Linear Algebra and Probability Overview

Intro to ML
Fall 2022
Alona Fyshe

Many of these slides are derived from Alex Thomo. Thanks!

More DALL-E 2 examples

- Plus Midjourney and StableDiffusion
- https://twitter.com/fabianstelzer/status/1561019187451011074

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

film still, portrait of an old man, wrinkles, dignified look, grey silver hair, peculiar nose, wise, eternal wisdom and beauty, incredible lighting and camera work, depth of field, bokeh, screenshot from a hollywood movie

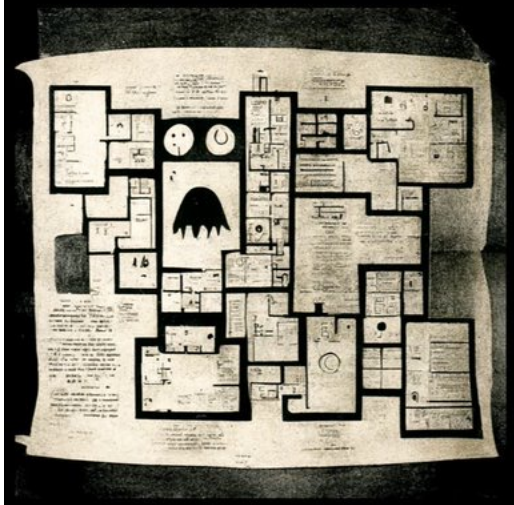
MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

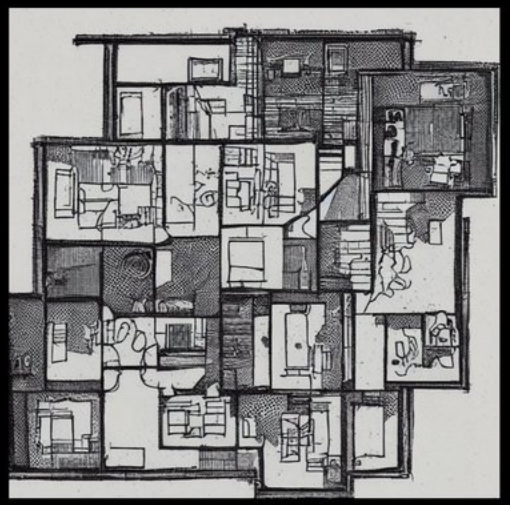
portrait of a man who looks exactly like super mario,
photography, portrait photograph

MIDJOURNEY



DALL-E 2

STABLEDIFFUSION



a spooky 1970s floor plan of a haunted house, worn paper, scary atmosphere, pain and regret

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

a human anatomical heart made of flowers, pastel, matte, masterpiece

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

beatles lego set, catalogue photograph

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

Pixar movie scene of a dark skull wizard fighting against Kermit the frog as a gladiator, incredible render, Presto

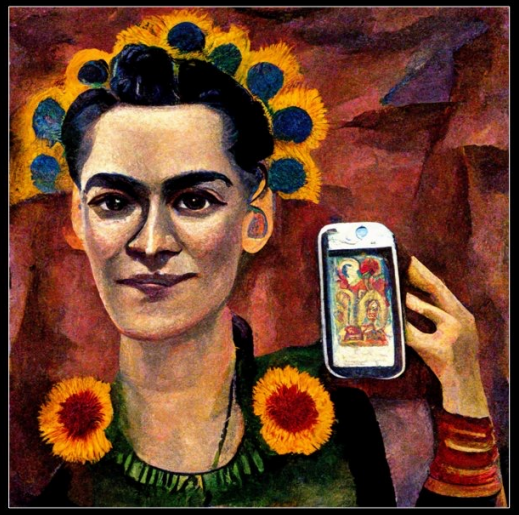
MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

beautiful landscape scene, studio ghibli, a wonderful alpine town with an ocean view, churces, medieval architecture

MIDJOURNEY



DALL-E 2

STABLEDIFFUSION

a woman holding an iphone in the air, talking a selfie, painting by Frida Kahlo

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

the word PROMPT, magnificent typography

MIDJOURNEY

DALL-E 2

STABLEDIFFUSION

Behind the scenes of shooting the moon landing, Hollywood studio, 1969, backstage photograph, astronaut actors, lighting

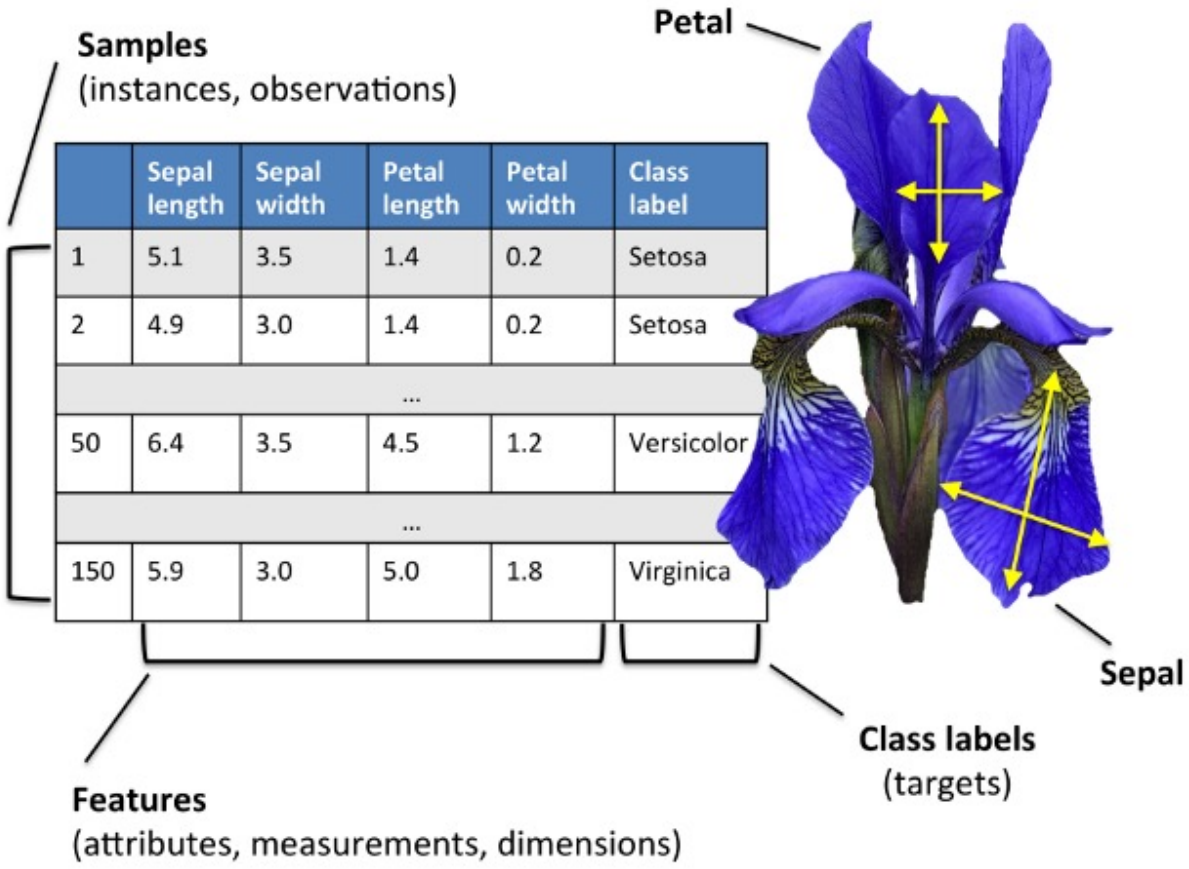
Now, onto the review

Resources website

- Linked to on eclass (first section)
- https://dem1995.github.io/machine-learning/

Before we begin

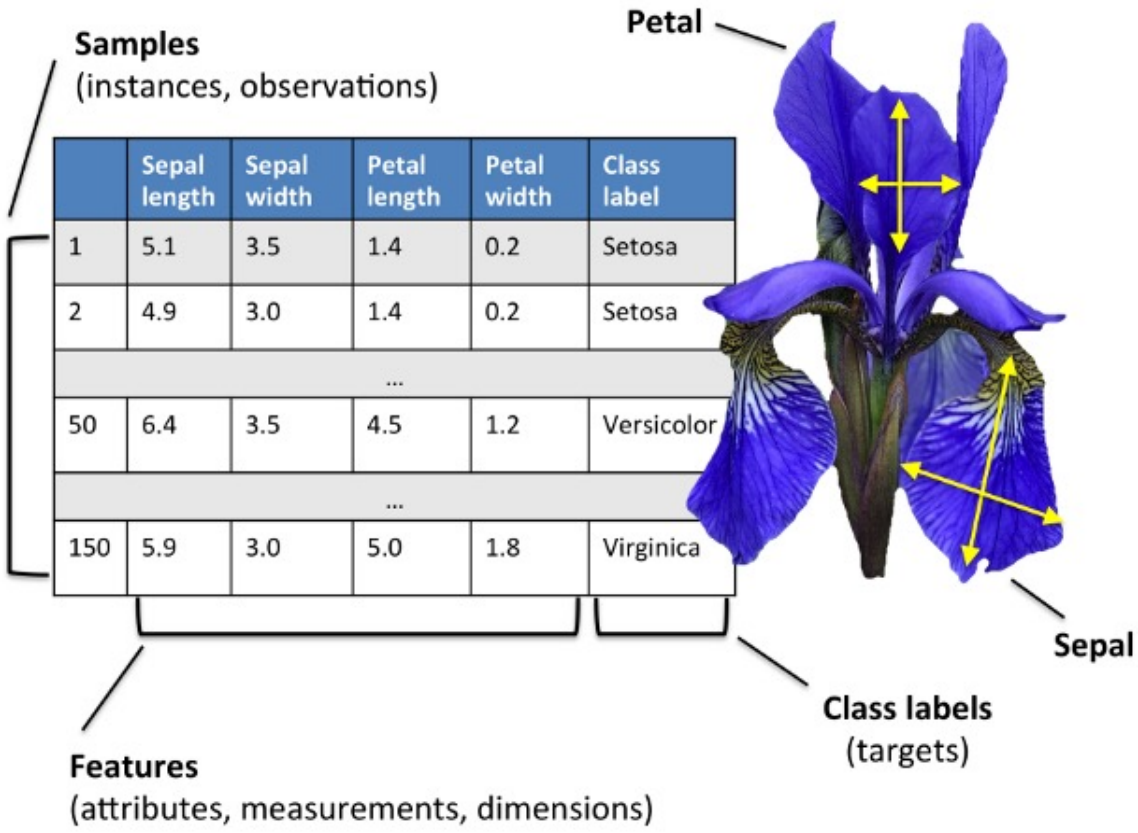
- To make this lecture a little more motivating, first we will talk about how we represent data in machine learning



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

Iris Dataset

- Four features, plus the class label
- Our task is to predict class label (flower type) from the 4 features
- To graph these feature vectors, we would need a 4D space
  - Difficult to visualize



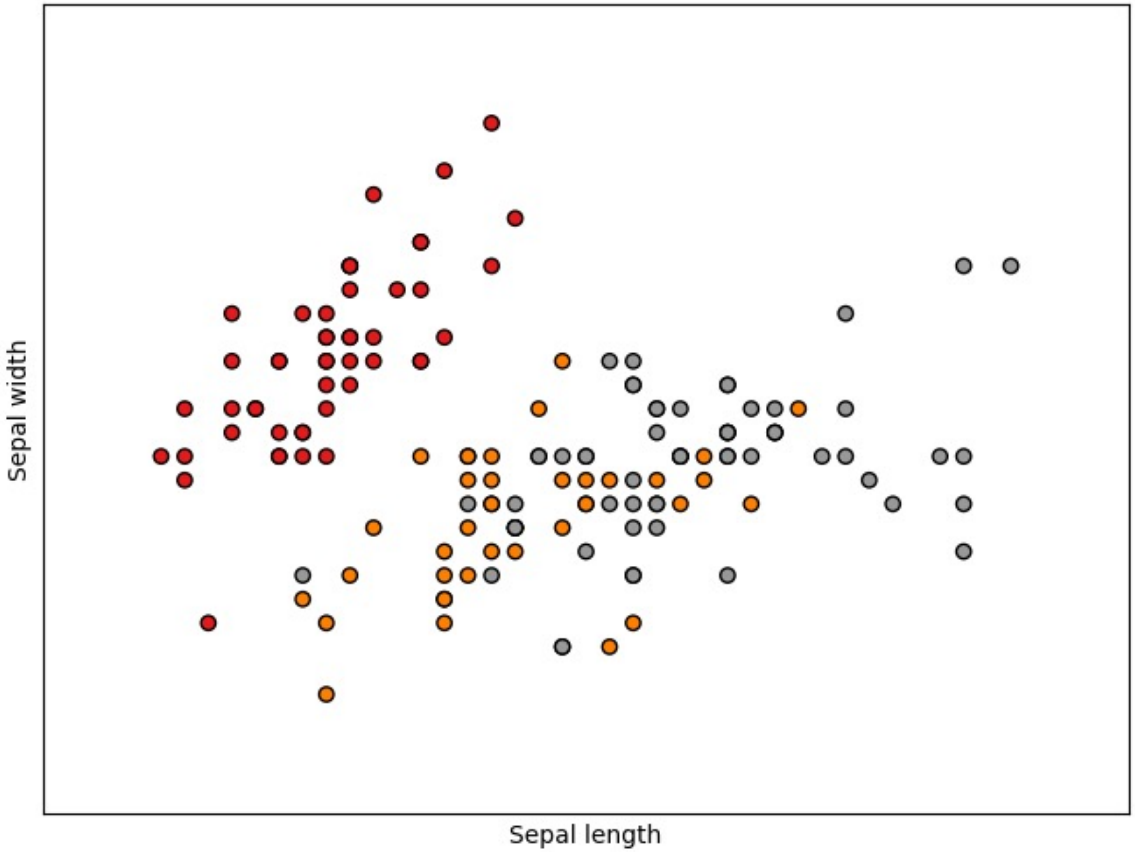
	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

https://medium.com/analytics-vidhya/exploratory-data-analysis-uni-variate-analysis-of-iris-data-set-690c87a5cd40

Dimensions as Features

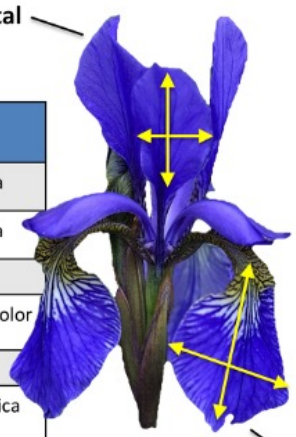
- We can use the dimensions of a vector to represent the values for different features in our data
  - E.g. the very famous Iris dataset

- In the figure →
  - X: sepal length
  - Y: sepal width
  - Color of dot: flower type



Samples
(instances, observations)

	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

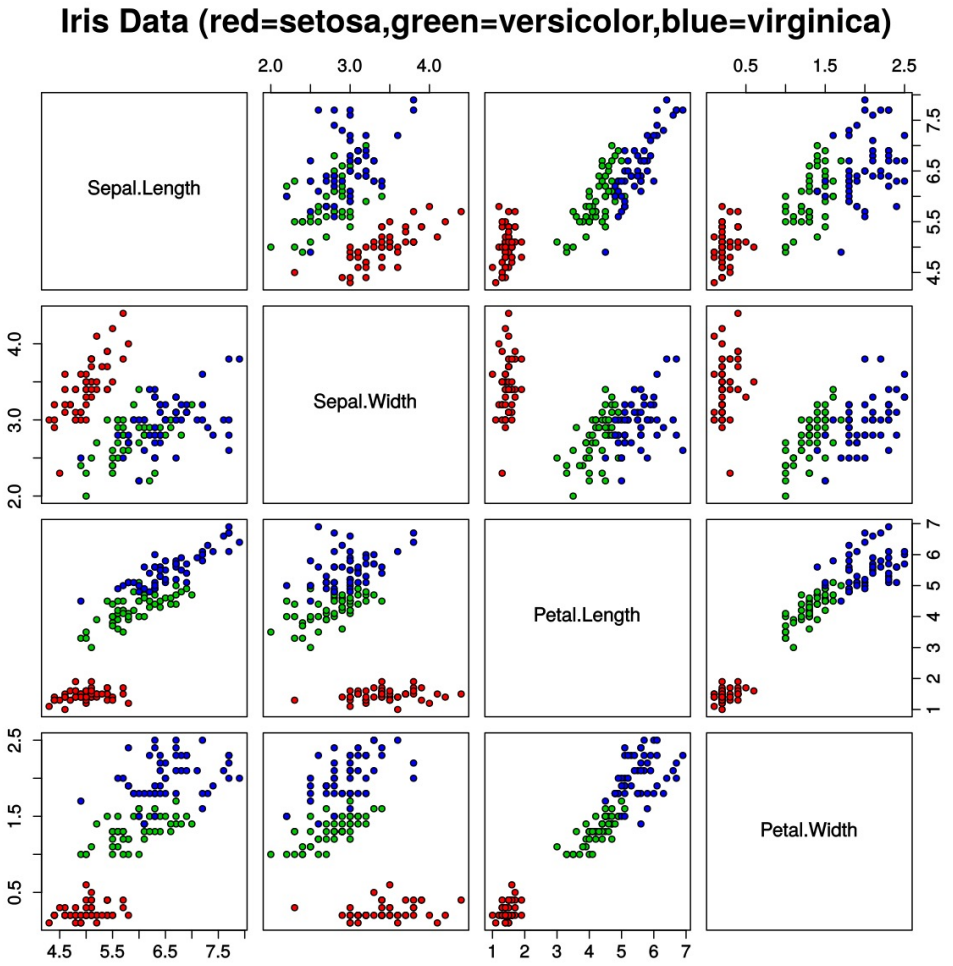


Petal

Sepal

Class labels
(targets)

Features
(attributes, measurements, dimensions)

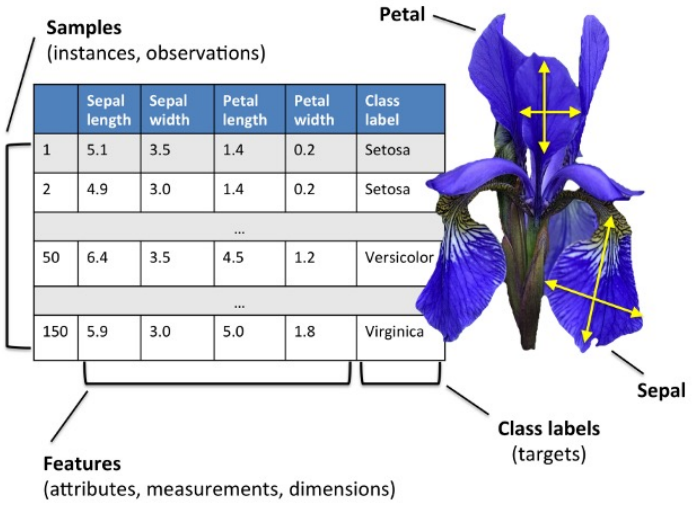


Source: https://en.wikipedia.org/wiki/Iris_flower_data_set

Iris Dataset

- Each of the 4 features are *continuous*
- The Class label is *discrete*

- How to represent class label?
- Unique integer values
  - (e.g. 1=Setosa, 2=Versicolor, 3=Virginica)
- One hot vector
  - [1, 0, 0] -> Setosa
  - [0, 1, 0] -> Versicolor
  - [0, 0, 1] -> Virginica



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
			...		
50	6.4	3.5	4.5	1.2	Versicolor
			...		
150	5.9	3.0	5.0	1.8	Virginica

https://medium.com/analytics-vidhya/exploratory-data-analysis-uni-variate-analysis-of-iris-data-set-690c87a5cd40

Discrete features

- Class label is an example of a discrete feature
  - As opposed to continuous features like length and width
- Features can also be discrete
  - E.g. number of petals
  - Hometown
  - Favorite movie
- Sometimes these features are ordinal (they have an ordering)
  - Number of petals
  - Maybe hometown (if ordered east to west could have some meaning)
  - Not favorite movie

Discrete features for ML

- When features are ordinal, it can make sense to represent them with integer numbers
- When features are categorical (i.e. non-ordinal) one hot vectors work better

- Why?

Some Notation

- A vector is a list of numbers
  - The number of dimensions is the length of the list
- A matrix is a table of numbers, so it has a length and a height
  - E.g. 5x2, 10x100
  - Convention is **R**ows x **C**olumns (I remember this as **R**oman **C**atholic)
- By this same logic, a vector is actually a matrix with length or height of 1
  - 6x1 is a column vector with 6 elements
  - 1x3 is a row vector with 3 elements

Some Notation

- Square brackets to denote boundaries of vectors/matrices
- Convention is for variable names that denote vectors to be $\boldsymbol{a}$ $\vec{a}$
  - Lowercase
  - Bold or have an arrow over them (not always adhered to if the context makes the form of the variable clear)
- Matrices
  - Uppercase
  - Plain font A

Some Notation

- Communicate the size of a matrix like this: $A \in \mathbb{R}^{n \times p}$
- The “R” is a symbol for real numbers (i.e. numbers that don’t need to be integers)

- Communicate the size of a vector like this: $\boldsymbol{a} \in \mathbb{R}^{p}$

- Transpose (T) means to swap rows for columns

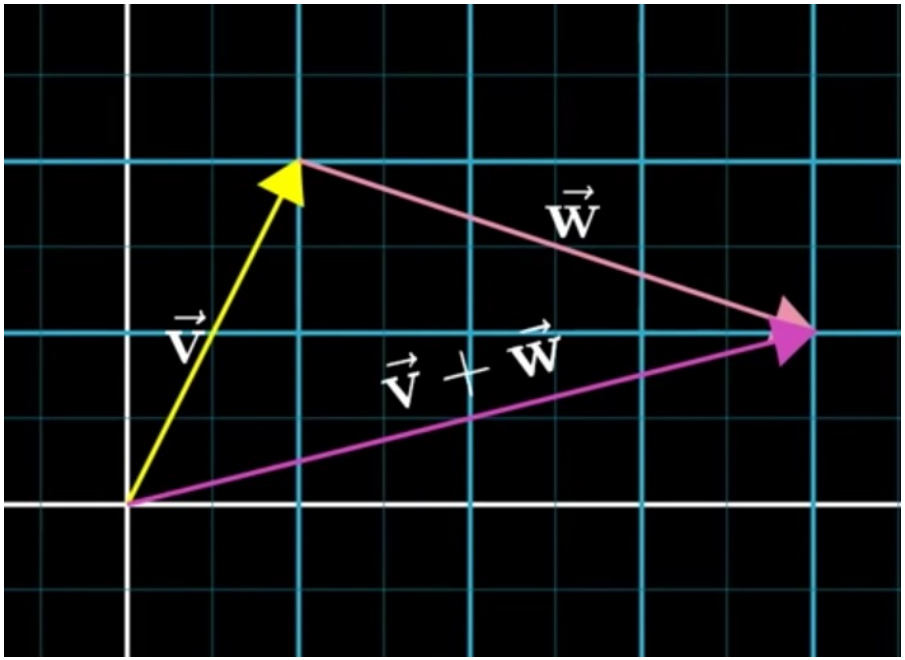
$$A \in \mathbb{R}^{n \times p}$$

$$A^{T} \in \mathbb{R}^{p \times n}$$

Example: text documents

- Representing text as a feature vector
- Example (nonsensical) text:
  - D1: brown cat brown cat dog cat mouse
  - D2: brown cat mouse mouse mouse
  - D3: dog brown brown cat meow
- Identify vocabulary (all words across all documents)
  - brown, cat, dog, mouse, meow (this is the feature order below)
- Features are the # of occurrences of each vocabulary word in doc.
  - D1: [2, 3, 1, 1, 0]
  - D2: [1, 1, 0, 3, 0]
  - D3: [2, 1, 1, 0, 1]

Vector Addition




```
v: [1,  2]
w: [3, -1]

v+w: [4,  1]
```

Pic from youtube playlist video 1

What does vector addition mean for our text dataset?

- Recall:
  - D1: [2, 3, 1, 1, 0]
  - D2: [1, 1, 0, 3, 0]

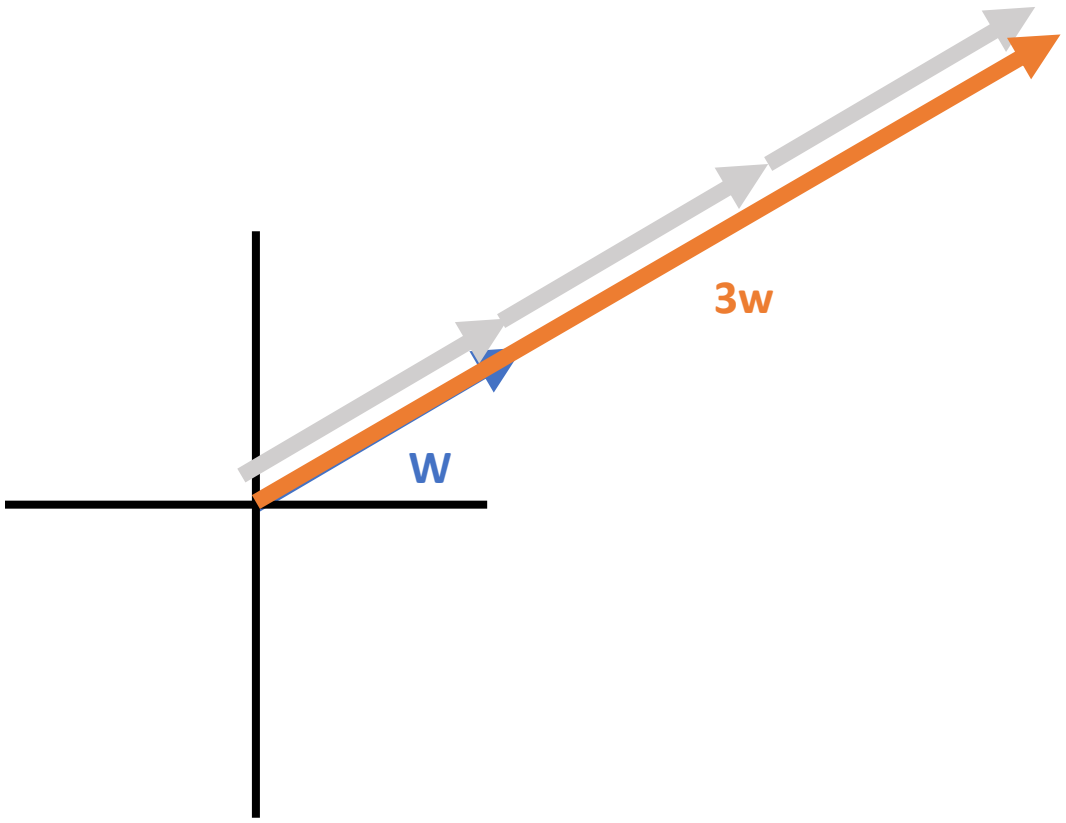
- What does it mean to have a new document A = D1 + D2?

- I.e. what document would give us a vector equivalent to A = D1 + D2?

Scalar multiplication for vectors

- Example

Scalar multiplication for vectors



What does scalar multiplication mean for our text dataset?

- Recall:
  - D1: [2, 3, 1, 1, 0]

- What does it mean to have document A = 2 * D1?

Inner product (dot product)

- Definition $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_i a_i b_i$
- E.g. 3-D vectors $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{3} a_i b_i$

$$= a_1 b_1 + a_2 b_2 + a_3 b_3$$

Inner product (dot product)

- This ends up being quite important in ML
- Corresponds to the weighted sum
- Many models make predictions using a weighted sum of the feature vector
- Another example: price vector multiplied by quantity vector

Inner product (dot product)

- Neat tricks with the inner product
- One hot vector times feature vector “selects” a particular element from the vector
- Example: a=[0, 1, 0], b=[7, 5, 8]

Inner product (dot product)

- Vectors can be squared
- E.g. b=[7, 5, 8], b^2 = ?

Length of a vector (Euclidean Norm)

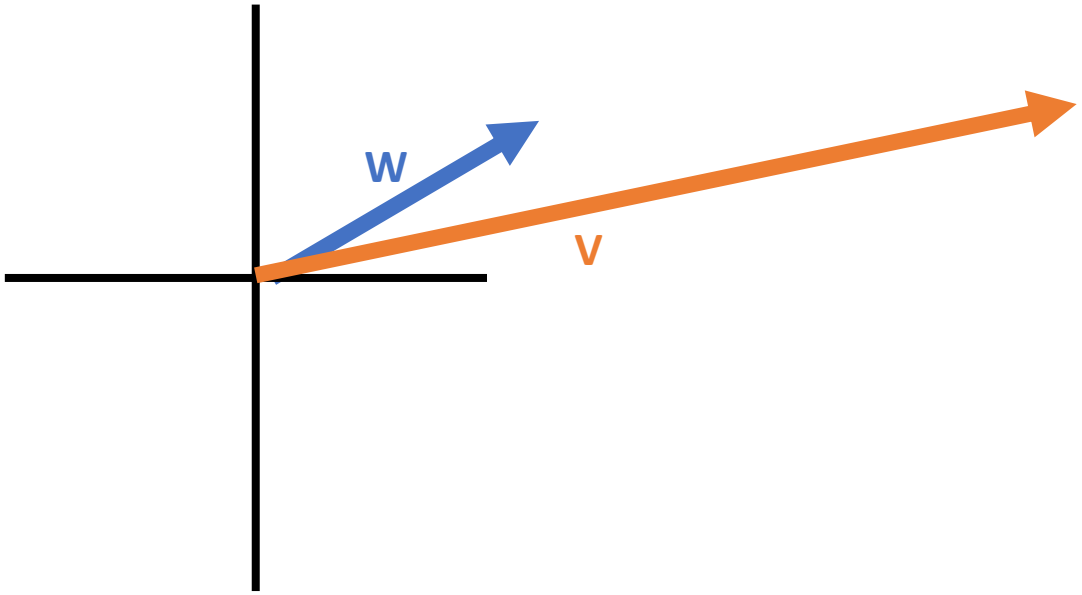
- Notation $\|a\|$
- Definition $\|a\| = \sqrt{a \cdot a}$

$$= \sqrt{\sum_i a_i^2}$$

- You may have seen this in the Pythagorean theorem (length of the hypotenuse)

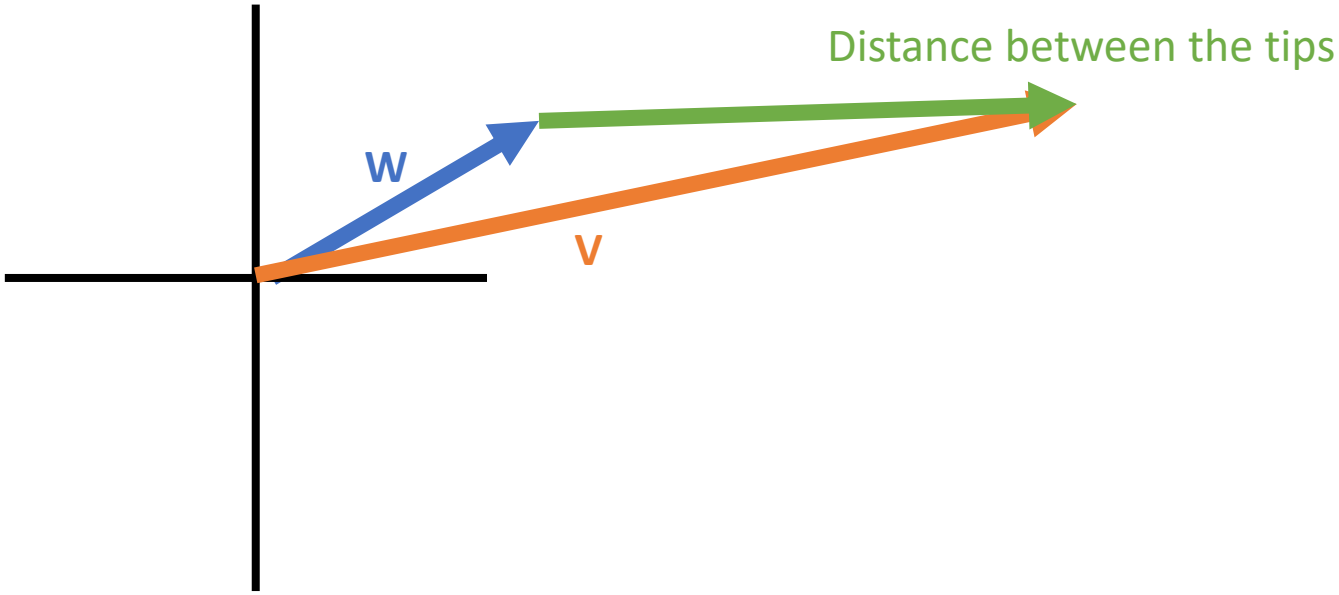
Vector Similarity

- It's often useful to compute the similarity between two vectors



Vector Similarity

- Definition one: Euclidean distance



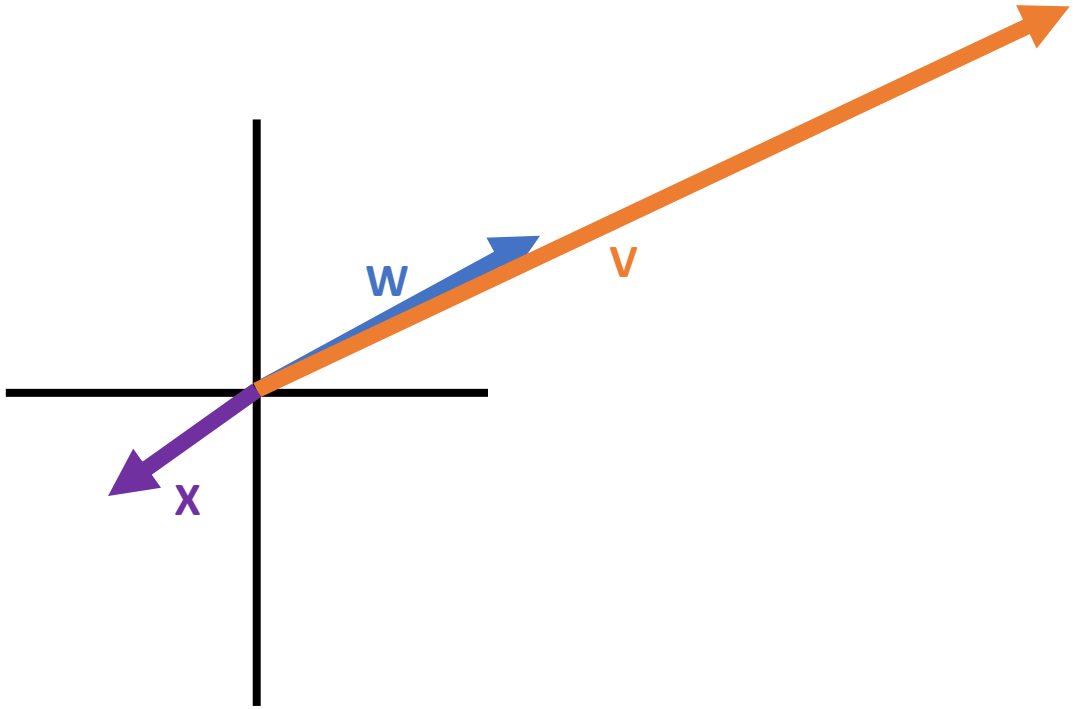
Vector similarity

- Euclidean Distance

$$\sqrt{(w-v)^2}$$

Vector similarity

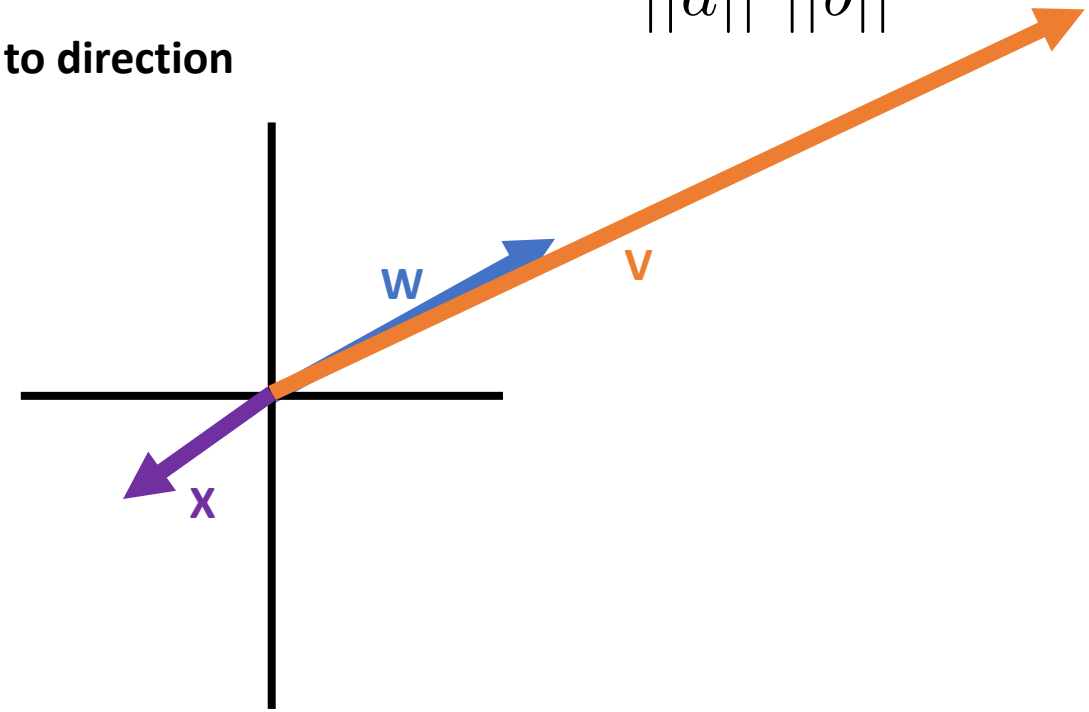
- Some problems with Euclidean Distance
- Here x and w are more similar than w and v
- Is that what we want?



Cosine Similarity

- Calculate the cosine of the angle between two vectors
  - Small angle -> very similar
  - Large angle -> very dissimilar
  - **Invariant to length, sensitive to direction**

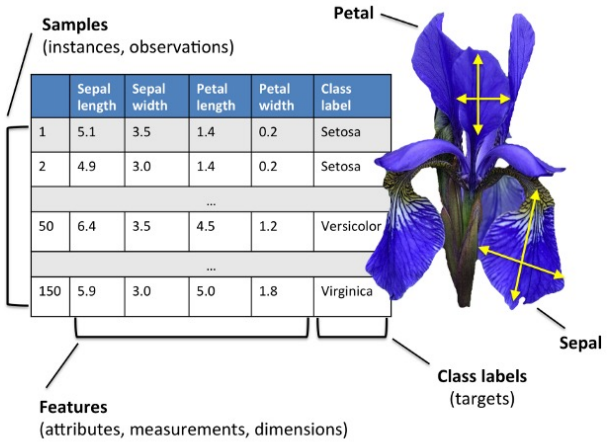
$$\frac{a \cdot b}{||a||\ ||b||}$$



Matrices

- Can be thought of as a function that transforms space

- In ML, our data is usually formatted into a matrix, where the rows correspond to data samples, and the columns correspond to the features



	Sepal length	Sepal width	Petal length	Petal width	Class label
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
...					
50	6.4	3.5	4.5	1.2	Versicolor
...					
150	5.9	3.0	5.0	1.8	Virginica

Matrix Multiplication

- E.g. A B
- Each *row* vector of A dot product with each *column* vector of B
  - Again, Roman Catholic to remember which is rows and which is cols
- Scalar appears in resulting matrix where the row and column intersect

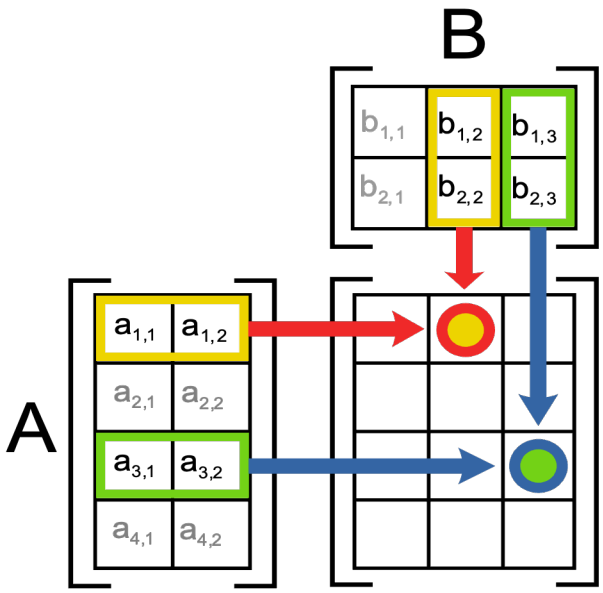
- The # cols of A must match # of rows in B

$$A \in \mathbb{R}^{n \times p}$$

$$B \in \mathbb{R}^{p \times m}$$

$$(AB) \in \mathbb{R}^{n \times m}$$

Matrix Multiplication



Matrix Multiplication

Example

Special Matrices

- Identity matrix (often denoted `I`)
  - Square matrix, All zeros, except for the diagonal elements are 1

$$\mathbf{I_n} = \underbrace{\left.\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}\right\}}_{n \text{ columns}} n \text{ rows}$$

Special Matrices

- Called the Identity because $IA = A$, for all matrices A

Special Matrices

- Inverse of a matrix $A(A^{-1}) = I$
- Only square matrices are invertible

- Finding the inverse is complex for large matrices
  - We won't worry about it, the computer can do it for us
- Some matrices are not invertible! (Singular)

Probability Overview

Many of these slides are derived from
Seyong Kim, Tom Mitchell, William Cohen, Eric Xing. Thanks!

Why do we care about probability?

- Helps us reason about how to make the best decision for cases were we need to generalize:

Temp	**Precip**	**Day**	**Clothes**	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**

Generalization

- Dealing with previously unseen cases
- Will she walk or drive?

Temp	Precip	Day	Clothes	
22	None	Fri	Casual	**Walk**
3	None	Sun	Casual	**Walk**
10	Rain	Wed	Casual	**Walk**
30	None	Mon	Casual	**Drive**
20	None	Sat	Formal	**Drive**
25	None	Sat	Casual	**Drive**
-5	Snow	Mon	Casual	**Drive**
27	None	Tue	Casual	**Drive**
24	Rain	Mon	Casual	**?**

We might plausibly make any of the following arguments:

- She's going to walk because it's raining today and the only other time it rained, she walked.
- She's going to drive because she has always driven on Mondays…

Random Variables

- Informally, X is a random variable if
  - X denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment
    - e.g. rolling a die

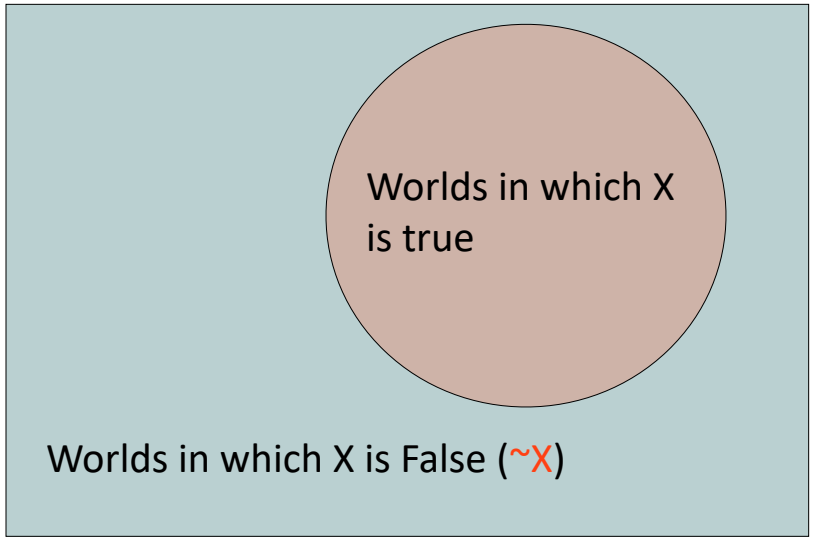
- Examples
  - X = True if a randomly drawn person from our class is female
    - binary
  - X = The hometown of a randomly drawn person from our class
    - multivalued
  - X = True if two randomly drawn persons from our class have same birthday
    - binary

Random Variables

- Define P(X) as "the fraction of possible worlds in which X is true" or "the fraction of times X holds, in repeated runs of the random experiment"
  - the set of possible worlds is called the sample space, S

Blue Rectangle: Sample space of all possible worlds (S)

Area = 1 (all possible things)



P(X) = Area of reddish oval

$0 < P(X) < 1$

A little formalism

More formally, we have

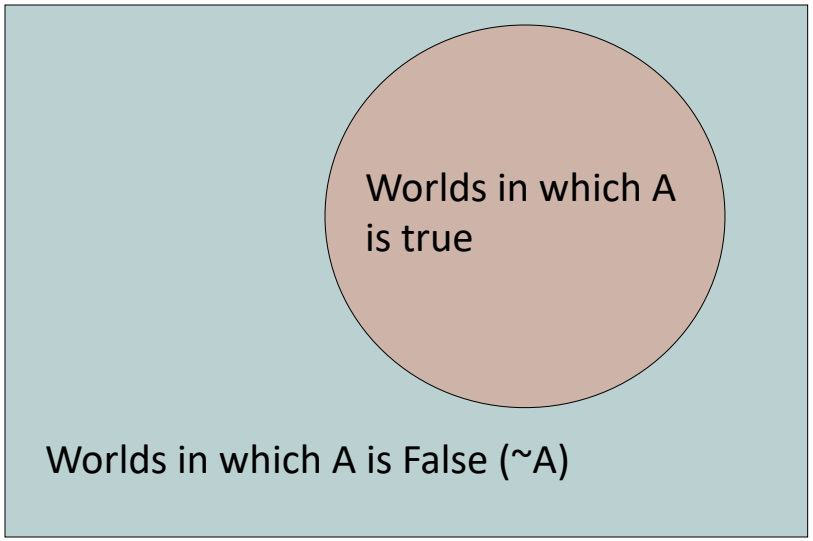
- a sample space **S** (e.g., set of students in our class)
  - aka the set of possible worlds

- a random variable is a function defined over the sample space
  - Handedness: S → { r, l} (binary, discrete)
  - Height: S → Real numbers (continuous)
- an event is a subset of S
  - e.g., the subset of S for which handedness = r
  - e.g., the subset of S for which (handedness=r) AND (eyeColor=blue)
- We are often interested in probabilities of specific events and of specific events conditioned on other specific events

The Axioms of Probability

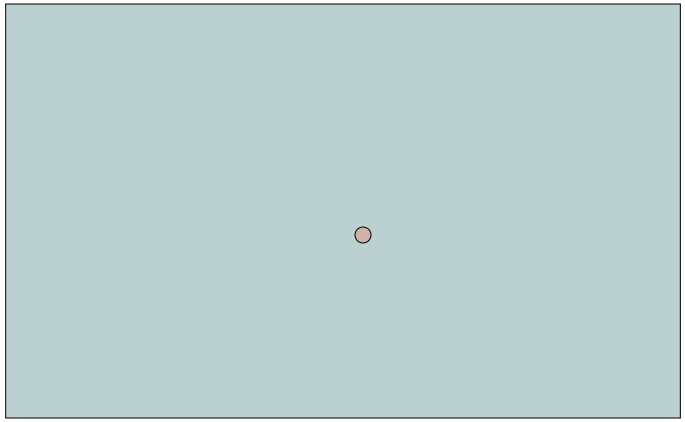
- Assume binary random variables A and B.
  - $0 <= P(A) <= 1$
  - $P(True) = 1$
  - $P(False) = 0$
  - $P(A \text{ or } B) = P(A) + P(B) – P(A \text{ and } B)$

Visualizing Probability Axioms



Interpreting the axioms

- P(A) = 0



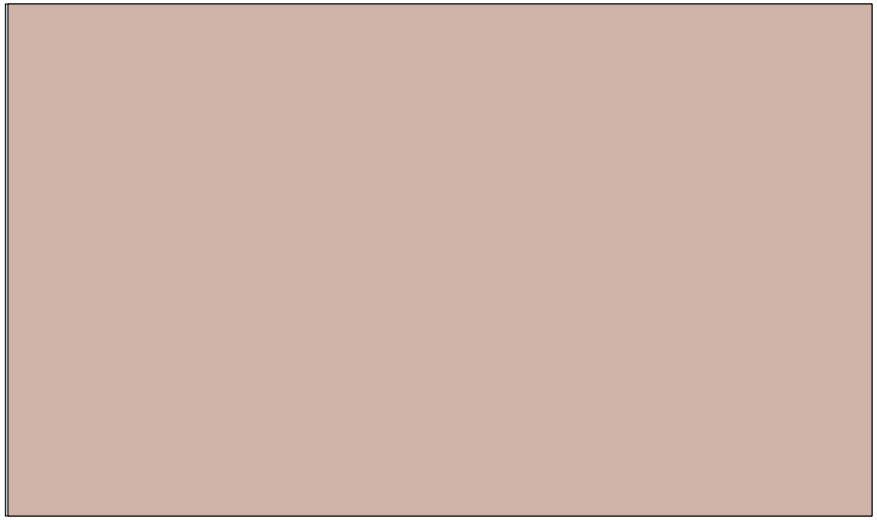
The area of A can't get any smaller than 0

And a zero area would mean no world could ever have A true

P(True) = 0

Interpreting the axioms

- P(A) = 1



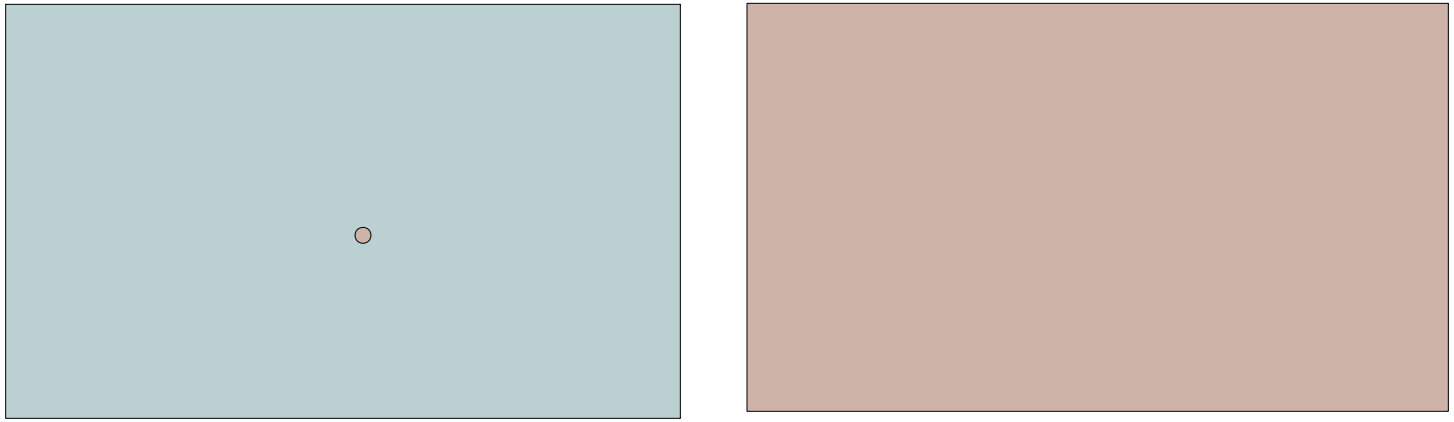
The area of A can't get any bigger than 1

And an area of 1 would mean all worlds will have A true

P(True) = 1

Interpreting the Axioms

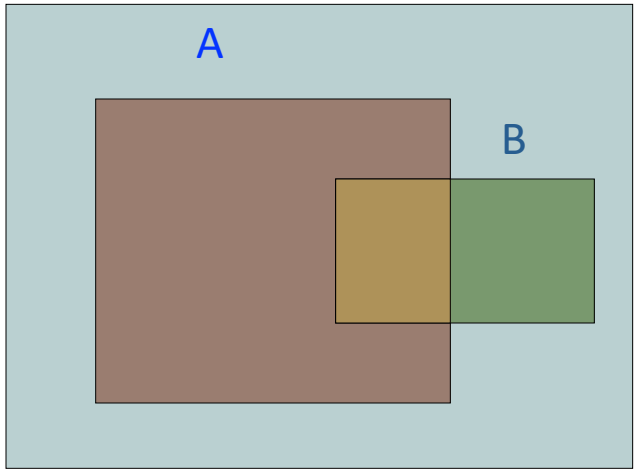
- 0 <= P(A) <= 1



Interpreting the axioms

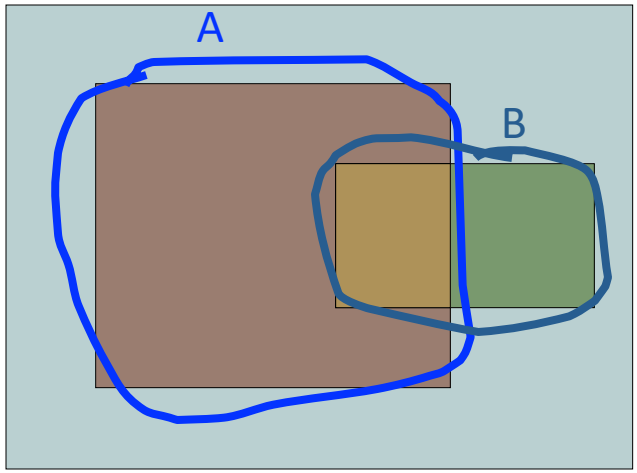
- P(A or B) = P(A) + P(B)

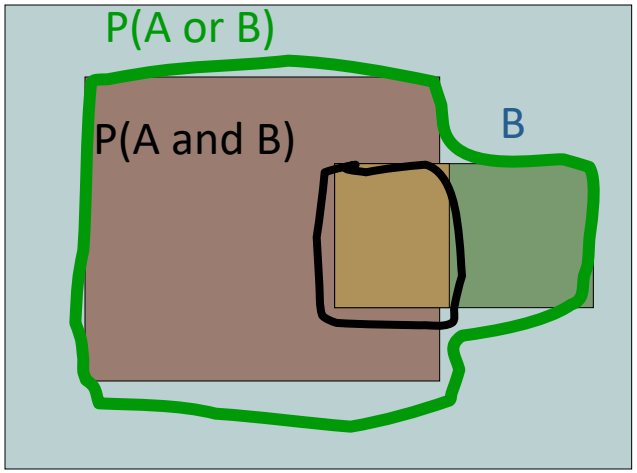
[WRONG! but why?]



Interpreting the axioms

- P(A or B) = P(A) + P(B) - P(A and B)





Simple addition and subtraction

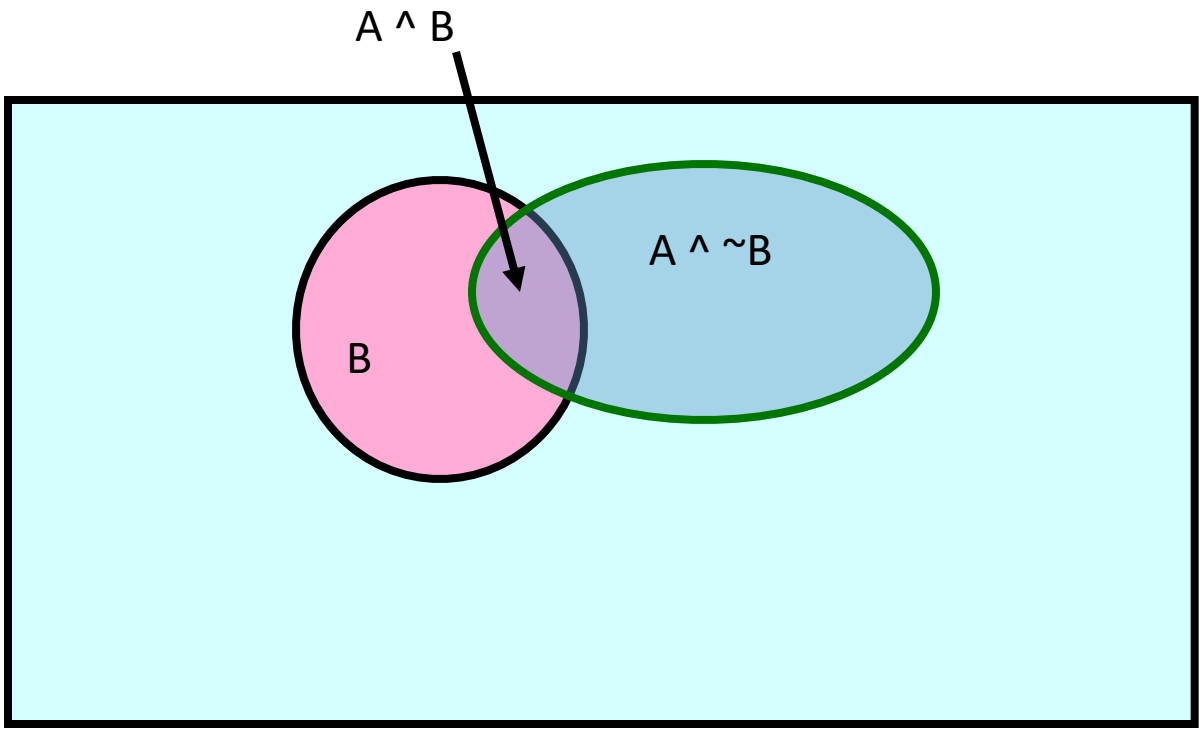
Another useful theorem

- 0 <= P(A) <= 1, P(True) = 1, P(False) = 0,
- P(A or B) = P(A) + P(B) - P(A and B)

➔ P(A) = P(A ^ B) + P(A ^ ~B)

Elementary Probability in Pictures

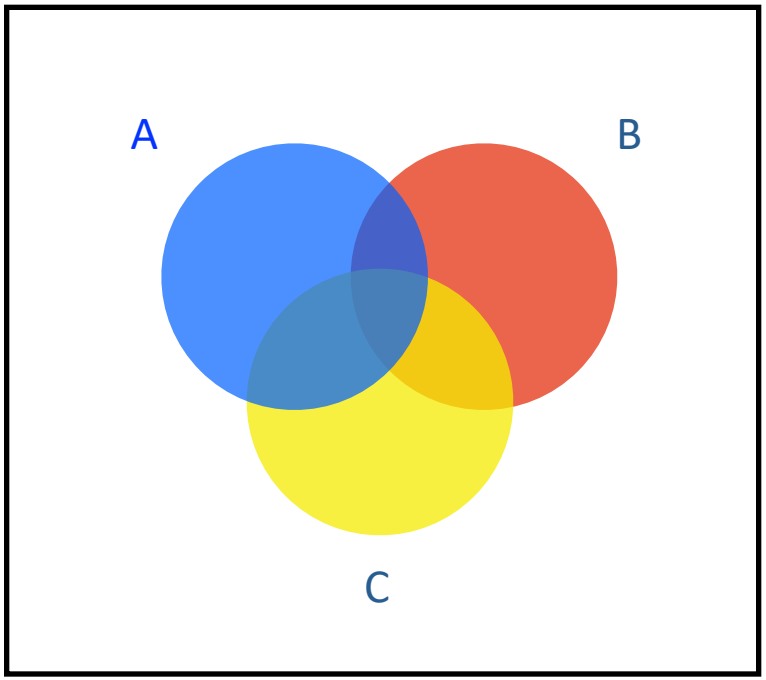
- P(A) = P(A ^ B) + P(A ^ ~B)



- P(A or B) = P(A ^ B) + P(A ^ ~B) + P(~A ^ B)

Extending the Axiom

- P(A or B or C) = ?



Multivalued Discrete Random Variables

- Suppose A can take on more than 2 values
- A is a random variable with arity k if it can take on exactly one value out of $\{v_1, v_2, \ldots v_k\}$
- *Example:* A=*{1,2,3….,20}:* good for 20-sided dice games
- Notation: let's write the event AHasValueOf*v* as "A=*v*"

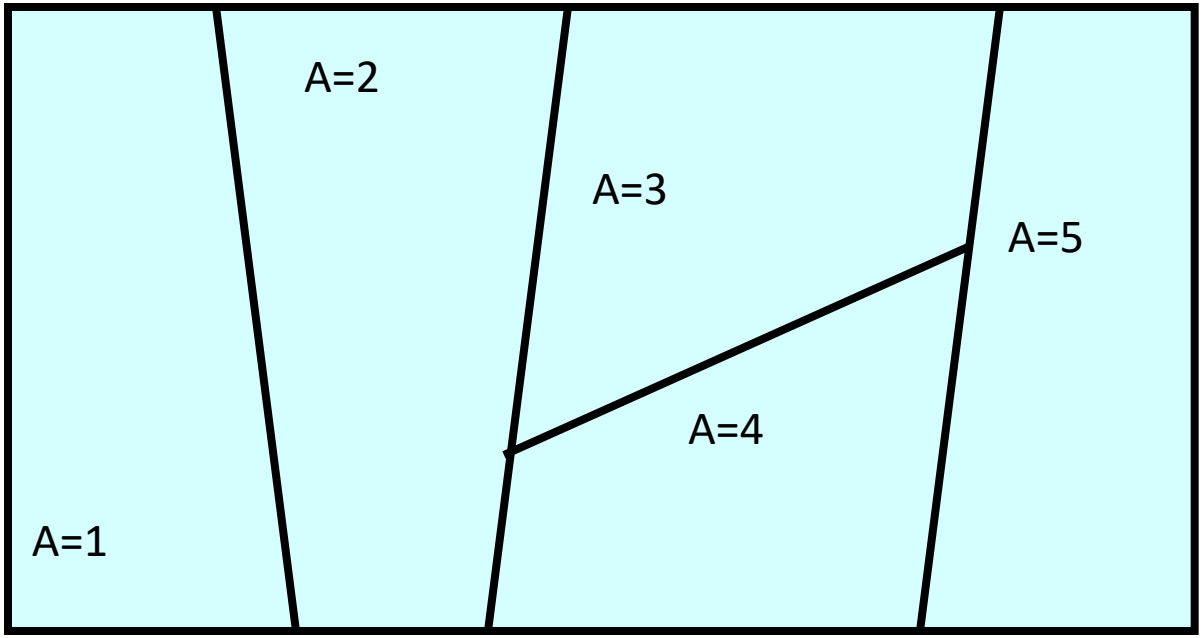
- Thus…

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \ \ldots \ \vee A = v_k)$$

Elementary Probability in Pictures

$$\sum_{j=1}^{k} P(A = v_j) = 1$$ (Law of total probability)

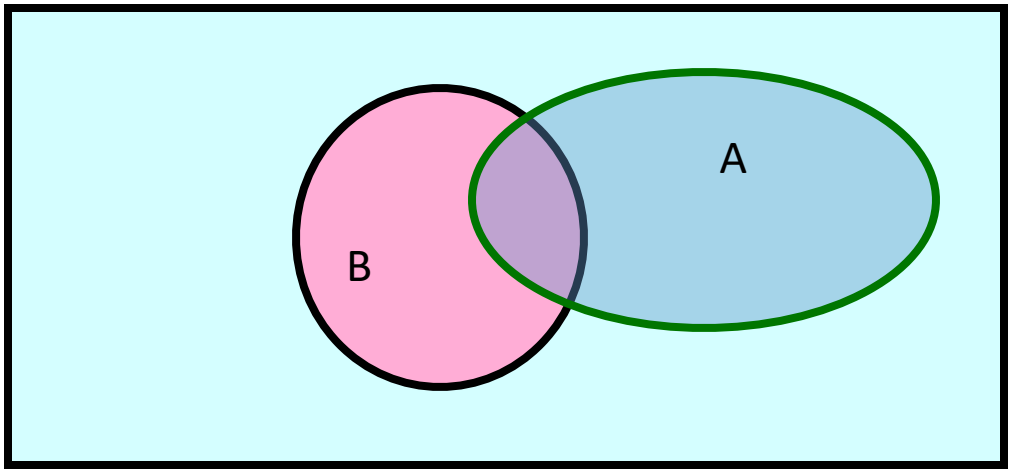


Definition of Conditional Probability

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

We say "probability of A given b"

Foundation for Bayes' Rule!



Definition of Conditional Probability

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

Corollary: The Chain Rule

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

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

Independent Events

- Definition: two events A and B are *independent* if:

  P(A and B)=P(A)*P(B)

- Intuition: knowing A tells us nothing about the value of B (and vice versa)
- From chain rule

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

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

- You frequently need to assume the independence of *something* to solve a learning problem.

Continuous Random Variables

- The discrete case: sum over all values of A is 1

$$\sum_{j=1}^{k} P(A = v_j) = 1$$

- The continuous case: infinitely many values for A and the *integral* is 1

$$\int_{-\infty}^{\infty} f_P(x)\,dx = 1$$

f(x) is a probability density function (pdf)

1. 0<=P(A) <= 1
2. Pr(True) = 1
3. P(A or B) = P(A) + P(B) - P(A and B)

also….

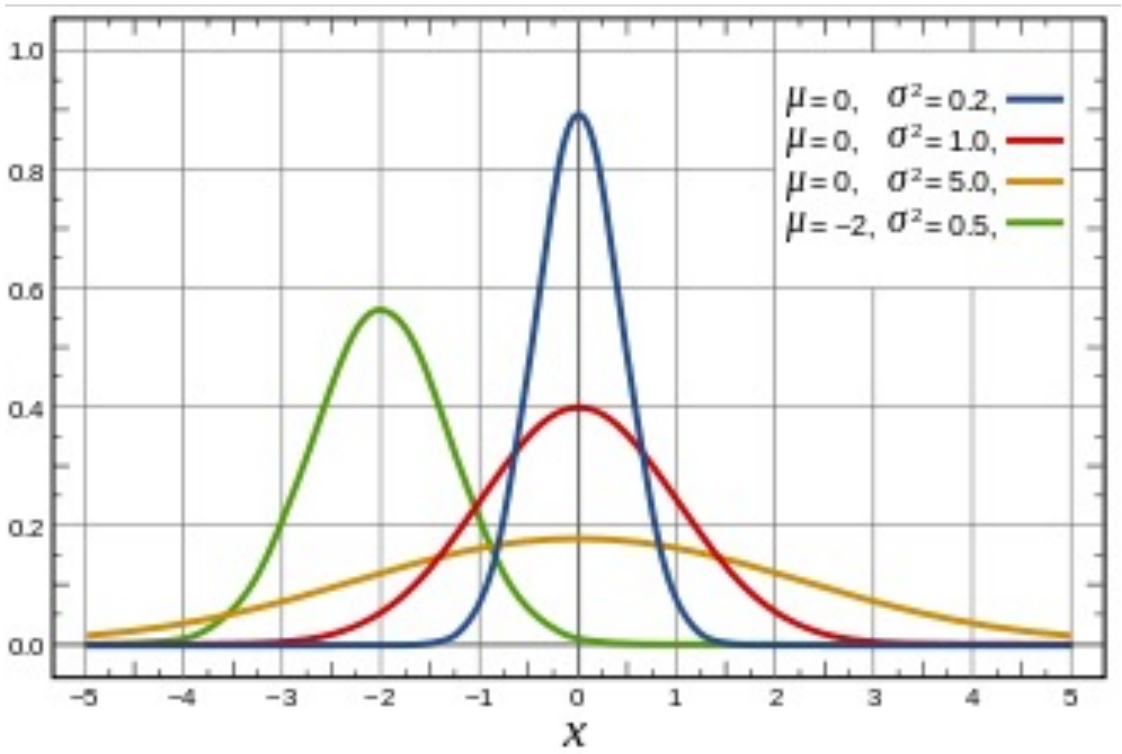
$$\forall x,\ f_P(x) \geq 0$$

Continuous Random Variables

Gaussian probability density with parameters

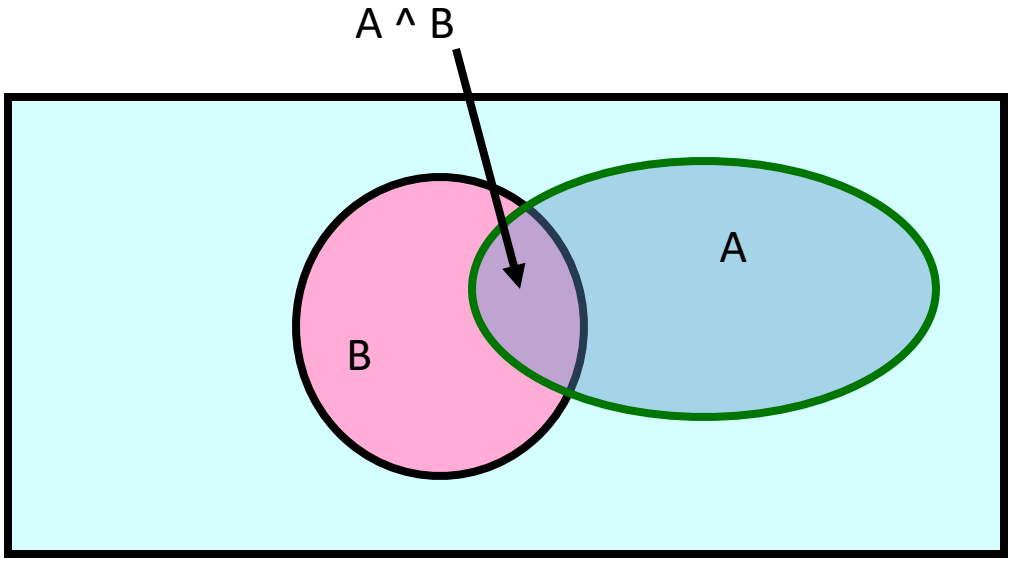
- mean μ
- standard deviation σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



Bayes Rule

- let's write two expressions for P(A ^ B)



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

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

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

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Other Forms of Bayes Rule

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

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

Applying Bayes Rule

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

A = you have the flu, B = you just coughed

Assume:
P(A) = 0.05

Also assume the following information is known to you
P(B|A) = 0.80
P(B| ~A) = 0.4

what is P(flu | cough) = P(A|B)?

Real life application of Bayes Rule!

- Covid vaccination efficacy

Half of the people in the hospital

with covid are vaccinated!

Given that a person is in the hospital, there's a 50% chance they are vaccinated.

$$P(\text{vacc} \mid \text{hospital}) = 0.50$$

Wow, the vaccine isn't very effective!

P (vacc | hospital)

P (not vacc | hospital)

What we know

P (hospital | vacc)

P (hospital | not vacc)

What we **want** to know

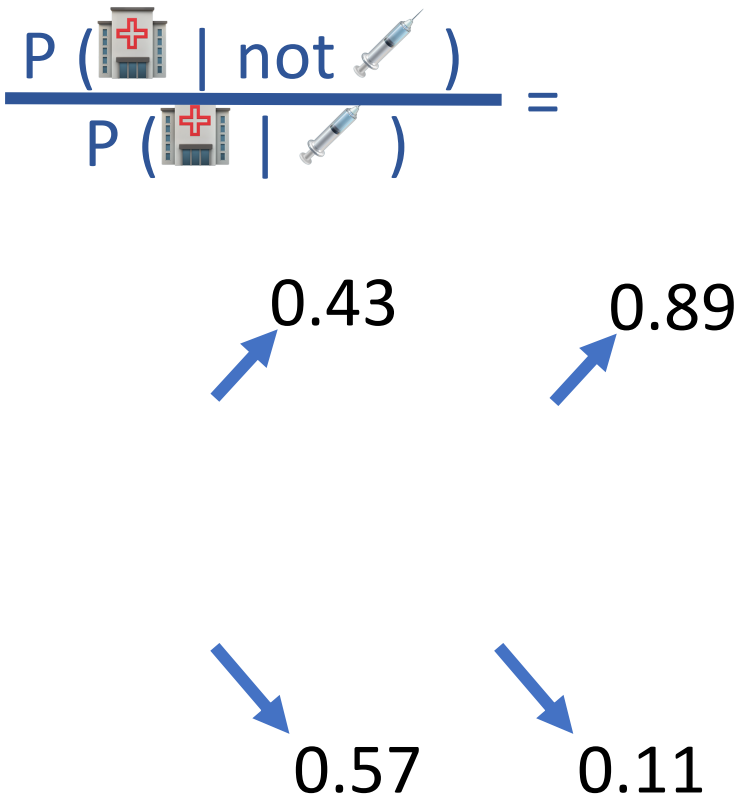
https://www.alberta.ca/stats/covid-19-alberta-statistics.htm#vaccine-outcomes

Bayes' Rule: $P(A|B) = \frac{P(B|A)*P(A)}{P(B)}$

$P(🏥 \mid 💉) = \frac{P(💉 \mid 🏥)\,P(🏥)}{P(💉)}$

$P(🏥 \mid \text{not } 💉) = \frac{P(\text{not } 💉 \mid 🏥)\,P(🏥)}{P(\text{not } 💉)}$

https://www.alberta.ca/stats/covid-19-alberta-statistics.htm#vaccinations



(These values were correct in Jan 2022)

Joint distribution

- Probability of >1 thing happening at the same time
  - Probability it will rain today and I forgot my umbrella
    - P(rain=true,umbrella=false)

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

The Joint Distribution

Example: Boolean variables A, B, C

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).

The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

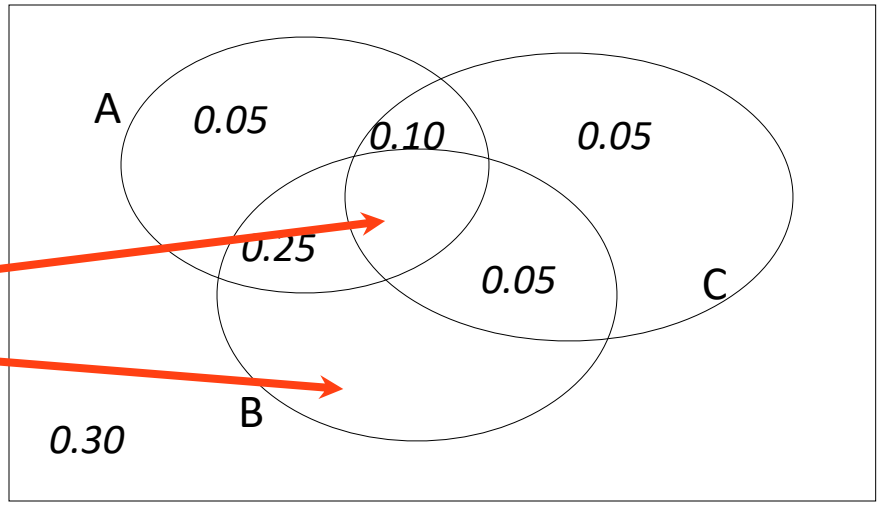
The Joint Distribution

Example: Boolean variables A, B, C

A	B	C	Prob
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10

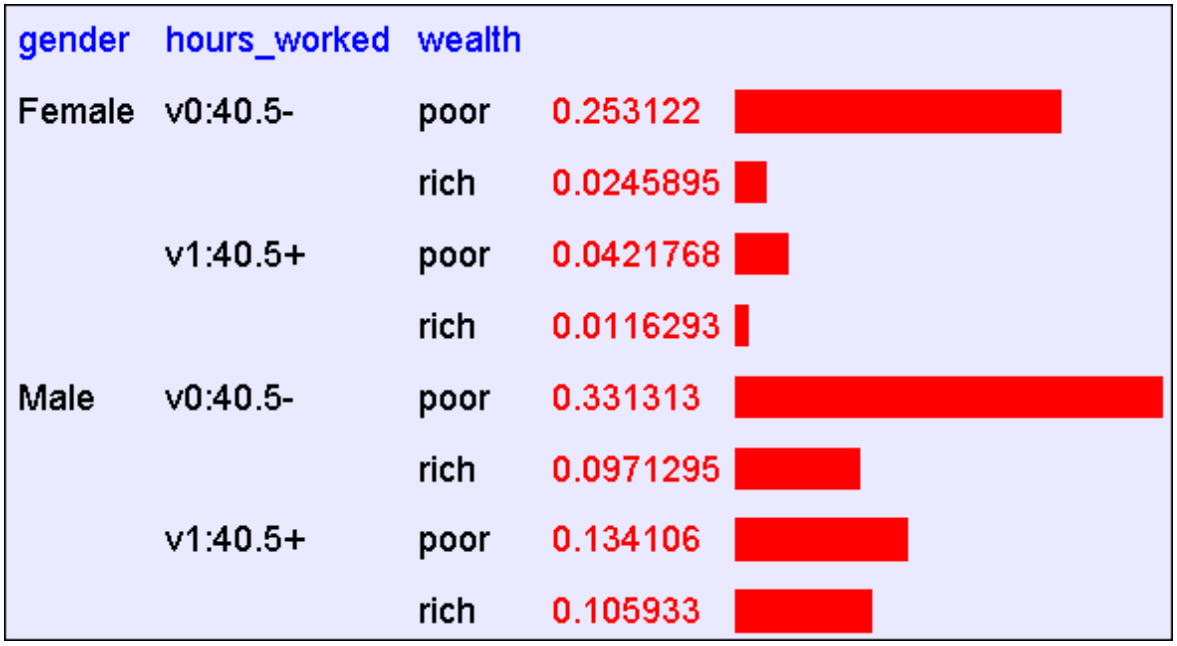
Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.



What goes here?

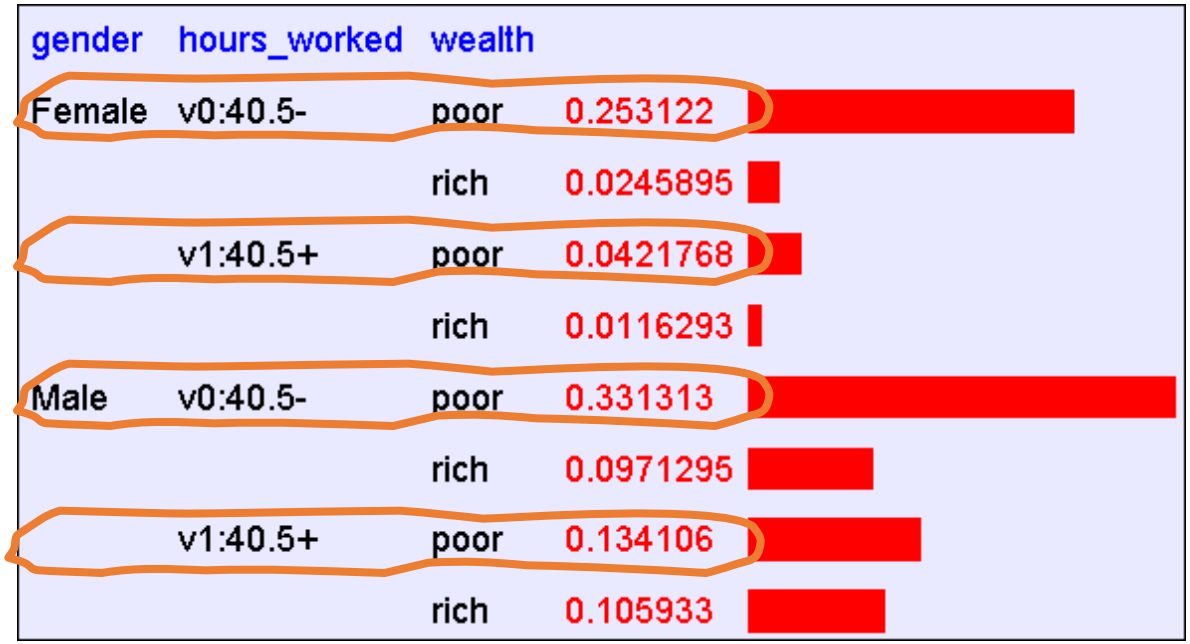
Joint Probability Distribution



gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Once you have the joint distribution, you can ask for the probability of any logical expression involving your attribute

Using the Joint Distribution



P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

Maximum Likelihood Estimation (MLE)

Rich vs Poor

What is the probability of a person being rich, given you know nothing else about that person?

3:2

Why 3/5?

We assume that the wealth of the people in our dataset *D* is independently distributed

θ = Probability of being rich = P(rich)

? = Probability of being poor = P(poor)

Why 3/5?

We assume that the wealth of the people in our dataset *D* is independently distributed

θ = Probability of being rich = P(rich)

? = Probability of being poor = P(poor)

D = { r, p, r, r, p} α_r = # rich α_p = # poor

$$P(D) = P(r \text{ and } p \text{ and } r \text{ and } r \text{ and } p)$$

A little math

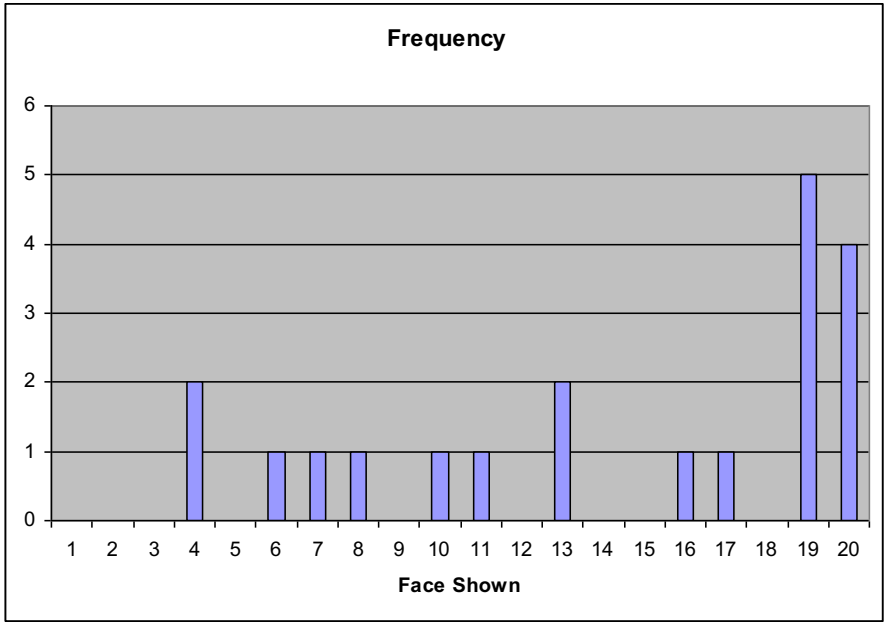
$$\underset{\theta}{\operatorname{argmax}}\ P(D) = (1-\theta)^{\alpha_F} * \theta^{\alpha_H}$$

That’s Maximum Likelihood Estimation (MLE)

It’s not always the best solution…

Issues with MLE estimate

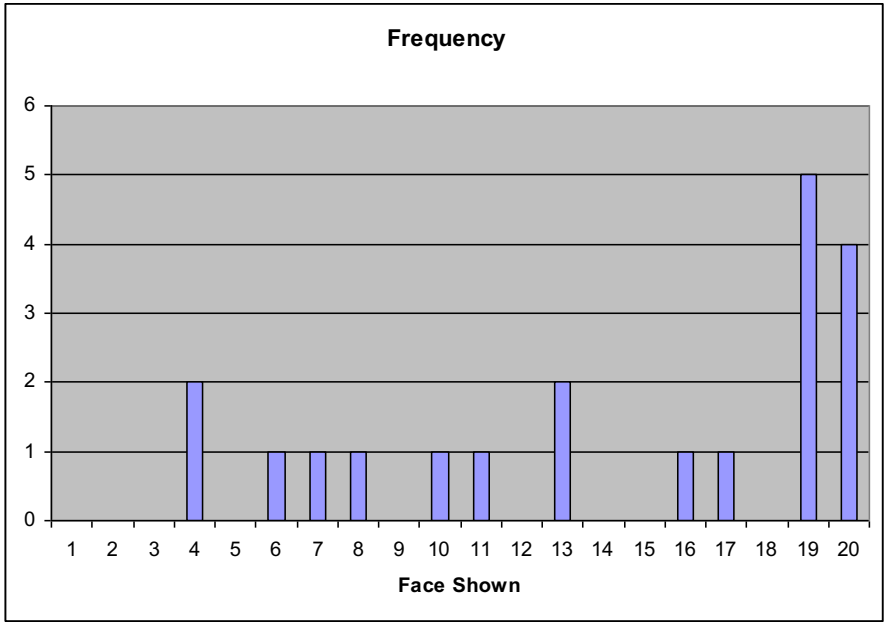
I bought a loaded 20-faced die (d20) on EBay…but it didn't come with any specs. How can I find out how it behaves?



1. Collect some data (20 rolls)
2. Estimate P(i)=CountOf(rolls of i)/CountOf(any roll)

Issues with MLE estimate

I bought a loaded 20-faced die (d20) on EBay…but it didn't come with any specs. How can I find out how it behaves?



P(1)=0

P(2)=0

P(3)=0

P(4)=0.1

…

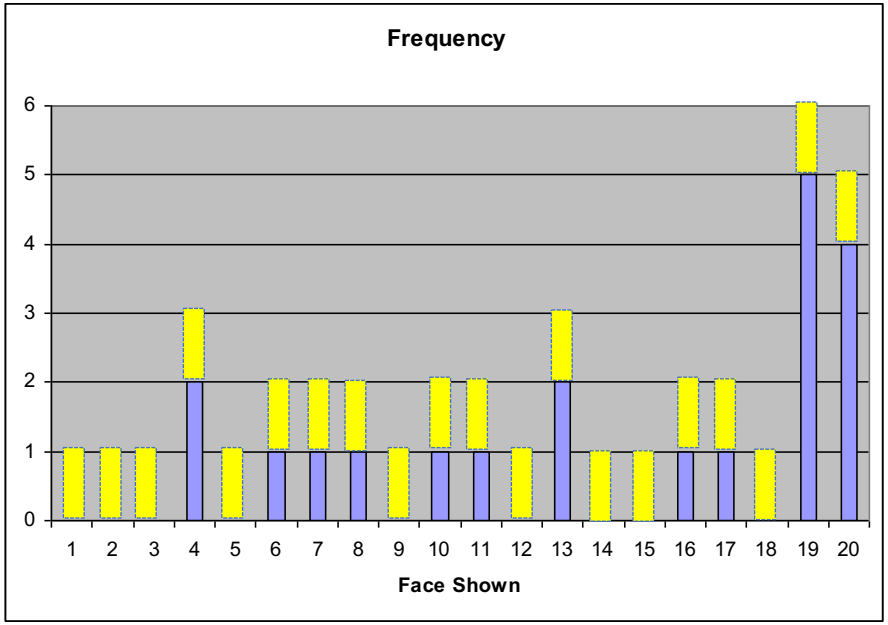
P(19)=0.25

P(20)=0.2

But: Do I really think it's *impossible* to roll a 1,2 or 3?

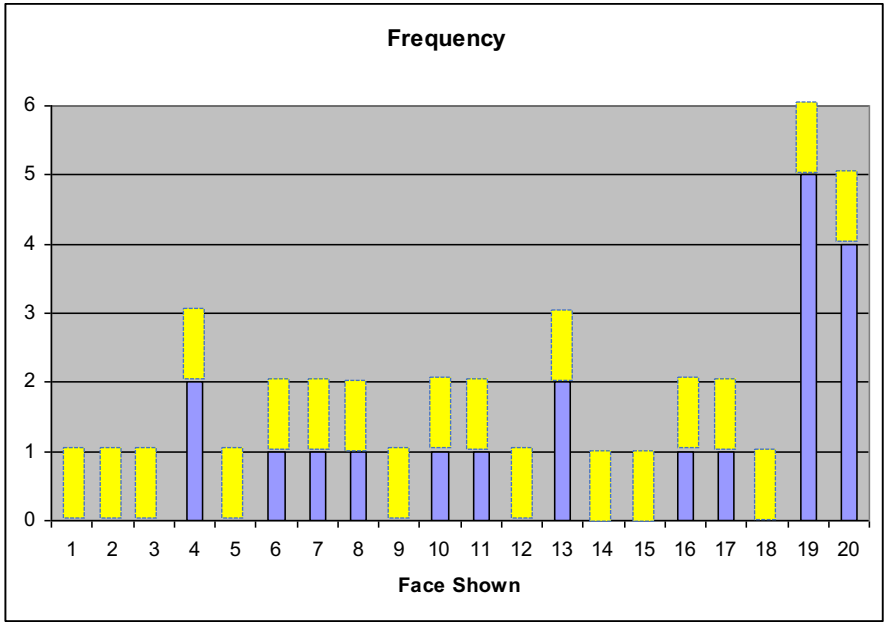
A better solution

I bought a loaded 20-faced die (d20) on EBay…but it didn't come with any specs. How can I find out how it behaves?



0. *Imagine* some data (20 rolls, each i shows up 1x)
1. Collect some data (20 rolls)
2. Estimate P(i)

A better solution?



P(1)=1/40

P(2)=1/40

P(3)=1/40

P(4)=(2+1)/40

…

P(19)=(5+1)/40

P(20)=(4+1)/40=1/8

MAP = maximum a posteriori estimate

$$\hat{P}(i) = \frac{CountOf(i) + 1}{CountOf(ANY) + CountOf(IMAGINED)}$$

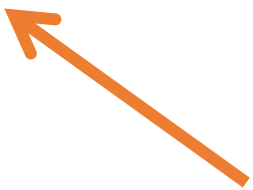
0.2 *vs.* 0.125 – really different! Maybe I should "imagine" less data?

What if we know that poor people are much more common than rich people?

We have a belief about θ

- $P(\theta|D) = P(D|\theta)*P(\theta)/P(D)$

$$\propto P(D|\theta)*P(\theta)$$



Now we can incorporate our belief about θ

This is a MAP (Maximum A Posteriori) Estimate

Conjugate Prior

- Our likelihood so far has been based on a Bernoulli distribution.
- Beta is a conjugate prior to Bernoulli
  - This means their pdfs play nice together
  - $P(D|\theta)*P(\theta)$ will be easy to deal with
  - Called the posterior likelihood

Beta/Binomial Distributions

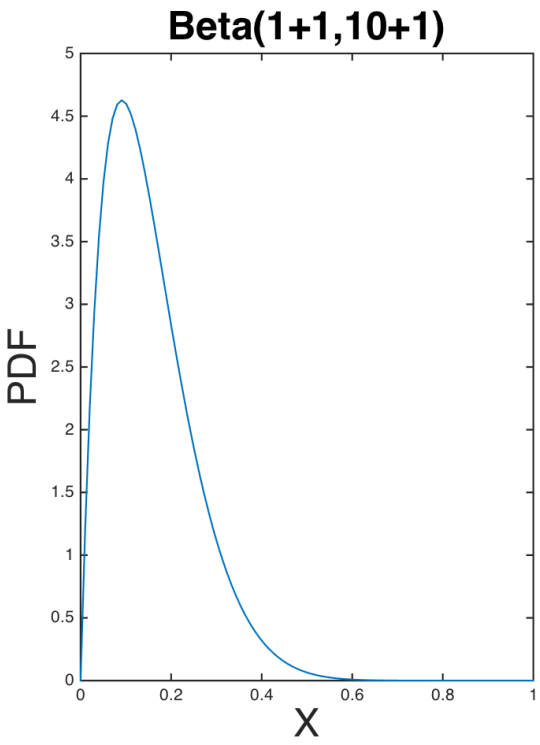
- Beta

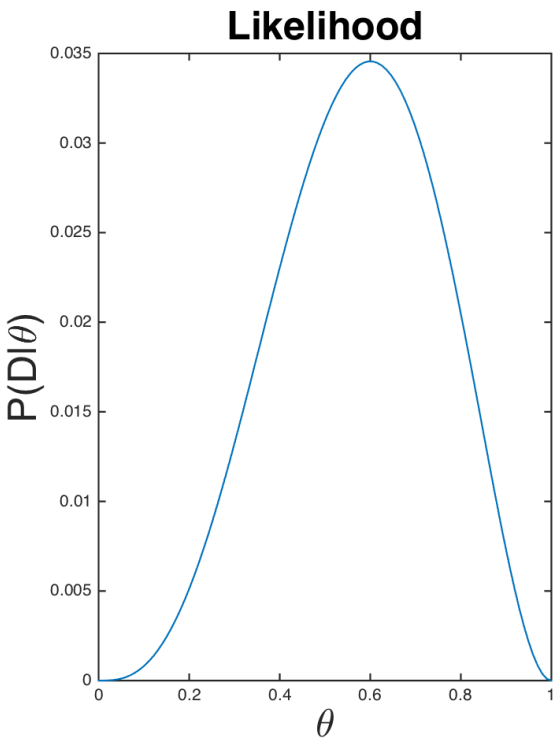
$$P(\theta) \propto (1-\theta)^{\beta_p - 1} * \theta^{\beta_r - 1}$$

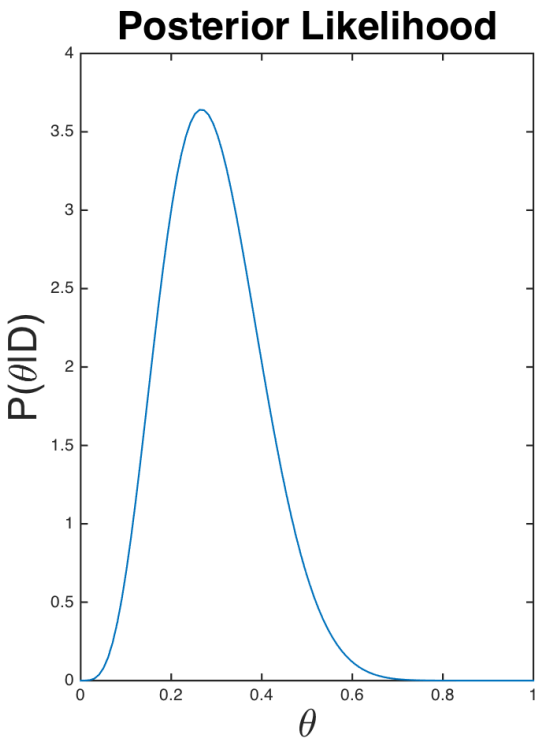
- Binomial

proportional to
(missing a constant)

$$P(\theta) = (1-\theta)^{\alpha_p} * \theta^{\alpha_r}$$


Beta(1+1,10+1)
PDF
X
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
0
0.2
0.4
0.6
0.8
1


Likelihood
P(DIθ)
θ
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0
0.2
0.4
0.6
0.8
1


Posterior Likelihood
P(θID)
θ
0
0.5
1
1.5
2
2.5
3
3.5
4
0
0.2
0.4
0.6
0.8
1

More math

$$P(\theta|D)P(\theta) \propto [(1-\theta)^{\alpha_p} * \theta^{\alpha_r}] * [(1-\theta)^{\beta_p - 1} * \theta^{\beta_r - 1}]$$

Exercises:

1. solve for θ
2. What if $\beta_p = 1 \quad \text{and} \quad \beta_r = 1$?

Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose θ that maximizes probability of observed data

$$\hat{\theta} = \arg\max_{\theta} P(\mathcal{D} \mid \theta)$$

- Maximum a Posteriori (MAP) estimate: choose θ that is most probable given **prior probability and the data**

$$\hat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D})$$

$$= \arg\max_{\theta} = \frac{P(\mathcal{D} \mid \theta) P(\theta)}{P(\mathcal{D})}$$

A wonderful tutorial:
http://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/bernoulli.pdf