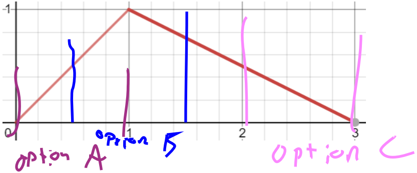
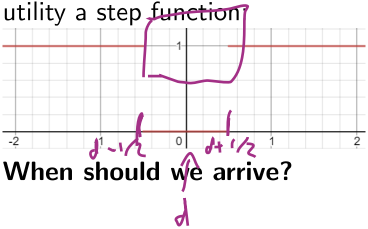
Oct 12 Probability Review

Suppose we're attending a 3-hour seminar (ugh), but at least snacks are served. We don't really care about the talk, but we want to be there for the second snacks arrive. Based on our prior knowledge of boring seminars, we have an idea that the probability of snack-arrival-time is the triangle below:



1. What's our expected loss?

We 're actually ok arriving anytime within 30 min of snack time. If we're a little early, we have a little more to chat about from the end of the last talk. If we're a little late, at least the snacks are still warm. This makes our utility a step function:



$L(d,x)$

$L = \begin{cases} 0 & \text{if } d-\frac{1}{2} < x < d+\frac{1}{2} \\ 1 & \text{else} \end{cases}$

When should we arrive?

2. How does this vary if we: ignore uncertainty, use the Bayes' decision, and/or have perfect information?

$E[\text{loss}] = \int \text{Loss times } f(x)$

"there is a 1hr window with no loss"

Announcements and To-Dos

Announcements:

1. AB pruning another example (with the α and β tracked): `https://youtu.be/xBXHtz4Gbdo`

2) Lectures can be streamed on Canvas.

3) Exam deadline w.

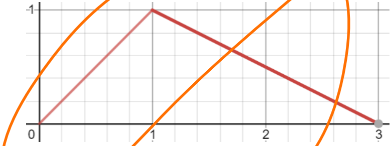
Last time we learned:

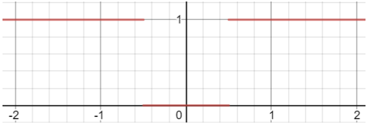
1. Finished up with EVs.

Our decisions

1. d_{Bayes} will be the "best" on-average decision, called the **Bayes' decision**. It minimizes *expected* loss. We will by trying to find this decision in some future algorithms.
2. We could also make other decisions: d_{pi} might be the decision we make with perfect information: *after* observing the "uncertain" elements.
3. We could also make other decisions: d_{iu} might be the decision we make ignoring the shape of the uncertainty and instead focusing on just its mean or average value.

When should we arrive?





d_{pi} asks: when do we arrive if we *are told* when snack time will be?

pm (–

P(snacktime = 34 min) = 1

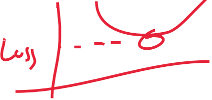
p(other times) = 0

choose d = 34 min

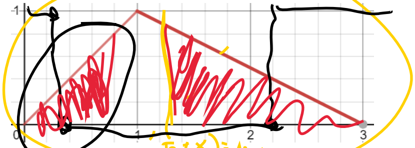
(or ~~19 – 49~~ 4–64 min)

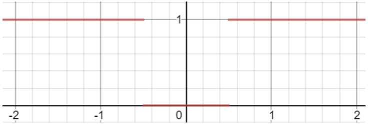
⇒ E[loss] = 0

When should we arrive?



$$\int_0^3 x p(x)\,dx \to \int_0^1 x(x)\,dx + \int_1^3 x\left(-\frac{x}{2}+\frac{3}{2}\right) dx = \frac{1}{3} + 1$$





d_{ui} is the best decision if we only know that the **average** snack time is 80 min in?

$E[\text{loss} \mid d = 4/3] =$

$\int \text{Loss Fn} \cdot p(x)\,dx$

Lose 1 if $x < 5/6$ $\Rightarrow P(t<.5) = \frac{1}{2}bh = \frac{1}{2}(5/6)(5/6)$

Lose 1 if $x > 11/6$ $\Rightarrow$

Choose $\boxed{d = 4/3}$

$(0, 5/6\ ~~1/3~~ - ~~7/3~~)\ 11/6$

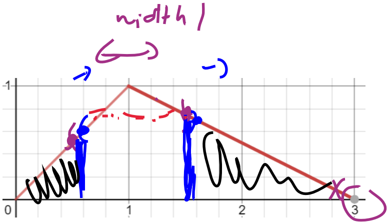
Loss function is:

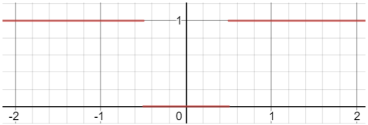
1 if $x < 11/3$

1 if $x > 7/3$

0 else.

When should we arrive?





d_{Bayes} is the best decision to minimize expected loss.

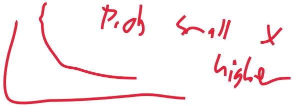
where is the window to hold the most gain

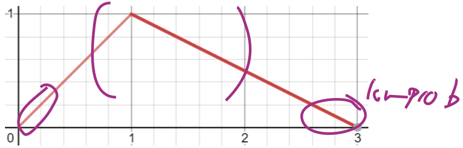
$$E[loss_d] = \int loss \cdot f(x)dx$$

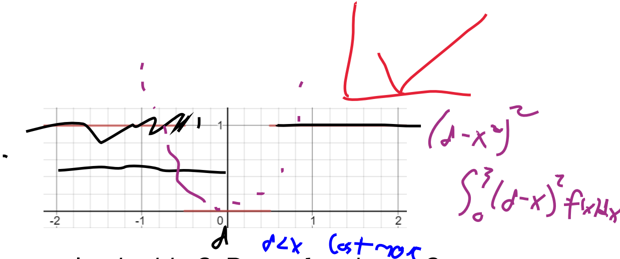
< Shuttles were 30min or more before 'd'

$$= \int_0^{d-1/2} f(x)dx + \int_{d+1/2}^{3} f(x)dx$$

When should we arrive?







$(d-x^2)^2$

$\int_0^3 (d-x)^2 f(x)dx$

What happens to d_{Bayes} if the cost of being too early *doubles*? Does d_{ui} change?

dBayes should move →

Loss

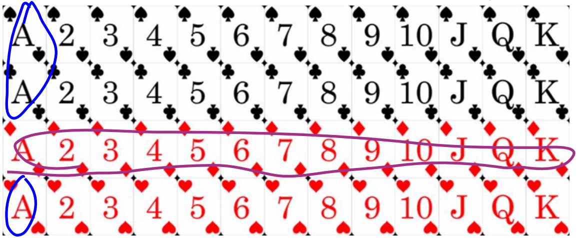
$$E[\text{loss}] = \int \text{loss} \cdot \text{outcomes} = \int_0^{d-1/2} \left(\frac{1}{2}\right) f(x)dx + \underbrace{\int_{d+1/2}^{3} (1)\, f(x)\, dx}_{P(X > d+1/2)\cdot 1}$$

$P(X < d-1/2)$

A Review of Probability

4 Aces

Suppose we have a standard 52-card played deck:



$13/52$

This has 4 suits, each of which contain the same 13 values.

Opening Soln'

Example: Suppose we draw a card from a traditional 52-card playing deck. What is the probability that the card is the $A\diamondsuit$? What is the probability that the card is either an A or a $\diamondsuit$?

V

suit / color

Opening Soln'

Example: Suppose we draw a card from a traditional 52-card playing deck. What is the probability that the card is the $A\diamondsuit$? What is the probability that the card is either an A or a $\diamondsuit$? Typically we

1. $P(\{A\diamondsuit\}) = \frac{1}{52}$
2. $P(\{A \cup \diamondsuit\}) = P(\{A\}) + P(\{\diamondsuit\}) - P(\{A \cap \diamondsuit\}) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$

inclusion - exclusion

→ 0 if "exclusive"

$$P(A\ \underline{\text{or}}\ B) = P(A) + P(B) - P(\text{both})$$

Opening Soln'

Example: Suppose we draw a card from a traditional 52-card playing deck. What is the probability that the card is the $A\diamondsuit$? What is the probability that the card is either an A or a $\diamondsuit$? (Sometimes we're interested in tracking multiple outcomes or multiple random variables at the same time! We call these *joint* or *conditional* probabilities and distributions. Let V denote the card value and C its color.)

1. **Definition:** The *joint probability* of $V = v$ and $C = c$ is the probability that both outcomes occur simultaneously, denoted $P(V = v, C = c)$.
2. **Definition:** The *conditional probability* of $V = v$ *given* $C = c$ is the probability that the card value is v if you already know that the card color was c. We denote this $P(V\underset{\uparrow}{|}C)$.
3. For example, $P(V = 6, C = \text{'red'}) = \overset{2}{\cancel{1}}/52$, but $P(V \underset{\nearrow}{=} 6|C = \text{'red'}) = 2/26$

On conditionals

Computing conditional probabilities often relies on revisiting some base rules:. The main two are that if we want to compute an AND statement of two probabilities, it tends to turn into multiplication. OR statements turn into addition. If the OR is exclusive, it's simple addition; if the OR is inclusive we have to worry about overlap!

1. One option for computing conditionals is to use their mathematical definition:

   (P(B))

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

   (A|B))

   $P(A$ $\underline{\text{given}}$ $B) = \dfrac{P(\text{both})}{P(B)}$

   provided that $P(B) > 0$.

2. This leads to the **Multiplication Rule:** for AND statements:
   - $P(A \cap B) = P\underline{(A|B)}P(B)$
   - $P(A \cap B) = P\underline{(B|A)}P(A)$

A conditional chain rule

Breaking complex AND statements into a long list of conditionals leads to what is sometimes called the probability *chain rule*. Suppose we have 3 random outcomes A, B and C.

C AND A & B

$$p(\boxed{A, B}, C) = P(C|\boxed{A, B})P(A, B) = P(C|A, B)P(B|\boxed{A})P(\boxed{A})$$

In general, we have D random variables $X_1, X_2, \dots X_D$:

$$P(X_1, X_2, \dots X_D) = P(\boxed{X_D}|X_1, X_2, \dots X_{D-1})P(X_{D-1}|X_1, X_2, \dots X_{D-2}) \cdots P(X_2|X_1)P(X_1)$$

lost one | all prior

· P(next-to-last | all prior)

· P(X_{D-2} | all prior...)

which we might "slice" shorthand as:

$$P(X_{1:D}) = P(X_D|X_{1:D-1})P(X_{D-1}|X_{1:D-2}) \cdots P(X_2|X_1)P(X_1)$$

Example What is the probability of being dealt a flush in poker (five cards)?

Example What is the probability of being dealt a flush in poker (five cards)?

↳ all 5 cards same suit

Solution: Two ways

1. Count all possible selections of five cards - $C(52, 5)$ - then count all possible selections of flushes: $C(13, 5)$ for the values on the flush and $C(4, 1)$ for the possible suits. Then

$$P(\textit{flush}) = \frac{C(13,5)C(4,1)}{C(52,5)}$$

Discrete

2. Do things *conditionally*:

P(card 2 & card #1 match)

= P(card 2 is suit of card 1 | card 1's suit) · P(card 1 has a suit)

$$P(\textit{all 5 cards same suit})$$
$$= P(\textit{cards 1-4 match suit AND card 5 matches that suit})$$
$$= P(\textit{cards 1-4 match suit})P(\textit{ card 5 matches that suit GIVEN cards 1-4 match suit})$$
$$= \cdots = \frac{52}{52}\boxed{\frac{12}{51}}\boxed{\frac{11}{50}}\underset{4}{\boxed{\frac{10}{49}}}\underset{5}{\boxed{\frac{9}{48}}}$$

Law of Total Probability

In order to turn complex probabilities into simpler computations, we often try to break down processes into lists of ANDs and *exclusive* ORs.

Definition: The *Law of Total Probability* tells us that if an event can happen in only a few possible *exclusive* scenarios, we can compute the probability of that event by only adding up those scenarios.

In full notation: if the only way E can happen is *one and exactly one* of the events $S_1, S_2, \ldots S_n$ (or those events are *exhaustive* and *exclusive*), then:

0 or Ø or ♡

$$P(E) = \sum_{i=1}^{n} \boxed{P(E \text{ and } S_i)} = P(E|S_i)p(S_i)$$

↑ multiplication / conditioning

This can also be used to compute the *marginal* probability of just one outcome if we're given a joint density $P(V = v, C = c)$:

$$P(V) = \sum_{c} P(V = v \text{ and } C = c) = P(V|C = c)p(C = c)$$

Bayes'

Bayes Theorem:

$$P(A|B) \overset{\text{defn. cond. Prob}}{=} \frac{P(A \cap B)}{P(B)} \overset{\text{multiplication}}{=} \frac{P(B|A)P(A)}{P(B)}$$

Example: An iconic Bayes' Theorem problem:

1. Let's assume 1% of people over the age of 40 have cancer.
2. Suppose we create a medical test such that 90% of the people with cancer will test positive (known as a false negative rate of 10%).
3. Suppose also that 8% of the people without cancer test positive under our test (a false positive rate of 8%).

What's the probability that a patient has cancer **given** that they test positive?

Bayes'

$P(C^c) = .99$

Example: An iconic Bayes' Theorem problem: $P(C) = .01$

1. Let's assume 1% of people over the age of 40 have cancer.
2. Suppose we create a medical test such that 90% of the people with cancer will test positive (known as a false negative rate of 10%). $P(+ | C) = .90$
3. Suppose also that 8% of the people without cancer test positive under our test (a false positive rate of 8%). $P(+ | C^c) = .08$

What's the probability that a patient has cancer **given** that they test positive? **Solution:** Let's use C for cancer and $+$ for testing positive. We have:

+ test happen in 2 exclusive ways:

$P(C) = 0.01$; $P(+|C) = .9$; $P(+|C^C) = .08$. We want:

true + $(+\cap C)$

$\boxed{P(C|+)} = \boxed{\frac{P(+|C)P(C)}{P(+)}} = \frac{P(+|C)P(C)}{\underbrace{P(+|C)P(C)}_{+\cap C} + \underbrace{P(+|C^C)P(C^C)}_{+\cap C^c}}$, all of which we were given:

false + $(+\cap C^c)$

$P(C|+) = \frac{0.9 \cdot 0.01}{0.9 \cdot 0.01 + 0.08 \cdot 0.99} = 0.10$ So there's a 10% chance you have cancer given that you test positive.

Probability Roundup

1. Probabilities must sum to 1 when considering all possible outcomes.

2. $P(A \textbf{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \textbf{ and } B)$.

3. $P(A \textbf{ given } B) = P(A|B) = \frac{P(A \textbf{ and } B)}{P(B)}$, and represents our thoughts about A after gaining the knowledge that event B definitely happened.

4. The *multiplication rule:* $P(A \textbf{ and } B) = P(A|B)P(B) = P(B|A)P(A)$

5. If events A and B are *independent*:
   $P(A \textbf{ and } B) = \underline{P(A)P(B)}$; $\quad P(A|B) = P(A)$; $\quad P(B|A) = P(B)$.

6. The Law of Total Probability

$$P(A) = P(A|E_1)P(E_1) + P(A|E_2)P(E_2) + \cdots + P(A|E_k)P(E_k)$$

7. **Bayes Theorem:** $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$

Naive Bayes

Example: Consider building an e-mail spam filter. We receive an e-mail with the words *buy*, *pills* and *deal*. Is this a SPAM e-mail, or valid (HAM)?

Definition: *Class conditional independence* is the assumption that the features of **x** are conditionally independent *given* **y**. This means that $P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$ We often make this naive assumption.

Under class conditional independence, we would be assuming that
$P(x = [\text{buy, pills, deal}]|y = \text{SPAM}) = P(\text{buy}|\text{SPAM})P(\text{pills}|\text{SPAM})P(\text{deal}|y = \text{SPAM})$
The Naive Bayes Classifier is just Bayes' theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

where we have y as the event of "it's spam" and x as the words in the e-mail.

Naive Bayes

We use some special vocabulary when we use Bayes theorem to describe the four pieces.

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

Definition: $P(x)$ and $P(y)$ are marginal distributions of x and y. Because we're trying to compute a final result on $P(y)$, the right-hand side $P(y)$ is called our prior distribution: it's what we know about y prior to observing any data x.

Definition: $P(x|y)$ is called the *likelihood*. Given a class y, how would we observe x? It typically comes from an assumed pdf of x.
For our classifier, it's the likelihood of a specific e-mail or set of words *given* that the e-mail is SPAM (or HAM). For a classification problem, it's thusly sometimes called the class-conditional probability.

Definition: $P(y|x)$ and $P(y)$ is the *posterior distribution*. It holds the (classification) probability that data y belongs to class c, given observation of its features x.

SPAM and HAM

Example: Suppose we get a new e-mail that's just the word "money." Given the following set of messages, compute $P(money|SPAM)$ and $P(money|HAM)$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

SPAM and HAM

Example: Suppose we get a new e-mail that's just the word "money." Given the following set of messages, compute $P(money|SPAM)$ and $P(money|HAM)$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Solution: For this small set of e-mails, these are just discrete values from counting the numbers of occurrences of each event. $P(money|SPAM) = \frac{|(both)|}{|(SPAM)|} = \frac{2}{3}$

$P(money|HAM) = \frac{|(both)|}{|(HAM)|} = \frac{1}{2}$

SPAM and HAM

Example, Cont'd:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

We want $P(SPAM|money)$; $P(HAM|money)$. We can get there with Bayes' Theorem:

$$P(SPAM|w) = \frac{P(w|SPAM)P(SPAM)}{P(w)}$$

and we can fully write out the denominator using LTP:

$$P(SPAM|w) = \frac{P(w|SPAM)P(SPAM)}{P(w|SPAM)P(SPAM) + P(w|HAM)P(HAM)}$$

SPAM and HAM

Example, Cont'd: Compute $P(SPAM|money)$ and $P(HAM|money)$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

SPAM and HAM

Example, Cont'd: Compute $P(SPAM|money)$ and $P(HAM|money)$:

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Solution:

$$P(SPAM|money) = \frac{P(money|SPAM)P(SPAM)}{P(money|SPAM)P(SPAM) + P(money|HAM)P(HAM)}$$

$$P(SPAM|money) = \frac{\frac{2}{3}\frac{3}{5}}{\frac{2}{3}\frac{3}{5} + \frac{1}{2}\frac{2}{5}} = \frac{2}{3}$$

Sanity check: 2 of the 3 times "money" appears is in a SPAM message!

Classifiers

So imagine we get an e-mail with "money" in it. According to our system, how do we *classify* it?

Since $P(SPAM|money) > P(HAM|money)$ (or more than half the time it's SPAM), it's probably reasonable to classify the e-mail as SPAM.

But we should always be worried about making mistakes, and estimating the probability that our classifier gets things wrong! Marking a SPAM message as HAM is annoying but not a big deal. Missing a crucial HAM e-mail because it got filtered out might be disastrous... is $P(SPAM|money) > .5$ the right threshold for this type of problem?

Bigger Better BSFs

Our initial math can filter messages based on the appearances of certain words, but we might want to do this with all of the word simultaneously. We can do that! (with some extra assumptions...)

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Suppose we receive an email with just the words {buy, nigeria}. Now we want to compute things like $P(SPAM|\{buy, nigeria\})$ and $P(HAM|\{buy, nigeria\})$. In turn, we'll have to compute the pieces like $P(\{buy, nigeria\}|HAM)$.

Data Science!

For pieces like $P(\{buy, nigeria\}|HAM)$ now we have a *set* of words: we could tally up e-mails that contain *both* words, but this might be very tedious to compute if we receive new emails.

Scary Example: How many distinct *sets* of words are in the sentence "This is not a spam e-mail about money from nigeria"?

Maybe we don't want to count all possible subsets of words from e-mails, so we might do all the math where we assume that word are *independent.* Is this a good assumption?

Data Science!

For pieces like $P(\{buy, nigeria\}|HAM)$ now we have a *set* of words: we could tally up e-mails that contain *both* words, but this might be very tedious to compute if we receive new emails.

Scary Example: How many distinct *sets* of words are in the sentence "This is not a spam e-mail about money from nigeria"?

Solution: This is the *Power set*! It's asking about all of the ways we can chop up the e-mail into subsets. For even that 10 word sentence that $2^{10} = 1024$ sets. Yikes!

Maybe we don't want to count all possible subsets of words from e-mails, so we might do all the math where we assume that word are *independent.* Is this a good assumption?

Data Science!

To compute $P(\{buy, nigeria\}|HAM)$ where we assume the words are *independent*, we have an **AND** statement before the conditional... we can just use multiplication!

$$P(\{buy, nigeria\}|HAM) = P(buy|HAM)P(nigeria|HAM)$$

Then we can reuse all of our work for individual words, and multiply the values for all of our individual words together to get the resulting probabilities for an e-mail. Note that the denominators are the same:
$P(HAM|email) = \frac{P(email|HAM)P(HAM)}{P(email)}$;
$P(SPAM|email) = \frac{P(email|SPAM)P(SPAM)}{P(email)}$
The process of making simplifications like independence to save from a computationally disastrous "exact" answer like all possible subsets is a major concern in Data Science.

Bigger Better BSFs

A final BSF: predict if {buy, nigeria} is SPAM.

TYPE:	ham	spam	spam	spam	ham
	work buy money	nigeria money viagra	fly buy nigeria	money buy fly	fly home nigeria

Bigger Better BSFs

A final BSF: predict if {buy, nigeria} is SPAM.

TYPE:	ham	spam	spam	spam	ham
	work	nigeria	fly	money	fly
	buy	money	buy	buy	home
	money	viagra	nigeria	fly	nigeria

Solution: $P(SPAM|\{buy, nigeria\})) = \frac{P(buy|SPAM)P(nigeria|SPAM)P(SPAM)}{P(\{buy,nigeria\})}$. From before, we have the *numerator* values as $\frac{2}{3}\frac{2}{3}\frac{3}{5} = 0.267$.

$P(HAM|\{buy, nigeria\}) = \frac{P(buy|HAM)P(nigeria|HAM)P(HAM)}{P(\{buy,nigeria\})}$. From before, we have the *numerator* values as $\frac{1}{2}\frac{1}{2}\frac{2}{5} = 0.1$.
Great idea: let's just compare those and ignore the denominators, since they're the same anyways! Since $0.267 > 0.1$, we predict the e-mail is SPAM.

Moving Forward

- This Week:
    1. nb day Friday
- Next time: AI meets Bayes’