CS 540 Introduction to Artificial Intelligence
Probability

Yingyu Liang
University of Wisconsin-Madison
Sept 14, 2021

Based on slides by Fred Sala

Probability: What is it good for?

- Language to express **uncertainty**

Probability is the math language to model uncertainty.

In AI/ML Context

- **Quantify predictions**

$[p(\text{lion}), p(\text{tiger})] = [0.98, 0.02]$

$[p(\text{lion}), p(\text{tiger})] = [0.01, 0.99]$

$[p(\text{lion}), p(\text{tiger})] = [0.43, 0.57]$

In AI/ML, probability theory is particularly useful as uncertainty is ubiquitous in this context, e.g., uncertainty in predictions; data distributions

Model Data Generation

- Model complex distributions

StyleGAN2 (Kerras et al ’20)

Outline

- Basics: definitions, axioms, RVs, joint distributions
- Independence, conditional probability, chain rule
- Bayes' Rule and Inference

Basics: Outcomes & Events

- Outcomes: possible results of an **experiment**
- **Events:** subsets of outcomes we're interested in

Ex: $\Omega = \underbrace{\{1, 2, 3, 4, 5, 6\}}_{\text{outcomes}}$

$\mathcal{F} = \underbrace{\{\emptyset, \{1\}, \{2\}, \ldots, \{1, 2\}, \ldots, \Omega\}}_{\text{events}}$

Key: an event is just a subset of the outcome space.

Basics: Outcomes & Events

- Event space can be smaller:

$$\mathcal{F} = \underbrace{\{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}}_{\text{events}}$$

- Two components always in it!

$$\emptyset, \Omega$$

We may consider a family of special events, not necessary all the subsets of the outcome space. But the event space must include the empty set and the full set.

Basics: Probability Distribution

- We have outcomes and events.
- Now assign probabilities $\text{For } E \in \mathcal{F},\ P(E) \in [0,1]$

Back to our example:

$$\mathcal{F} = \underbrace{\{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\}}_{\text{events}}$$

$$P(\{1,3,5\}) = 0.2, P(\{2,4,6\}) = 0.8$$

Basics: **Axioms**

- Rules for probability:
  - For all events $E \in \mathcal{F}, P(E) \geq 0$
  - Always, $P(\emptyset) = 0, P(\Omega) = 1$
  - For disjoint events, $P(E_1 \cup E_2) = P(E_1) + P(E_2)$

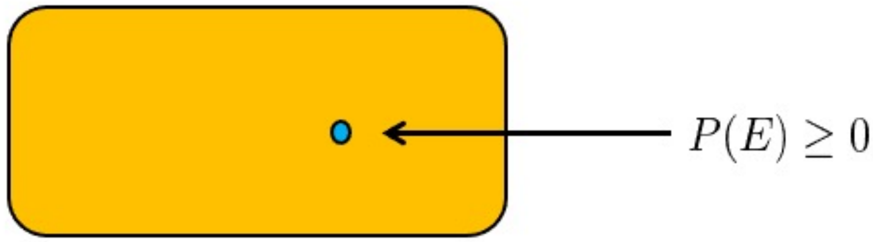
- Easy to derive other laws. Ex: non-disjoint events

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

A probability (distribution) is a function mapping from the event space to real numbers, ie, assign a value to each event in the event space. The assignment needs to satisfy the axioms.
(The slide show the axioms for the finite event space. We have slightly more complicated axioms for infinite event space.)

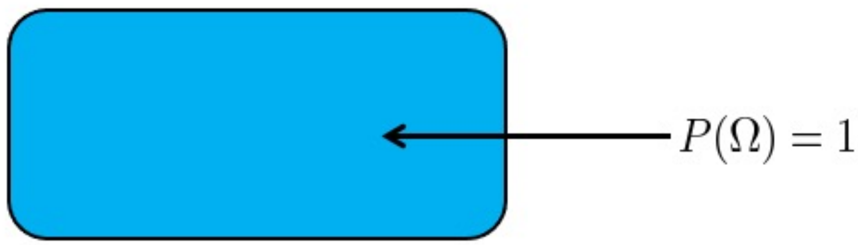
Visualizing the Axioms: I

- **Axiom 1:** $E \in \mathcal{F}, P(E) \geq 0$



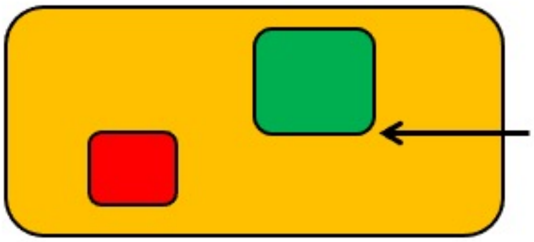
Visualizing the Axioms: **II**

- **Axiom 2:** $P(\emptyset) = 0, P(\Omega) = 1$



Visualizing the Axioms: **III**

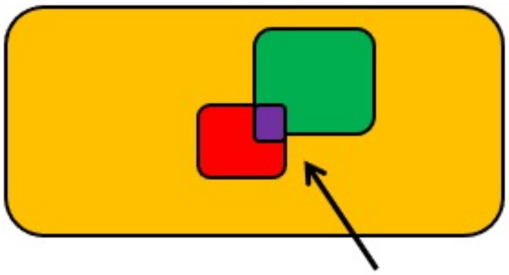
- **Axiom 3: disjoint** $P(E_1 \cup E_2) = P(E_1) + P(E_2)$



$P(E_1 \cup E_2) = P(E_1) + P(E_2)$

Visualizing the Axioms

- Also, other laws:



$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

Break & Quiz

- **Q 1.1**: There are exactly 3 candidates for a presidential election. We know X has a 30% chance of winning, B has a 35% chance. What's the probability that C wins?
- A. 0.35
- B. 0.23
- C. 0.333
- D. 0.8

Break & Quiz

- **Q 1.1**: There are exactly 3 candidates for a presidential election. We know X has a 30% chance of winning, B has a 35% chance. What's the probability that C wins?
- **A. 0.35**
- B. 0.23
- C. 0.333
- D. 0.8

1 - 30% - 35%

Break & Quiz

- **Q 1.2**: What’s the probability of selecting a black card or a number 6 from a standard deck of 52 cards?
- A. 26/52
- B. 4/52
- C. 30/52
- D. 28/52

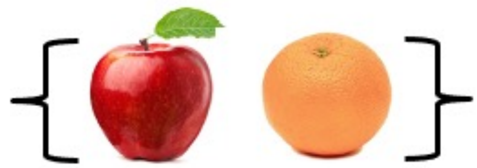
Break & Quiz

- **Q 1.2**: What's the probability of selecting a black card or a number 6 from a standard deck of 52 cards?
- A. 26/52
- B. 4/52
- C. 30/52
- **D. 28/52**

#black cards: 52/2=26
#card 6 that are not black: 2

Basics: **Random Variables**

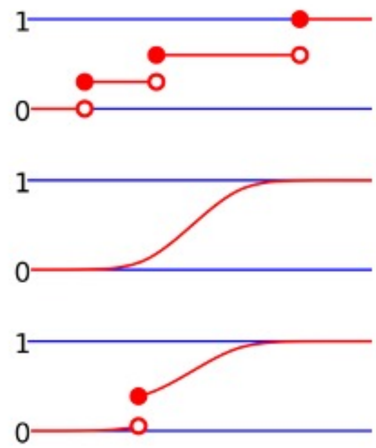
- Really, functions
- Map outcomes to real values $X : \Omega \to \mathbb{R}$



- Why?
  - So far, everything is a set.
  - Hard to work with!
  - Real values are easy to work with

Random numbers are also functions, mapping from the outcome space to real numbers, ie, for each outcome we have a real value for the random variable.

Basics: **CDF & PDF**



Wiki CDF

- Can still work with probabilities:

$$P(X=3) := P(\{\omega : X(\omega) = 3\})$$

- Cumulative Distribution Func. (CDF)

$$F_X(x) := P(X \leq x)$$

- Density / mass function $p_X(x)$

Basics: **Expectation & Variance**

- Another advantage of RVs are ``summaries''
- Expectation: $E[X] = \sum_a a \times P(x = a)$
  - The "average"
- Variance: $Var[X] = E[(X - E[X])^2]$
  - A measure of spread
- Higher moments: other parametrizations

Basics: **Joint Distributions**

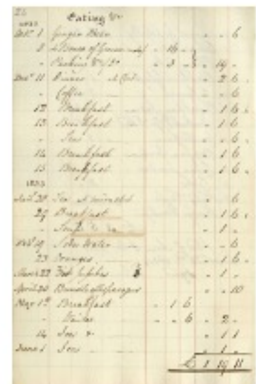
- Move from one variable to several
- Joint distribution: $P(X = a, Y = b)$
  - Why? Work with **multiple** types of uncertainty

Basics: **Marginal** Probability

- Given a joint distribution $P(X = a, Y = b)$
  - Get the distribution in just one variable:

$$P(X = a) = \sum_b P(X = a, Y = b)$$

  - This is the "marginal" distribution.



Basics: **Marginal** Probability

$$P(X=a)=\sum_b P(X=a,Y=b)$$

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

$$[P(\text{hot}),P(\text{cold})]=[\tfrac{195}{365},\tfrac{170}{365}]$$

Probability Tables

- Write our distributions as tables

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

- # of entries? 6.
  - If we have n variables with k values, we get k^n entries
  - **Big!** For a 1080p screen, 12 bit color, size of table: $10^{7490589}$
  - No way of writing down all terms

If we have n variables, then the joint probability table is of an n-dim array. If each variable has k values, then the number of entries in this table is k*k*…*k = k^n.

Independence

- Independence between RVs:

$$P(X,Y) = P(X)P(Y)$$

- Why useful? Go from k^n entries in a table to $\sim kn$
- Collapses joint into **product** of marginals

If all n variables are independent, then we only need to write down the individual tables for each individual variable, and can compute the joint probability by the definition of independence.

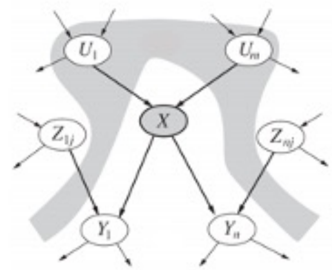
Conditional Probability

- For when we know something,

$$P(X = a | Y = b) = \frac{P(X = a, Y = b)}{P(Y = b)}$$

- Leads to **conditional independence**

$$P(X, Y | Z) = P(X|Z)P(Y|Z)$$



Credit: **Devin Soni**

Key definition; used a lot in AI/ML, such as Bayes' rule.

Random variables X and Y are independent conditioned on random variable Z, if their joint probability (conditioned on Z) is the product of their marginal probabilities (conditioned on Z).
We can think of P(|Z) as a new probability distribution.

Chain Rule

- Apply repeatedly,

$$P(A_1, A_2, \ldots, A_n)$$
$$= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1)\ldots P(A_n|A_{n-1}, \ldots, A_1)$$

- Note: still big!
  - If some **conditional independence**, can factor!
  - Leads to **probabilistic graphical models**

Break & Quiz

Q 2.1: Back to our joint distribution table:

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

What is the probability the temperature is hot given the weather is cloudy?

A. 40/365

B. 2/5

C. 3/5

D. 195/365

Break & Quiz

Q 2.1: Back to our joint distribution table:

	Sunny	Cloudy	Rainy
hot	150/365	40/365	5/365
cold	50/365	60/365	60/365

What is the probability the temperature is hot given the weather is cloudy?

A. 40/365
B. 2/5
C. 3/5
D. 195/365

P(hot|cloudy) = P(hot, cloudy) / P(cloudy) = (40/365) / (40/365 + 60/365) = 2/5

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

A. 0.3
B. 0.06
C. 0.24
D. 0.2

Break & Quiz

Q 2.2: Of a company's employees, 30% are women and 6% are married women. Suppose an employee is selected at random. If the employee selected is a woman, what is the probability that she is married?

A. 0.3
B. 0.06
C. 0.24
D. 0.2

P(married | woman) = P(married woman) / P(woman) = 6% / 30% = 0.2

Reasoning With Conditional Distributions

- Evaluating probabilities:
  - Wake up with a sore throat.
  - Do I have the flu?
- One approach: $S \to F$
  - Too strong.
- **Inference**: compute probability given evidence $P(F|S)$
  - Can be much more complex!

Probabilistic reasoning: first is to convert a decision making problem into a conditional probability problem

Using **Bayes' Rule**

- Want: $P(F|S)$
- **Bayes' Rule:** $P(F|S) = \frac{P(F,S)}{P(S)} = \frac{P(S|F)P(F)}{P(S)}$
- Parts:
  - $P(S) = 0.1$ Sore throat rate
  - $P(F) = 0.01$ Flu rate
  - $P(S|F) = 0.9$ Sore throat rate among flu sufferers

So: $P(F|S) = 0.09$

Bayes' rule: followed from two applications of the definition of conditional probability.

Useful: when P(F|S) is hard to reason about but P(S|F) is easy. For example, inferring the disease from the symptoms is hard, but inferring the symptoms from the disease is easy (by looking at statistics).

Using Bayes' Rule

- Interpretation $P(F|S) = 0.09$
  - Much higher chance of flu than normal rate (0.01).
  - Very different from $P(S|F) = 0.9$
    - 90% of folks with flu have a sore throat
    - But, only 9% of folks with a sore throat have flu

- Idea: **update** probabilities from

evidence

Bayesian **Inference**

- Fancy name for what we just did. Terminology:

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

- *H* is the hypothesis
- *E* is the evidence

Bayesian **Inference**

- Terminology:

$$P(H|E) = \frac{P(E|H)\color{red}{P(H)}}{P(E)} \longleftarrow \textbf{Prior}$$

- Prior: estimate of the probability **without** evidence

Bayesian Inference

- Terminology:

Likelihood

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

- Likelihood: probability of evidence **given a hypothesis**.

Bayesian Inference

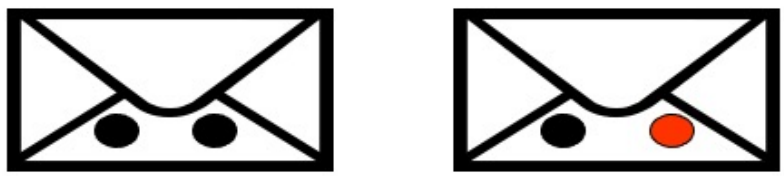
- Terminology:

$$\underbrace{P(H|E)}_{\text{Posterior}} = \frac{P(E|H)P(H)}{P(E)}$$

- Posterior: probability of hypothesis **given evidence**.

Two Envelopes Problem

- We have two envelopes:
  - E_1 has two black balls, E_2 has one black, one red
  - The **red** one is worth \$100. Others, zero
  - Open an envelope, see one ball. Then, can switch (or not).
  - You see a black ball. **Switch?**



Two Envelopes Solution

- **Let's solve it.**

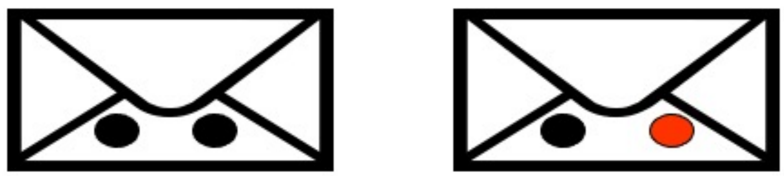
$$P(E_1|\text{Black ball}) = \frac{P(\text{Black ball}|E_1)P(E_1)}{P(\text{Black ball})}$$

- **Now plug in:**

$$P(E_1|\text{Black ball}) = \frac{1 \times \frac{1}{2}}{P(\text{Black ball})}$$

$$P(E_2|\text{Black ball}) = \frac{\frac{1}{2} \times \frac{1}{2}}{P(\text{Black ball})}$$

So switch!



Here E_1 denotes the event that the opened envelop is the envelop with two black balls, and E2 the event that it's the one with one black and one red. "Black ball" denotes the event that you see a black ball in the opened envelop.

We first convert the decision making problem into the problem of computing the conditional probabilities P(E_1 | Black ball) and P(E_2 | Black ball).
By Bayes' rule, we can compute the two. The former is larger than the latter. So it's more likely that the opened envelop contains two black balls and we should switch.

Break & Quiz

Q 3.1: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

A. 5/104
B. 95/100
C. 1/100
D. 1/2

Break & Quiz

Q 3.1: 50% of emails are spam. Software has been applied to filter spam. A certain brand of software can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%. Now if an email is detected as spam, then what is the probability that it is in fact a nonspam email?

A. **5/104**
B. 95/100
C. 1/100
D. 1/2

We first convert the problem into the problem of computing the conditional probability P(nonspam | detected as spam).

By Bayes' rule, we have
P(nonspam | detected as spam)
= P(detected as spam | nonspam) P(nonspam) / P(detected as spam)
= 5% * (1-50%) / (50% * 99% + 50% * 5%)
= 5/104

Break & Quiz

Q 3.2: A fair coin is tossed three times. Find the probability of getting 2 heads and a tail

A. 1/8
B. 2/8
C. 3/8
D. 5/8

Break & Quiz

Q 3.2: A fair coin is tossed three times. Find the probability of getting 2 heads and a tail

A. 1/8
B. 2/8
C. 3/8
D. 5/8

The sequence can be HHT, HTH, THH. Each case has a probability 1/8, so in total 3/8.