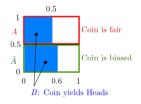
CS70: Lecture 18.

Bayes' Rule, Mutual Independence, Collisions and Collecting

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

Bayes and Biased Coin



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

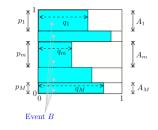
$$\begin{aligned} ⪻[A] = 0.5; Pr[\bar{A}] = 0.5 \\ ⪻[B|A] = 0.5; Pr[B|\bar{A}] = 0.6; Pr[A \cap B] = 0.5 \times 0.5 \\ ⪻[B] = 0.5 \times 0.5 + 0.5 \times 0.6 = Pr[A]Pr[B|A] + Pr[\bar{A}]Pr[B|\bar{A}] \\ ⪻[A|B] = \frac{0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.6} = \frac{Pr[A]Pr[B|A]}{Pr[A]Pr[B|A] + Pr[\bar{A}]Pr[B|\bar{A}]} \\ &\approx 0.46 = \text{fraction of } B \text{ that is inside } A \end{aligned}$$

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])

Bayes: General Case

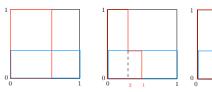


Pick a point uniformly at random in the unit square. Then

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

Conditional Probability: Pictures

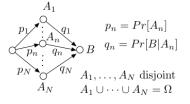
Illustrations: Pick a point uniformly in the unit square



- 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. $Pr[B|A] = b_1 < Pr[B|\overline{A}] = b_2$. Note: $Pr[B] \in (b_1, b_2)$.

Bayes Rule

Another picture:



$$Pr[A_n|B] = \frac{p_n q_n}{\sum_m p_m q_m}.$$

Why do you have a fever?



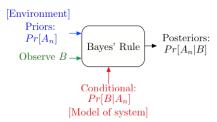
$$\textit{Pr}[\text{Flu}|\text{High Fever}] = \frac{0.15 \times 0.80}{0.15 \times 0.80 + 10^{-8} \times 1 + 0.85 \times 0.1} \approx 0.58$$

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

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

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

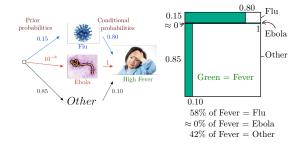
Bayes' Rule Operations



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

Why do you have a fever?

Our "Bayes' Square" picture:



Note that even though Pr[Fever|Ebola] = 1, one has

 $Pr[Ebola|Fever] \approx 0.$

This example shows the importance of the prior probabilities.

Thomas Bayes



Source: Wikipedia.

Why do you have a fever?

We found

 $Pr[Flu|High Fever] \approx 0.58$, $Pr[Ebola|High Fever] \approx 5 \times 10^{-8}$, $Pr[Other|High Fever] \approx 0.42$

One says that 'Flu' is the Most Likely a Posteriori (MAP) cause of the high

'Ebola' is the Maximum Likelihood Estimate (MLE) of the cause: it causes the fever with the largest probability.

Recall that

$$p_m = Pr[A_m], q_m = Pr[B|A_m], Pr[A_m|B] = \frac{p_m q_m}{p_1 q_1 + \dots + p_M q_M}$$

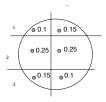
- ► MAP = value of m that maximizes $p_m q_m$.
- ▶ MLE = value of m that maximizes q_m .

Independence

Recall:

A and B are independent $\Leftrightarrow Pr[A \cap B] = Pr[A]Pr[B]$ $\Leftrightarrow Pr[A|B] = Pr[A].$

Consider the example below:

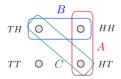


 $\begin{array}{l} (A_2,B) \text{ are independent: } Pr[A_2|B] = 0.5 = Pr[A_2]. \\ (A_2,\bar{B}) \text{ are independent: } Pr[A_2|\bar{B}] = 0.5 = Pr[A_2]. \\ (A_1,B) \text{ are not independent: } Pr[A_1|B] = \frac{0.1}{0.5} = 0.2 \neq Pr[A_1] = 0.25. \end{array}$

Pairwise Independence

Flip two fair coins. Let

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



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

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

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

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

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

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

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

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

Proof:

Left as an exercise!

Example 2

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

Then

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

Also,

 A_1 and $A_3 \cap A_5$ 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,

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

This leads to a definition

Balls in bins

One throws m balls into n > m bins.



Mutual Independence

Definition Mutual Independence

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

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

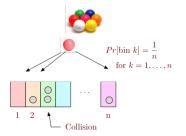
(b) More generally, the events $\{A_j, j \in J\}$ are mutually independent if

$$Pr[\cap_{k\in K}A_k] = \prod_{k\in K}Pr[A_k]$$
, 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.

Balls in bins

One throws m balls into n > m bins.



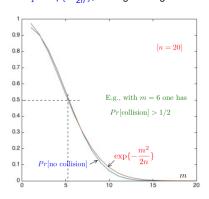
Theorem:

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

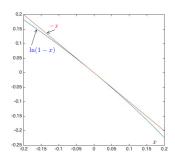
Balls in bins

Theorem:

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



Approximation



$$\exp\{-x\} = 1 - x + \frac{1}{2!}x^2 + \dots \approx 1 - x$$
, 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.

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

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

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

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

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

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

With 11 = 000, one in as

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

If m = 60, we find that

$$\label{eq:pressure} \textit{Pr}[\text{no collision}] \approx \exp\{-\frac{m^2}{2n}\} = \exp\{-\frac{60^2}{2\times365}\} \approx 0.007.$$

If m = 366, then Pr[no collision] = 0. (No approximation here!)

The Calculation.

 A_i = no collision when *i*th ball is placed in a bin.

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

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

Product rule:

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

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

Hence,

$$\ln(Pr[\text{no collision}]) = \sum_{k=1}^{m-1} \ln(1 - \frac{k}{n}) \approx \sum_{k=1}^{m-1} (-\frac{k}{n})^{\binom{n}{n}}$$

$$= -\frac{1}{n} \frac{m(m-1)}{2}^{\binom{n}{n}} \approx -\frac{m^2}{2n}$$

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

(†)
$$1+2+\cdots+m-1=(m-1)m/2$$
.

Checksums!

Consider a set of *m* files.

Each file has a checksum of b bits.

How large should b be for Pr[share a checksum] $< 10^{-3}$?

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

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,

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

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

Collect all cards?

Thus,

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

Hence,

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

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

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

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.

$$Pr[A_m] = (1 - \frac{1}{n}) \times \dots \times (1 - \frac{1}{n})$$

$$= (1 - \frac{1}{n})^m$$

$$(Pr[A_n]) = mln(1 - \frac{1}{n}) \times m \times (\frac{1}{n})$$

$$ln(Pr[A_m]) = mln(1 - \frac{1}{n}) \approx m \times (-\frac{1}{n})$$

 $Pr[A_m] \approx exp\{-\frac{m}{n}\}.$

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

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: $ln(1-\varepsilon) \approx -\varepsilon$.

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:

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

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

Plug in and get

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