ILLINOIS INSTITUTE OF TECHNOLOGY

Mathematics Review

CS-585

Natural Language Processing

Derrick Higgins

Slides based in part on material from:

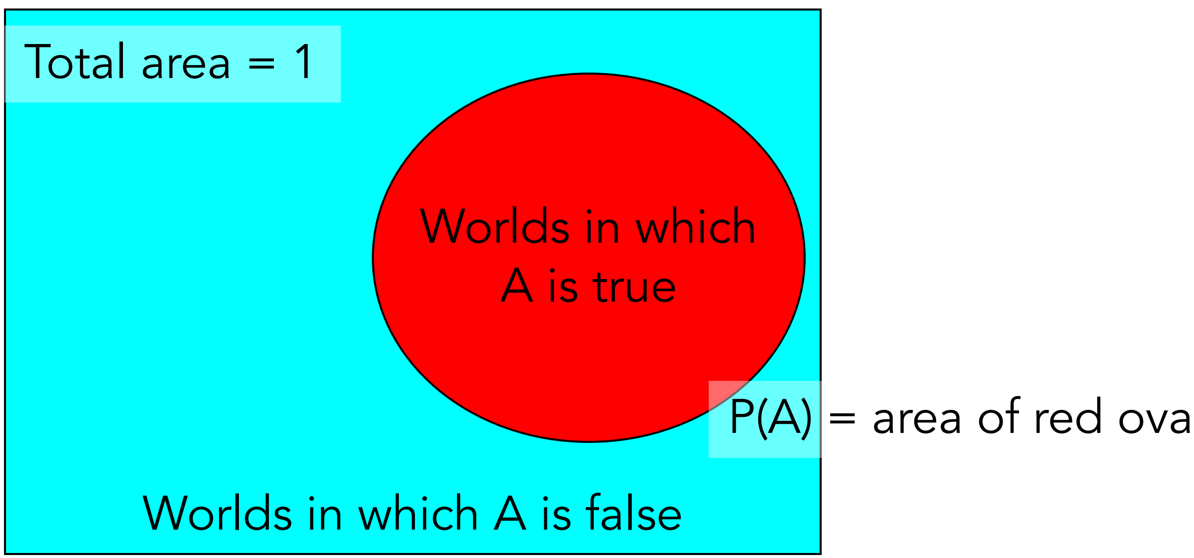
- *Artificial Intelligence: A Modern Approach, 2nd Edition*
  Russell & Norvig (Prentice-Hall: 2003)
- Slides by Patrick Nichols (MIT)

PROBABILITY THEORY REVIEW

Probability: Intuitive

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

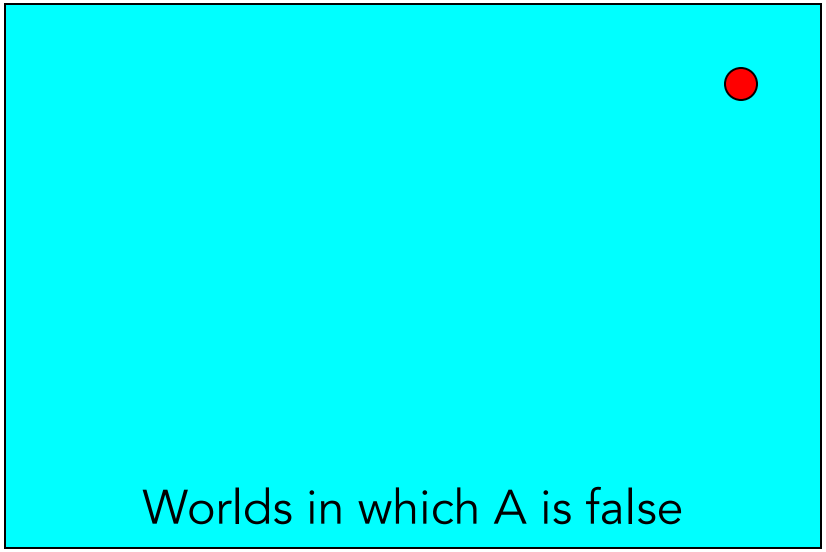
Event Space of all possible worlds



Probability: Axioms

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

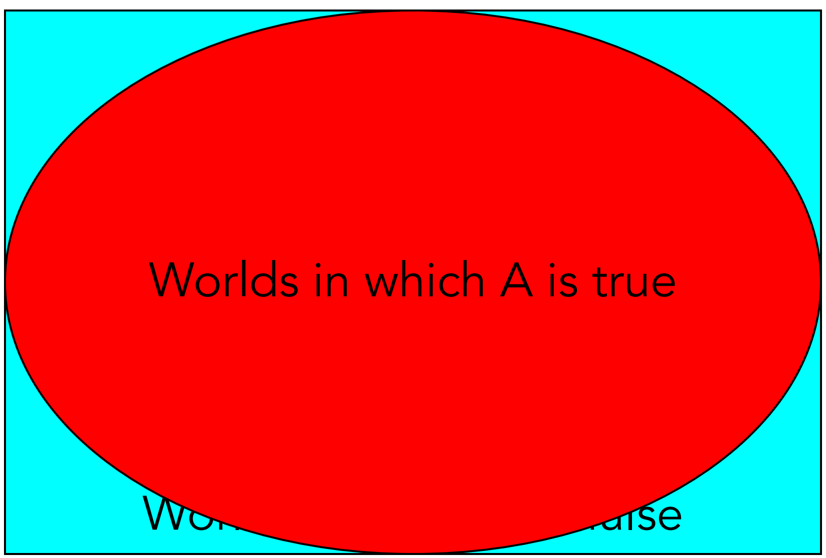
Probability: Axioms

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

Red oval can't get smaller than 0

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

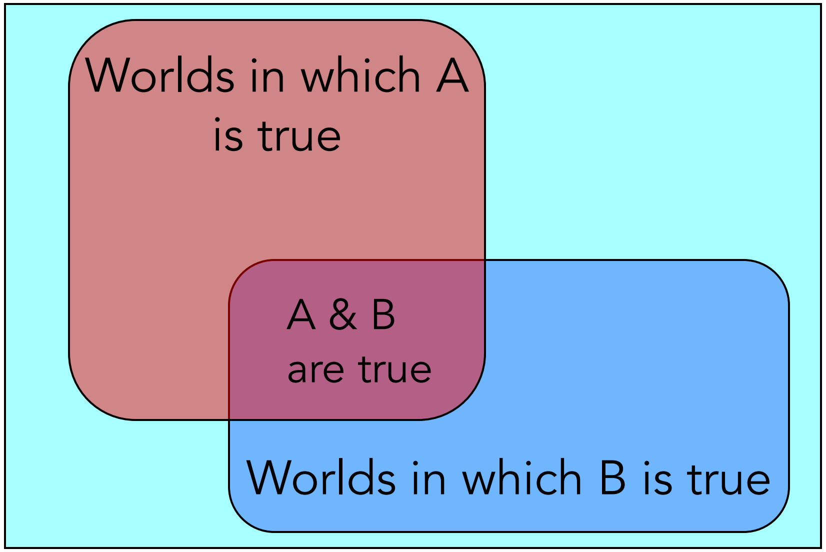
Probability: Axioms

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

Red oval can't get larger than 1

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

Probability: Axioms

- $0 \leq P(A) \leq 1$
- $P(\mathit{true}) = 1$
- $P(\mathit{false}) = 0$
- 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

Axioms:

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

We can show that:

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

And furthermore:

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

Multivalued Random Variables

Suppose *A* can take on more than 2 values

- e.g., *POS* is one of {noun,verb,adjective,adverb}

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

Thus:

- $P(A=vi \; \& \; A=vj) = 0 \quad \text{if } i \neq j$
- $P(A=v1 \text{ or } A=v2 \text{ or } \ldots \text{ or } A=vk) = 1$

Easy Facts About Multivalued RVs

Axioms:

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

Recall:

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

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

- And therefore:

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

More Facts About Multivalued RVs

Axioms:

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

Recall:

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

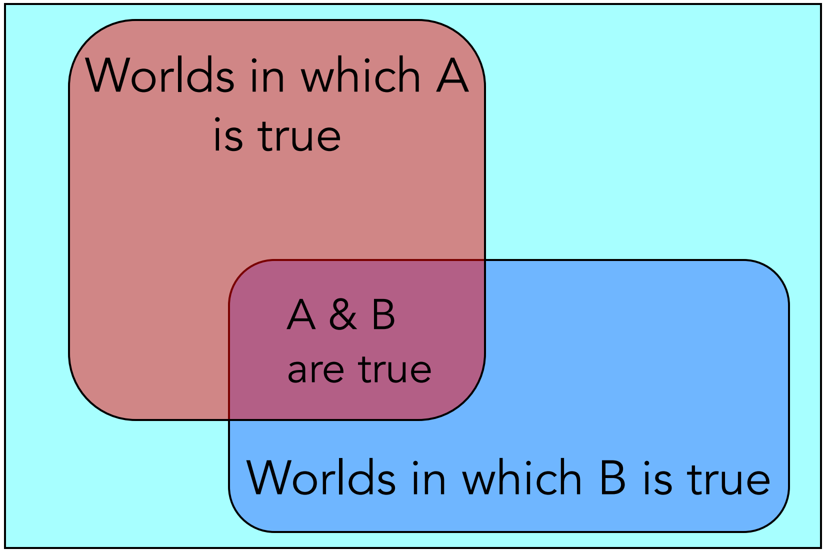
$$P(B \wedge [X = v1 \vee \cdots \vee X = vi]) = \sum_{j=1}^{i} P(B \wedge X = vj)$$

- And therefore:

$$P(B) = \sum_{j=1}^{k} P(B \wedge X = vj)$$

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:

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

- Thus, the *Chain Rule:*

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

$$P(A1 \wedge A2 \wedge \cdots \wedge An) = P(A1 \mid A2 \cdots An)P(A2 \mid A3 \cdots An) \cdots P(An)$$

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:

$$P(\boldsymbol{\varphi}) = \boldsymbol{\Sigma}_{\omega:\omega \models \varphi} P(\boldsymbol{\omega})$$

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: $P(\boldsymbol{\varphi}) = \boldsymbol{\Sigma}_{\omega:\omega \models \varphi} P(\boldsymbol{\omega})$

- 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: $P(\boldsymbol{\varphi}) = \boldsymbol{\Sigma}_{\omega:\omega \models \varphi} P(\boldsymbol{\omega})$

- 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:

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

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

$$= 0.4$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	$\boxed{.108}$	$\boxed{.012}$	.072	.008
¬ cavity	$\boxed{.016}$	$\boxed{.064}$	.144	.576

- Denominator can be viewed as a normalization constant **a**

P(*Cavity* | *toothache*) = **a** P(*Cavity*,*toothache*)

= **a** [P(*Cavity*,*toothache*,*catch*) + P(*Cavity*,*toothache*,~*catch*)]

= **a** [$<0.108,0.016>$ + $<0.012,0.064>$]

= **a** $<0.12,0.08>$ = $<0.6,0.4>$

General idea: compute distribution on query variable (*Cavity*) by fixing evidence variables (*Toothache*) and summing over hidden variables (*Catch*)

Inference by enumeration

Typically, we want **P(Y | E = e)**:

posterior joint distribution of the query variables **Y**
given specific values **e** for the evidence variables **E**

Let the hidden variables be **H** = **X** \ (**Y** ∪ **E**)

Then we can just sum over the hidden variables and normalize:

$$\mathbf{P}(\mathbf{Y} \mid \mathbf{E} = \mathbf{e}) = \boldsymbol{\alpha}\mathbf{P}(\mathbf{Y},\mathbf{E} = \mathbf{e}) = \boldsymbol{\alpha} \left[\sum_{\mathbf{h}} \mathbf{P}(\mathbf{Y},\mathbf{E}= \mathbf{e}, \mathbf{H} = \mathbf{h}) \right]$$

Terms are atomic events, because **Y** U **E** U **H** = **X**

Inference by enumeration

- Obvious problems:

1. Worst-case time complexity $O(d^n)$ where d is the largest arity
2. Space complexity $O(d^n)$ to store the joint distribution
3. How to find the numbers for $O(d^n)$ entries in the joint distribution?

Independence

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

$$P(A|B) = P(A)$$

- Equivalently

$$P(B|A) = P(B)$$

- That is, the probability we give A (or B) is not affected by finding out that B (or A)

Independence Facts

If A and B are independent boolean RVs:

- **P(A & B)** = P(A | B) P(B) = **P(A) P(B)**
- **P(~A|B)** = 1-P(A|B) = 1-P(A) = **P(~A)**
- **P(A|~B)** = P(A & ~B) / P(~B)

  = P(~B | A)P(A) / P(~B)

  = P(~B)P(A) / P(~B)

  = **P(A)**

Multivalued Independence

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

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

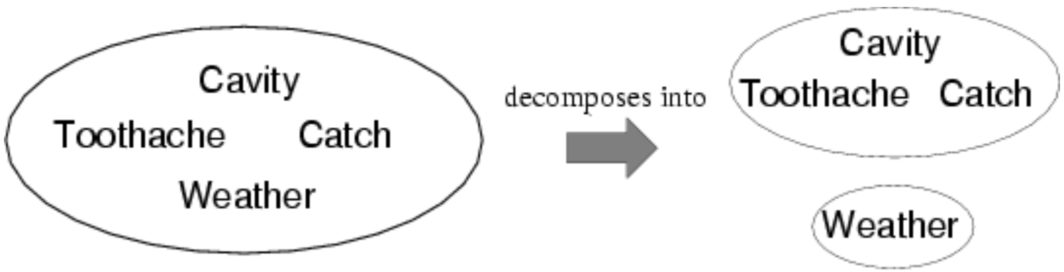
- From which we can show, for example:

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

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

Independence

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



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

  - 16 entries reduced to 10
  - 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

- $P(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries

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

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

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

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

- *Catch* is conditionally independent of *Toothache* given *Cavity*:

  $P(Catch \mid Toothache, Cavity) = P(Catch \mid Cavity)$

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

Conditional independence

- Get the full joint distribution using chain rule:
-

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

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

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

= **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*) **P**(Cavity)

Based on only 2 + 2 + 1 = 5 independent parameters

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*.

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.

Conditional independence

For boolean random variables,

- A is conditionally independent of B given C iff:

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

$$P(A|\sim B,C) = P(A|C)$$

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** slim!

Bayes' Theorem

Bayes' rule:

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

- In distribution form

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

- Useful for assessing diagnostic probability from causal probability:

$$P(Cause|Effect) = P(Effect|Cause)\, P(Cause) \,/\, P(Effect)$$

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

  = (1/2 * 1/2) / (1/2*1/2 +
  2/3*1/2)

  = (1/4) / (1/4 + 1/3)

  = (1/4) / (7/12)

  = 3/7

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

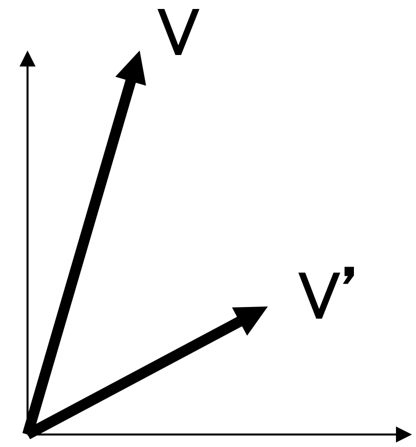
- **Matrices** are sets of numbers organized into rows and columns (2-dimensional)
- Each row has the same dimension, and each column has the same dimension
- An MxN matrix has M rows and N columns
- A vector is an Nx1 matrix
- **Tensors** are like matrices, but in higher dimensions

$$A = \begin{array}{cc} \xrightarrow{\text{rows}} & \\ \begin{bmatrix} a & b \\ c & d \end{bmatrix} & \Big\downarrow \text{columns} \end{array}$$

Vector Interpretation

- Think of a vector as a line in 2D or 3D
- Think of a matrix as a transformation on a line or set of lines

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix}$$



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

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

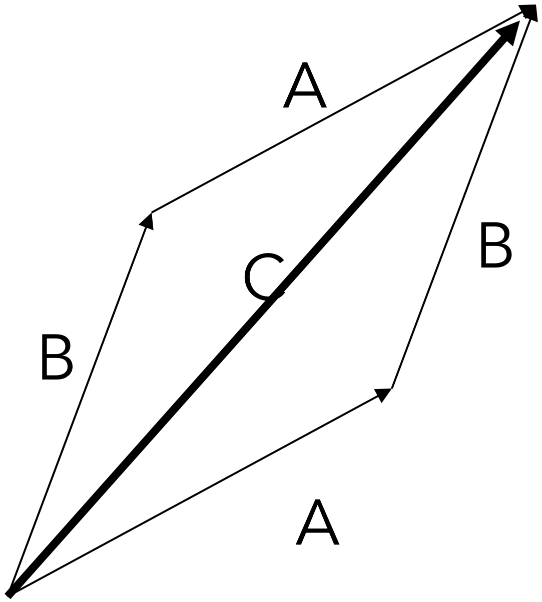
The magnitude is the square root of the dot product of a vector with itself

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

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

Vectors: Dot Product

- Interpretation: the dot product measures to what degree two vectors are aligned



A+B = C
(use the head-to-tail method to combine 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!

Affine Transformations

- For machine learning, we often want to compute a linear function of input features:

$$y = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n + b$$

- The input features can be collected into a vector *x*, and the linear coefficients into another vector *w*. Then the linear transformation can be represented as a matrix multiplication plus a constant intercept term:

$$y = \vec{w}^T \vec{x} + b$$

- This type of linear operation is called an affine transformation

Affine Transformations

- We can subsume the intercept term by concatenating a [1] to our feature vector *x*, and [*b*] to the weight vector *w*:

$$x' = [x_0, x_1, \cdots, x_n, 1]^T$$
$$w' = [w_0, w_1, \cdots, w_n, b]^T$$
$$y = \vec{w}'^T \vec{x}'$$

- So affine transformations can be represented as matrix multiplication.

$$\vec{y} = WX$$

- Note that due to the associativity of matrix multiplication, successive affine transformations can always be represented as a single affine (linear) transformation

$$\vec{y} = W_0 W_1 \cdots W_n X$$
$$W \stackrel{\text{def}}{=} W_0 W_1 \cdots W_n$$
$$\vec{y} = WX$$

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)