Review of Basic Probability and Statistics Concepts

CompSci 590-07
Duke University
Ronald Parr

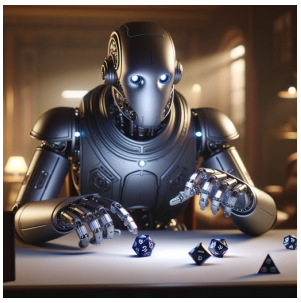
1

Goals Of These Slides

- Revisit a topic most of you have seen already to
  - Refresh your memories
  - Synchronize notation
  - Provide a reference to which you can check back
- Provide context – for AI

2

Why does AI need uncertainty?



- Reason: Sh*t happens
- Actions don't have deterministic outcomes

- Can logic be the "language" of AI???
- Problem: General logical statements are almost always false

- Truthful and accurate statements about the world would seem to require an endless list of *qualifications*
- How do you start a car?
- Call this "The Qualification Problem"

3

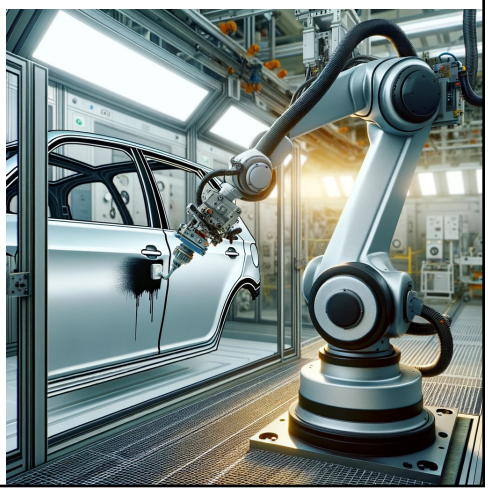
The Qualification Problem

- Is this a real concern?
- YES!
- Systems that try to avoid dealing with uncertainty tend to be brittle.
- Plans fail
- Shortest path to goal isn't great if it doesn't ***really*** get you to the goal

4

When can we (mostly) ignore qualifications?

- When environment is highly engineered/controlled
- Objects moving in free space
- Carefully controlled factories
- When replanning is relatively cheap



5

What are probabilities mathematically?

- Probabilities are defined over random variables
- Random variables represented with capitals: X,Y,Z
- RVs take on values from a finite domain: d(X), d(Y), d(Z)
- We use lower case letters for values from domains
  - X=x asserts: RV X has taken on value x
  - P(x) is shorthand for P(X=x)

11

Kolmogorov's axioms of probability

- $0 \leq P(a) \leq 1$
- P(true) = 1; P(false)=0
- P(a or b) = P(a) + P(b) – P(a and b)
- Subtract to correct for double counting

- Sufficient to completely specify probability theory for discrete variables
- Continuous variables need *density functions*

Nota bene: We don't use these, but they are included to emphasize how simply probability is axiomatized

12

Notation

- P(XY) = P(X,Y) = a joint probability distribution over all settings of X and Y
  (potentially a **table** with a large number of entries)

- P(xy) = P(x,y) = P(x AND y) = P(x ^ y) = P(X=x,Y=y),P(X=x AND Y=y)=P(X=x ^ Y=x) = **a single number** corresponding the probability that both X=x and Y=y

- P(Xy) = a table with one entry for each value of X when y is true – not a distribution

- P(X=false)=P($\bar{x}$)=P(¬x) = P(~x) = a single number for case where X is a binary variable takes value false (or zero)

13

Why Probabilities Are Messy

- Probabilities are not truth-functional
- Computing P(a and b) requires the joint distribution
  - It is not, in general, a function of P(a) and P(b)
  - It is not, in general, a function of P(a) and P(b)
  - It is not, in general, a function of P(a) and P(b)
- This fact led to many approximation methods such as certainty factors and fuzzy logic briefly getting traction (Why?)

14

Working With Joint Distributions

- A joint distribution is an assignment of probabilities to every possible atomic event
- We can define all other probabilities in terms of the joint probabilities by *marginalization*:

$$P(a) = P(a \wedge b) + P(a \wedge \neg b)$$

$$P(a) = \sum_{e_i \in e(a)} P(e_i)$$

15

Example

- P(cold ∧ headache) = 0.4
- P(¬cold ∧ headache) = 0.2
- P(cold ∧ ¬ headache) = 0.3
- P(¬ cold ∧ ¬ headache) = 0.1

- What are P(cold) and P(headache)?

16

Independence

- If A and B are independent: P(A ∧ B) = P(A)P(B)

- P(cold ∧ headache) = 0.4
- P(¬cold ∧ headache) = 0.2
- P(cold ∧ ¬ headache) = 0.3
- P(¬ cold ∧ ¬ headache) = 0.1

- Are cold and headache independent?

17

Conditional Probabilities

- Ordinary probabilities for random variables: *unconditional* or *prior* probabilities
- P(a|b) = P(a AND b)/P(b)
- This tells us the probability of a **given that we know *only* b**
- If we know c and d, we can't use P(a|b) directly (without additional assumptions)
- Annoying, but solves the qualification problem…

20

Connidtional Probability **Solves** the Qualification Problem

- P(car_starts|turn_key)
- Probability of car starting given that we know ***only*** that key has been turned
- The conditioning bar indicates that the probability is defined with respect to a particular state of knowledge, *not as an absolute thing*

21

Condition with Bayes's Rule

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

$$P(A \mid B)P(B) = P(B \mid A)P(A)$$

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

Note that we will usually call Bayes's rules "Bayes Rule"

22

Let's Play Doctor

- P(cold) = 0.7, P(headache) = 0.6
- P(headache|cold) = 0.57
- What is P(cold|headache) using Bayes Rule?

$$P(c \mid h) = \frac{P(h \mid c)P(c)}{P(h)}$$

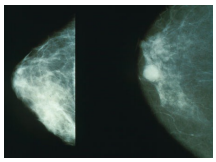
$$= \frac{0.57 * 0.7}{0.6} = 0.66$$

- IMPORTANT: Not always symmetric

23

Another Example

- From: http://opinionator.blogs.nytimes.com/2010/04/25/chances-are/ (attributed to Gerd Gigerenzer)
- "…The probability that one of these women has breast cancer is 0.8 percent. If a woman has breast cancer, the probability is 90 percent that she will have a positive mammogram. If a woman does ***not*** have breast cancer, the probability is 7 percent that she will still have a positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she actually has breast cancer?"
- 95/100 U.S. doctors answered ~75%



Source: Wikipedia

24

Expectation

- Most of us use expectation in some form when we compute averages
- What is the average value of a fair die roll?
- (1+2+3+4+5+6)/6 = 3.5
  (we divide by 6 because all outcomes are equally likely)
- Is it possible for all children to be above average?

25

Expectation in General

- Suppose we have some RV X
- Suppose we have some function f(X)
- What is the expected value of f(X)?

$$\underset{x}{E} f(x) = \sum_{x} P(X) f(X)$$

26

Linearity of Expectation

- Suppose we have f(X) and g(Y).
- What is the expected value of f(X)+g(Y)?

$$\begin{aligned}
\underset{XY}{E} f(X)+g(Y) &= \sum_{XY} P(X \wedge Y)(f(X)+g(Y)) \\
&= \sum_{XY} P(X \wedge Y) f(X) + \sum_{XY} P(X \wedge Y) g(Y) \\
&= \sum_{X} \sum_{Y} P(X \wedge Y) f(X) + \sum_{Y} \sum_{X} P(X \wedge Y) g(Y) \\
&= \sum_{X} f(x) \sum_{Y} P(X \wedge Y) + \sum_{Y} g(Y) \sum_{X} P(X \wedge Y) \\
&= \sum_{X} f(x) P(X) + \sum_{Y} g(Y) \sum_{X} P(X \wedge Y) \\
&= \underset{X}{E} f(X) + \underset{Y}{E} g(Y)
\end{aligned}$$

27

Sample (Empirical) Mean

- Suppose we observe $X_1 \ldots X_n$
- Assume these are independently drawn, and indentically distributed (IID)
- What is our estimate for E(X)?

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Why is this "right"?

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{nE(X)}{n} = E(X)$$

Also…

28

Properties of the Sample Mean

- Is unbiased and consistent
  - Unbiased: Expected value = true value (shown on previous slide)
  - Consistent: Converges to true value
- Minimizes expected squared loss on the training data

29

Variance

- Expected, squared deviation from the mean
- "How much we trust the mean"

$$Var(X) = E[(X - E(X))^2]$$
$$= E(X^2) - E(X)^2$$

Nota bene: We will typically assume that Var(X) is finite.

35

Properties of Variance

$$\sigma^2(X) = Var(X) = E[(X - E(X))^2]$$

$Var(aX) = ???$ $\quad a^2 Var(X)$

$Var(aX + b) = ???$ $\quad a^2 Var(X)$

$Var(X + Y) = ???$ $\quad Var(X) + Var(Y) + 2E[(X - E(X))(Y - E(Y))]$

If X,Y are independent: $\quad Var(X) + Var(Y)$

36

Covariance

$$Var(X+Y) = Var(X) + Var(Y) + 2E[(X - E(X))(Y - E(Y))]$$

Covariance captures the leftover:

$$Cov(X,Y) = Cov(Y,X) = E[(X - E(X))(Y - E(Y))]$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

If X,Y are independent, cov(X,Y)= 0

37

Standard Deviation

$$\sigma(X) = SD(X) = \sqrt{Var(X)}$$

- Sometimes more natural than variance:

$$\sigma(aX) = a\sigma(X)$$

- Often not, for X,Y independent (compare with variance):

$$\sigma(X+Y) = \sqrt{\sigma^2(X) + \sigma^2(Y)}$$

38

Sample Variance

- Suppose we observe $X_1 \ldots X_n$
- Assume these are independently drawn, and identically distributed (IID)
- What is our estimate for V(X)?

$$\bar{\sigma}^2 = \frac{\sum_{i=1}^{n} (\bar{X} - X_i)^2}{n-1}$$

- Note: n-1 is necessary for it to be unbiased

39

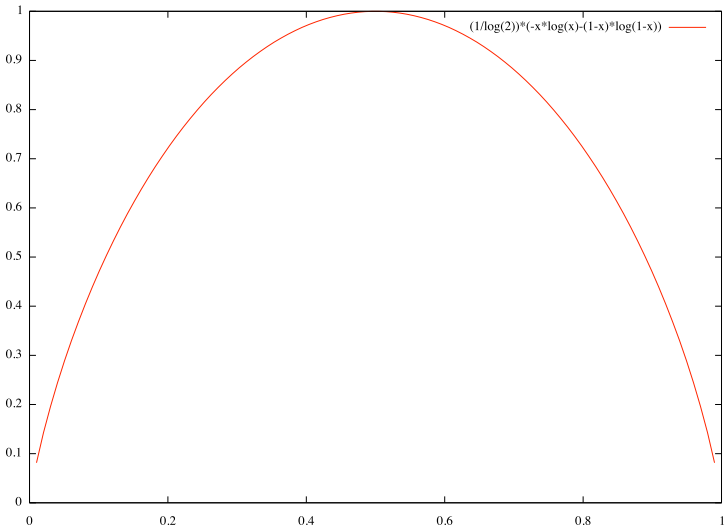
Entropy

$$H(P) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

For an event space with n events p_i = prob of event i

40

Entropy of a Coin



41

Is Entropy Bad?

- Reducing entropy is often thought of a good thing, but…
- Maximizing entropy can sometimes imply making the weakest assumptions – maximum entropy solution is considered the "most conservative" one
- Entropy can increase by changing your event space
  - In RL, some methods measure the entropy of a stochastic policy, i.e., the entropy in the distribution over actions
  - Atari games have 18 possible actions (9 joystick positions x 2 button positions)
  - For games that don't use the button, actions with or without the button are identical
  - Entropy increases if you consider the button, even though the true policy space doesn't change

42

Statistician humor

- Statistician 1: I hear that you met Bob.
- Statistician 2: Yes – he's a jerk.
- Statistician 1: Change in entropy: 0.

- What did statistician 1 just say?

46

Requirements on Continuous Distributions

- p(x)>1 *is* possible so long as:

$$\int_x p(x)dx = 1$$

- Don't confuse p(x) and P(X=x)
- P(X=x) for any x is 0!

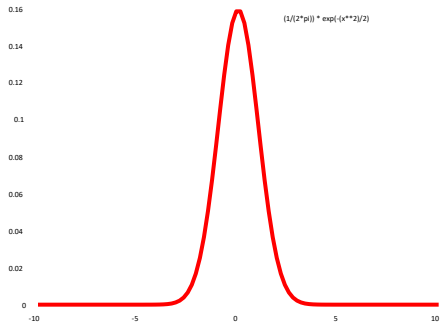
$$P(x \in A) = \int_A p(x)dx$$

47

Continuous Random Variables

- Domain is some interval, region, or union of regions
- Uniform case: Simplest to visualize
  (event probability is proportional to area)
- Non-uniform case visualized with extra dimension

Gaussian (normal/bell) distribution:



48

Sloppy Comment about Continuous Distributions

- In many cases, you can generalize what you know about discrete distributions to continuous distributions, replacing "P" with "p" and "Σ" with "$\int$"
- Proper treatment of this topic requires measure theory and is beyond the scope of the class

49