Basic Probability Review

CPS 271

Ron Parr

Probability: Who needs it?

- Learning without probabilities is possible
  - Version spaces
  - Explanation based learning
- Learning almost always involves
  - Noise in data
  - Prediction about the future
- Learning systems that don't use probability in some way tend to be very, very brittle

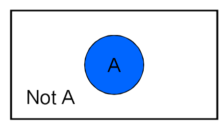
Probabilities

- Natural way to represent uncertainty
- People have intuitive notions about probabilities
- Many of these are wrong or inconsistent
- Most people don't get what probabilities mean
- Finer details of this question still debated

Relative Frequencies

- Probabilities defined over events
- Space of all possible events is "event space"

Event space:



- Think: Playing blindfolded darts with the Venn diagram..

Understanding Probabilities

- Probabilities have dual meanings
  - Relative frequencies (frequentist view)
  - Degree of belief (Bayesian view)
- Neither is entirely satisfying
  - No two events are truly the same (reference class problem)
  - Statements should be grounded in reality in some way

Why probabilities are good (despite the difficulties)

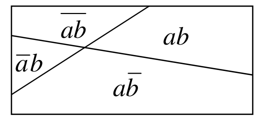
- Subjectivists: probabilities are degrees of belief
- Are all degrees of belief probability?
  - AI has used many notions of belief:
    - Certainty Factors
    - Fuzzy Logic
- Can prove that a person who holds a system of beliefs inconsistent with probability theory can be tricked into accepted a sequence of bets that is guaranteed to lose

Probabilities over discrete events
(and the horror of common notation)

- Probabilities defined over sets of random variables
- RVs usually represented with capitals: X,Y,Z
- Use lower case letters for values from domains
- X=x asserts that the random variable X has taken on value x
- P(x) is shorthand for P(X=x)

Event spaces for discrete RVs

- 2 variable case



- Important: Event space grows exponentially in number of random variables
- Components of event space = atomic events

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:

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

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

- AKA: **Sum rule**, marginalization

Why Probabilities Are Messy

- Probabilities are not truth-functional
- To compute P(a and b), need joint distribution
  - sum out all of the other events from distribution
  - In general, it is not a function of P(a) and P(b)
  - In general, it is not a function of P(a) and P(b)
  - In general, it is not a function of P(a) and P(b)
  - This fact led to many approximations methods such as certainty factors and fuzzy logic (Why?)

Independence

- RVs A and B are independent iff:
  - P(AB)=P(A)P(B)
- Independence:
  - Make things computationally easy
  - Makes things boring
    - From an algorithmic standpoint
    - From a predictive standpoint
  - Is almost never true
  - Is approximately true for “unrelated” events

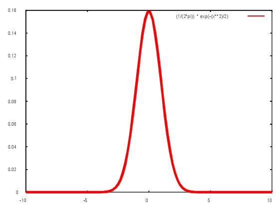
Kolmogorov’s axioms of probability

- 0<=P(a)<=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
- This is sufficient to specify probability theory for discrete variables
- Continuous variables need density functions

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:



Updating Kolmogrov's Axioms

- Use lower case for probability density
- Use end of the alphabet for continuous vars
- For discrete events: $0 \le P(a) \le 1$
- For densities: $0 \le p(x)$

- Is p(x)>1 possible???

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!

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

Cumulative Distributions

- When distribution is over numbers, we can ask:
  - P(X>=c) for some c
  - P(X<c) for some c
  - P(a<=X<=b) for some, a and b
- Solve by
  - Summation
  - Integration
- Cumulative sometimes called
  - CDF
  - Distribution function

Sloppy Comment about Continuous Distributions

- In many, many cases, you can generalize what you know about discrete distributions to continuous distributions by replace "p" with "P" and "Σ" with "∫"

- Proper treatment of this topic requires measure theory and is beyond the scope of the text and class

Conditional Probabilities

- Ordinary probabilities = unconditional or prior probabilities

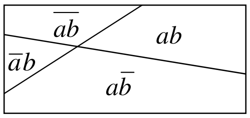
- P(a|b) = probability of a **given that we know *only* b**

- If we know c and d, we can't use P(a|b) directly
  (annoying, but important detail!)

- P(a|a)=1

Conditional Probability

- P(b|a) = Probability of event b given that event a is true



- Idea: In what fraction of a event space is b also true?

P(B|A) = P(AB)/P(A)

Definition of Conditional Probability

- Following geometric intuitions from previous slide
  - P(B|A) = P(AB)/P(A)
  - P(A|B) = P(AB)/P(B)
- Also known as the **product rule**:
  - P(B|A)P(A) = P(AB)=P(BA)
  - P(A|B)P(B) = P(AB)=P(BA)

Condition with Bayes's Rule

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

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

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

Why Bayes's Rule is Cool

- Solves the "inverse probability" problem
- Diagnosis:
  - Often we know: P(Symptoms|Disease) from data
  - Want: P(D|S) to diagnose patients
- Sensing:
  - Know: P(Observation|Reality)
  - Want: P(R|O)
- Learning:
  - Know: P(Data|Hypothesis about source model)
  - Want: P(H|D)

Expectation

$$E(X) = \sum_X XP(X)$$

- Matches some colloquial notions of average
- "Mean"
- Arithmetic mean (uniform weights)
- For continuous random variables:

$$E(X) = \int_X Xp(X)dX$$

Nota bene: We will be assuming that E(X) is finite.

Properties of Expectation

$$E(f(X)) = \sum_X f(X)P(X)$$

$E(aX) = ???$ $aE(X)$

$E(aX+b) = ???$ $aE(X)+b$

$E(X+Y) = ???$ $E(X)+E(Y)$

$E(XY) = ???$ If X,Y are independent: $E(X)E(Y)$

Variance

- Hard to define in words
- "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.

Properties of Variance

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

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

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

$Var(X + Y) = ???$

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

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

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

Standard Deviation

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

- Even harder to define in English
- Sometimes more natural than variance:

$$SD(aX) = \quad aSD(X)$$

- Often not, for X,Y independent:

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

Sample Mean

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

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

- Why?

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

Also…