Alex Psomas: Lecture 15.

Bayes' Rule, Mutual Independence, Collisions and Collecting

1. Conditional Probability
2. Independence
3. Bayes' Rule
4. Balls and Bins
5. Coupons

Conditional Probability: Review

Recall:

- $Pr[A|B] = \frac{Pr[A \cap B]}{Pr[B]}$.
- Hence, $Pr[A \cap B] = Pr[B]Pr[A|B] = Pr[A]Pr[B|A]$.
- A and B are *positively correlated* if $Pr[A|B] > Pr[A]$,

  i.e., if $Pr[A \cap B] > Pr[A]Pr[B]$.
- A and B are *negatively correlated* if $Pr[A|B] < Pr[A]$,

  i.e., if $Pr[A \cap B] < Pr[A]Pr[B]$.
- A and B are *independent* if $Pr[A|B] = Pr[A]$,

  i.e., if $Pr[A \cap B] = Pr[A]Pr[B]$.
- Note: $B \subset A \Rightarrow A$ and B are positively correlated.
  $(Pr[A|B] = 1 > Pr[A])$
- Note: $A \cap B = \emptyset \Rightarrow A$ and B are negatively correlated.
  $(Pr[A|B] = 0 < Pr[A])$

Monty Hall

3 closed doors. Behind one of the doors there is a prize (car). The others have goats.

You pick a door. Say door number 1

I open door 2 or door 3. One of the two that I **know** doesn't have the prize. Say it was door 2

I ask: Would you like to change your door to number 3?

Question: What should you do in order to maximize the probability of winning?

Monty Hall

Change!!!!

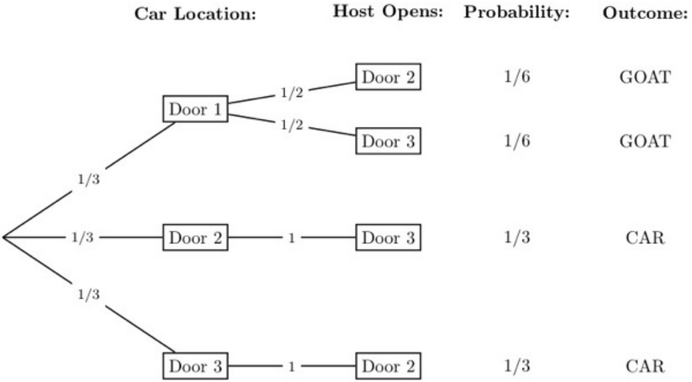
What is the probability that the prize is in door 3? $\frac{2}{3}$!

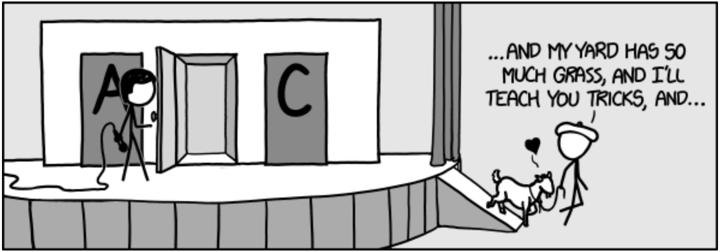
How does that make any sense????

Say the original door where the prize is random. So each door has probability $\frac{1}{3}$.

You pick door 1. What's the probability that it's in either 2 or 3? $\frac{2}{3}$

The door I opened wasn't random! I knew it didn't have a prize!! Therefore, switching, is like getting to pick two doors at the beginning!


Car Location:
Host Opens:
Probability:
Outcome:
Door 1
Door 2
Door 3
1/3
1/3
1/3
1/2
1/2
1
1
Door 2
Door 3
Door 3
Door 2
1/6
1/6
1/3
1/3
GOAT
GOAT
CAR
CAR


A
C
...AND MY YARD HAS SO MUCH GRASS, AND I'LL TEACH YOU TRICKS, AND...

Balls in bins

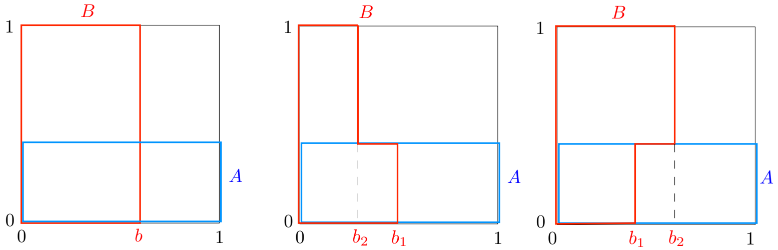
I throw 5 (indistinguishable) balls in two bins. What is the probability that the first bin is empty?

1. Approach 1: There are 6 outcomes: $(5,0)$, $(4,1)$, $(3,2)$, $(2,3)$, $(1,4)$, $(0,5)$. Probability that the first bin is empty is $\frac{1}{6}$
2. Approach 2: I pretend I can tell the balls apart. There are 2^5 outcomes: $(1,1,1,1,1)$, $(1,1,1,1,2)$, $\ldots(2,2,2,2,2)$. $(x,1,x,x,x)$ means that the second ball I threw landed in the first bin.
   Probability that the first bin ie empty is $\frac{1}{2^5}$. The fact that I can tell them apart shouldn't change the probability.

Well... I guess probability is wrong...
Or...... Could one of the approaches be wrong???
Approach 1 is WRONG! Why did we divide by $|\Omega|$???
Why??????? Noooooooooooooooooooooooo

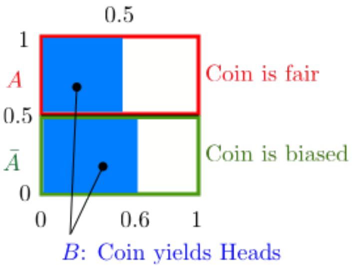
Conditional Probability: Pictures

Illustrations: Pick a point uniformly in the unit square



- Left: A and B are independent. $Pr[B] = b; Pr[B|A] = b$.
- Middle: A and B are positively correlated.
  $Pr[B|A] = b_1 > Pr[B|\bar{A}] = b_2$. Note: $Pr[B] \in (b_2, b_1)$.
- Right: A and B are negatively correlated.
  $Pr[B|A] = b_1 < Pr[B|\bar{A}] = b_2$. Note: $Pr[B] \in (b_1, b_2)$.

Bayes and Biased Coin



Pick a point uniformly at random in the unit square. Then

$$Pr[A] = 0.5; Pr[\bar{A}] = 0.5$$

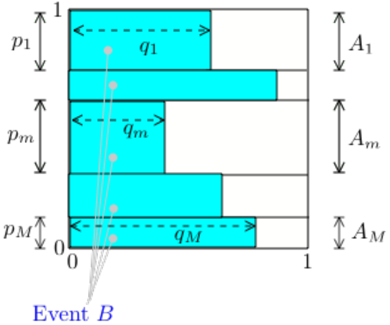
$$Pr[B|A] = 0.5; Pr[B|\bar{A}] = 0.6; Pr[A \cap B] = 0.5 \times 0.5$$

$$Pr[B] = 0.5 \times 0.5 + 0.5 \times 0.6 = Pr[A]Pr[B|A] + Pr[\bar{A}]Pr[B|\bar{A}]$$

$$Pr[A|B] = \frac{0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.6} = \frac{Pr[A]Pr[B|A]}{Pr[A]Pr[B|A] + Pr[\bar{A}]Pr[B|\bar{A}]}$$

$$\approx 0.46 = \text{fraction of } B \text{ that is inside } A$$

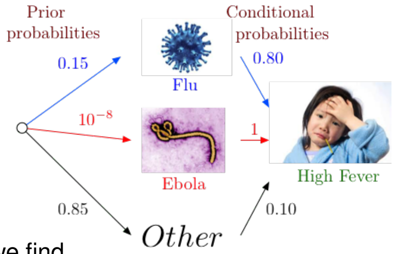
Bayes: General Case



Pick a point uniformly at random in the unit square. Then

$$Pr[A_m] = p_m, m = 1, \ldots, M$$
$$Pr[B|A_m] = q_m, m = 1, \ldots, M; Pr[A_m \cap B] = p_m q_m$$
$$Pr[B] = p_1 q_1 + \cdots p_M q_M$$
$$Pr[A_m|B] = \frac{p_m q_m}{p_1 q_1 + \cdots p_M q_M} = \text{ fraction of } B \text{ inside } A_m.$$

Why do you have a fever?



Using Bayes' rule, we find

$$Pr[\text{Flu}|\text{High Fever}] = \frac{0.15 \times 0.80}{0.15 \times 0.80 + 10^{-8} \times 1 + 0.85 \times 0.1} \approx 0.58$$

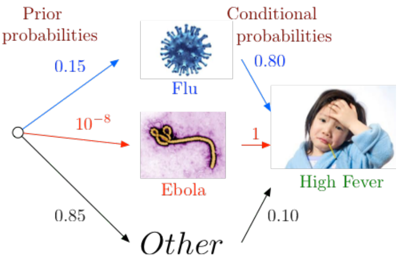
$$Pr[\text{Ebola}|\text{High Fever}] = \frac{10^{-8} \times 1}{0.15 \times 0.80 + 10^{-8} \times 1 + 0.85 \times 0.1} \approx 5 \times 10^{-8}$$

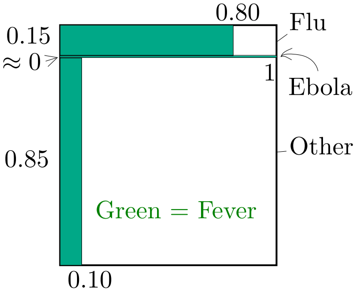
$$Pr[\text{Other}|\text{High Fever}] = \frac{0.85 \times 0.1}{0.15 \times 0.80 + 10^{-8} \times 1 + 0.85 \times 0.1} \approx 0.42$$

The values $0.58, 5 \times 10^{-8}, 0.42$ are the posterior probabilities.

Why do you have a fever?

Our “Bayes’ Square” picture:





58% of Fever = Flu

≈ 0% of Fever = Ebola

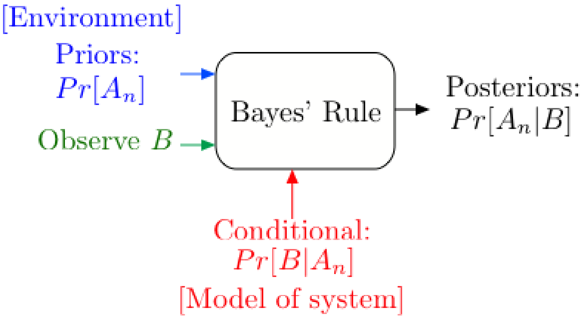
42% of Fever = Other

Note that even though $Pr[\text{Fever}|\text{Ebola}] = 1$, one has

$$Pr[\text{Ebola}|\text{Fever}] \approx 0.$$

This example shows the importance of the prior probabilities.

Bayes' Rule Operations



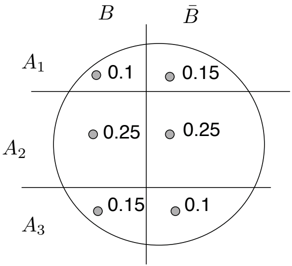
Bayes' Rule is the canonical example of how information changes our opinions.

Independence

Recall :

$$\begin{aligned} &A \text{ and } B \text{ are independent} \\ &\Leftrightarrow Pr[A \cap B] = Pr[A]Pr[B] \\ &\Leftrightarrow Pr[A|B] = Pr[A]. \end{aligned}$$

Consider the example below:

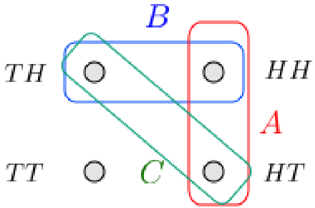


(A_2, B) are independent: $Pr[A_2|B] = 0.5 = Pr[A_2]$.
$(A_2, \bar{B})$ are independent: $Pr[A_2|\bar{B}] = 0.5 = Pr[A_2]$.
(A_1, B) are not independent: $Pr[A_1|B] = \frac{0.1}{0.5} = 0.2 \neq Pr[A_1] = 0.25$.

Pairwise Independence

Flip two fair coins. Let

- $A =$ 'first coin is H' $= \{HT, HH\}$;
- $B =$ 'second coin is H' $= \{TH, HH\}$;
- $C =$ 'the two coins are different' $= \{TH, HT\}$.



A, C are independent; B, C are independent;

$A \cap B, C$ are not independent. ($Pr[A \cap B \cap C] = 0 \neq Pr[A \cap B]Pr[C]$.)

A did not say anything about C and B did not say anything about C, but $A \cap B$ said something about C!

Example 2

Flip a fair coin 5 times. Let A_n = 'coin n is H', for $n = 1, \ldots, 5$.

Then,

$$A_m, A_n \text{ are independent for all } m \neq n.$$

Also,

$$A_1 \text{ and } A_3 \cap A_5 \text{ are independent.}$$

Indeed,

$$Pr[A_1 \cap (A_3 \cap A_5)] = \frac{1}{8} = Pr[A_1]Pr[A_3 \cap A_5]$$

. Similarly,

$$A_1 \cap A_2 \text{ and } A_3 \cap A_4 \cap A_5 \text{ are independent.}$$

This leads to a definition ....

Mutual Independence

Definition Mutual Independence

(a) The events $A_1, \ldots, A_5$ are mutually independent if

$$Pr[\cap_{k \in K} A_k] = \Pi_{k \in K} Pr[A_k], \text{ for all } K \subseteq \{1, \ldots, 5\}.$$

(b) More generally, the events $\{A_j, j \in J\}$ are mutually independent if

$$Pr[\cap_{k \in K} A_k] = \Pi_{k \in K} Pr[A_k], \text{ for all finite} K \subseteq J.$$

Example: Flip a fair coin forever. Let A_n = 'coin n is H.' Then the events A_n are mutually independent.

Mutual Independence

Theorem

(a) If the events $\{A_j, j \in J\}$ are mutually independent and if K_1 and K_2 are disjoint finite subsets of J, then

$$\cap_{k \in K_1} A_k \text{ and } \cap_{k \in K_2} A_k \text{ are independent.}$$

(b) More generally, if the K_n are pairwise disjoint finite subsets of J, then the events

$$\cap_{k \in K_n} A_k \text{ are mutually independent.}$$

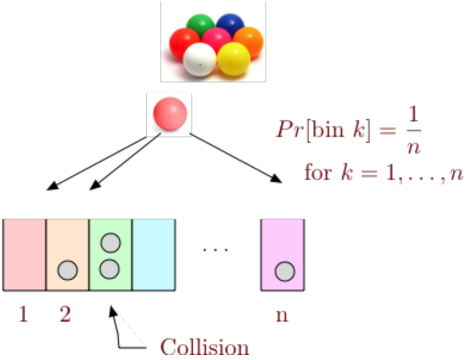
(c) Also, the same is true if we replace some of the A_k by $\bar{A}_k$.

Balls in bins

One throws m balls into $n > m$ bins.

Balls in bins

One throws m balls into $n > m$ bins.



Theorem:
$Pr[\text{no collision}] \approx \exp\{-\frac{m^2}{2n}\}$, for large enough n.

The Calculation.

A_i = no collision when ith ball is placed in a bin.

$Pr[A_1] = 1$

$Pr[A_2|A_1] = 1 - \frac{1}{n}$

$Pr[A_3|A_1, A_2] = 1 - \frac{2}{n}$

$Pr[A_i|A_{i-1} \cap \cdots \cap A_1] = (1 - \frac{i-1}{n})$.

no collision = $A_1 \cap \cdots \cap A_m$.

Product rule:
$Pr[A_1 \cap \cdots \cap A_m] = Pr[A_1]Pr[A_2|A_1] \cdots Pr[A_m|A_1 \cap \cdots \cap A_{m-1}]$

$$\Rightarrow Pr[\text{no collision}] = \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right).$$

$$\Rightarrow Pr[\text{no collision}] = \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right).$$

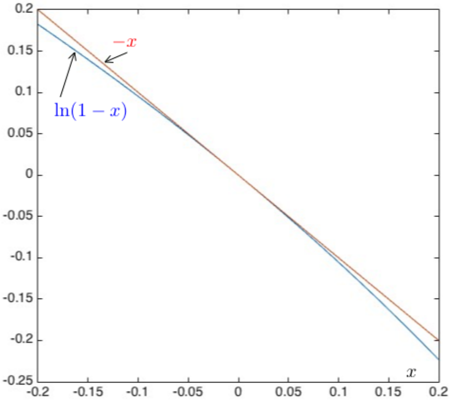
Hence,

$$\begin{aligned} \ln(Pr[\text{no collision}]) &= \sum_{k=1}^{m-1} \ln(1 - \frac{k}{n}) \approx \sum_{k=1}^{m-1} (-\frac{k}{n}) \ ^{(*)} \\ &= -\frac{1}{n} \frac{m(m-1)}{2}^{(\dagger)} \approx -\frac{m^2}{2n} \end{aligned}$$

$^{(*)}$ We used $\ln(1-\varepsilon) \approx -\varepsilon$ for $|\varepsilon| \ll 1$.

$^{(\dagger)}$ $1 + 2 + \cdots + m - 1 = (m-1)m/2$.

Approximation

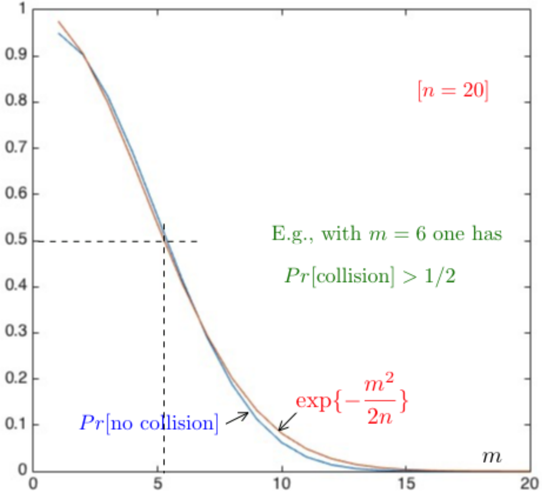


$$\exp\{-x\} = 1 - x + \frac{1}{2!}x^2 + \cdots \approx 1 - x, \text{ for } |x| \ll 1.$$

Hence, $-x \approx \ln(1-x)$ for $|x| \ll 1$.

Balls in bins

Theorem:
$Pr[\text{no collision}] \approx \exp\{-\frac{m^2}{2n}\}$, for large enough n.



Balls in bins

Theorem:
$Pr[\text{no collision}] \approx \exp\{-\frac{m^2}{2n}\}$, for large enough n.

In particular, $Pr[\text{no collision}] \approx 1/2$ for $m^2/(2n) \approx \ln(2)$, i.e.,

$$m \approx \sqrt{2\ln(2)n} \approx 1.2\sqrt{n}.$$

E.g., $1.2\sqrt{20} \approx 5.4$.

Roughly, $Pr[\text{collision}] \approx 1/2$ for $m = \sqrt{n}$. ($e^{-0.5} \approx 0.6$.)

The birthday paradox

Today's your birthday, it's my birthday too..

Probability that m people all have different birthdays?
With $n = 365$, one finds

$Pr[\text{collision}] \approx 1/2$ if $m \approx 1.2\sqrt{365} \approx 23$.

If $m = 60$, we find that

$$Pr[\text{no collision}] \approx \exp\{-\frac{m^2}{2n}\} = \exp\{-\frac{60^2}{2 \times 365}\} \approx 0.007.$$

If $m = 366$, then $Pr[\text{no collision}] = 0$. (No approximation here!)

The birthday paradox

n	$p(n)$
1	0.0%
5	2.7%
10	11.7%
20	41.1%
23	50.7%
30	70.6%
40	89.1%
50	97.0%
60	99.4%
70	99.9%
100	99.99997%
200	99.9999999999999999999999999998%
300	$(100 - (6\times10^{-80}))$%
350	$(100 - (3\times10^{-129}))$%
365	$(100 - (1.45\times10^{-155}))$%
366	100%
367	100%

Checksums!

Consider a set of m files.
Each file has a checksum of b bits.
How large should b be for $Pr[\text{share a checksum}] \leq 10^{-3}$?

Claim: $b \geq 2.9\ln(m) + 9$.

Proof:

Let $n = 2^b$ be the number of checksums.
We know $Pr[\text{no collision}] \approx \exp\{-m^2/(2n)\} \approx 1 - m^2/(2n)$.
Hence,

$$
\begin{aligned}
& Pr[\text{no collision}] \approx 1 - 10^{-3} \Leftrightarrow m^2/(2n) \approx 10^{-3} \\
& \Leftrightarrow 2n \approx m^2 10^3 \Leftrightarrow 2^{b+1} \approx m^2 2^{10} \\
& \Leftrightarrow b + 1 \approx 10 + 2\log_2(m) \approx 10 + 2.9\ln(m).
\end{aligned}
$$

Note: $\log_2(x) = \log_2(e)\ln(x) \approx 1.44\ln(x)$.

Coupon Collector Problem.

There are n different baseball cards.
(Brian Wilson, Jackie Robinson, Roger Hornsby, ...)

One random baseball card in each cereal box.

Theorem: If you buy m boxes,

(a) Pr[miss one specific item] $\approx e^{-\frac{m}{n}}$

(b) Pr[miss any one of the items] $\leq ne^{-\frac{m}{n}}$.

Coupon Collector Problem: Analysis.

Event A_m = 'fail to get Brian Wilson in m cereal boxes'

Fail the first time: $(1 - \frac{1}{n})$
Fail the second time: $(1 - \frac{1}{n})$
And so on ... for m times. Hence,

$$
\begin{aligned}
Pr[A_m] &= (1 - \frac{1}{n}) \times \cdots \times (1 - \frac{1}{n}) \\
&= (1 - \frac{1}{n})^m \\
ln(Pr[A_m]) &= m \ln(1 - \frac{1}{n}) \approx m \times (-\frac{1}{n}) \\
Pr[A_m] &\approx \exp\{-\frac{m}{n}\}.
\end{aligned}
$$

For $p_m = \frac{1}{2}$, we need around $n \ln 2 \approx 0.69n$ boxes.

Collect all cards?

Experiment: Choose m cards at random with replacement.

Events: E_k = 'fail to get player k' , for k = 1, . . . , n

Probability of failing to get at least one of these n players:

$$p := Pr[E_1 \cup E_2 \cdots \cup E_n]$$

How does one estimate p? Union Bound:

$$p = Pr[E_1 \cup E_2 \cdots \cup E_n] \leq Pr[E_1] + Pr[E_2] \cdots Pr[E_n].$$

$$Pr[E_k] \approx e^{-\frac{m}{n}}, k = 1, \ldots, n.$$

Plug in and get

$$p \leq ne^{-\frac{m}{n}}.$$

Collect all cards?

Thus,

$$Pr[\text{missing at least one card}] \leq ne^{-\frac{m}{n}}.$$

Hence,

$$Pr[\text{missing at least one card}] \leq p \text{ when } m \geq n\ln(\frac{n}{p}).$$

To get $p = 1/2$, set $m = n\ln(2n)$.

E.g., $n = 10^2 \Rightarrow m = 530; n = 10^3 \Rightarrow m = 7600$.

Summary.

Bayes' Rule, Mutual Independence, Collisions and Collecting

Main results:

- Bayes' Rule: $Pr[A_m|B] = p_m q_m/(p_1 q_1 + \cdots + p_M q_M)$.
- Product Rule:
  $Pr[A_1 \cap \cdots \cap A_n] = Pr[A_1]Pr[A_2|A_1] \cdots Pr[A_n|A_1 \cap \cdots \cap A_{n-1}]$.
- Balls in bins: m balls into $n > m$ bins.

$$Pr[\text{no collisions}] \approx \exp\{-\frac{m^2}{2n}\}$$

- Coupon Collection: n items. Buy m cereal boxes.

  $Pr[\text{miss one specific item}] \approx e^{-\frac{m}{n}}$; $Pr[\text{miss any one of the items}] \leq ne^{-\frac{m}{n}}$.

Key Mathematical Fact: $\ln(1-\varepsilon) \approx -\varepsilon$.