ORIE 4742 - Info Theory and Bayesian ML

Lecture 1: Probability Review

January 23, 2020

Sid Banerjee, ORIE, Cornell

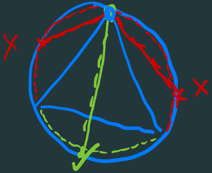
“probability theory is common sense reduced to calculation”

not quite. . .

Bertrand's problem

given an equilateral triangle inscribed in a circle, and a random chord, what is the probability the chord is longer than the side of the triangle?

pick random endpoint (fixing one end)

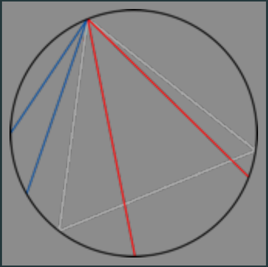


$\mathbb{P}[\text{chord} \geqslant \text{side}] = 1/3$

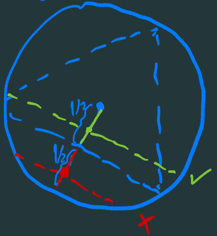
not quite. . .

Bertrand's ~~problem~~ paradox

given an equilateral triangle inscribed in a circle, and a random chord, what is the probability the chord is longer than the side of the triangle?



pick any radius and random center

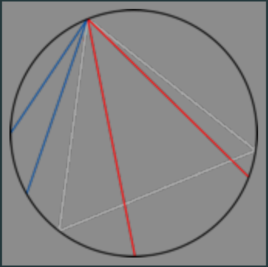


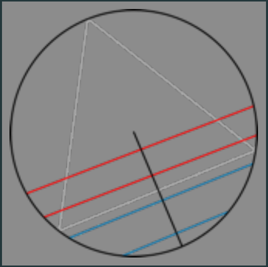
$\mathbb{P}[\text{chord} > \text{side}] = 1/2$

not quite. . .

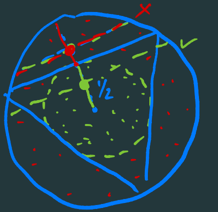
Bertrand's problem

given an equilateral triangle inscribed in a circle, and a random chord, what is the probability the chord is longer than the side of the triangle?





pick random center in ⊙

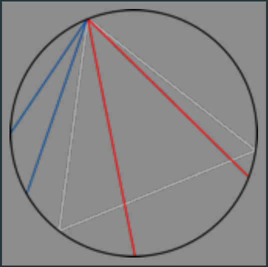


$\mathbb{P}[\text{chord} > \text{side}] = 1/4$

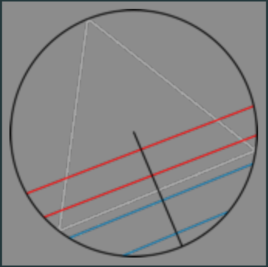
not quite. . .

Bertrand's problem

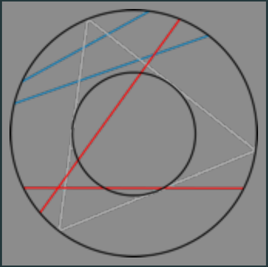
given an equilateral triangle inscribed in a circle, and a random chord, what is the probability the chord is longer than the side of the triangle?



$p = 1/3$



$p = 1/2$

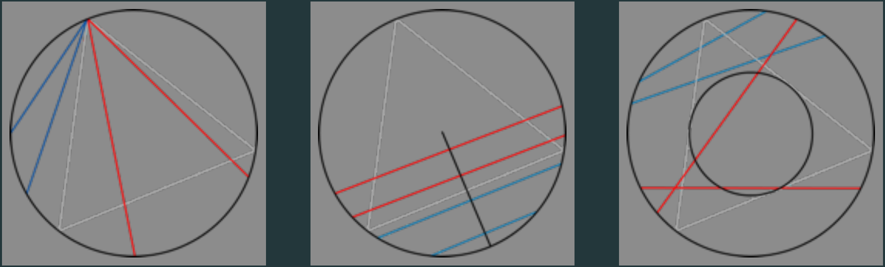


$p = 1/4$

not quite. . .

Bertrand's problem

given an equilateral triangle inscribed in a circle, and a random chord, what is the probability the chord is longer than the side of the triangle?



the moral (for this course. . . and for life)

be very precise about defining experiments/random variables/distributions

also see Wikipedia article on Bertrand's paradox

the essentials

reading assignment

Bishop: chapter 1, sections 1.2 - 1.2.4

Mackay: chapter 2 (less formal, but much more fun!)

things you must know and understand

- random variables (rv) and cumulative distribution functions (cdf)
- conditional probabilities and Bayes rule
- expectation and variance of random variables
- independent and mutually exclusive events
- basic inequalities: union bound, Jensen, Markov/Chebyshev
- common rvs (Bernoulli, Binomial, Geometric, Gaussian (Normal))

random variables and cdf

sample space, random variable

random experiment: outcome cannot be predicted in advance.

sample space Ω: the set of all possible outcomes of the experiment

random variable: any function from $\Omega \to \mathbb{R}$ (random vector:$\Omega \to \mathbb{R}^d$)

example: flip two coins, and let $X = \#$ of heads $(\mathbb{P}[\text{heads}] = h)$

$\Omega = \{ HH, HT, TH, TT \}$

prob.	h^2	$h(1-h)$	$(1-h)h$	$(1-h)^2$
X :	2	1	1	0

cumulative distribution function

ALERT!!

always try to think of probability and rvs through the cdf

for any rv X (discrete or continuous), its probability distribution is defined by its cumulative distribution function (cdf)

$$F(x) = \mathbb{P}[X \leq x]$$

using the cdf we can compute probabilities

$$\mathbb{P}[a < X \leq b] = F(b) - F(a)$$

visualizing a cdf

The plot of a cdf obeys 3 essential rules + one convention

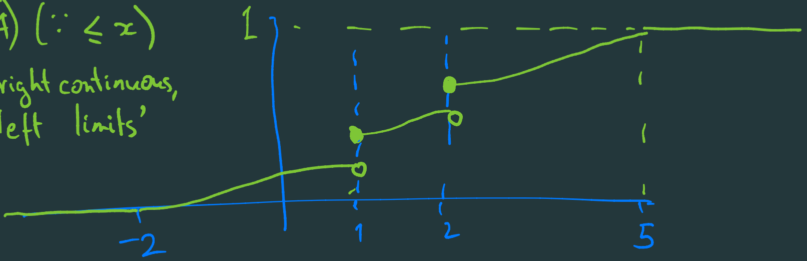
Example: consider an rv $\in [-2, 5]$ with a **jumps** at 1 and 2

1) $F(x) \in [0,1]$, 2) $\lim_{x \to -\infty} F(x) = 0$, $\lim_{x \to \infty} F(x) = 1$

3) $F(x)$ is non-decreasing

4) $(\because \leq x)$

'right continuous, left limits'



discrete random variables

for a discrete random variable taking values in $\mathbb{N}$, another characterization is its probability mass function (pmf) $p(\cdot)$

$$p(x) = \mathbb{P}[X = x]$$

- any pmf $p(x)$ has the following properties:

$$p(x) \in [0, 1] \, \forall \, x \in \mathbb{N} \qquad , \qquad \sum_{x \in \mathbb{N}} p(x) = 1$$

- the pmf $p(\cdot)$ is related to the cdf $F(\cdot)$ as

$$F(x) = \quad \sum_{y \leq x} p(y)$$

$$p(x) = \quad F(x) - F(x-1)$$

continuous random variables

for a continuous random variable taking values in $\mathbb{R}$, another characterization is its probability density function (pdf) $f(\cdot)$

$$\mathbb{P}[a < X \leq b] = \int_a^b f(x)\,dx$$

- any pdf $f(x)$ has the following properties:

$$f(x) \geq 0 \,\forall\, x \in \mathbb{R} \qquad , \qquad \int_{-\infty}^{\infty} f(x)dx = 1$$

- ALERT!! It is not true that $f(x) = \mathbb{P}[X = x]$. In fact, for any x,

$$\mathbb{P}[X = x] = 0 \qquad \left(\neq f(x) \right)$$

continuous random variables

thus, for continuous rv X with pdf $f(\cdot)$ and cdf $F(\cdot)$, we have

$$\mathbb{P}[a < X \leq b] = F(b) - F(a) = \int_a^b f(x)dx$$

now we can go from one function to the other as

$$F(x) = \int_{-\infty}^{x} f(x)dx$$

$$f(x) = \frac{d}{dx} F(x)$$

(assuming differentiable ...)

expectations and independence

expected value (mean, average)

let X be a random variable, and $g(\cdot)$ be any real-valued function

- If X is a discrete rv with $\Omega = \mathbb{Z}$ and pmf $p(\cdot)$, then

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

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

$$\left(\text{Eg. } g(x) = (x - \mathbb{E}[X])^2 \right)$$

$$\Rightarrow \mathbb{E}[g(X)] = \text{Var}(X)$$

- If X is a continuous rv with $\Omega = \mathbb{R}$ and pdf $f(\cdot)$, then

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\, dx$$

variance and standard deviation

- **Definition**: $Var(X) = \mathbb{E}\Big[\underbrace{\big(X - \overset{\overset{\text{a number}}{\downarrow}}{\mathbb{E}[X]}\big)^2}_{g(x)}\Big]$ Std-deviation $\sigma(X) = \sqrt{Var(x)}$
- (More useful formula for computing variance)

$$
\begin{aligned}
Var(X) &= \mathbb{E}\left[\left(X - \mathbb{E}[X]\right)^2\right] \\
&= \mathbb{E}\left[\left(X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\right)\right] \\
&= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\
&= \underbrace{\mathbb{E}[X^2] - \mathbb{E}[X]^2}_{\geq 0}
\end{aligned}
$$

Side-fact

$$\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$$

Why? because $g(x) \geq 0$

Universal property!!

independence

what do we mean by “random variables X and Y are independent”?
(denoted as $X \perp\!\!\!\perp Y$; similarly, $X \not\perp\!\!\!\perp Y$ for ‘not independent’)

intuitive definition: knowing X gives no information about Y

formal definition: $\mathbb{P}[X \leq x, Y \leq y] = \underbrace{F(x)}_{\mathbb{P}[X \leq x]} \cdot \underbrace{F(y)}_{\mathbb{P}[Y \leq y]} \quad \forall\, x, y \in \mathbb{R}$

centering

- One measure of ~~in~~dependence between rv is their covariance

$$Cov(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] \qquad \text{(formal definition)}$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \qquad \text{(for computing)}$$

independence and covariance

how are independence and covariance related?

- X and Y are independent, then they are uncorrelated

  in notation: $X \perp\!\!\!\perp Y \quad \Rightarrow \quad Cov(X, Y) = 0$

- however, uncorrelated rvs can be dependent

  in notation: $Cov(X, Y) = 0 \quad \not\Rightarrow \quad X \perp\!\!\!\perp Y$

- $Cov(X, Y) = 0 \Rightarrow X \perp\!\!\!\perp Y$ only for multivariate Gaussian rv (this though is confusing; see this Wikipedia article)

linearity of expectation

for any rvs X and Y, and any constants $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

note 1: no assumptions! (in particular, does not need independence)

linearity of expectation

for any rvs X and Y, and any constants $a, b \in \mathbb{R}$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

note 1: no assumptions! (in particular, does not need independence)

note 2: does not hold for variance in general

for general X, Y

$Var(aX + bY) =$

when X and Y are independent

$Var(aX + bY) =$

using linearity of expectation

the TAs get lazy and distribute graded assignments among n students uniformly at random. On average, how many students get their own hw?

using linearity of expectation

the TAs get lazy and distribute graded assignments among n students uniformly at random. On average, how many students get their own hw?

Let $X_i = \mathbb{1}_{\left[\text{student i gets her hw}\right]}$ (indicator rv)

$N =$ number of students who get their own hw $= \sum_{i=1}^{10} X_i$

then we have:

$$\begin{aligned}\mathbb{E}[N] &= \mathbb{E}[\sum_{i=1}^{n} X_i] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} \mathbb{P}[X_i = 1] = \sum_{i=1}^{n} \frac{1}{n} = 1\end{aligned}$$

useful probability inequalities

inequality 1: The Union Bound

Let $A_1, A_2, \ldots, A_k$ be events. Then

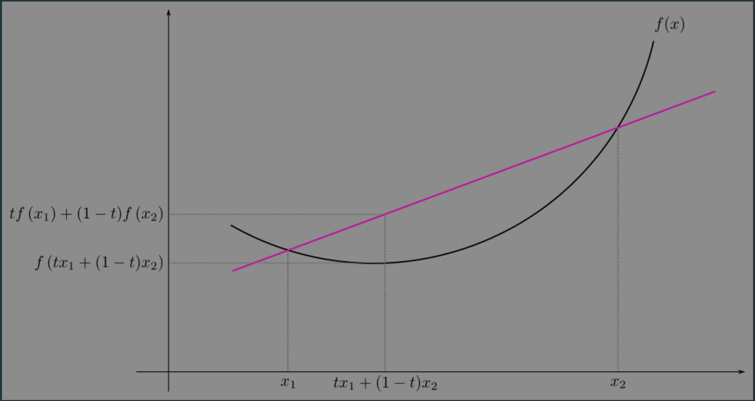
$$P(A_1 \cup A_2 \cup \cdots \cup A_k) \leq (P(A_1) + P(A_2) + \cdots + P(A_k))$$

inequality 2: Jensen's Inequality

If X is a random variable and f is a convex function, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Proof sketch (plus way to remember)



inequality 3: Markov and Chebyshev's inequalities

Markov's inequality

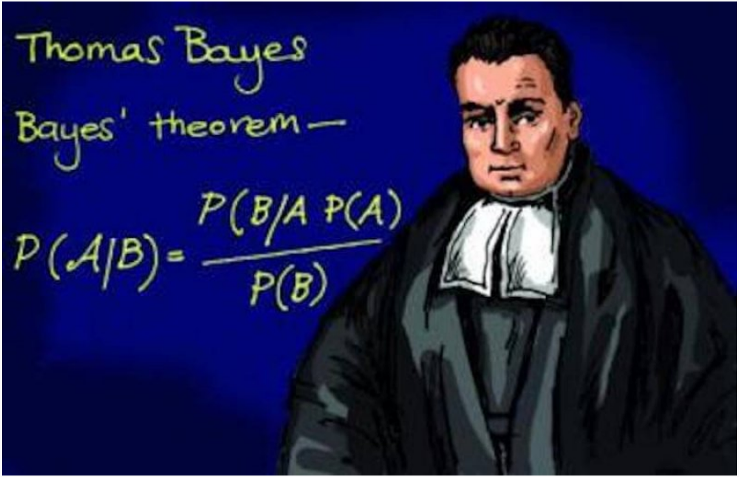
For any rv. $X \geq 0$ with mean $\mathbb{E}[X]$, and for any $k > 0$,

$$\mathbb{P}\left[X \geq k\right] \leq \frac{\mathbb{E}[X]}{k}$$

Chebyshev's inequality

For any rv. X with mean $\mathbb{E}[X]$, finite variance $\sigma^2 > 0$, and for any $k > 0$,

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq k\sigma\right] \leq \frac{1}{k^2}$$



conditioning and Bayes' rule

marginals and conditionals

let X and Y be discrete rvs taking values in $\mathbb{N}$. denote the joint pmf:

$$p_{XY}(x,y) = \mathbb{P}[X = x, Y = y]$$

marginalization: computing individual pmfs from joint pmfs as

$$p_X(x) = \sum_{y \in \mathbb{N}} p_{XY}(x,y) \qquad p_Y(y) = \sum_{x \in \mathbb{N}} p_{XY}(x,y)$$

conditioning: pmf of X given $Y = y$ (with $p_Y(y) > 0$) defined as:

$$\mathbb{P}[X = x | Y = y] \triangleq p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}$$

more generally, can define $\mathbb{P}[X \in \mathcal{A} | Y \in \mathcal{B}]$ for sets $\mathcal{A}, \mathcal{B} \in \mathbb{N}$

see also this visual demonstration

the basic ‘rules’ of Bayesian inference

let X and Y be discrete rvs taking values in $\mathbb{N}$, with joint pmf $p(x,y)$

product rule

for $x, y \in \mathbb{N}$, we have: $p_{XY}(x,y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$

sum rule

for $x \in \mathbb{N}$, we have: $p_X(x) = \sum_{y\in\mathbb{N}} p_{X|Y}(x|y)p_Y(y)$

and most importantly!

Bayes rule

for any $x, y \in \mathbb{N}$, we have:

$$p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{\sum_{x\in\mathbb{N}} p_{Y|X}(y|x)p_X(x)}$$

see also this video for an intuitive take on Bayes rule

Bayesian inference: example

Mackay's three cards

We have three cards $C1, C2, C3$, with $C1$ having faces Red-Blue, $C2$ having faces Blue-Blue; and $C3$ having faces Red-Red.

A card is randomly drawn and placed on a table – its upper face is Red. What is the colour of its lower face?

Bayesian inference: example

$C1 =$ Red-Blue, $C2=$ Blue-Blue; $C3=$ Red-Red. A card is randomly drawn, and has upper face Red. What is the colour of its lower face?

Let $X \in \{C1, C2, C3\}$ be the identity of drawn card, $Y_b \in \{b, r\}$ be the color of bottom face, and $Y_t \in \{b, r\}$ be the color of top face. Then:

$$\begin{aligned}\mathbb{P}[Y_b = b | Y_t = b] = \mathbb{P}[X = C2 | Y_t = b] &= \frac{\mathbb{P}[Y_t = b | X = C2]\mathbb{P}[X = C2]}{\mathbb{P}[Y_t = b]} \\ &= \frac{1 \times (1/3)}{(1/2) \times (1/3) + 1 \times (1/3) + 0 \times (1/3)} = 2/3\end{aligned}$$

ALERT!!

always write down the probability of everything

Bayesian inference: example

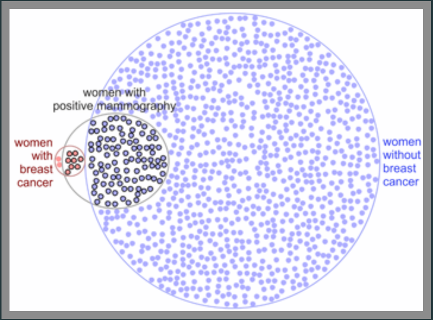
Eddy's mammogram problem

The probability a woman at age 40 has breast cancer is 0.01. A mammogram detects the disease 80% of the time, but also mis-detects the disease in healthy patients 9.6% of the time. If a woman at age 40 has a positive mammogram test, what is the probability she has breast cancer?

Bayesian inference: example

Eddy's mammogram problem

The probability a woman at age 40 has breast cancer is 0.01. A mammogram detects the disease 80% of the time, but also mis-detects the disease in healthy patients 9.6% of the time. If a woman at age 40 has a positive mammogram test, what is the probability she has breast cancer?



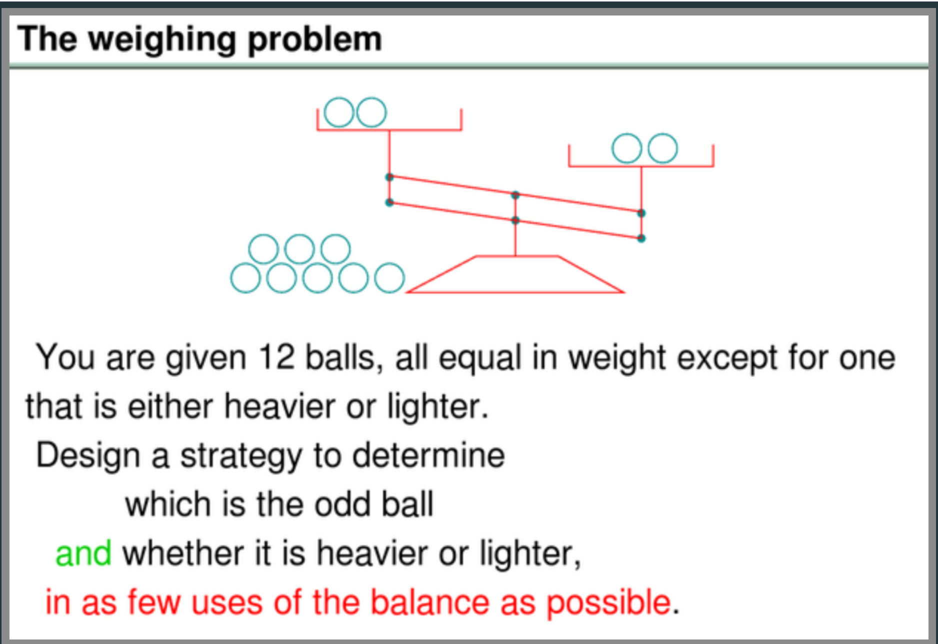
credit: Micallef et al.

see also this video for more about the odds ratio

quantifying information content

how much ‘information’ does a random variable have?

Mackay's weighing puzzle



The weighing problem

You are given 12 balls, all equal in weight except for one that is either heavier or lighter.

Design a strategy to determine

which is the odd ball

and whether it is heavier or lighter,

in as few uses of the balance as possible.