Conditional Probability: Review

Recall:

- \( P[A|B] = \frac{P[A \cap B]}{P[B]} \)
- \( A \) and \( B \) are positively correlated if \( P[A|B] > P[A] \), i.e., if \( P[A \cap B] > P[A]P[B] \).
- \( A \) and \( B \) are negatively correlated if \( P[A|B] < P[A] \), i.e., if \( P[A \cap B] < P[A]P[B] \).
- \( A \) and \( B \) are independent if \( P[A|B] = P[A] \), i.e., if \( P[A \cap B] = P[A]P[B] \).
- Note: \( B \subseteq A \Rightarrow A \) and \( B \) are positively correlated. \( P[A|B] = 1 > P[A] \)
- Note: \( A \cap B = \emptyset \Rightarrow A \) and \( B \) are negatively correlated. \( P[A|B] = 0 < P[A] \)

Conditional Probability: Pictures

Illustrations: Pick a point uniformly in the unit square

\[
\begin{align*}
\text{Left: } A \text{ and } B \text{ are independent. } P[B] = b; P[B|A] = b. \\
\text{Middle: } A \text{ and } B \text{ are positively correlated. } P[B|A] = b_1 > P[B|\bar{A}] = b_2. \text{ Note: } P[B] \in (b_1, b_2). \\
\text{Right: } A \text{ and } B \text{ are negatively correlated. } P[B|A] = b_1 < P[B|\bar{A}] = b_2. \text{ Note: } P[B] \in (b_1, b_2).
\end{align*}
\]

Bayes: General Case

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

\[
\begin{align*}
P[A_m] &= \rho_m, m = 1, \ldots, M \\
P[B|A_m] &= \omega_m, m = 1, \ldots, M, P[A_m \cap B] = \rho_m \omega_m \\
P[B] &= \rho_1 \omega_1 + \cdots + \rho_M \omega_M \\
P[A_m|B] &= \frac{\rho_m \omega_m}{\sum_m \rho_m \omega_m} = \text{fraction of } B \text{ inside } A_m
\end{align*}
\]
Why do you have a fever?

Using Bayes’ rule, we find

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

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

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

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

One says that ‘Flu’ is the Most Likely a Posteriori (MAP) cause of the high fever.
‘Ebola’ is the Maximum Likelihood Estimate (MLE) of the cause: it causes the fever with the largest probability.

Recall that

\[ p_m = P(A_m), q_m = P(B|A_m), P[A_m|B] = \frac{p_m q_m}{p_1 q_1 + \cdots + p_M q_M} \]

Thus,

- MAP = value of \( m \) that maximizes \( p_m q_m \).
- MLE = value of \( m \) that maximizes \( q_m \).

Why do you have a fever?

We found

\[ P(\text{Flu}|\text{High Fever}) = 0.58, \]
\[ P(\text{Ebola}|\text{High Fever}) \approx 5 \times 10^{-8}, \]
\[ P(\text{Other}|\text{High Fever}) \approx 0.42 \]

Thomas Bayes

A Bayesian picture of Thomas Bayes.


Bayes’ Rule Operations

Bayes’ Rule is the canonical example of how information changes our opinions.
Mutual Independence

Definition Mutual Independence

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

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

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

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

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

Pairwise Independence

Flip two fair coins. Let

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

If $A$ did not say anything about $C$ and $B$ did not say anything about $C$, then $A \cap B$ would not say anything about $C$.

Independence

Recall:

$n$ and $B$ are independent

$\iff Pr[A \cap B] = Pr[A]Pr[B]$  

$\iff Pr[A|B] = Pr[A].$

Consider the example below:

$(A_2, B)$ are independent: $Pr[A_2|B] = 0.5 = Pr[A_2].$

$(A_2, \overline{B})$ are independent: $Pr[A_2|\overline{B}] = 0.5 = Pr[A_2].$

$(A_1, B)$ are not independent: $Pr[A_1|B] = \frac{1}{8} \neq Pr[A_1] = 0.25.$

This leads to a definition ....

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

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

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

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

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

Proof:

See Notes 25, 2.7.

Example 2

Flip a fair coin 5 times. Let $A_n = \text{‘coin } n \text{ is H’}$, for $n = 1, \ldots, 5$. Then,

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

Also,

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

Indeed,

$$Pr[A_1 \cap (A_2 \cap A_5)] = \frac{1}{8} = Pr[A_1]Pr[A_2 \cap A_5].$$

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.
Balls in bins

One throws \(m\) balls into \(n > m\) bins.

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

In particular, \(\Pr[\text{no collision}] \approx 1/2\) for \(m = \sqrt{n}\). i.e.,
\[
m \approx \sqrt{2 \ln(2) n} \approx 1.2 \sqrt{n}.
\]
Roughly, \(\Pr[\text{collision}] \approx 1/2\) 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
\[
\Pr[\text{collision}] \approx 1/2 \text{ if } m \approx 1.2 \sqrt{365} \approx 23.
\]
If \(m = 60\), we find that
\[
\Pr[\text{no collision}] \approx \exp\left(-\frac{m^2}{2n}\right) = \exp\left(-\frac{60^2}{2 \times 365}\right) \approx 0.007.
\]
If \(m = 366\), then \(\Pr[no \text{ collision}] = 0.\) (No approximation here!)
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) $\Pr[\text{miss any one of the items}] \leq ne^{-\frac{m}{n}}$.

Thus,

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

Hence,

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

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

**Claim:**

One random baseball card in each cereal box.

Let $H$ be the number of baseball cards in the $m$ cereal boxes.

Hence,

$$\Pr[H \leq 102] = \Pr[H \leq 530] = \Pr[H \leq 7600].$$

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

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 $\Pr[\text{no collision}] = \exp\left(-\frac{m^2}{2n}\right) \approx 1 - \frac{m^2}{2n}$.

Hence,

$$\Pr[\text{no collision}] \approx 1 - 10^{-3} \Rightarrow m^2/(2n) \approx 10^{-3} \Rightarrow 2n \approx m^2 10^{-3} \Rightarrow m^2 10^3 \Rightarrow m = 2^{10} \Rightarrow \log_2(m) \approx 10 + 2\log_2(10).$$

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

Collect all cards?

**Experiment:** Choose $m$ cards at random with replacement.

Events: $E_k = \text{fail to get player } k$, for $k = 1, \ldots, n$

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

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

Summary.

Bayes’ Rule, Mutual Independence, Collisions and Collecting

Main results:

- **Bayes’ Rule:** $\Pr[A|B] = \frac{\Pr[A] \Pr[B|A]}{\Pr[B] \Pr[A]}$.
- **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 $n$ bins.

$$\Pr[\text{no collisions}] = \exp\left(-\frac{m^2}{2n}\right).$$

- **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: $\ln(1 - \epsilon) \approx -\epsilon$. 

Collect all cards?

Thus,

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

Coupon Collector Problem: Analysis.

Event $A_m = \text{fail to get Brian Wilson in } m \text{ cereal boxes}$

Fail the first time: $(1 - \frac{1}{n})$

Fail the second time: $(1 - \frac{1}{n})^2$

And so on ... for $m$ times. Hence,

$$\Pr[A_m] = (1 - \frac{1}{n}) \times \cdots \times (1 - \frac{1}{n}) = (1 - \frac{1}{n})^m.$$ 

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

$$\Pr[A_m] \approx \exp\left(-\frac{m}{n}\right).$$

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