Tufts

Lecture 10:
Probability Theory

Artificial Intelligence (CS 131)

1

A Problem Involving Games

- Two players put money in on a game of chance
- First one to certain number of points, P, wins the money
- Game is interrupted before either wins, however:
  - Player one has $n < P$ points
  - Player two has $m < P$ points
- Who wins?

2

Paccioli's Solution

- *Summa de Arithmetica, Geometria, Proportioni et Proportionalita* (1487)
- Each gets proportion given by *points so far*, relative to total points by all players
- Player with n points gets:
  $n \;/\; (n + m)$
- Player with m points gets:
  $m \;/\; (n + m)$

According to Paccioli, any other solution is simply "preposterous"

3

Problems for Paccioli: Small Samples

- What if we've only played one round?
- What if we've played a few rounds, but there are *many more* left still to go?
- Is it fair that I win all or most of the money if I've only won a few games by luck at the start?

4

Tartaglia's Solution

- *Trattato generale di numeri e misure* (1556)
- Winner so far gets extra share in proportion to ***current lead***, relative to total needed to win
- So if player with n points leads, and P points complete the game, that player gets own half of the money, plus extra bonus, taken out of losing player's share:

$$(n - m) \;/\; P$$

The Trouble with Tartaglia: Ignoring Likely Outcomes

- Suppose we play to 100 points and the situation is:
  - Player 1: 80, Player 2: 70?
  - Player 1:99, Player 2: 89?
  - Player 1: 15, Player 2: 5?
- What's the difference between these situations?
- Depending upon the situation, the current lead may or may not tell us much about how *likely* we are to win in the future

The Birth of Probability Theory

- This "Problem of Points" (among other things) led to creation of probability theory
- A realization that it didn't really matter how many games you had *already* won
- Instead, we want to think about the chance that you are going to win in the future
- What was wanted:
  1. An exact measure of how likely something is
  2. A system for calculating with such likelihood measurements

Creators of Probability Theory

Blaise Pascal (1623-62)

Pierre de Fermat (1629-95)

Christiaan de Huygens (1601-65)

Sources of Uncertainty

- Stochastic environments
  - players in games who play randomly, random action-effects
- Imprecise models
  - complex systems (weather, real-life games), unknowns and errors in descriptions (Mars Polar Lander)
- Noisy data
  - limited range, obscuring weather, defective sensors
- Limitless exceptions to our knowledge
  - "Mammals have live young," "Republicans are military hawks"
- Many more…

9

Reasoning Under Uncertainty

- Even simple exceptions to rules cause problems for apparently logical reasoning patterns:

"Usually, mammals bear live young"
"Usually, warm-blooded animals are mammals"

"Usually, warm-blooded animals bear live young"

"Usually, people who join the Army are male"
"Usually, female ROTC members join the Army"

"Usually female ROTC members are male"

10

Probability and Uncertainty

- Precise framework for exact reasoning under uncertainty
- Allows us to:

1. Combine multiple pieces of evidence
2. Update our beliefs as new evidence comes in
3. Predict what is likely going to happen in the future
4. Diagnose what is likely to have happened in the past, given the present circumstances

11

Basic Elements of Probability Theory

- Random variable: thing with uncertain outcomes, and its own domain of values, which could be:
  - Boolean (True/False)
  - Discrete (countable domains)
  - Continuous (real-numbered domains)
- Outcome: particular setting of a value for some variable
  - Example: $\text{die}_1 = 3$
- Event: a combination of outcomes
  - Example: $\text{die}_1 = 3 \wedge \text{die}_2 = 2 \wedge \text{die}_3 = 6$

12

Random Variables & Outcomes

- Notation:
  - X, Y, Z : individual variables
  - x, y, z : outcome values from the domain of each variable
- We may use more informative names:
  - StudentXGrade: grade for Student "X"
  - A, B: possible grade values
- Atomic event: particular outcome for a set of variables
  - Perhaps a single event: StudentXGrade = A
  - Perhaps a combination:

    StudentXGrade = A ∧ StudentYGrade = B

Basic Logical Notation in Play

- Logical symbols:
  - NOT: ¬ A (Event A did not occur)
  - AND: A ∧ B (Both A and B occur)
  - OR: A ∨ B (Either A or B, or both, occur)
  - SOME: ∃x… (Exists at least one x such that…)
  - ALL: ∀x… (All x are such that…)
- Basic logical facts:
  - Non-contradiction: A ∧ ¬A == False
  - Excluded Middle: A ∨ ¬A == True

Properties of Basic Events

- Mutually exclusive: At *most* one event can be true
- Exhaustive: At *least* one event must be true
- Thus, one (and only one) of the following has to be true, assuming two variables with two values each:

  StudentXGrade = A ∧ StudentYGrade = A

  StudentXGrade = A ∧ StudentYGrade = B

  StudentXGrade = B ∧ StudentYGrade = A

  StudentXGrade = B ∧ StudentYGrade = B

Probability Distributions

- For any variable and outcome, we have the unconditional probability that it is true:

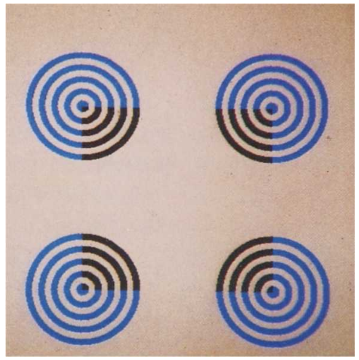
  P(StudentXGrade = A) = 0.78
- Distribution: collection of all probabilities for a variable:

  P(StudentXGrade) = {0.78, 0.22} [A, B]
- We can then calculate the probability of more complex events based on the basic distributions:

  P(StudentXGrade = A ∧ StudentYGrade = B) = ?
- Where do these basic numbers actually *come from?*

Subjective Probability

- One view: probabilities describe what we *believe* to be true about the world
  - *How confident* we are that something will/won't happen
  - Can be based on a number of things (observation of human behavior, expert opinion, polls, etc.)
- *Question*: what do we do if our beliefs are *different* than those of others?



Objective Probability

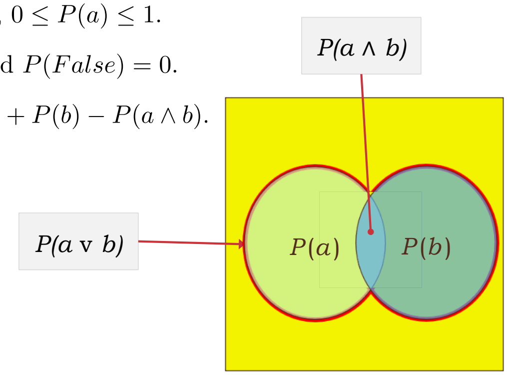
- Another view: probability distributions reflect *reality*
  - Events really do happen with some set probability
  - If we have an *accurate* distribution, then our numbers match with how the world is in and of itself
- *Question*: what is the mechanism that makes all of this work?

Frequentist Probability

- Another idea: probabilities are based on actual, *empirical observation* of events over time
  - We track how many times different outcomes occur, *relative to* the total number of events seen so far
- *Potential Issues*:
  - Requires many observations
  - Unrepeatable/very rare events are hard to put meaningful numbers on
  - Objectivist view thinks this is just an attempt to learn what is really there in the first place

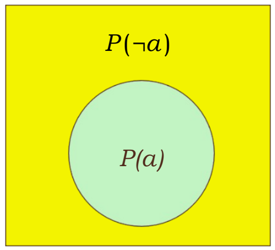
Basic Axioms of Probability

1. For any event a, $0 \leq P(a) \leq 1$.
2. $P(True) = 1$ and $P(False) = 0$.
3. $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$.



Using the Axioms

$$P(a \vee \neg a) = P(a) + P(\neg a) - P(a \wedge \neg a) \qquad \text{[Axiom 3]}$$
$$P(True) = P(a) + P(\neg a) - P(False) \qquad \text{[Logic]}$$
$$1 = P(a) + P(\neg a) \qquad \text{[Axiom 2]}$$
$$P(\neg a) = 1 - P(a)$$



Joint Probability

- For random variables {X1, X2, …, Xn} joint distribution gives probability of *each possible combination* of outcomes
- Can be written as a table of values:

	StudentXGrade = A	StudentXGrade = B
StudentYGrade = A	0.72	0.08
StudentYGrade = B	0.18	0.02

Note: if the table is supposed to represent a *proper* joint distribution, then the numbers must all sum to 1.0

Marginal Probability

	StudentXGrade = A	StudentXGrade = B
StudentYGrade = A	0.72	0.08
StudentYGrade = B	0.18	0.02

- Given the joint distribution we can get the probability of any outcome by marginalizing: summing all values for which the outcome of interest is true
- For example, probability student X gets an A:

$$P(StudentXGrade = A) \;=\; (0.72 + 0.18) \;=\; 0.9$$

- Probability that *either* X or Y gets an A:

$$P(StudentXGrade = A \;\vee\; StudentYGrade = A)$$
$$= \;(0.72 + 0.18 + 0.008) \;=\; 0.98$$

Size of the Joint Distribution

- How many numbers do we need to specify the full joint distribution in each of the following cases:

1. Three variables as follows:
   - StudentGeneration = {First, Other}
   - StudentYear = {1, 2, 3, 4}
   - StudentGrade = {A, B, C, D, F}

$$(|SGen| \times |SYr| \times |SGrade| - 1)$$
$$= (2 \times 4 \times 5 - 1) = 39$$

2. A set of n binary (2-valued) variables?

$$(|X|_1 \times |X|_2 \times \cdots \times |X|_n - 1) = 2^n - 1$$

3. A set of n m-ary (m-valued) variables?

$$(|X|_1 \times |X|_2 \times \cdots \times |X|_n - 1) = m^n - 1$$

Conditional Probability

- One of the most important things we do with probabilities involves *reasoning from evidence*
  - We don't always care about the prior (simple) probability that something is true no matter what
- Rather, we want to know how likely it actually is given what else we know…
  - *Given* a storm warning, how likely is a hurricane?
  - *Given* the cards on the table, how likely is my opponent to have a better poker hand than me?

The Product Rule for Conditional Probability

- We define the conditional probability of event a, given event b, using the basic rule:

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

- Equivalently, we have the product rule for joint probability (in two different, equivalent forms):

$$P(a \wedge b) = P(a\,|\,b)P(b)$$
$$P(a \wedge b) = P(b\,|\,a)P(a)$$

  - ***Why*** are these equivalent? (Follows from basic logic and the basic definition given above.)

Calculating Conditional Probability

- Given the following joint distribution:

	StudyHrs = 3-5	StudyHrs = 1-2	StudyHrs = 0
Grade = A	0.12	0.08	0.02
Grade = B	0.16	0.14	0.04
Grade = C	0.06	0.08	0.06
Grade = F	0.01	0.05	0.18

1. What is the probability of getting an A if we study 3–5 hours?
2. What is the probability of getting a B *or higher* if we study that same amount?
3. What is the probability that a student who studies *less* than 3–5 hours *doesn't* get an A?

Calculating Conditional Probability

	StudyHrs = 3-5	StudyHrs = 1-2	StudyHrs = 0
Grade = A	0.12	0.08	0.02
Grade = B	0.16	0.14	0.04
Grade = C	0.06	0.08	0.06
Grade = F	0.01	0.05	0.18

1. What is the probability of getting an A if we study for 3–5 hours?

$$P(A\,|\,3\text{-}5hrs) = \frac{\boxed{P(A \wedge 3\text{-}5hrs)}}{\boxed{P(3\text{-}5hrs)}}$$

- The first value needed comes directly from the joint distribution, while the second comes by marginalizing

$$P(A\,|\,3\text{-}5hrs) = \frac{0.12}{(0.12 + 0.16 + 0.06 + 0.01)}$$
$$= \frac{0.12}{0.35} \approx 0.343$$

Calculating Conditional Probability

	StudyHrs = 3-5	StudyHrs = 1-2	StudyHrs = 0
Grade = A	0.12	0.08	0.02
Grade = B	0.16	0.14	0.04
Grade = C	0.06	0.08	0.06
Grade = F	0.01	0.05	0.18

2. What is the probability of getting a B or higher if we study for 3–5 hours?

$$P(A \lor B \,|\, 3\text{-}5hrs) = \frac{\boxed{P((A \lor B) \land 3\text{-}5hrs)}}{\boxed{P(3\text{-}5hrs)}}$$

- Here, we get *both* values needed by marginalizing

$$P(A \lor B \,|\, 3\text{-}5hrs) = \frac{(0.12 + 0.16)}{(0.12 + 0.16 + 0.06 + 0.01)}$$

$$= \frac{0.28}{0.35} = 0.8$$

29

Calculating Conditional Probability

	StudyHrs = 3-5	StudyHrs = 1-2	StudyHrs = 0
Grade = A	0.12	0.08	0.02
Grade = B	0.16	0.14	0.04
Grade = C	0.06	0.08	0.06
Grade = F	0.01	0.05	0.18

3. What is the probability that a student who studies *less than* 3–5 hours *doesn't* get an A?

$$P(\neg A \,|\, \neg 3\text{-}5hrs) = \frac{\boxed{P(\neg A \land \neg 3\text{-}5hrs)}}{\boxed{P(\neg 3\text{-}5hrs)}}$$

- Again, we get both values needed by marginalizing

$$P(\neg A \,|\, \neg 3\text{-}5hrs) = \frac{0.55}{0.65} \approx 0.846$$

30

Normalization constants

- Note *same divisor* in these 2 calculations:

$$P(A \,|\, 3\text{-}5hrs) = \frac{P(A \land 3\text{-}5hrs)}{P(3\text{-}5hrs)}$$

$$P(A \lor B \,|\, 3\text{-}5hrs) = \frac{P((A \lor B) \land 3\text{-}5hrs)}{P(3\text{-}5hrs)}$$

- This prior probability, of a study-time event of 3–5 hours, is called a normalization constant
- Since it is the same for *all* probabilities conditional on that same event, sometimes abbreviated as simply a constant factor alpha (α)

$$P(A \,|\, 3\text{-}5hrs) = \alpha \, P(A \land 3\text{-}5hrs)$$

- Why is this noteworthy?

31