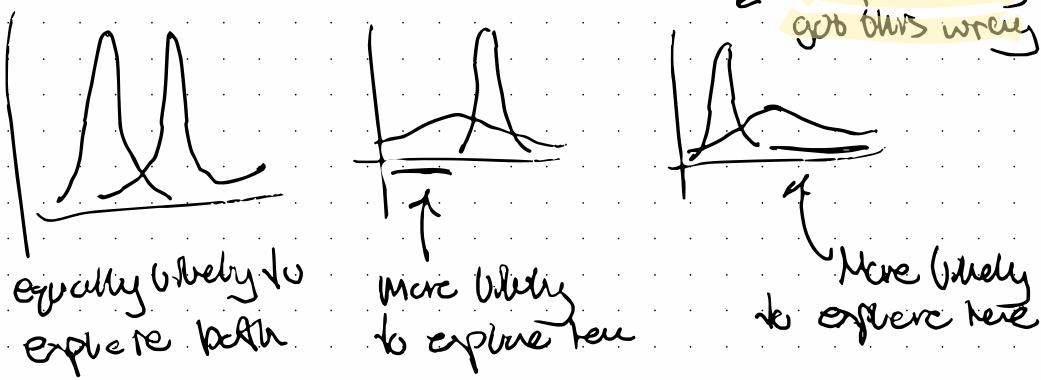
How do we choose a & b?

$a = 1$? uniform, all outcomes
$b = 1$? likely

Bayesian A/B testing

How does sampling from Beta(a', b') help us?

↳ organically solve multi armed bandit



↳ Adjusts accordingly. (Tompson sampling explore higher, exploit low ver)

Online Nature of Bayesian A/B Testing

Traditional AB-T.

↳ Approximations

↳ Can't "peek" @ results

↳ Calculate sample sizes

↳ difficult to communicate

↳ lots of unintuitive 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/brew when you want.

Thresholds for Bayesian A/B T.

Re $P(\mu_1 > \mu_2)$ 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 pdfs $= P(\mu_1, \mu_2) = P(\mu_1) * P(\mu_2)$
↑ works due to independence.

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

$$P(\mu_2 > \mu_1) = \sum_{i=0}^{a_2-1} \frac{B(a_1+i, b_1+b_2)}{(b_2+i)\, B(1+i, b_2)\, B(a_1, b_1)}$$

↑ Quite the monster

or...

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

Equation is too complicated, see lecture notes.

How does bayesian bandit work?

→ Bandit

↳ p ← probability
↳ a, b = (1, 1) ← priors

3 actions:
- pull
- sample
- update

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

Sample = sample from $B(a, b)$ $\quad x$ = 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 inputting data.

Solution: use simulation

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

instead of initializing a & b as 1

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

as described in previous lectures

AB is chosen based on which sample gives higher p: `'A' if BanditA.sample() > BanditB.sample() else 'B'`

↳ relevant bandit gets view

if click then increment click for that bandit

→ over time, should converge to bandit w/ better CTR (done bwe)

Question: how do we do this after the fact?

↳ if 50:50 then both distributions will be developed.

user_id	impression	reg	pu
0	1	0	0
1	1	1	1
2	1	0	0
3	1	1	0
4	1	1	0

↳ for each column & case, make bandit

↳ rows & 1s

$a = 1 + 1s$

$b = 1 + (\text{rows} - 1s)$

↳ update priors for each row.

↳ test & control groups.

very good!

Excercise: Compare strategies

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

- Bayesian Bandit
- ε-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 likelihood

↳ Think bayes, Bayesian Methods for Hackers