Bayesian Thinking for Applied Machine Learning

2019 Lviv Data Science Summer School at Ukrainian Catholic University

Instructor: Max Sklar
July, 2019

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Class Survey: Why take this course?

Other reasons to take this course

Understanding “first principles” of data science can be practical - and even essential when your usually toolbox of tricks fails to deliver.

In your work, and in your life, you are going to need to form beliefs about the world, and take action under uncertain conditions.

There is currently a “reproducibility crisis” throughout the sciences, and clear thinking on drawing conclusions and taking action from data is the only way out!

Goals for This Class

Be able to understand, mediate, and when necessary adjudicate disagreements between people when it comes to interpreting the meaning of probabilities, data and models.

Understand the Bayesian Framework for solving problems.. And it's limitations

Be fluent in the nature of probability and uncertainty

Be fluent in Bayes Rule and Bayesian Thinking - As applied in Machine Learning

Understand the significance and meaning Markov Chain Monte Carlo, which finally unlocked the potential of Bayes Rule in recent years.

Understand the challenges of causality modeling.

Get a taste of probabilistic programming in python

Be a more interesting person than you were at the start of the class.

Format of the Class

Lecture class, going through derivations, explanations, and case studies both in history, and in my personal work at Foursquare

Includes some class discussion about the big issues around the nature of data and probability, and ultimately what this means for machine learning.

Interrupt with questions, especially on key terms. Move back and forth between technical/formal + intuitive/informal

Finally, a few python demos - some of my other code which implements Bayesian ideas, as well as an introduction to pymc and probabilistic programming.

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

I’m Max Sklar.

I am a Data Scientist, Machine Learning Engineer, and New Product Developer at Foursquare

I currently live in Brooklyn, NY

Statue of liberty

I currently live in Brooklyn, NY

First time in Ukraine

But one of my favorite places in New York City is a 24-hour Ukrainian diner called Veselka

More on this LATER

Got my first taste of Machine Learning at NYU

Information Systems Program, 2009-2011

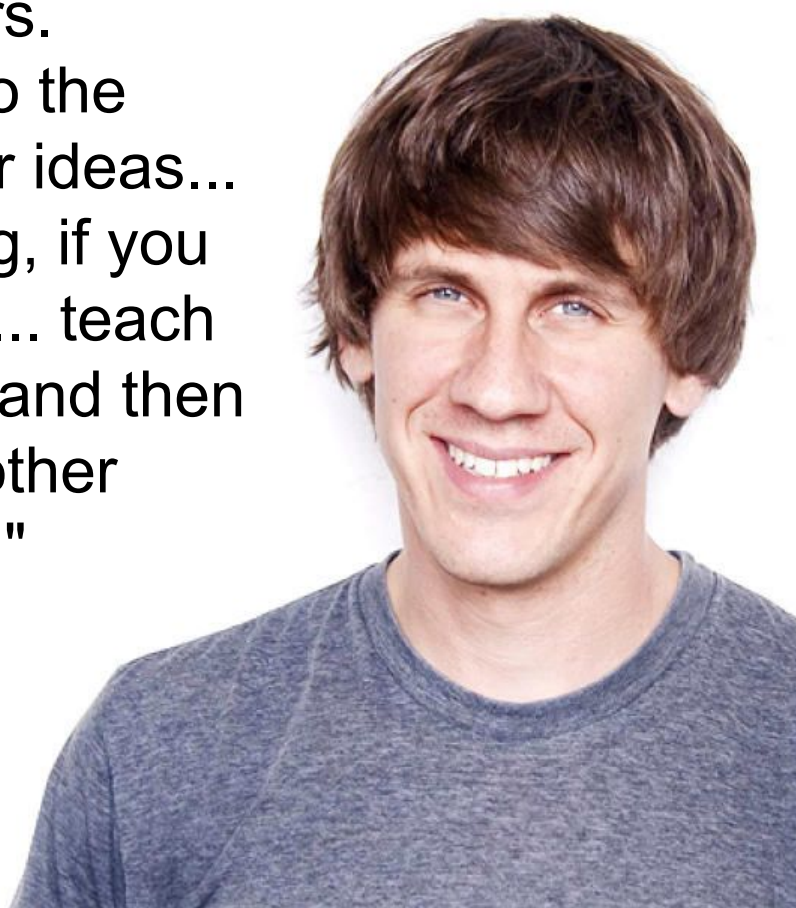
Machine Learning with Yann Lecun

Data Mining at Stern, also Natural Language Processing

Gravitating Towards these problems, also interested in Location and Local Search

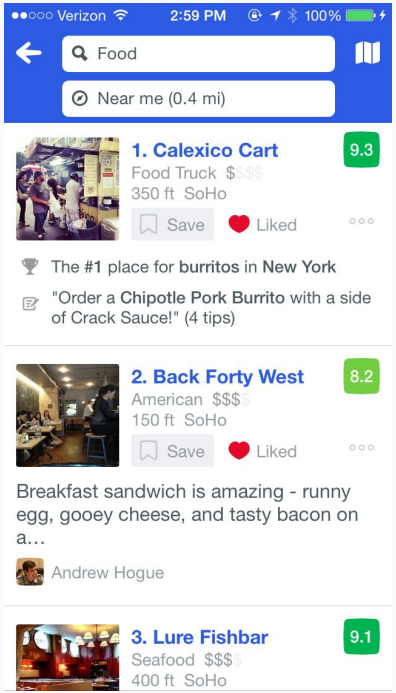
"Don’t let people tell you that your ideas can’t work… Don’t listen to the haters. There’s always haters. Don’t listen to the people who want to shoot down your ideas... If you’re passionate about something, if you have an idea that no one’s done yet... teach yourself that those ideas don’t work and then iterate on top of them. But don’t let other people push you around in that way!"

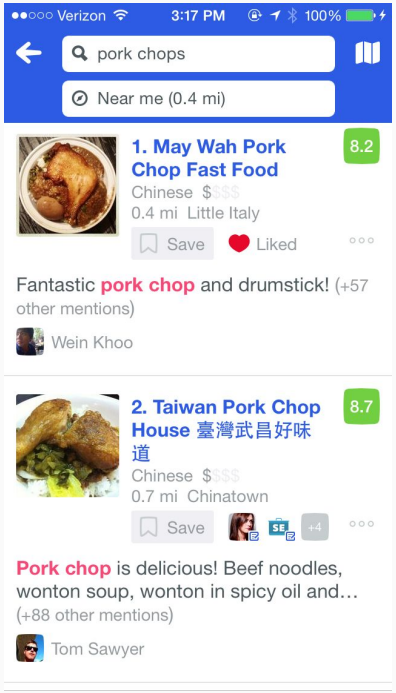
Dennis Crowley, 2011



My Mission at Foursquare: Build the best Local Recommendation Engine in the World

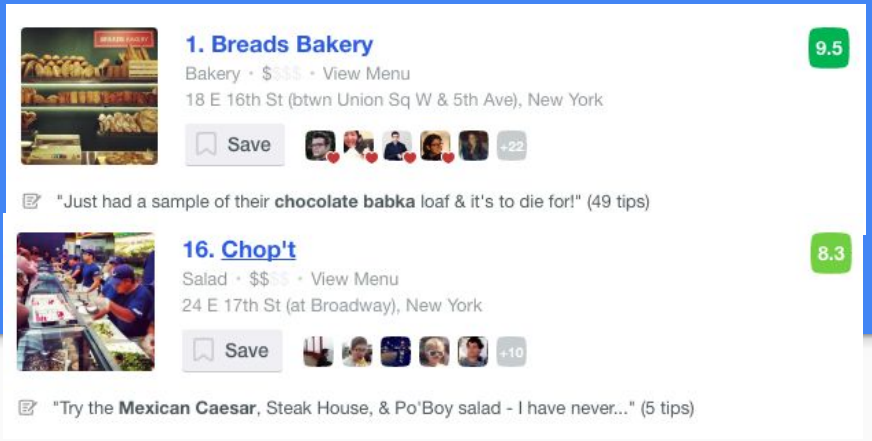
2011

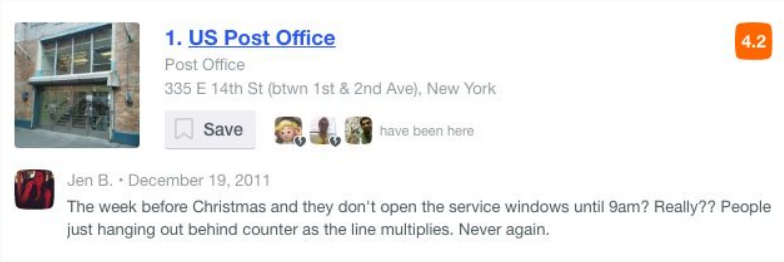




Ratings

2012 - 2014

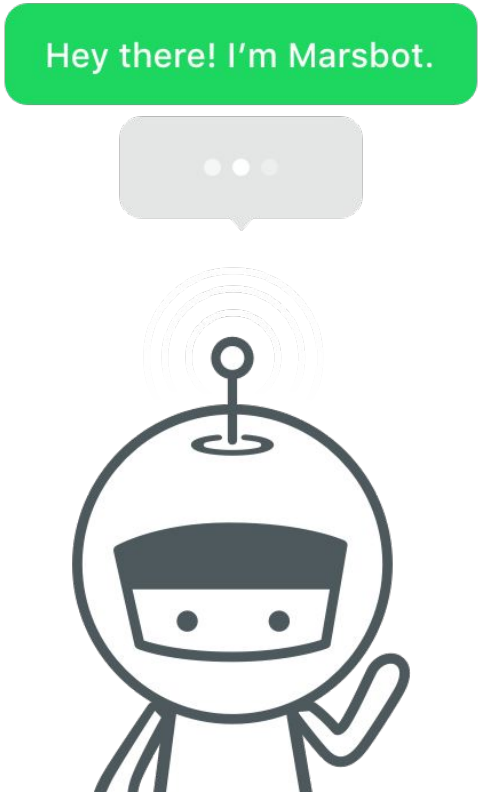




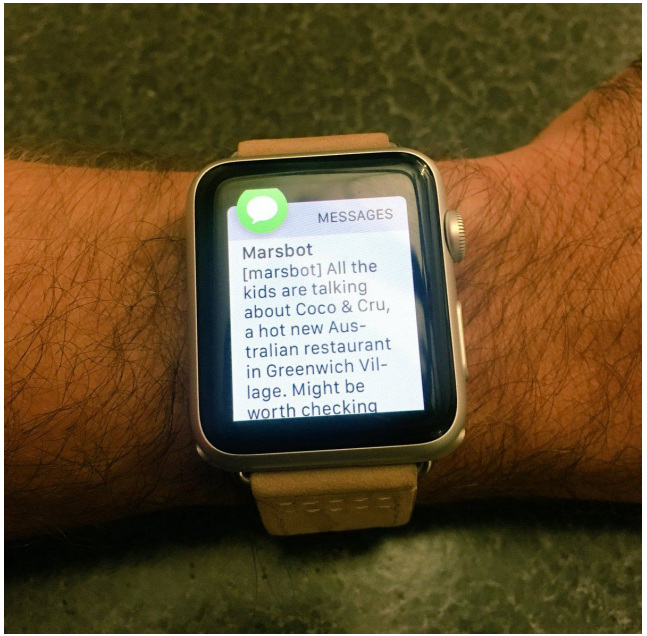
Marsbot: An SMS-Based Personality for Local Recommendations

Marsbot is a character in your pocket that learns your daily habits and routines, uses that to build a profile of you and texts you local recommendations.

2015 - 2016



Example of a push recommendations:



Wed, Jun 1, 11:47 AM

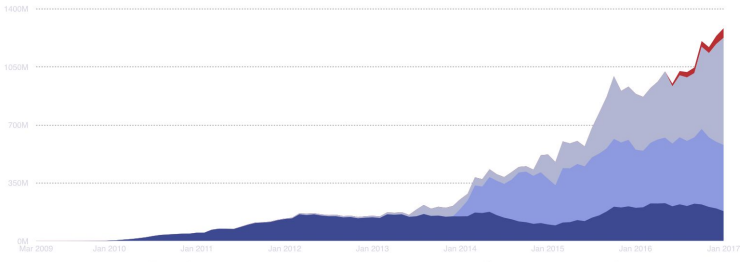
[marsbot] After a meal at La Bagel Delight, some people like to grab a cup of coffee at Hungry Ghost nearby. (http://4sq.com/1T3aHzj)

2017 - Started work on Foursquare's Attribution Product

Main purpose: Use the panel to measure an ad's ability to drive people into stores and locations.

Lift = percent increase in visits with ad

Causality Modeling - more on this later!



11B
HISTORICAL CHECK-INS

1.2 B
OBSERVED DATA POINTS

8B
MONTHLY API CALLS

Metrics are global. Checkins include all explicit checkins, passive checkins are all canonical pcheckins, and user/partner content include tips, edits, likes, ratings, photos, pingbacks, and Pinpoint snaps, and SDK visits are visits from SDK partners. Pilgrim Actives are users (SDK and inapp) that have location on. Total venues are both commercial and private. Checkins are cumulative. API calls are for the most current month.

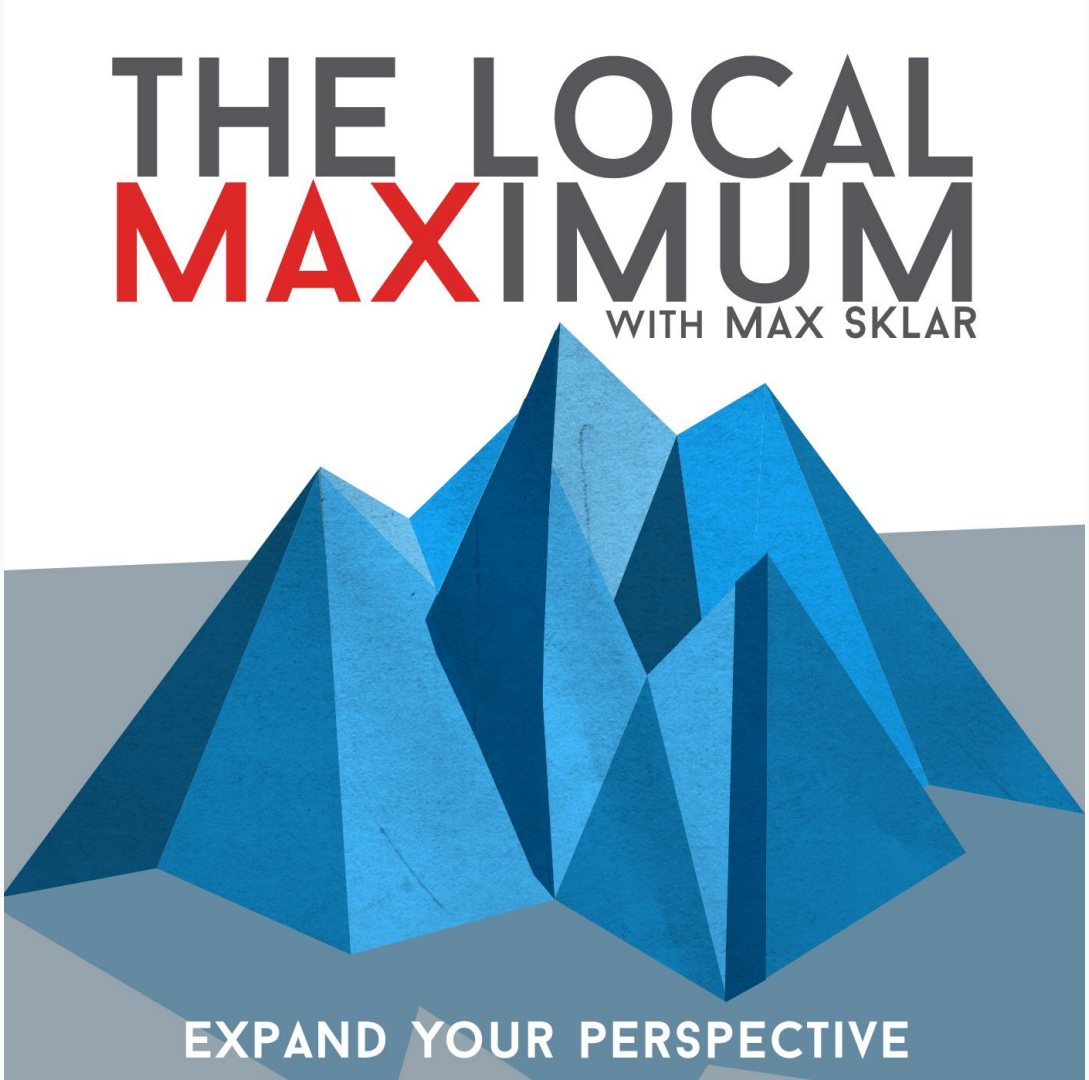
Local Maximum Podcast

Started in February of 2018

Interview engineers, data sciences, entrepreneurs, and authors.

Talk about issues important to the tech community in particular, but to the wider world as well.

Listen to sample (1:00)



Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Episode 21: Probability, Belief, and The Truth

I covered these 3 concepts, and discovered that while their definitions seem straightforward, there is a lot of nuance and misunderstanding.

Group discussion (groups of 5):

- Define Probability
- Define Belief
- Define Truth

The Nature of Truth and Facts

Big word: “epistemology” - theory of knowledge - how do we know what’s true?

We act on many beliefs - how do we justify them?

René Descartes (1596 - 1650): “I think, therefore I am”

The Nature of Truth and Facts

Questions without “estimated” answers: Number of Water Molecules in the Atlantic Ocean

Can we separate action and decisions from beliefs?

What forces ensure that our beliefs align with the “true” causal connections?
- Natural Selection?
- Natural Language (describing things imprecisely but practically)
- Formal Language (describing things precisely but not always practically)
- Formal Logic (documenting true statements, weeding out contradictions)

Ludwig Von Mises (b 1881 Lviv, d 1973 New York)
Free Market Economist. Author of Human Action.
Description from Wikipedia: Man **acts** to dispel feelings of uneasiness, but can only succeed in acting if he **comprehends causal connections** between the ends that he wants to satisfy, and available means.

Argument By Analogy

This might be the most common form of consciously forming beliefs.

Take something you've seen in the past, make an analogy to the situation you're facing now, and draw a conclusion.

Example of argument by analogy in Curb Your Enthusiasm:
https://www.youtube.com/watch?v=gO3Y_IlPyXc

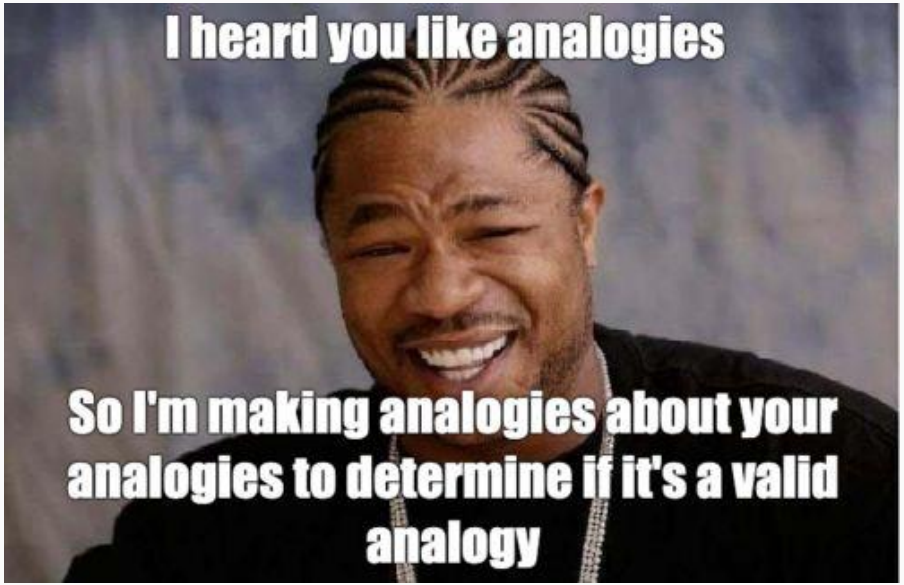
Argument By Analogy

Is argument by analogy valid?

Example: Other human beings are conscious just like I am. One dice roll vs others.

In general yes: Only requires a single assumption.. That there exists some regularity in the universe. Or in a photo - some pixels tell us more information about other pixels.

In particular: there are BAD ANALOGIES. So the argument becomes - is this a good analogy or a bad analogy. So we make an analogy ON the analogy.

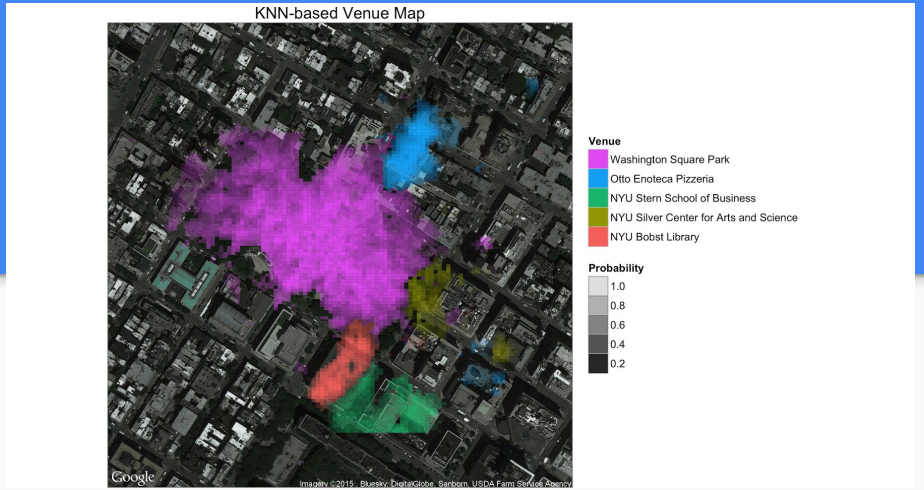


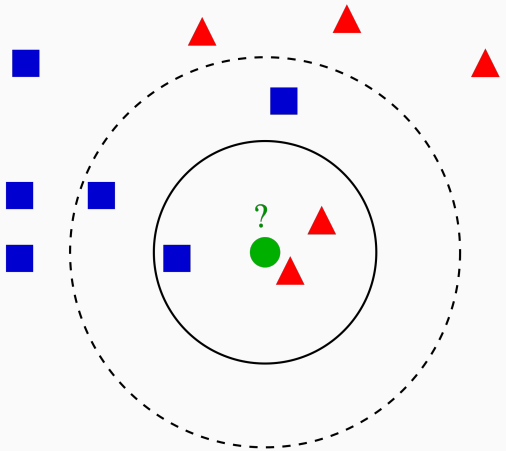
Argument By Analogy

What does this have to do with machine learning?

Arguably, every machine learning algorithm is an argument by analogy. It assumes that what you've seen in the past generalizes (in some way) to what you will see in the future.

The best example of ML by analogy:
K-nearest-neighbors





Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
11) Complicated Posteriors: The MCMC Approach
12) The Problem of Causality and Attribution
13) Intro to Probabilistic Programming: PyMC3
14) Projects and Applications

Expanding to Probabilities

Instead of formally knowing facts, we keep in mind several possible answers - along with their relative likelihood.

Benefit: Can make better decisions

Drawback: More mental overhead, or from a machine perspective, more memory and code complexity.

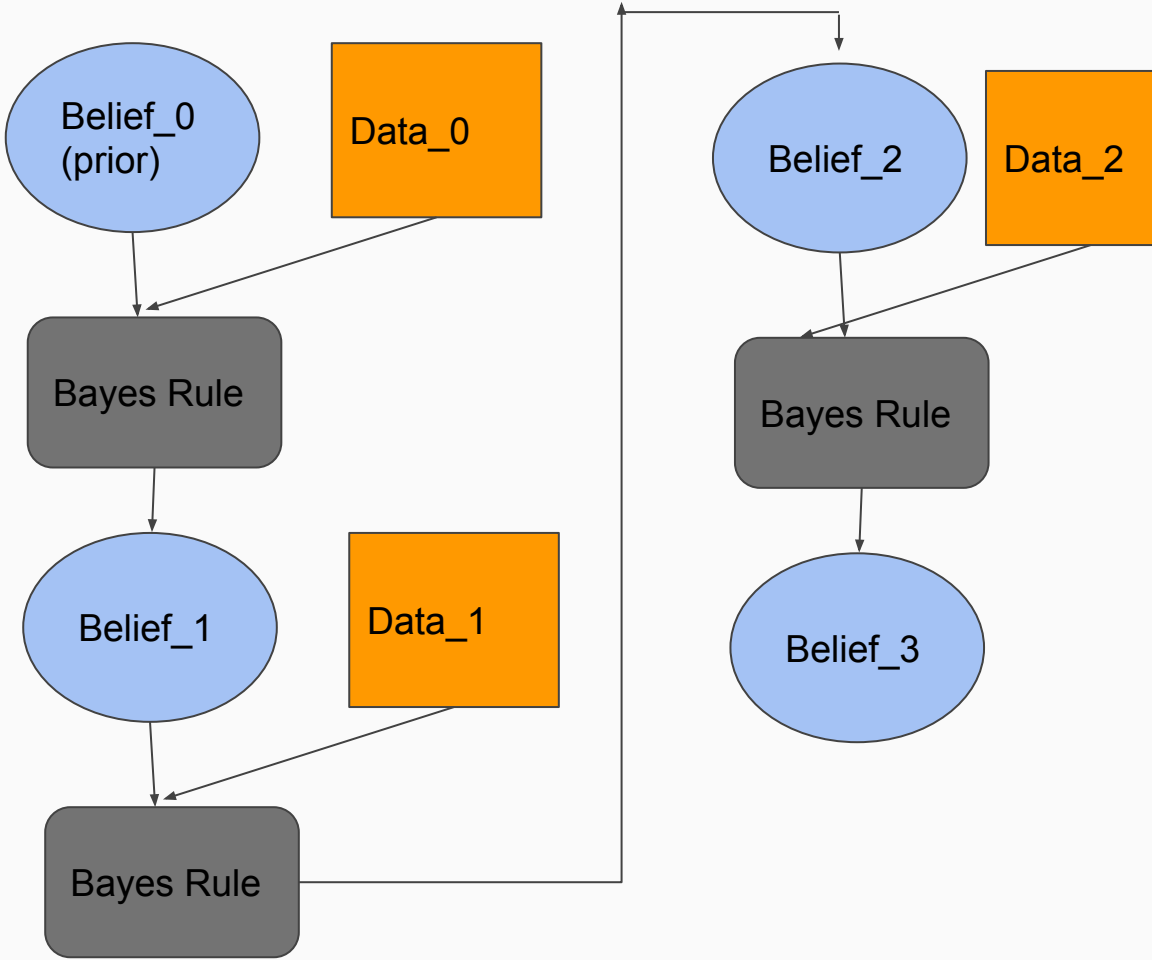
Conclusion: Not all Beliefs should be considered as probabilities, but upgrading a belief from a singular statement to a probabilistic one can bring value if you do it strategically!

Different Interpretations of Probability

Objective Probability	Subjective Probability
Probability is an inherent property of a system - the long-run ratio of repeated trials	Probability is a statement of degree of belief in different potential truths
Preferred by Frequentists	Preferred by Bayesians
Applies well under controlled circumstances: repeated experiments (perfect analogies), games of chance, quantum physics	Applies well under unique events, imperfect analogies, different observers

Finally, we get to Bayesian Inference

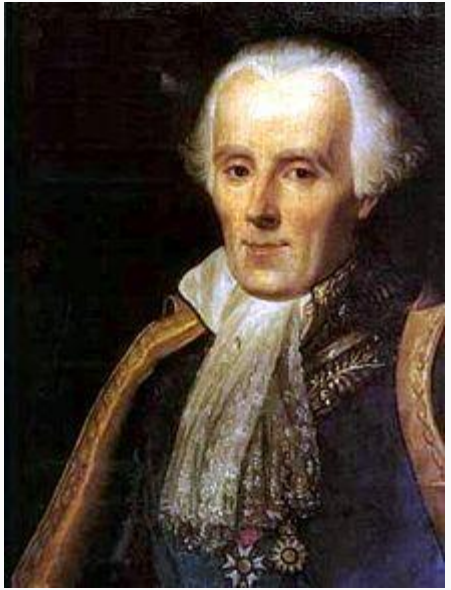
Bayesian Inference is a model for updating subjective beliefs (probabilities) in light of new data



Finally, we get to Bayesian Inference

Thomas Bayes: Presbyterian minister in Scotland, also mathematician.
- Work on “Bayes Rule” published after death.
- Motivation?

Thomas Bayes: 1701 - 1761



Pierre Simon Laplace
1749 - 1827

Finally we get to Bayesian Inference

We have multiple hypotheses (H_1, H_2, H_3, ..etc), and we want to know what the relative probability of each one is.

Derivations on the board: 2 hypothesis version, several hypothesis version

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

$$P(H|D) \propto P(D|H)P(H)$$

In English: The Probability of the Hypothesis given the data is proportional to the likelihood of the data under that hypothesis times the prior of that hypothesis

Finally we get to Bayesian Inference

We have multiple hypotheses (H_1, H_2, H_3, ..etc), and we want to know what the relative probability of each one is.

Derivations on the board: 2 hypothesis version, several hypothesis version

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

Prior (Belief before seeing data)

Posterior (Belief after seeing data)

$$P(H|D) \propto P(D|H)P(H)$$

Likelihood - probability of seeing the data while assuming the hypothesis in question is true

In English: The Probability of the Hypothesis given the data is proportional to the likelihood of the data under that hypothesis times the prior of that hypothesis

Review of the Terms

P(H|D) = probability of the hypothesis after seeing the data = posterior
P(D|H) = probability of the data given the hypothesis is true
P(H) = probability of the hypothesis before seeing the data = prior

P(D) = probability of seeing the data in general (assuming the hypothesis is selected using your prior) = a normalization constant

Canonical Example: Medical Testing

Problem: you are tested for a disease that randomly affects 7.5 million people in the world (assume out of a 7.5 billion world population). The test is 99% accurate in either direction. Your test result comes out positive - what's the probability that you have the disease?

Hypothesis Set = {Have Disease, Don't have Disease}

Data = "Tested Positive"

Run Bayes rule, both from equation and in Max's tabular form

More on the Normalizing Constant + Continuous Hypothesis Space

The Hypothesis Space H could be discrete, or continuous (i.e. a PDF across all real numbers). In this case, the answer P(H|D) is a PDF value (it's actual probability is zero)

The normalizing constant P(D) is tricky.

In Discrete Hypothesis Spaces, it is a sum. In Continuous Hypothesis Spaces, it is a integral.

Without it, you're answers are "unnormalized" probabilities, which means they don't sum up (or integrate up) to 1 but are still correct in terms of ratios to each other. As we'll see, this is often the only information you need.

$$P(D) = \sum_{h \in H} P(D|H)$$

$$P(D) = \int_{h \in H} P(D|H)$$

Now that we see how Bayes Rule works, what are some of the Limitations?

1) We need both a prior and data for Bayes Rule to work. There is no objective way to come up with a Prior. We'll discuss priors further in this course
2) Two people in agreement on the application of Bayes rule, and in agreement on data could have different priors that lead to different conclusions. (Do posteriors converge?)
3) The Hypothesis Space is incomplete - either on purpose to simplify the model, or because it's because it's a hypothesis we haven't considered.
4) The Normalizing constant can be intractable (curse of Dimensionality)
    a) A Finite Hypothesis space with an astronomical number of possibilities (code breaking)
    b) An infinite (continuous or discrete) case without a tractable way to integrate
    c) Foreshadow the role of Markov Chain Monte Carlo

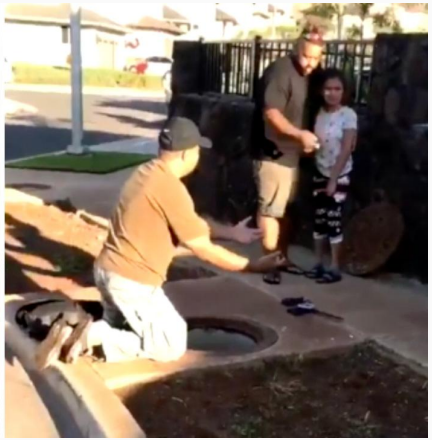
Flat Earthers

- Started out at a satire (claiming the Earth is actually flat)
- But then some people on the internet started believing it.
- They will not change their mind despite evidence
- Derive their posterior from their (100% prior on Earth being flatr) on the board.
- Question for Class: Is this a valid application of Bayes rule?

Hawaii Missile Scare
An “Estimated” Posterior

February 2018 - State official accidentally sent this alarm out to all residents of Hawaii.

Episode 1 of The Local Maximum: Can we estimate the probability of an actual incoming Ballistic Missile given the alert?
- On board, tabular, class input on probabilities.



Interview Question: Bayesian Coin (10 mins)

I have 2 coins in a bag: Coin A is fair, and Coin B is double sided (both sides are head). I pull one coin out, but I don’t examine it, so I don’t know whether I have coin A or coin B.

I flip my coin once: it lands on heads. What is your posterior belief as to whether I have coin A or B

Level 2: I flip the coin N times, and each time it lands on heads. What is the probability - in terms of N that I have each coin (they should add to 1)

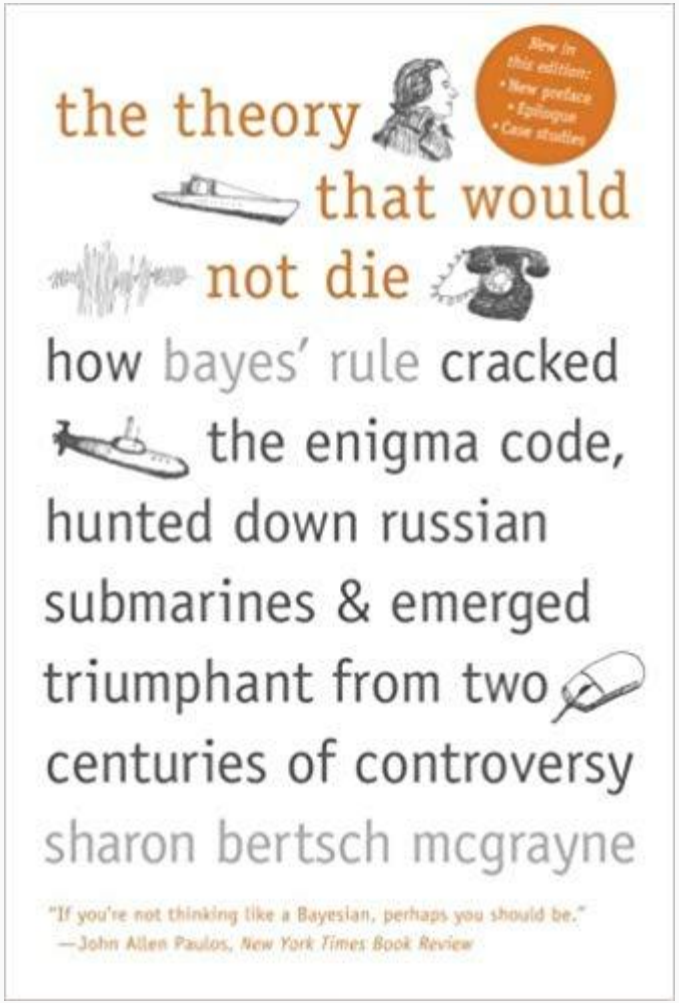
Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Main Source

The Theory That Would Not Die

By: Sharon Bertsch Mcgrayne



Frequentist Statistics Dominates

Compare 2 Hypotheses with Significance Testing:

- If the null hypothesis is true, what's the probability that I get a value at least this extreme? (p-value)
- Not reliant on Priors, answers come directly from the data
- Includes Confidence Bounds
- Common set of standards and "tests" developed
- "Subjective" Bayesian methods considered less scientific.
- Focus on repeatable experiments

Ronald Fisher
1890 - 1960
Significance Testing, Experimental Design

Bayesian Secrets

Codebreaking in Bletchley Park, UK during World War 2.

German Enigma Machine - they knew how it worked, but there were an astronomical amount of settings.

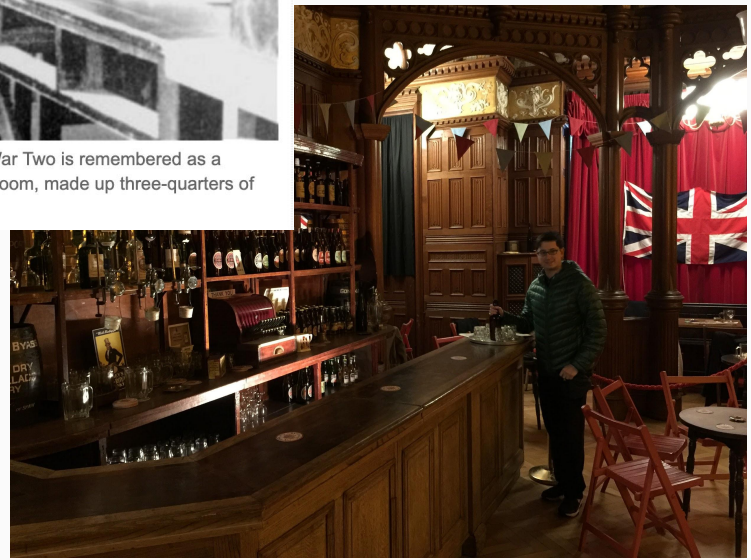
Solution is narrowed down through hunches and raw calculation in a Bayesian manner (what are phrases we expect in the unencrypted German codes)

Highly confidential for many years, making it difficult to reach academia as a good case study.

The extraordinary female codebreakers of Bletchley Park

As a new book reveals the unknown stories of the Bletchleyettes, Sarah Rainey speaks to two survivors about love, loss and a life of secrecy at The Park

The infamous codebreaking HQ at Bletchley Park during World War Two is remembered as a mostly male endeavour. But women, pictured here in the typists' room, made up three-quarters of the workforce Photo: Getty Images



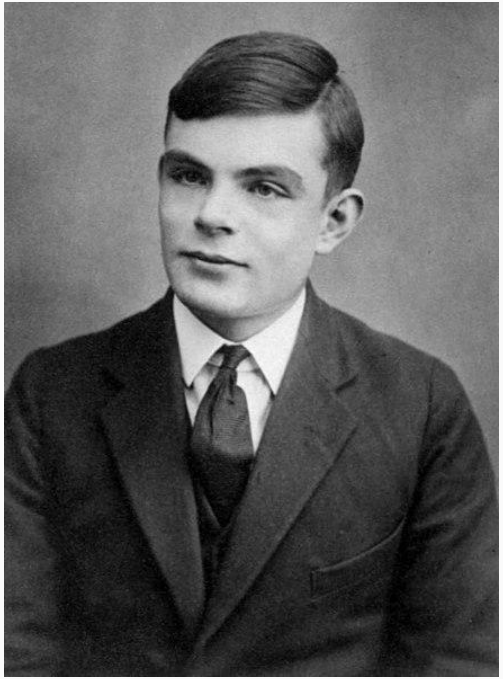
Bayesian Secrets

Codebreaking in Bletchley Park, UK during World War 2.

German Enigma Machine - they knew how it worked, but there were an astronomical amount of settings.

Solution is narrowed down through hunches and raw calculation in a Bayesian manner (what are phrases we expect in the unencrypted German codes)

Highly confidential for many years, making it difficult to reach academia as a good case study.



Alan Turing
1912 - 1954
Laid Foundation of Modern Theoretical AI and Computer Science

Picked Up By Industry - Life Insurance 1940s - 1960s

Arthur Bailey (actuary) noticed that the life insurance practices were unsound according to academic standards.

Ended up showing that they were actually using Bayesian methods (priors)

Also - estimating the probability of an event that has never occurred before (in this case a commercial airline collision)

Group Thought Project (5 minutes)

In groups of 5, brainstorm some strategies you could use to estimate the probability of an event that has never before occurred.

In the book, the case is one of two commercial airlines colliding (they correctly predicted it would happen).

You can think of other - theoretical or real life examples - when listing strategies.

Group discussion.

1966: US Navy Palomares Bayesian Search

US Navy lost a hydrogen bomb off the coast of Spain in an accident.

Approach: **Bayesian Search**

Basic Idea: Came up with several hypothesis on what happened to the bomb. Mapped to probabilities on a grid. Update those probabilities when a grid tile is searched - always search the most likely area.

Result: Bomb was found quickly

Bayesian vs Frequentist Wars

Despite applications, academia was slow to pick up on Bayesian methods.
- Didn't want to revise existing standards
- Non-standard Bayesian Posteriors were difficult to sample from in the 20th century.
- Fundamental disagreements on interpretation of Probabilities

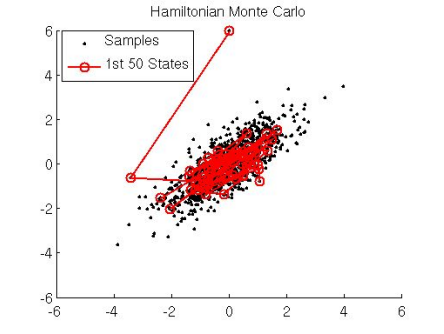
Bayesian vs Frequentist - An Uneasy Peace?

In the 21st century, there's been a shift to discovering what works on a practical level, and Bayesianism gets a seat at the table.

Often, the two methods come up with similar answers.

Frequentist methods still taught in schools and in entrenched industry standards.

As we will soon learn, machine learning methods (like gradient descent, MCMC) has made Bayesian methods much more practical in the 21st century

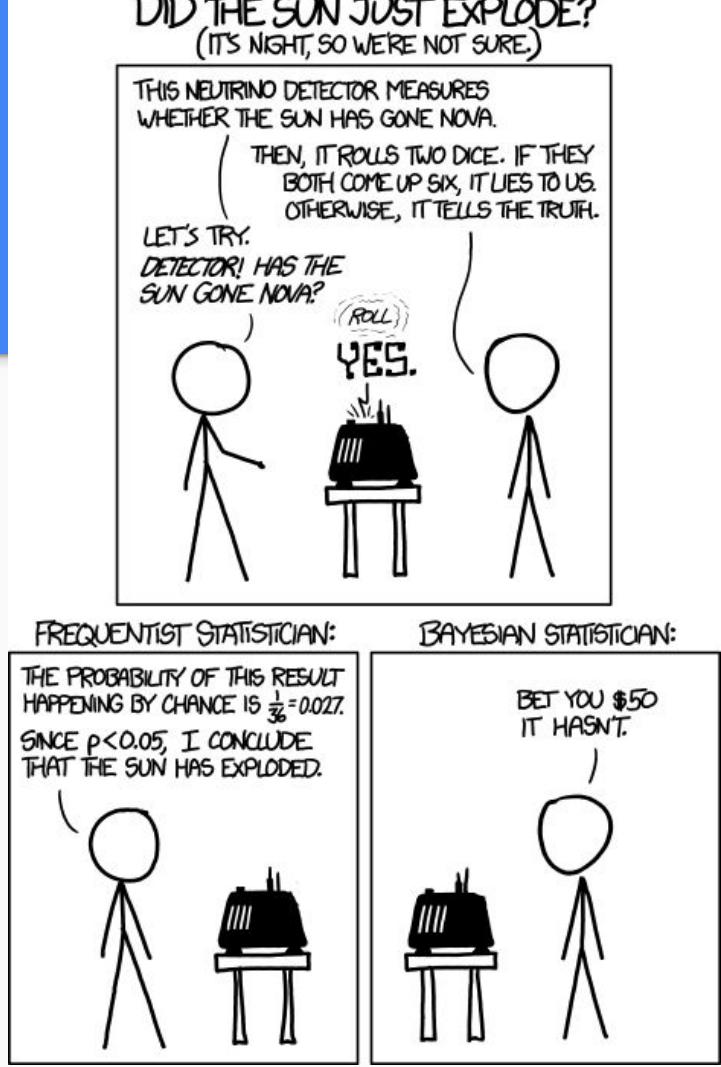


Some Lingering Issues with Frequentist Statistics

Great methodology for repeated experiments & time tested heuristics…

Beware the P!
(http://whatdoesthequantsay.com/2014/04/26/warning-beware-the-p)
P-value significance testing misapplies.
What is P-Hacking?

Episode 22 of The Local Maximum: Bayes Rulez, Death to
P-Hacking



Reproducibility Crisis in Science

FOOTNOTES TO PLATO JUNE 28, 2018

Thomas Bayes

Thomas Bayes and the crisis in science

David Papineau argues that it is crucial for scientists to start heeding the lessons of Thomas Bayes

DAVID PAPINEAU

JULY 5, 2018

Beware those scientific studies—most are wrong, researcher warns

by Ivan Couronne

Seafood is one of many food types that have been linked with lower cancer risks

A few years ago, two researchers took the 50 most-used ingredients in a cook book and studied how many had been linked with a cancer risk or benefit, based on a variety of studies published in scientific journals.

What does Bayesian Inference have going for it? Machine Learning Applications

Posterior Distributions can be turned into Loss/Energy functions, and then you can:
1) Use hill climbing, or gradient descent to find a Maximum Likelihood Estimate (this is the most likely hypothesis given the data)
2) Using Markov Chain Monte Carlo methods, you can sample from the posterior distribution

- If the hypothesis is a real number, this allows us to graph the posterior easily.

Conjugate Priors provided a great shortcut in the past

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Priors and Probabilities: A Dark Art?

Can we ever agree on a prior?

There are 2 types of priors:
- **Informative**: Based on Past data, or beliefs
- **Uninformative**: I want to start as “open minded” as possible

In either case, there can be disagreements on what to do...

Informative Priors

1) I'm thinking of a random number between zero and one. (Unless you know something about my psychology, you should aim for uninformative)
2) I just got a coin in change from the grocery store. What's the probability of it landing on heads?

   In case #2, we should have a belief that the coin is probably fair, or close to fair, given our previous experience with coins in our society. Especially random ones that come as change. However, it's not impossible for the coin to be weighted (as it lands) or double sided..

   Different people will quantify this experience differently.

Informative Priors

Informative Priors by analogy and data:

I want to find a prior distribution over my hypothesis space H.

I've run similar experiments with the same hypothesis space in the past.

I can therefore estimate the prior over the hypothesis space H based on past (similar but not identical) experiments.

This is also a Bayesian problem, and actually requires a HYPERPRIOR over the distribution of the distribution of priors. BUT there's usually more room for error in a HYPERPRIOR than a prior.

Example on board: Single Card, Deck of Cards, Multiple Decks of Cards

Indifference Principle

Probably the most important basic rule-of-thumb.

If there are several outcomes, and there is no information to distinguish these outcomes from one another (they are symmetric) - then these outcomes should all be assigned the same probability.

Examples:
Coin flip (½)
Dice roll (⅙)
Roulette Play (1/38)

Uninformative Priors

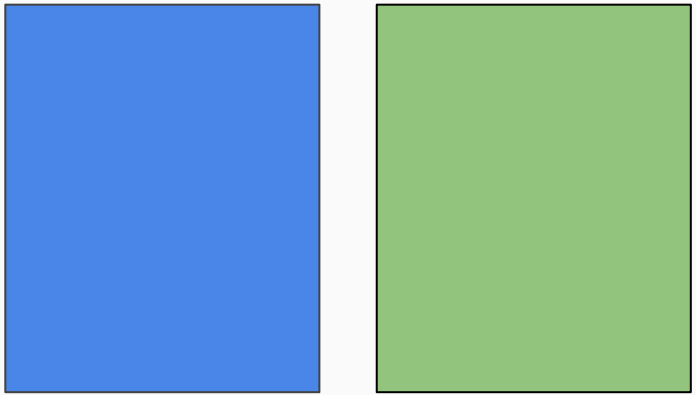
Indifference Principle

If you have a finite, discrete space of Hypotheses, and no reason to favor one over the other, you should weight them all equally.

For example: my opponent rolled a pair of dice: die A and die B. What's my prior over the result of die A?

New evidence - the sum is greater than 6. What's my posterior over the result of die A?

Now it breaks down



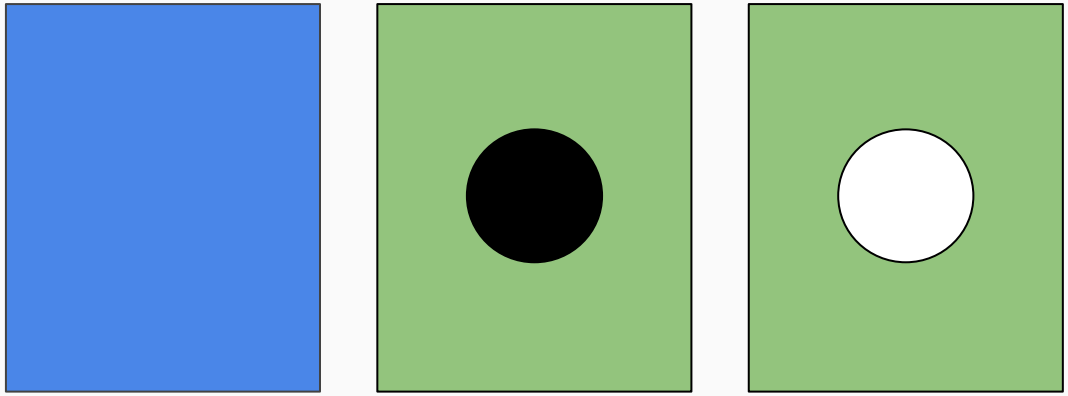
We know there are 2 types of cards: blue and green

You don’t know what the deck looks like other than that these are the two allowable cards, but you’re betting partner has the same information that you do.

In the next draw from the deck, your prior on whether you get blue or green should be 50%.

Your prior on the proportion of cards that are blue/green (which is still discrete because it’s a finite deck) should be symmetrical.

Now it breaks down



Now there are 3 types of cards: still blue and green, but some green cards have a black circle, and some have a white circle.

Do you assign:
50% blue, 25% green-black, 25% green-white

33% all around?

There's actually no agreed answer to this - the indifference principle no longer helps

Another Breakdown of Indifference

Suppose I went to a library and picked 1000 books out of the fiction section at random. Then I looked at all the whole numbers mentioned in those books (actually in the text, not in the table of contents or anything) and recorded the first digit of each of those numbers.

There are 9 possible digits (0 can’t be a leading digit): 1, 2, 3, 4, 5, 6, 7, 8, 9

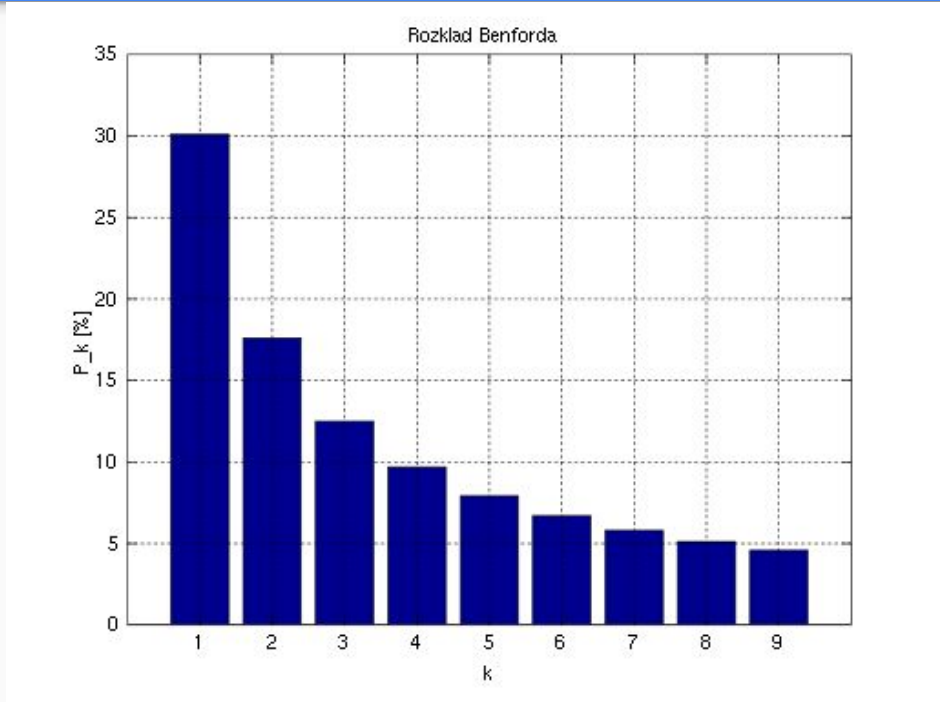
The indifference principle suggests that we should see them in equal amounts (1/9)

Benford's Law

But we don't!

Actually smaller numbers should be favored over larger numbers. The leading digit of 1 dominates if we plot numbers on a log scale.

Benford's law shows the distribution we actually get

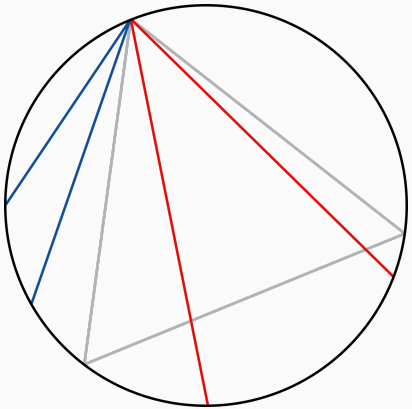


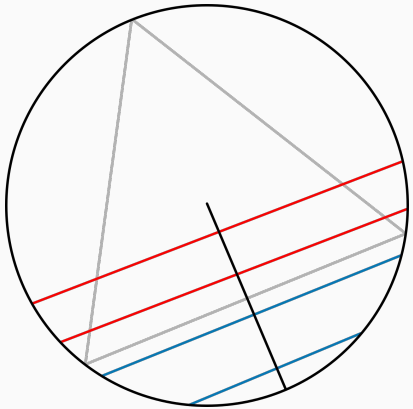
It gets worse

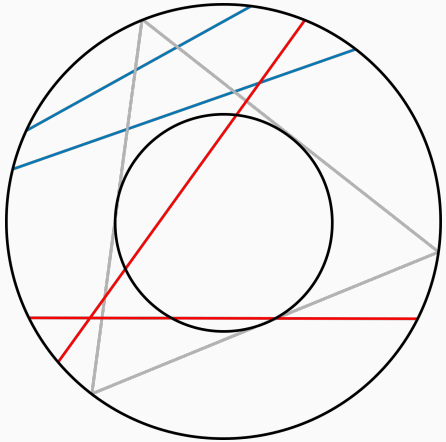
Continuous Spaces have lots of Gotchyas

My Hypothesis space is a Chord on a Circle. What's the expected Length?

Leads to Bertrand's Paradox - there are at least 3 reasonable ways to do it







Continuous Spaces have lots of Gotchyas

Random Probability, Random number Between Zero and 1.

A Uniform Distribution Sounds Reasonable…

Until you look at what happens in multiple dimensions! The prior starts to become very informative - very bias against corners! (More when we look at the Dirichlet Distribution)

Continuous Spaces have lots of Gotchyas

My Hypothesis Space is the set of positive real numbers.

Can you have such a uniform distribution?
- Not integrable, but actually you can still run this through Bayes Rule (we'll see this when we do the gamma-poisson distribution)
- Better to choose a distribution that discounts enormous values - a very light exponential decay
- Alternatively, if the positive number represents the Odds form of a probability p/(1-p) then you can just do a uniform distribution over p (Exercise: derive & graph this)

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Common Hypothesis Spaces

Go over on white board:
Probability Space (2 dimensional categorical)
Categorical Space (discrete values)
Odds
Log Odds
Minus Log Odds

Ask yourself: am I in Log Space?

Multidimentional Vector
Finite Permutation Space (n!)

Normal Distribution as Energy Function

Log Odds, Energy Function (remove normalizing constant)

What does the Gaussian/Normal distribution look like when transformed into an Energy Function?

Proof that the Maximum Likelihood estimate of the Gaussian is equal to the mean.

Mention Yann Lecun's Paper on Energy Functions.. Log Likelihood is only one of them.
http://yann.lecun.com/exdb/publis/pdf/lecun-06.pdf

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Conjugate Priors are the greatest thing!

What is a conjugate Prior?

In general, Bayesian inference models can produce very complex posterior distributions.

BUT - in some cases the the posterior distribution is the same type of distribution (but with different parameters) as the prior distribution

Allows Bayesian Inference without much computation, also intuitive

Example: Beta Binomial (Whiteboard)

Thomas Bayes original problem: we have probability value p between 0 and 1.

Our prior over p is the uniform distribution between 0 and 1 (draw on board)

We draw N samples from p (either 0s or 1s) and then update our probability distribution. c_0 = count of 0s, c_1 = count of 1s. c_0 + c_1 = N

What's the Posterior? Why is it called Beta Binominal

Smoothing (whiteboard)

Compute the Expected Value of the Beta Binomial

Explain Smoothing, how the prior parameters are equivalent to smoothing parameters, prevent division by zero.

Example: Superuser Football @ Foursquare

Show slides and explain the use-case

- Combines the Beta-Binomial Smoothing + Log Space Idea

More Conjugate Prior Distributions

- Dirichlet-Multinomial (Separate Slides) - Good for ratios of counts, probability simplex.
- Gamma-Poisson Prior - good for rates of occurance of random events
- Normal with Known Variance

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Using the Posterior with MLE

MLE: Maximum likelihood estimate.

For example, suppose you have a machine learning model with M parameters. Your Hypothesis Space is $\mathbb{R}^N$

Your posterior is over the Hypothesis Space of possible parameters, but you want to select a SINGLE MODEL. Then, you probably want the MLE.

Using the Posterior with MLE

Take -log(p) of your PDF, remove any constant factors for simplicity.

Gradient Descent
Hill Climbing (when gradient isn't available)
In the other direction, Newton's Method (Second Order Methods)

If the function is convex - convex optimization algorithms can be even faster

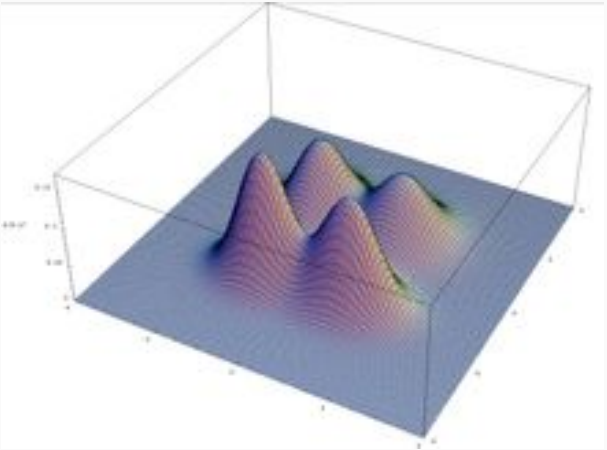
One point of complexity: what about multi-modal Distributions?

MLE with Multi-Modal Distributions

You can get stuck.. In a LOCAL MAXIMUM (or local minimum in minus log space)

Potential Solutions:

- Run it multiple times at different starting points to find many local maxima and search for the best
- Hill-Climbing: Increase jump distance when getting stuck



Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

The Substitution Cypher

We pick a permutation of 27 characters (26 letters + space), and encode a piece of text. Can you crack the code without knowing the key?

Yes! Using Bayesian Inference, Language Models, and Multi-Hill climbing

Local Maximum Episode 4 - Codebreaking, Bizarro Harry Potter, and the Proper Way to Gerrymander

Python Code:
https://github.com/maxsklar/LocalMaximum/tree/master/SubstitutionCipher

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Bayesian Approach to Machine Learning

Hypothesis Space = The Space of models, or model parameters

Priors: Regularization (usually so parameters don’t get too large)

Posteriors: Used to choose a model (or randomly draw models).

Example: Logistic Regression

Derive Binary Logistic Regression on the white board

Show that Gaussian Prior = Ridge Regression

Show that Laplace Prior = Lasso Regression

Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications

Hierarchical And Mixture Models

So far, the models have been simple.

Hierarchical models we've seen before - particularly in Dirichlet - a hierarchy of models, hyperparameters

Mixture models: combine several submodels

Examples:

Multinomial Mixture Model (Simple)

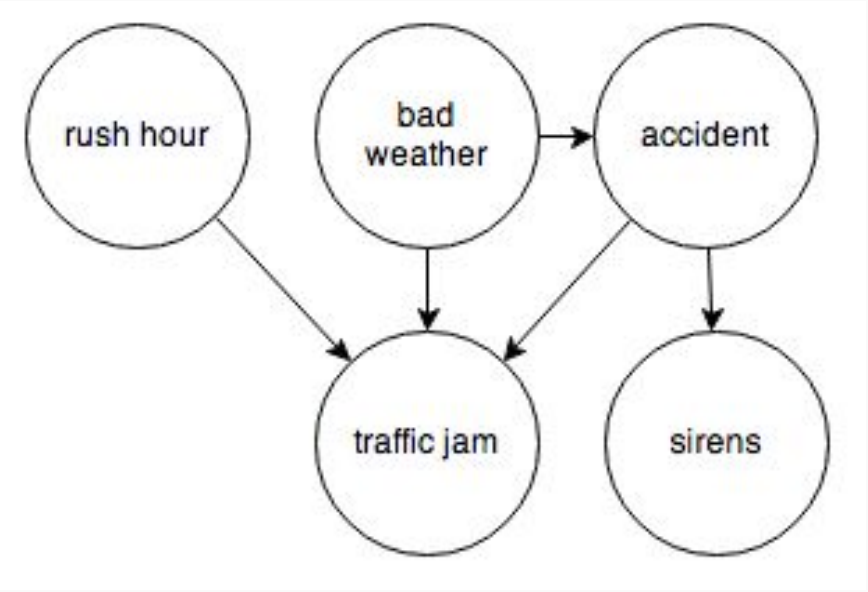
Latent Dirichlet Allocation

Gaussian Mixture Model

Bayesian Networks

Build the causality structure into your models.

Solve sub-problems



Course Roadmap

0) Introduction and Goals
1) Max's Bio and Background
2) Probability, Belief, and the Truth
3) Introduction to Subjective Probability and Bayes Rule
4) History of Bayesian and Frequentist Statistics
5) More About Priors
6) Ratios and Log Space
7) Conjugate Priors
8) Complicated Posteriors: The MLE Approach
9) Examples: The Substitution Cypher
10) Bayesian Interpretation of Linear and Logistic Regression
12) Hierarchical and Mixture Models
13) Complicated Posteriors: The MCMC Approach
14) The Problem of Causality and Attribution
15) Intro to Probabilistic Programming: PyMC3
16) Projects and Applications