ILLINOIS INSTITUTE OF TECHNOLOGY

Mathematics Review

CS-585

Natural Language Processing

Sonjia Waxmonsky

Slides based in part on material from:

- *Artificial Intelligence: A Modern Approach, 2nd Edition*
  Russell & Norvig (Prentice-Hall: 2003)
- Slides by Patrick Nichols (MIT), Derrick Higgins (IIT)

PROBABILITY THEORY REVIEW

Probability: Vocabulary

Some concepts we will cover today:

- *Complement*
- *Conditional probability*
- *Prior probability*
- *Posterior probability*
- *Random variable (binary vs multi-valued)*
- *Independent variables*
- *Conditional independence*
- *Chain rule*
- *Atomic event*
- *Bayes Rule*

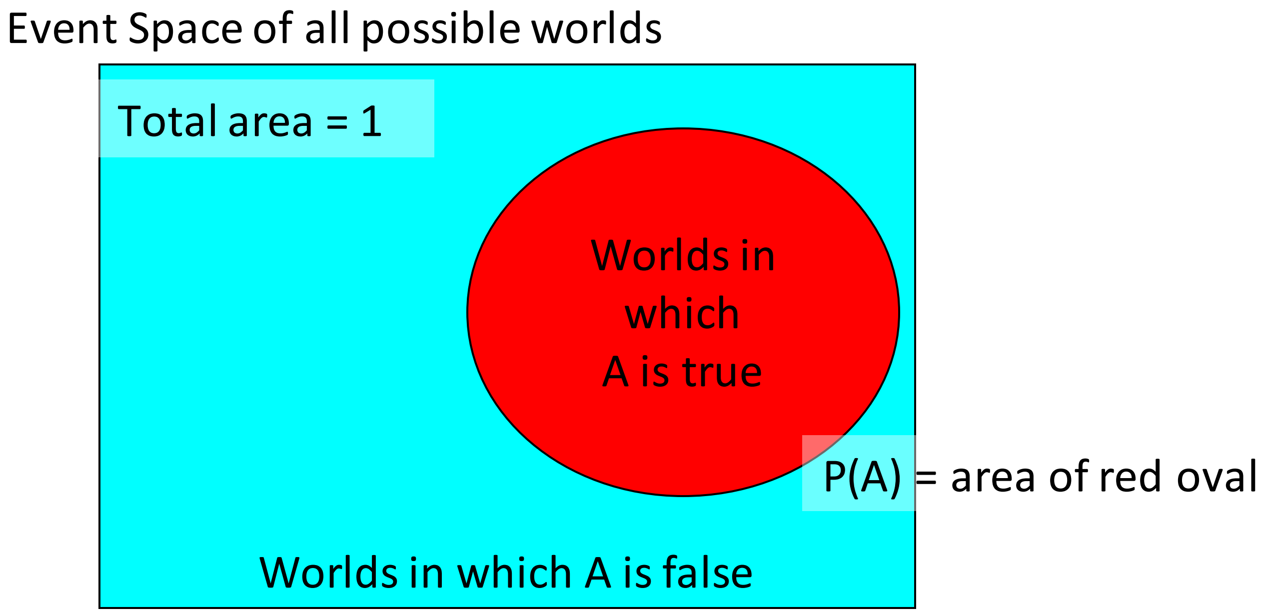
Probability and NLP

Why do we need probability for modern NLP?

- Based on what we can observe, what is the *most likely* label or outcome of multiple possibilities?
  - *What is the best interpretation of "Time flies like an arrow"?* [Parsing, part-of-speech tagging]
  - *How many stars would the writer of this review give to this movie?* [Sentiment analysis]
  - *What is the best response to "Who is the vice president"?* [Question answering]

Probability: Intuitive

- P(*A*) denotes "fraction of possible worlds (given what I know) in which *A* is true"

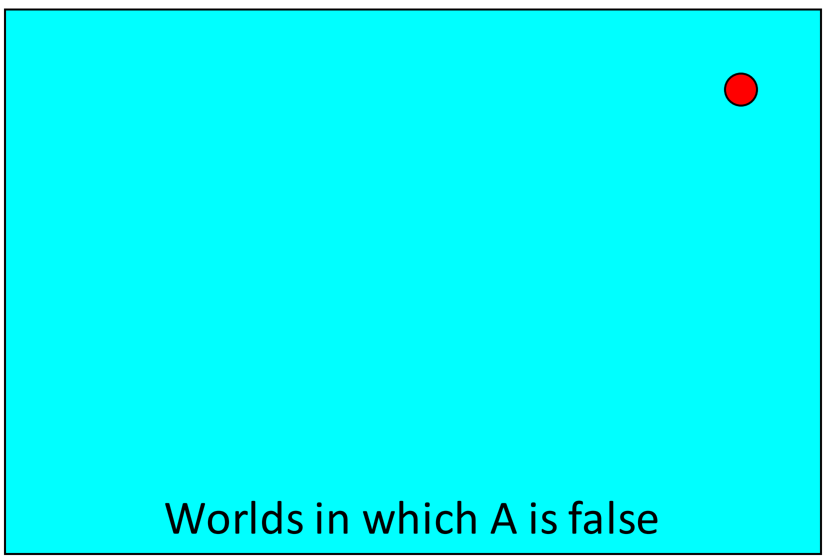


Probability: Axioms

- $0 \le P(A) \le 1$
- $P(\mathit{true}) = 1$
- $P(\mathit{false}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$

Probability: Axioms

- **$0 \leq P(A)$** ≤ 1
- $P(true) = 1$
- **$P(false) = 0$**
- $P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$

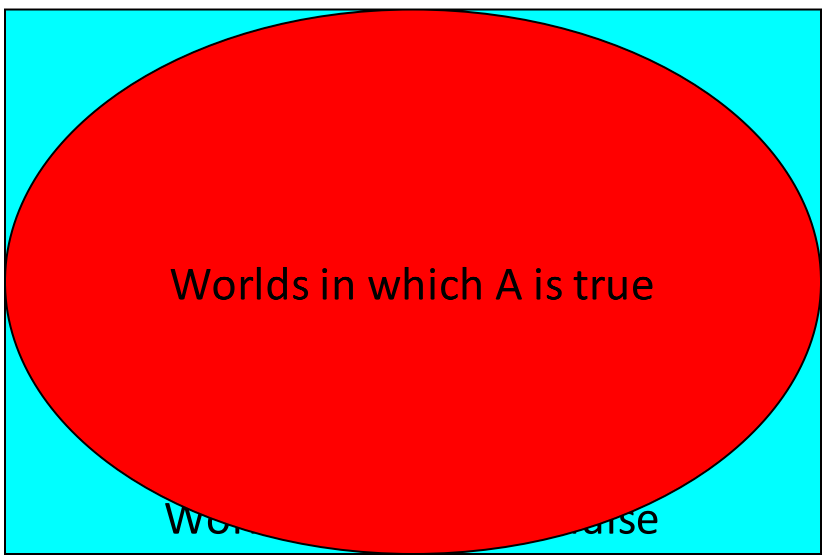


Red oval can’t get smaller than 0

Area of 0 means that A is true in **no** possible worlds…

Probability: Axioms

- $0 \leq$ **$P(A) \leq 1$**
- **P(*true*) = 1**
- P(*false*) = 0
- P(*A* or *B*) = P(*A*) + P(*B*) - P(*A* & *B*)

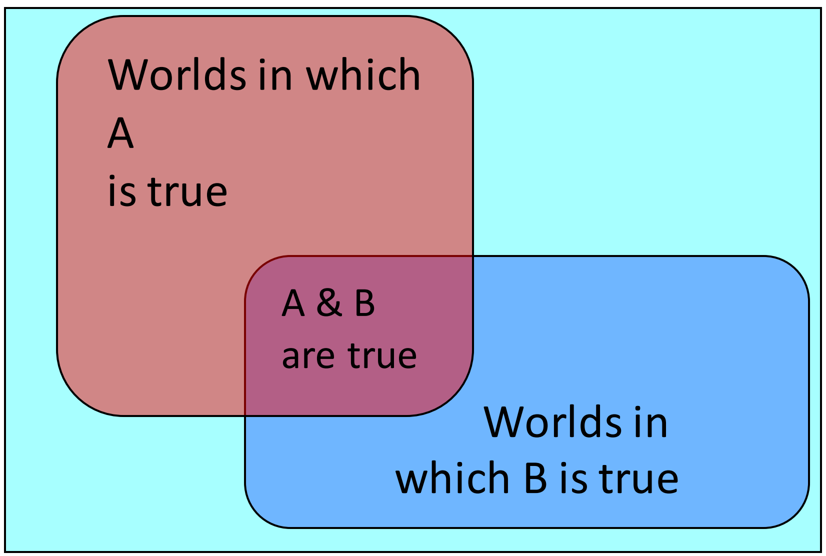


Red oval can't get larger than 1

Area of 1 means that A is true in **all** possible worlds…

Probability: Axioms

- $0 \le P(A) \le 1$
- $P(\mathit{true}) = 1$
- $P(\mathit{false}) = 0$
- P(A or B) = P(A) + P(B) - P(A & B)



Size of union is sum of sizes
minus size of intersection

Some Provable Facts

<u>Axioms</u>:

- $0 \leq P(A) \leq 1$
- $P(\mathit{true}) = 1$
- $P(\mathit{false}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$

We can show that:

- $P(\sim A) = P(\text{not } A) = 1 - P(A)$

And furthermore:

- $P(A) = P(A \,\&\, B) + P(A \,\&\, \sim B)$

Here P(~A) is the *complement* of A

Multivalued Random Variables

Suppose A can take on more than 2 values

Example:
Part of Speech (*POS):* {noun, verb, adjective, adverb}

Call A a **random variable with arity k** if A can take on one of **k** different values in some set {v1, v2, ..., vk}

Thus:

- $P(A=\mathrm{v}i \;\&\; A=\mathrm{v}j) = 0$ if $i \neq j$
- $P(A=\mathrm{v}1 \text{ or } A=\mathrm{v}2 \text{ or } \ldots \text{ or } A=\mathrm{v}k) = 1$

Easy Facts About Multivalued RVs

Axioms:

- $0 \le P(A) \le 1$; $P(true) = 1$; $P(false) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$

Recall:

- $P(A=vi \,\&\, A=vj) = 0$ if $i \ne j$; $P(A=v1 \text{ or } A=v2 \text{ or } \dots \text{ or } A=vk) = 1$
- We can show that:

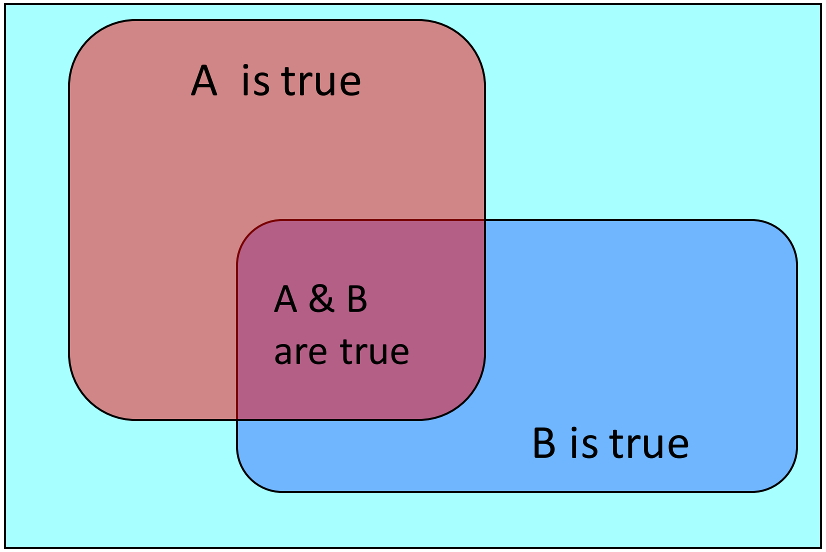
$$P(A = v1 \vee A = v2 \vee \cdots \vee A = vi) = \sum_{j=1}^{i} P(A = vj)$$

- And therefore:

$$P(A = v1 \vee \cdots \vee A = vk) = \sum_{j=1}^{k} P(A = vj)$$

Conditional Probability

- P(*A*|*B*) = "probability of *A* **given** *B*" = fraction of possible worlds with *B* true that also have *A* true



P(Headache) = 0.1
P(Flu) = 0.02
P(Headache|Flu) = 0.5

"Headaches are rare, Flu is much rarer, but if you have the Flu, you have a 50-50 chance of having a headache."

Conditional Probability

- *Formal definition:*

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

- *This gives us:*

$$P(A \wedge B) = P(A \mid B)P(B)$$

Chain Rule

From Conditional Probability:

$$P(A \wedge B) = P(A|B)\, P(B)$$

Thus, we have:

$$\begin{aligned} P(A \wedge B \wedge C) &= P(A|B \wedge C) P(B \wedge C) \\ &= P(A|B \wedge C)\, P(B|C)\, P(C) \end{aligned}$$

Generalizing:

$$P(A1 \wedge A2 \wedge \ldots \wedge An) = P(A1|A2 \wedge \ldots \wedge An)\, P(A2|A3 \wedge \ldots \wedge An)\, P(An)$$

Atomic Events

- Atomic event: A complete specification of the state of the world about which the agent is uncertain

  E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are 4 distinct atomic events:

  Cavity = false & *Toothache = false*
  Cavity = false & *Toothache = true*
  Cavity = true & *Toothache = false*
  Cavity = true & *Toothache = true*

- Atomic events are mutually exclusive and exhaustive

Prior probability

- Prior or unconditional probabilities of propositions

  e.g., P(*Cavity* = true) = 0.1 and P(*Weather* = sunny) = 0.72 correspond to belief prior to arrival of any (new) evidence

- Probability distribution gives values for all possible assignments:

  P(*Weather*) = <0.72,0.1,0.08,0.1> (normalized, i.e., sums to 1)

- Joint probability distribution for a set of random variables gives the probability of every atomic event on those random variables

  P(*Weather,Cavity*) = a 4 × 2 matrix of values:

Weather =	sunny	rainy	cloudy	snow
Cavity = true	0.144	0.02	0.016	0.02
Cavity = false	0.576	0.08	0.064	0.08

- Every question about a domain can be answered by the joint distribution

Inference

- Generally: Given some information about the probability distribution, determine the probability of some proposition ϕ

- ϕ = *Cavity*
- ϕ = *Cavity & Toothache*
- ϕ = *~Study & (GoodGrade* **or** *GoodJob)*

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$$

Inference by enumeration

- Start with the joint probability distribution:

	toothache		$\neg$ *toothache*	
	catch	$\neg$ *catch*	*catch*	$\neg$ *catch*
cavity	.108	.012	.072	.008
$\neg$ *cavity*	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$

- P(toothache or cavity) =

0.108 + 0.012 + 0.016 + 0.064 +0.072 + 0.008= 0.28

Inference by enumeration

- Start with the joint probability distribution:

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Can also compute conditional probabilities:

P(~*cavity* | *toothache*) ?

Inference by enumeration

- Start with the joint probability distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

- Can also compute conditional probabilities:

$$P(\sim cavity \mid toothache) = \frac{P(\sim cavity \ \& \ toothache)}{P(toothache)}$$

$$= \frac{0.016+0.064}{0.108 + 0.012 + 0.016 + 0.064}$$

$$= 0.4$$

Independence

Two boolean random variables A and B are said to be **independent** if and only if

$$\mathbf{P(A|B) = P(A)}$$

That is, the probability we give A is not affected by learning the probability of B

QUESTION: If $P(A|B) = P(A)$ can we show :

$$\mathbf{P(B|A) = P(B)}$$

Independence

Two boolean random variables A and B are said to be **independent** if and only if

$$\mathbf{P(A|B) = P(A)}$$

That is, the probability we give A is not affected by learning the probability of B

QUESTION: If P(A)=P(A|B) can we show P(B)=P(B|A) ?

$$\mathbf{P(A \wedge B)} = P(A|B)\,P(B) = \mathbf{P(A)\,P(B)}$$

$$\mathbf{P(B)} = \cancel{P(A \wedge B)/P(B)} = \mathbf{P(B|A)}$$

$P(A \wedge B)/P(A)$ ← correction

Independence

If A and B are independent boolean RVs then:

- **P(A | B) = P(A)** *(by definition)*
- **P(A ∧ B)** = P(A | B) P(B) = **P(A) P(B)**
- **P(B | A) = P(B)**

QUESTION:

If A and B are independent boolean variables, then are their complements ~A and ~B also independent? That is, can we prove the following?

- **P(A|B) = P(A)** if and only if **P(~A | ~B) = P(~A)**

Independence

If A and B are independent boolean RVs, can we show their complements are independent?

Have: $P(A)P(B) = P(A\&B)$

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

$P(\sim A\&\sim B)$ $= 1-P(A \text{ or } B)$

$$=1- P(A) - P(B) + P(A)P(B)$$
$$= (1-P(A))(1-P(B))$$
$$= \mathbf{P(\sim A)P(\sim B)}$$

Multivalued Independence

- For multivalued RVs A and B, A is independent of B iff

$$\forall u,v : P(A=u \mid B=v) = P(A=u)$$

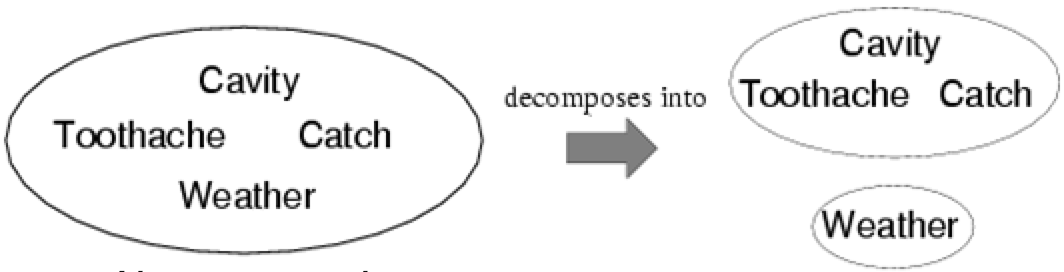
- From which we can show, for example:

$$\forall u,v : P(A=u \wedge B=v) = P(A=u)P(B=v)$$

$$\forall u,v : P(B=v \mid A=u) = P(B=v)$$

Independence

- So, suppose our domain knowledge allows us to make certain ***independence assumptions*** on our random variables:



P(*Toothache, Catch,*

= **P**(*Toothache, Catch, Cavity*) **P**(*Weather*)

  - 16 entries reduced to 10 , **Why**?
  - For *n* independent biased coins, $O(2^n) \rightarrow O(n)$

- Absolute independence powerful but rare…
  - Dentistry is a large field with hundreds of variables, none of which are really independent of each other. **What to do?**

Conditional independence

- "If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache."

- We can then say *Catch* is conditionally independent of *Toothache* given *Cavity:*

  (1) $\mathbf{P}(catch \mid toothache, cavity) = \mathbf{P}(catch \mid cavity)$

- The same independence holds if I haven't got a cavity:

  (2) $\mathbf{P}(catch \mid toothache, \neg cavity) = \mathbf{P}(catch \mid \neg cavity)$

- Equivalent statements:
  $\mathbf{P}(Toothache \mid Catch, Cavity) = \mathbf{P}(Toothache \mid Cavity)$
  $\mathbf{P}(Toothache, Catch \mid Cavity) = \mathbf{P}(Toothache \mid Cavity)\, \mathbf{P}(Catch \mid Cavity)$

Conditional independence

For boolean random variables,

- A is conditionally independent of B given C iff:

  P(A|B,C) = P(A|C)

  P(A|~B,C) = P(A|C)

For multivalued random variables,

- A is conditionally independent of B given C iff:

$$\forall u,v,w : P(A=u \mid B=v \wedge C=w) = P(A=u \mid C=w)$$

Inference with Conditional Probabilities

- S = stiff neck, M = meningitis
- $P(S|M) = 0.8$, $P(S) = 0.2$, $P(M) = 0.0001$

- Suppose you wake up with a stiff neck - since 80% of the time, meningitis is associated with a stiff neck, you probably have meningitis and should rush to the hospital!!

- Is this correct reasoning?

Inference with Conditional Probabilities

- S = stiff neck, M = meningitis
- $P(S|M) = 0.8$, $P(S) = 0.2$, $P(M) = 0.0001$

- $$\begin{aligned} P(M|S) &= P(M \,\&\, S) / P(S) \\ &= P(S|M)P(M) / P(S) \\ &= (0.00008) / 0.2 \\ &= 0.0004 \end{aligned}$$
- The risk is higher, but still **very** small!

Bayes' Theorem

Bayes' rule:

$$P(A \mid B) = P(B \mid A)\, P(A) \,/\, P(B)$$

- In distribution form

$$\mathbf{P}(Y|X) = \mathbf{P}(X|Y)\, \mathbf{P}(Y) \,/\, \mathbf{P}(X) = \alpha \mathbf{P}(X|Y)\, \mathbf{P}(Y)$$

- Useful for assessing diagnostic probability from causal probability:

$$P(\text{Cause}|\text{Effect}) = P(\text{Effect}|\text{Cause})\, P(\text{Cause}) \,/\, P(\text{Effect})$$

Bayes, Thomas (1783) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418.**

Bayes' Rule and Gambling

- Suppose there are two sealed envelopes, one ("Win") with $1, 2 red beads, and 2 black beads; the other with no money, 1 red bead, and 2 black beads.

- I draw an envelope at random and offer to sell it to you. How much should you be willing to pay?

Bayes' Rule and Gambling

- I draw an envelope at random and offer to sell it to you. How much should you be willing to pay?

- Now, you are allowed to see one (randomly drawn) bead from the selected envelope:
  - If it is black, how much should you be willing to pay?
  - If it is red, how much should you be willing to pay?

Bayes' Rule and Gambling

- If the bead is black…

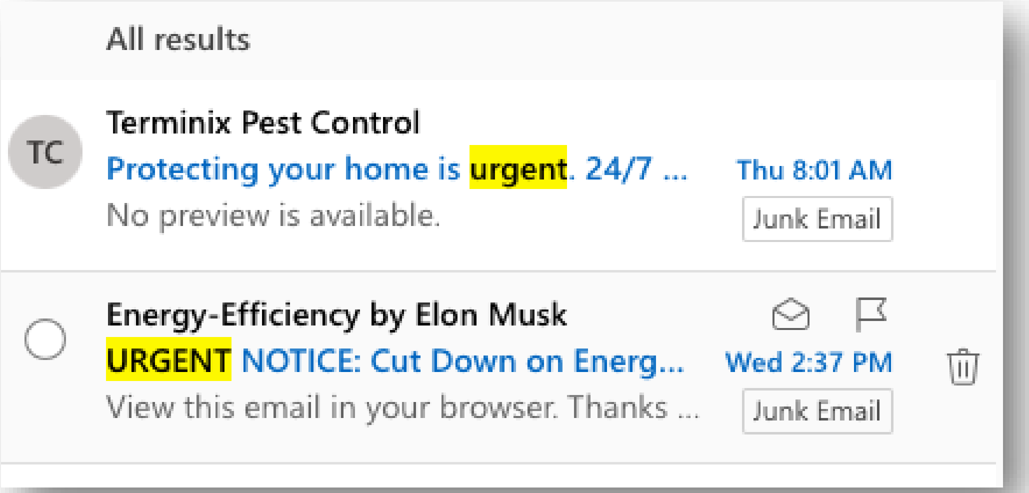
- P(Black) = (1/2*1/2 + 2/3*1/2) = 7/12

- **P(Win|Black) = P(Black|Win)P(Win) / P(Black)**

  = (1/2 * 1/2) / (7/12)
  = 3/7

Bayes' Rule and Text

U: "Urgent" in text
S: Email is Spam

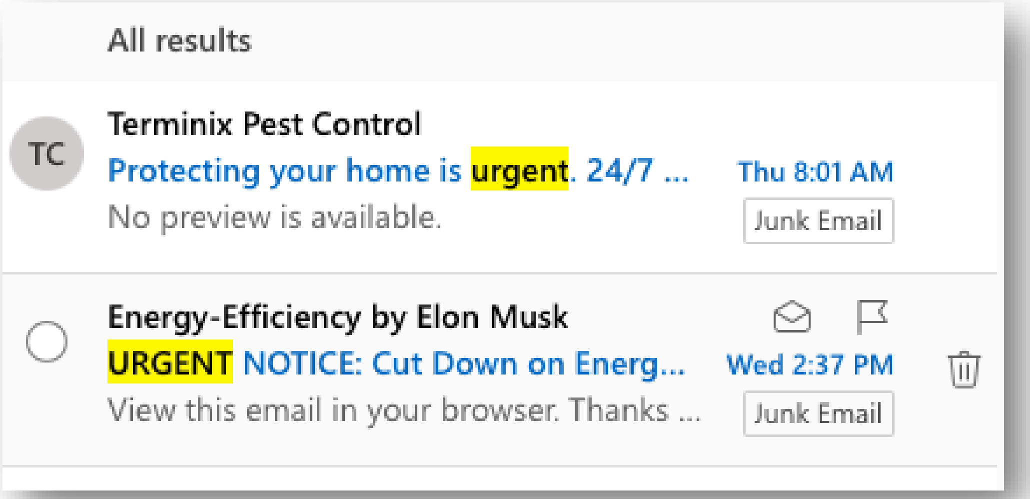


$P(S) = 60\%$
$P(U \mid S) = 10\%$
$P(U \mid \sim S) = 1\%$

What should we do if we receive an "Urgent" email?

Bayes' Rule and Text

U: "Urgent" in text
S: Email is Spam



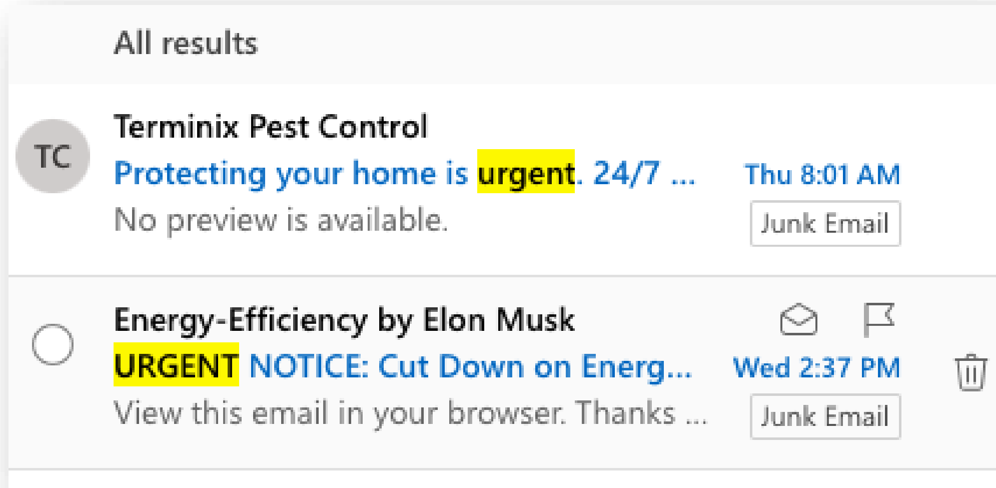
P(S) = 60%
P(U| S) = 10%
P(U|~S) = 1%

What should we do if we receive an "Urgent" email?

P(U) = P(U|S)P(S)+P(U|~S)P(~S) = 0.06+0.004 = 0.064
P(S|U) = P(U|S)P(S) / P(U)
= 0.06 / 0.064 = **0.9375**

Bayes' Rule and Text

U: "Urgent" in text
S: Email is Spam



P(S) = 60% ← **Prior Prob**
P(U| S) = 10%
P(U|~S) = 1%

What should we do if we receive an "Urgent" email?

P(U) = P(U|S)P(S)+P(U|~S)P(~S) = 0.06+0.004 = 0.064
P(S|U) = P(U|S)P(S) / P(U)
= 0.06 / 0.064 = **0.9375** ← **Posterior Prob**

LINEAR ALGEBRA REVIEW

Scalars, Vectors, Matrices and Tensors

- Scalars are the numbers we know and love.

$$a = 1$$
$$b = e$$
$$c = -0.3$$

- Vectors are arrays of numbers – elements of $\mathbb{R}^n$
- They are typically written in a column (column vector)

$$\vec{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Scalars, Vectors, Matrices and Tensors

- **Matrices** are sets of numbers organized into rows and columns (2-dimensional)
- Each row has the same dimension, and each column has the same dimension
- An MxN matrix has M rows and N columns
- A vector is an Nx1 matrix
- **Tensors** are like matrices, but in higher dimensions

$$\begin{array}{ccc} & \text{rows} & \\ A = & \begin{bmatrix} a & b \\ c & d \end{bmatrix} & \text{columns} \end{array}$$

Vectors: Dot Product

$$a \cdot b = a^T b = [a_1 \quad a_2 \quad a_3] \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

Think of the dot product as a matrix ***multiplication***

$$\|a\|^2 = a^T a = a_1^2 + a_2^2 + a_3^2$$

The ***magnitude*** is the square root of the dot product of a vector with itself

$$a \cdot b = \|a\| \|b\| \cos(\theta)$$

The dot product is also related to the angle between the two vectors

Norms

A norm is a way of measuring the magnitude of a vector

Specifically, a norm must satisfy

$$f(x) = 0 \Rightarrow x = 0$$

$$f(x + y) \leq f(x) + f(y)$$

$$f(\alpha x) = |\alpha| f(x)$$

L_1 Norm: $\|x\|_1 = \sum_i |x_i|$

L_2 Norm: $\|x\|_2 = \sqrt{\sum_i (x_i)^2}$

L_0 Norm: $\|x\|_0 = \sum_i 1 - \delta_{(x_i)(0)}$

L_∞ Norm: $\|x\|_\infty = \max_i |x_i|$

Matrix Operations

- Addition, Subtraction, Multiplication

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$$

Just add elements

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} - \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a-e & b-f \\ c-g & d-h \end{bmatrix}$$

Just subtract elements

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$$

Multiply each row by each column

Multiplication

- Is AB = BA? Maybe, but maybe not!

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & \ldots \\ \ldots & \ldots \end{bmatrix} \quad \begin{bmatrix} e & f \\ g & h \end{bmatrix}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} ea+fc & \ldots \\ \ldots & \ldots \end{bmatrix}$$

- Heads up: multiplication is NOT ***commutative***!

Transpose of a Matrix

- Swap rows and columns
- The transpose of a column vector is a row vector, and vice-versa

$$A = \begin{bmatrix} a & b & c \\ d & e & f \end{bmatrix}$$

$$A^T = \begin{bmatrix} a & d \\ b & e \\ c & f \end{bmatrix}$$

$$\vec{v} = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$$\vec{v}^T = \begin{bmatrix} 1 & 0 & -1 \end{bmatrix}$$

Inverse of a Matrix

- Identity matrix:
  AI = A
- Some matrices have an inverse, such that:
  $AA^{-1} = I$
- Inversion is tricky:
  $(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$

  Derived from non-commutativity property

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Inverse of a Matrix

$$\begin{bmatrix} a & b & c & 1 & 0 & 0 \\ d & e & f + 0 & 1 & 0 \\ g & h & i & 0 & 0 & 1 \end{bmatrix}$$

1. Append the identity matrix to A
2. Subtract multiples of the other rows from the first row to reduce the diagonal element to 1
3. Transform the identity matrix as you go
4. When the original matrix is the identity, the identity has become the inverse!

Orthogonality

- (Non-zero) vectors are **orthogonal** if their dot product is zero (geometrically, perpendicular)

$$\text{Orthogonal}(\vec{x}, \vec{y}) \stackrel{\text{def}}{=} \vec{x}^T \vec{y} = 0$$

- **Orthonormal**: orthogonal with unit norm
- An **orthogonal matrix** is one with mutually *orthonormal* rows and columns
- For an orthogonal matrix A:

$$A^{-1} = A^T$$

Other concepts

- Determinant (of a matrix)
- Trace (of a matrix)
- Eigendecomposition (of a matrix)
- Pseudoinverse (of a matrix)

The matrix Cookbook:
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf