Machine Learning Foundations
Probability & Information Theory

Quantifying Uncertainty and Building A.I. Systems that Reason Well Despite It

Jon Krohn, Ph.D.

jonkrohn.com/talks

github.com/jonkrohn/ML-foundations

Machine Learning Foundations
Probability & Information Theory

Slides: jonkrohn.com/talks

Code: github.com/jonkrohn/ML-foundations

Stay in Touch:

jonkrohn.com to sign up for email newsletter

linkedin.com/in/jonkrohn

jonkrohn.com/youtube

twitter.com/JonKrohnLearns

The Pomodoro Technique

Rounds of:

- 25 minutes of work
- with 5 minute breaks

Questions best handled at breaks, so save questions until then.

When people ask questions that have already been answered, do me a favor and let them know, politely providing response if appropriate.

Except during breaks, I recommend attending to this lecture only as topics are not discrete: Later material builds on earlier material.

POLL

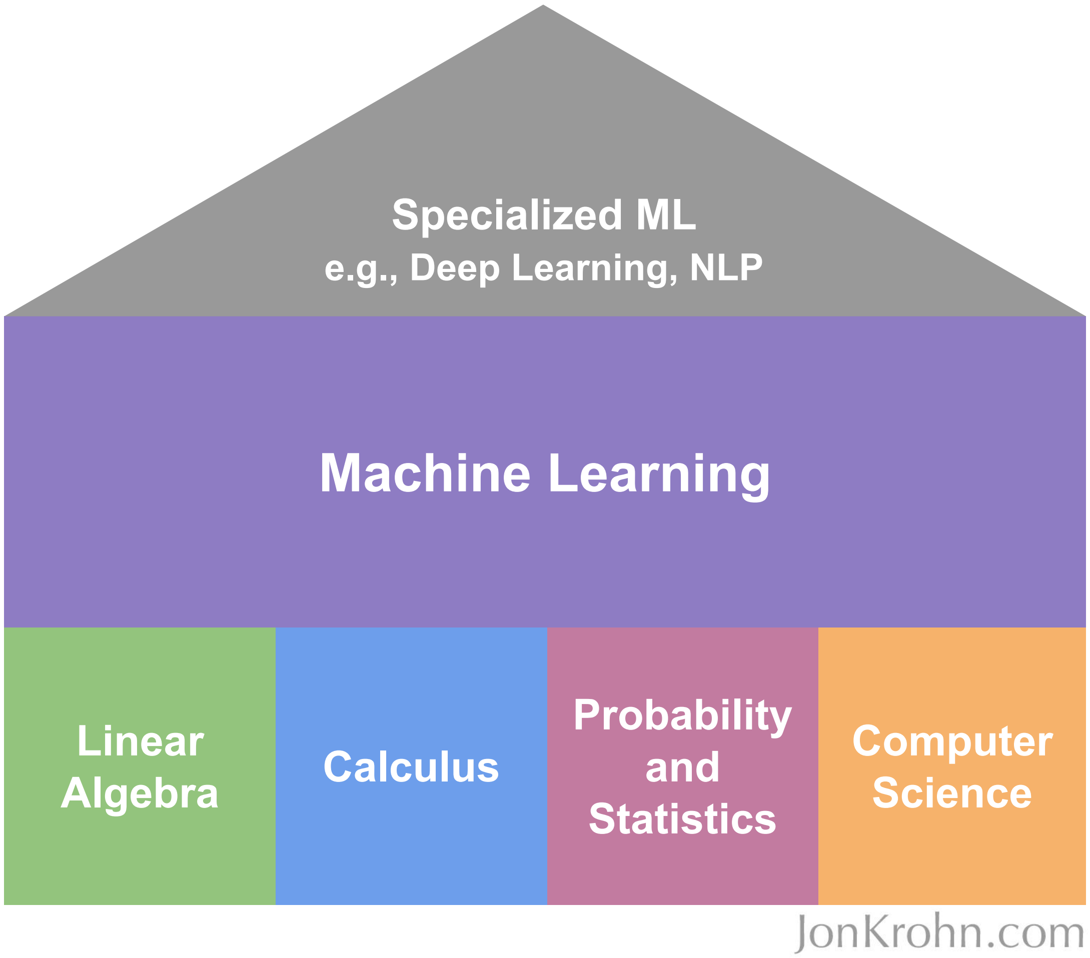
What is your level of familiarity with Probability Theory?

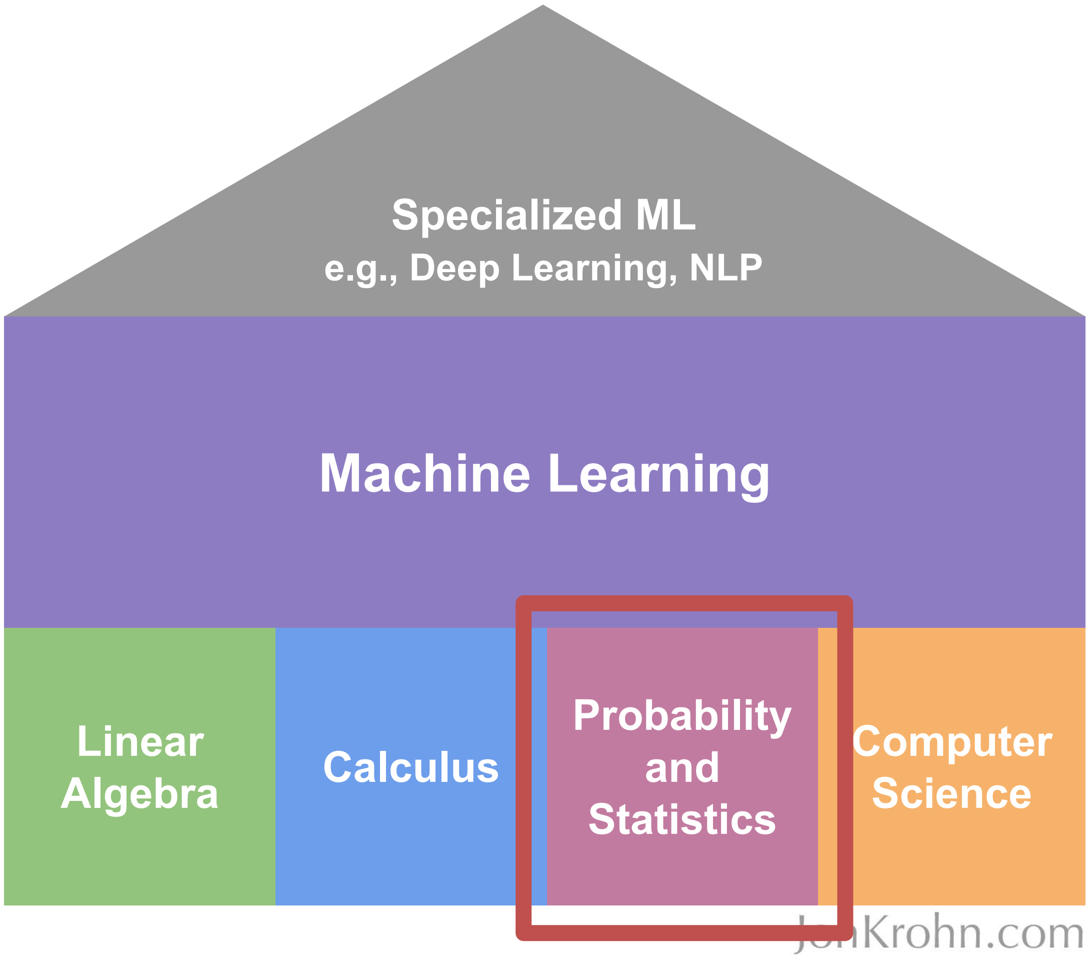
- Little to no exposure
- Some understanding of the theory
- Deep understanding of the theory
- Deep understanding of the theory and experience applying probability theory or statistical models with code

POLL

What is your level of familiarity with Machine Learning?

- Little to no exposure, or exposure to theory only
- Experience applying machine learning with code
- Experience applying machine learning with code and some understanding of the underlying theory
- Experience applying machine learning with code and strong understanding of the underlying theory


Specialized ML
e.g., Deep Learning, NLP
Machine Learning
Linear Algebra
Calculus
Probability and Statistics
Computer Science
JonKrohn.com



Specialized ML
e.g., Deep Learning, NLP
Machine Learning
Linear Algebra
Calculus
Probability and Statistics
Computer Science
JonKrohn.com

ML Foundations Series

Probability & Information Theory **builds upon** and is **foundational for**:

1. Intro to Linear Algebra
2. Linear Algebra II: Matrix Operations
3. Calculus I: Limits & Derivatives
4. Calculus II: Partial Derivatives & Integrals
5. **Probability & Information Theory**
6. **Intro to Statistics**
7. Algorithms & Data Structures
8. **Optimization**

Probability & Information Theory

1. Intro to Probability
2. Distributions in Machine Learning
3. Information Theory

Probability & Information Theory

1. **Intro to Probability**
2. Distributions in Machine Learning
3. Information Theory

Segment 1: Intro to Probability

- What Probability Theory Is
- A Brief History: Frequentists vs Bayesians
- Applications of Probability to Machine Learning
- Random Variables
- Discrete vs Continuous Variables
- Probability Mass and Probability Density Functions
- Expected Value
- Measures of Central Tendency: Mean, Median, and Mode
- Quantiles: Quartiles, Deciles, and Percentiles
- The Box-and-Whisker Plot
- Measures of Dispersion: Variance, Standard Deviation, and Standard Error
- Measures of Relatedness: Covariance and Correlation
- Marginal and Conditional Probabilities
- Independence and Conditional Independence
- Bayes' Rule

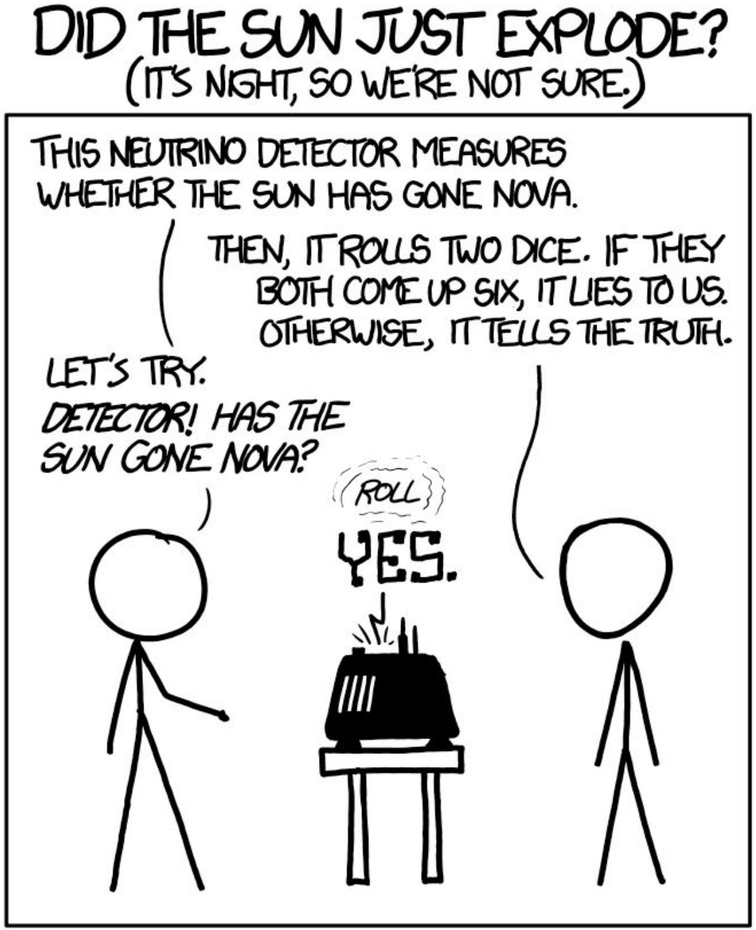
A Brief History of Probability

- Earliest known use: Arab mathematicians (8th-13th c.)
  - Largely related to cryptographic communications
  - **Al-Kindi** (9th c.): first known to make statistical inference
- Later further developed by Europeans to study games of chance
  - 16th c.: Italian polymath Gerolamo Cardano
  - 17th c.: Frenchmen Pierre de Fermat and Blaise Pascal
- Largely combinatorial up to this point in history
  - E.g., working with integers of count data
- Modern probability theory:
  - Mostly devised in 20th c. (e.g., Soviet Kolmogorov, Austrian von Mises)
  - Allows us to work with continuous, real (e.g., float) values
  - Underpins frequentist stats, Bayesian stats, and machine learning

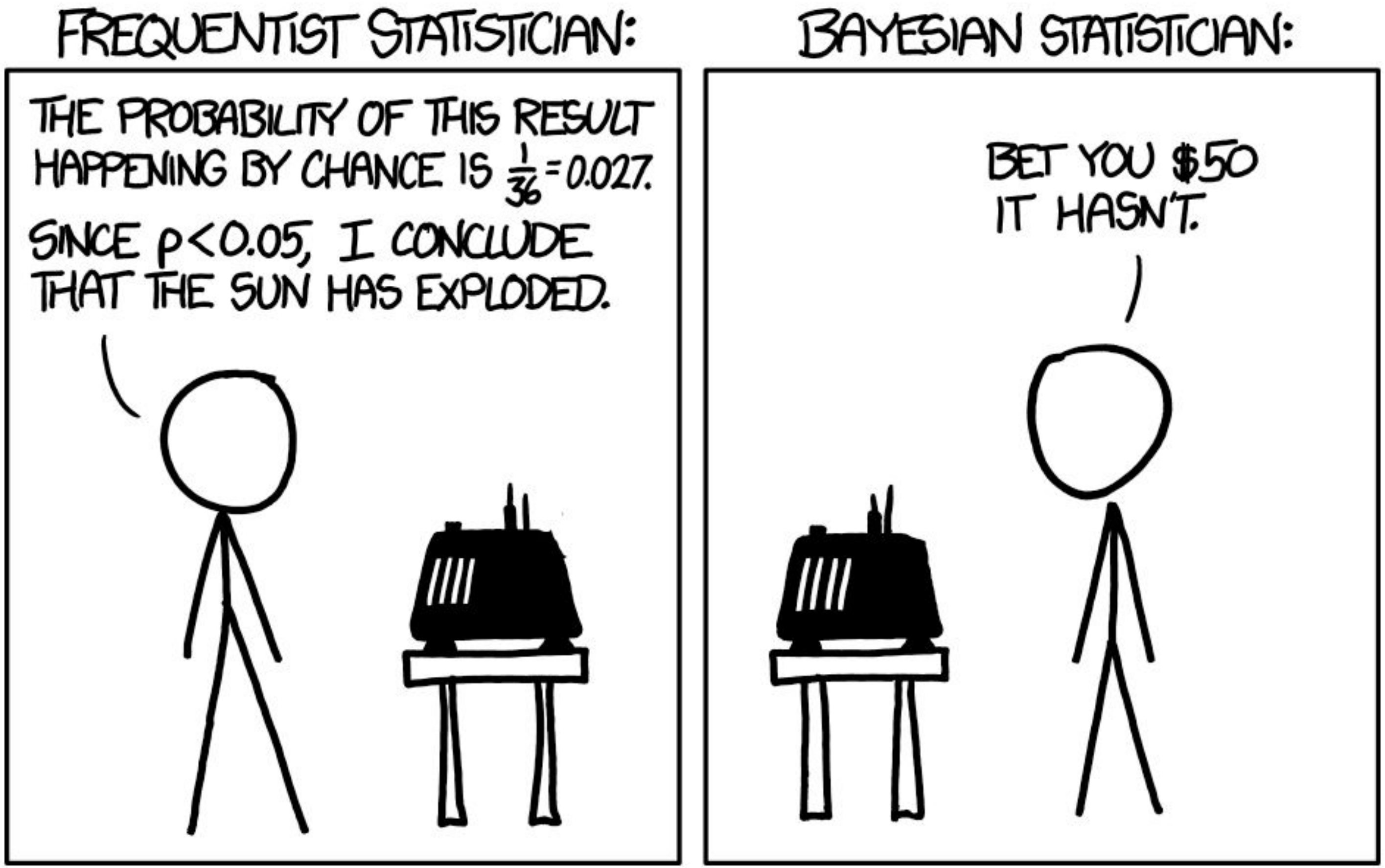
Image in public domain

What Probability Theory Is

- Mathematical study of processes that include uncertainty
- Probabilities expressed over range of 0 (will not happen) to 1 (will happen)
- Enables models of future non-deterministic events based on historical data
    - **Statistics** (*Intro to Stats*)
        - Quantifies confidence in inferences based on probabilistic events
        - Provides framework for supporting or rejecting hypotheses
    - **Machine learning** (entirety of *ML Foundations* series)
        - Modeling approach that scales to large, high-dimensional data
- Key concepts:
    - **Law of large numbers** (*Hands-on code demo*: `5-probability.ipynb`)
    - **Random variables** (*Segment 1*)
    - **Probability distributions** (*Segments 1 & 2*)
    - **Central limit theorem** (*Segment 2*)



imgs.xkcd.com/comics/frequentists_vs_bayesians.png



imgs.xkcd.com/comics/frequentists_vs_bayesians.png

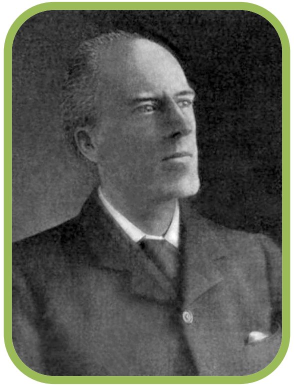
Bayesian Statistics

- Can incorporate prior knowledge from, e.g., experimental results, beliefs
- How lay people think about probabilities: “There’s 80% chance it’ll rain today.”
- English philosopher (and minister) **Thomas Bayes**
  - Devised particular case of “Bayes’ theorem” in 1763
  - (Notes published post-mortem by Richard Price)
- French polymath **Pierre-Simon Laplace**
  - In late 18th / early 19th c.
  - Expanded to probability and statistical problems
- Drawbacks:
  - Beliefs are icky to some
  - Generally computationally expensive

Image in public domain

Frequentist Statistics

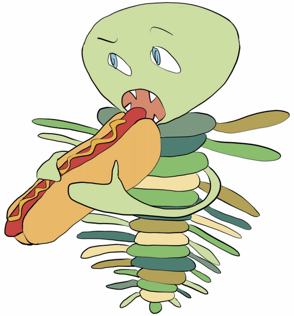
- Focus on "objective" probabilities
- "On 100 days exactly like today, it would rain on 80 of them."
- Arbitrary threshold of "less than 5% chance result occurs by chance"
- Discussed as early as 1837 by Siméon Denis Poisson
- Expanded in 19th c. by many (incl. J.S. Mill, John Venn, George Boole)
- (Sir) **R.A. Fisher** (declined Sir) **Karl Pearson** developed much of modern statistical techniques (*Intro to Statistics*) in 20th c.
- Only statistical approach taught to most in 20th c.
- Generally computationally inexpensive
- Drawbacks:
    - Not designed for large feature sets (inputs)
    - 5% threshold too high for large sample sizes
    - Prior probabilities ignored



Images in public domain

Applications of Probability to ML

- Bayesian stats has today become a type of ML used where:
    - Sample sizes tend to be not very large
    - Typically have evidence for priors (initial parameter values)
- Probability concepts ubiquitous in AI, incl. ML (focus of *Prob. & Info. Thy.*):
    - Uncertainty typically involved in mapping inputs to outputs
    - Output probabilities: "98% chance image is of a hot dog"
    - Some models are stochastic (non-deterministic)
    - With stats, can confidently compare model performances



Applications of Probability to ML

Why can't most A.I. systems be certain and deterministic?

1. The process being modeled is itself stochastic, e.g.:
   - Games of chance
   - Human behavior in general
   - Stock market in particular
2. Model inputs are not comprehensive, e.g.:
   - Car crash inevitable around curve
   - Candidate for role has offer from another employer
3. Model is incomplete, e.g.:
   - Computational complexity of perfect solution is astronomical
   - Modeling approach for solving problem perfectly is unknown
   - Building perfect model is unreasonably expensive

Generally, all three of the above are true.

Random Variable

- Variable whose value is determined by a process that has uncertainty
- Notation:
  - Scalar: plain type, e.g., h for height
  - In italics if a particular value (a.k.a., a particular state), e.g.:
    - $h_1 = 172$

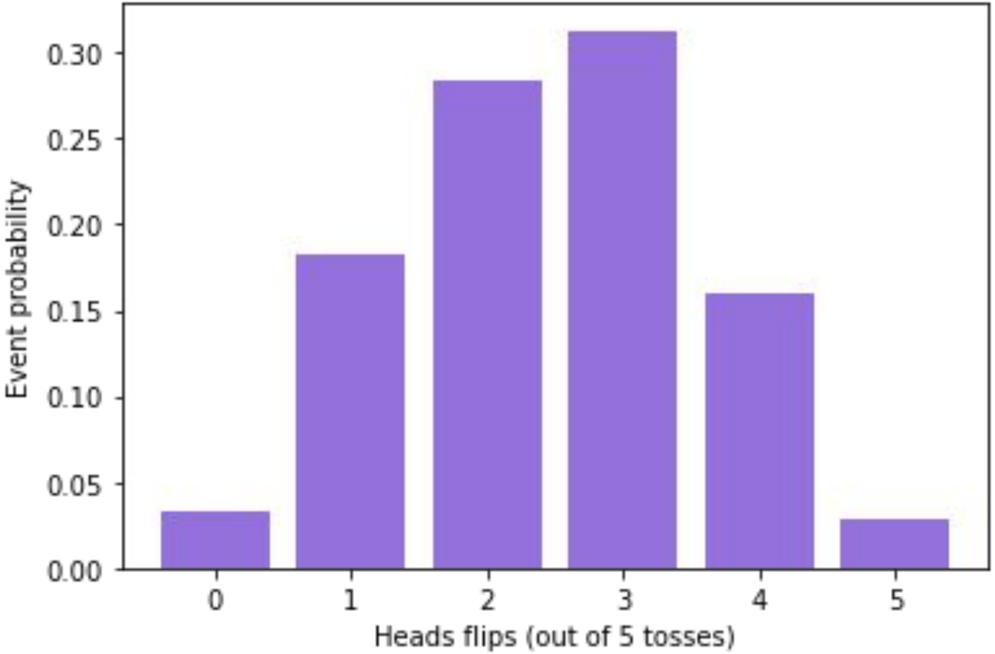
Random Variable

Two varieties:

1. **Discrete**:
   - Countable number of states (can be finite or infinite)
   - Could be category (e.g., “heads”, “tails”)
   - Could be integer (e.g., result of rolling a die)
2. **Continuous**:
   - Real value (represented by float in computing)
   - E.g.: height, speed, temperature

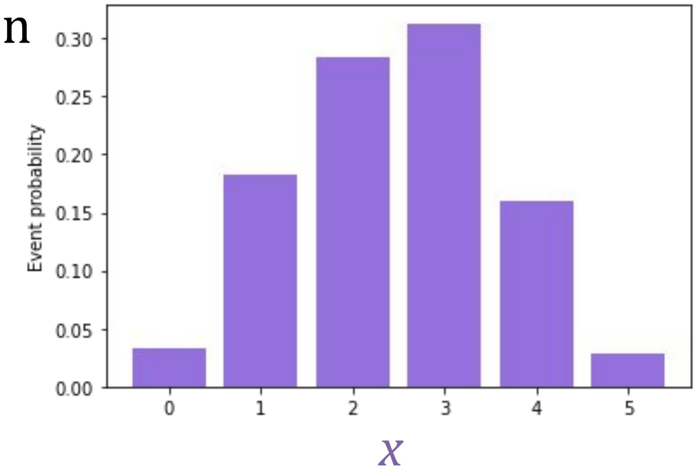
Probability Distribution

Describes likelihood of random variable taking on its possible values:



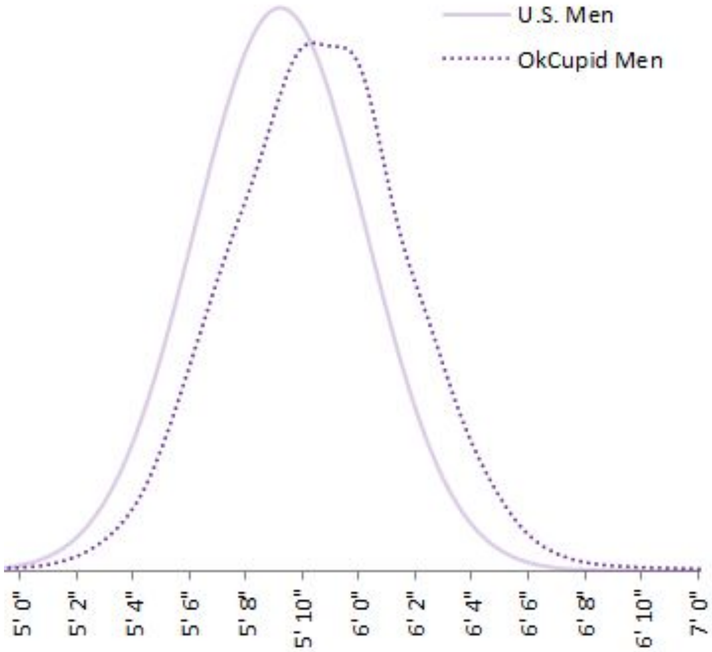
Probability Mass Function (PMF)

- Describes *mass* of probability distribution of a <u>discrete</u> random variable
- Notation:
  - Capitalized, italicized P
  - Distinguish PMFs by random variable: $P(\mathrm{x})$, $P(\mathrm{y})$, etc.
  - Probability of a particular state x: $P(x)$ or $P(\mathrm{x} = x)$ or $x \sim P(\mathrm{x})$
- E.g.: $P(\mathrm{x} = x) = P(\mathrm{x} = 2) = 0.3125$
  - Can be derived from theory or observation
- Three essential properties of $P(\mathrm{x})$:
  - Every possible value of x within domain
  - Each $P(x)$ can only range from 0 to 1
  - Sum of all $P(x)$ must equal 1
    - (This is called **normalization**)

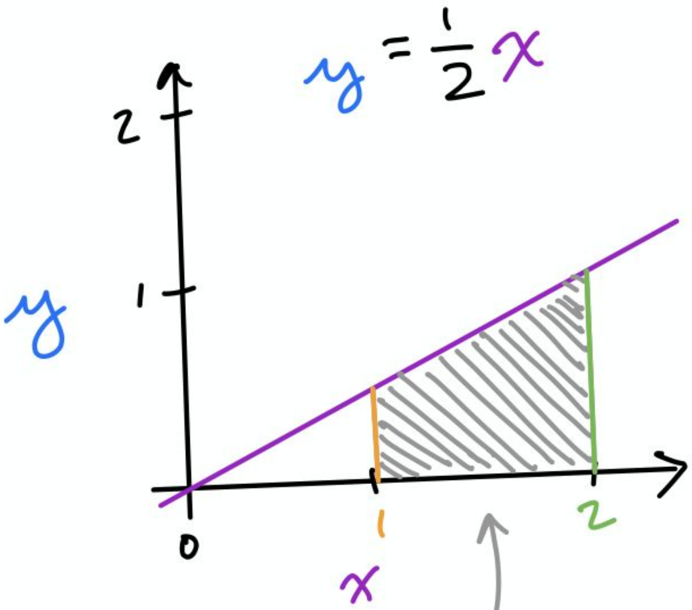


Probability Density Function (PDF)

- PMF analogue for continuous random variable
- Notation:
  - Lower-case, italicized p
  - Like PMFs, distinguish by $p(\mathrm{x})$, $p(\mathrm{y})$, etc.
- Three essential properties of $p(\mathrm{x})$:
  - Like PMFs, every possible value in domain
  - Every $p(x)$ must be ≥ 0
  - $\int p(x)dx = 1$
- Probability that x is between points a and b:
  - $\int_{[a,\, b]} p(x)dx$



theblog.okcupid.com/the-big-lies-people-tell-in-online-dating-a9e3990d6ae2



$$\int \frac{1}{2}x\,dx = \frac{1}{2}\left(\frac{x^2}{2}\right) + C = \frac{x^2}{4} + C$$

$$\text{At } x = 1,\ \frac{x^2}{4} + C = \frac{1^2}{4} + C = \frac{1}{4} + C$$

$$\text{At } x = 2,\ \frac{x^2}{4} + C = \frac{2^2}{4} + C = 1 + C$$

$$\int_1^2 \frac{1}{2}x\,dx = (1 + C) - \left(\frac{1}{4} + C\right) = \frac{3}{4}$$

Exercises

Would a PMF or PDF be better-suited to describing:

1. Residential property values?
2. Likelihood of each NFL team winning the Super Bowl?
3. Duration of commute from Greenwich, CT to midtown Manhattan?

Solutions

1. Continuous prices: PDF
2. Discrete teams: PMF
3. Continuous durations: PDF

Expected Value

The long-term average of some random variable x.

If x is discrete: $$\mathbb{E} = \sum_{x} x P(x)$$

If x is continuous: $$\mathbb{E} = \int x p(x) dx$$

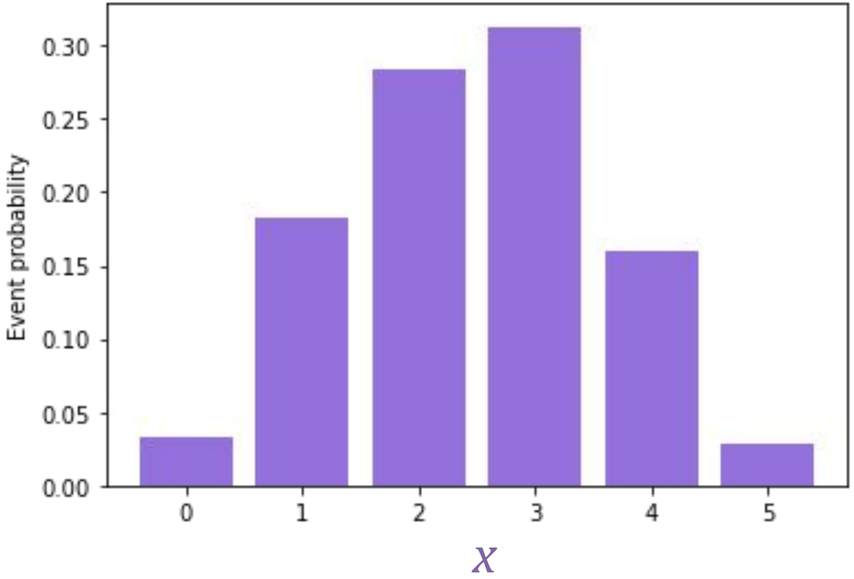
Expected Value

$$\mathbb{E} = \sum_x x P(x)$$

$P(0) = P(5) = 1/32 \cong 0.031$

$P(1) = P(4) = 5/32 \cong 0.16$

$P(2) = P(3) = 10/32 \cong 0.31$



$$\mathbb{E} = (1/32)0 + (5/32)1 + (10/32)2 + (10/32)3 + (5/32)4 + (1/32)5 = 2.5$$

Hands-on code demo

Joint Probability Distribution

- Probability distributions can represent the probability of multiple random variables simultaneously
- Probability both $\mathrm{x} = x$ and $\mathrm{y} = y$ is: $P(\mathrm{x} = x, \mathrm{y} = y)$
- E.g.:
  - P(flip 1 = *heads*, flip 2 = *heads*) = 0.25
  - P(card value = *ace*, card color = *red*) = 2/52 = 1/26 ≅ 0.038
  - p(height = *180-190cm*, weight = *20-30kg*) = 0

Marginal Probability

"Sum Rule" for discrete variables:

$$\forall x \in \mathrm{x},\ P(\mathrm{x}=x) = \sum_{y} P(\mathrm{x}=x, \mathrm{y}=y)$$

		y (★ rating)				
		1	2	3	4	margin
x	hot dog	2	5	25	3	35/76 = .46
	burger	7	6	9	14	36/76 = .47
	pizza	4	1	0	0	5/76 = .07

Marginal Probability

"Sum Rule" for discrete variables:

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

Integrate for continuous variables:

$$p(x) = \int p(x, y)\, dy$$

Conditional Probability

Probability of an outcome given another outcome occurred:

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

$P(\mathrm{x} = x) > 0$ because otherwise nothing to be conditional of.

Examples:

- P(flip 1 = *heads*, flip 2 = *heads*) = 0.25
- P(flip 2 = *heads* | flip 1 = *heads*) = 0.25/0.5 = 0.5
- Without replacement:
    - P(card 1 = *ace*, card 2 = *ace*) = 4/52 x 3/51 = 12/2652 = 1/221
    - P(card 2 = *ace* | card 1 = *ace*) = (1/221) / (4/52) ≅ 0.059

Exercises

		y (★ rating)			
		1	2	3	4
	hot dog	2	5	25	3
x	burger	7	6	9	14
	pizza	4	1	0	0

1. Calculate the marginal probability of y in the fast food example.
2. What is P(card type = *face*, card color = *black*) ?
3. Without replacement, what is P(card 2 = *face* | card 1 = *face*) ?

Solutions

1. $P(\text{1 star}) = 13/76 \cong 0.17$
   $P(\text{2 star}) = 12/76 \cong 0.16$
   $P(\text{3 star}) = 34/76 \cong 0.45$
   $P(\text{4 star}) = 17/76 \cong 0.22$

2. $P(\text{Jack, queen, or king of spades or clubs})$: $6/52 \cong 0.115$

3:

$P(\text{card 1} = \textit{face}, \text{card 2} = \textit{face}) = (12/52)(11/51) = 132/2652 = 11/221$

$P(\text{card 2} = \textit{face} \mid \text{card 1} = \textit{face}) = (11/221) / (12/52) \cong 0.216$

Chain Rule of Probabilities

We already know: $$P(y \mid x) = \frac{P(y, x)}{P(x)}$$

$$P(y, x) = \frac{P(y, x)}{P(x)} P(x) = P(y \mid x) P(x)$$

$$\frac{P(x)}{P(x)} = 1$$

Chains can be longer, e.g.:

$$P(z, y, x) = P(z \mid y, x) P(y \mid x) P(x)$$

Independent Random Variables

$$\mathrm{x} \perp \mathrm{y}$$

"in the set of"

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x}=x, \mathrm{y}=y) = p(\mathrm{x}=x)p(\mathrm{y}=y)$$

"for each"

E.g.:

- Probability of throwing heads and drawing an ace
- Probability of throwing heads on two consecutive tosses

Conditional Independence

$$\mathrm{x} \perp \mathrm{y} \mid \mathrm{z}$$

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},$$
$$p(\mathrm{x}=x, \mathrm{y}=y \mid \mathrm{z}=z) = p(\mathrm{x}=x \mid \mathrm{z}=z)\, p(\mathrm{y}=y \mid \mathrm{z}=z)$$

E.g.:

- Probability of throwing heads on two consecutive tosses, if two possible coins could be used: regular or two-headed
- At Olympics: probability of wrestler winning gold and weightlifter winning gold, if both come from country with doping scandal

Probability & Information Theory

1. Intro to Probability
2. **Distributions in Machine Learning**
3. Information Theory

Segment 2: Distributions in ML

- Uniform
- Gaussian: Normal and Standard Normal
- The Central Limit Theorem
- Log-Normal
- Exponential and Laplace
- Binomial and Multinomial
- Poisson
- Mixture Distributions
- Preprocessing Data for Model Input

Hands-on code demo

Preprocessing Data for Model Input

- Most popular statistical models are "parametric", meaning they assume normally-distributed inputs:
  - **Box-Cox** transformation adjusts toward normal
- Standard normal is ideal in ML:
  - Subtract mean (adjusts μ to 0)
  - Divide by standard deviation (adjusts σ to 1)
  - (In neural network architecture, we can pass inputs through batch normalization layer)
- Encode binary variables as `0` or `1`

Exercises

1. Which distribution is best-suited to representing the weight of year-old babies:
   a. Gaussian
   b. Log-normal
   c. Poisson
2. ...the number of puppies at doggy day care?
   a. Gaussian
   b. Multinomial
   c. Poisson
3. ...the height of adults?
   a. Gaussian
   b. Multinomial
   c. Mixture

Solutions

1. a
2. c
3. c

Probability & Information Theory

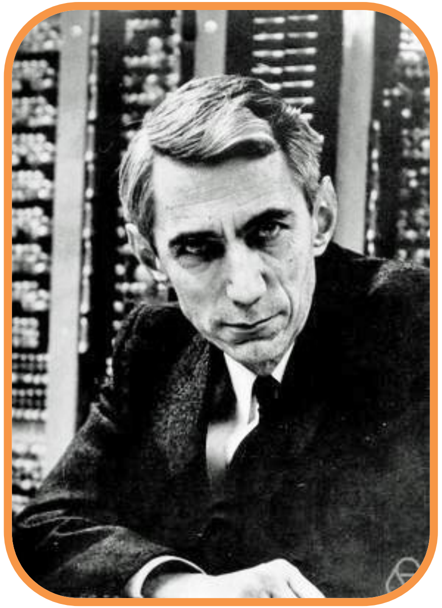
1. Intro to Probability
2. Distributions in Machine Learning
3. **Information Theory**

Segment 3: Information Theory

- What Information Theory Is
- Self-Information
- Nats, Bits and Shannons
- Shannon and Differential Entropy
- Kullback-Leibler Divergence
- Cross-Entropy

What Information Theory Is

- Field of applied mathematics
- While prob. thy. facilitates uncertain statements and reasoning despite uncertainty, info. thy. quantifies uncertainty in a signal, e.g., a distribution
- American engineer **Claude Shannon** (1916-2001)
  - Proposed the field ("father")
  - Developed many early theories, papers, books
- Critical to fields of:
  - Electrical engineering
  - Computer science
- Applications across many fields, incl.:
  - Biology
  - Physics
  - Machine learning



commons.wikimedia.org/wiki/File:ClaudeShannon_MFO3807.jpg

Self-Information

The essential concept of information theory is:

- Likelier events have less information content than rarer ones
- E.g.: message that sun rose this morning has no informational value

The associated equation, for **self-information**, is: $I(x) = -\log P(x)$

- Quantifying informational content:
  - If event is guaranteed (i.e., $P(x)=1$), $I(x)=0$
  - Less likely an event, the greater $I(x)$
  - Independent events are additive:
    - If one head flip has $I(x)$, two heads flips has $2I(x)$

Hands-on code demo

Next Subject: *Intro to Statistics*

Apply probability theory to:

- Quantify differences between distributions
- Quantify relatedness of distributions
- Confidently reject or approve hypotheses (e.g., select an ML model) despite uncertainty

POLL *with Multiple Answers Possible*

What other topics interest you most?

- Linear Algebra
- Calculus
- More Probability Theory
- More Information Theory
- Introductory Stats (Frequentist)
- Bayesian Stats
- Computer Science (e.g., algorithms, data structures)
- Machine Learning Basics
- Advanced Machine Learning, incl. Deep Learning
- Something Else

Stay in Touch

jonkrohn.com to sign up for email newsletter

linkedin.com/in/jonkrohn

youtube.com/c/JonKrohnLearns

twitter.com/JonKrohnLearns

NEBULA

PLACEHOLDER
FOR:

5-Minute Timer

PLACEHOLDER
FOR:

10-Minute Timer

PLACEHOLDER FOR:

15-Minute Timer