Probabilistic Reasoning

Doug Downey, Northwestern EECS 348

Spring 2013

Limitations of logic-based agents

- Qualification Problem
  - Action's preconditions can be complex
  - Action(Grab, t) => Holding(t)
    ….unless gold is slippery or nailed down or too heavy or our hands are full or…
- Brittleness
  - One contradiction in KB => KB entails everything

Limitations of logic-based agents

- Qualification Problem
  - Action's preconditions can be
  - Action(Grab, t) => Holding(t)
    ….unless gold is slippery or nailed down or too heavy or our hands are full or…

P(success) = 0.97

- Brittleness
  - One contradiction in KB => KB entails everything

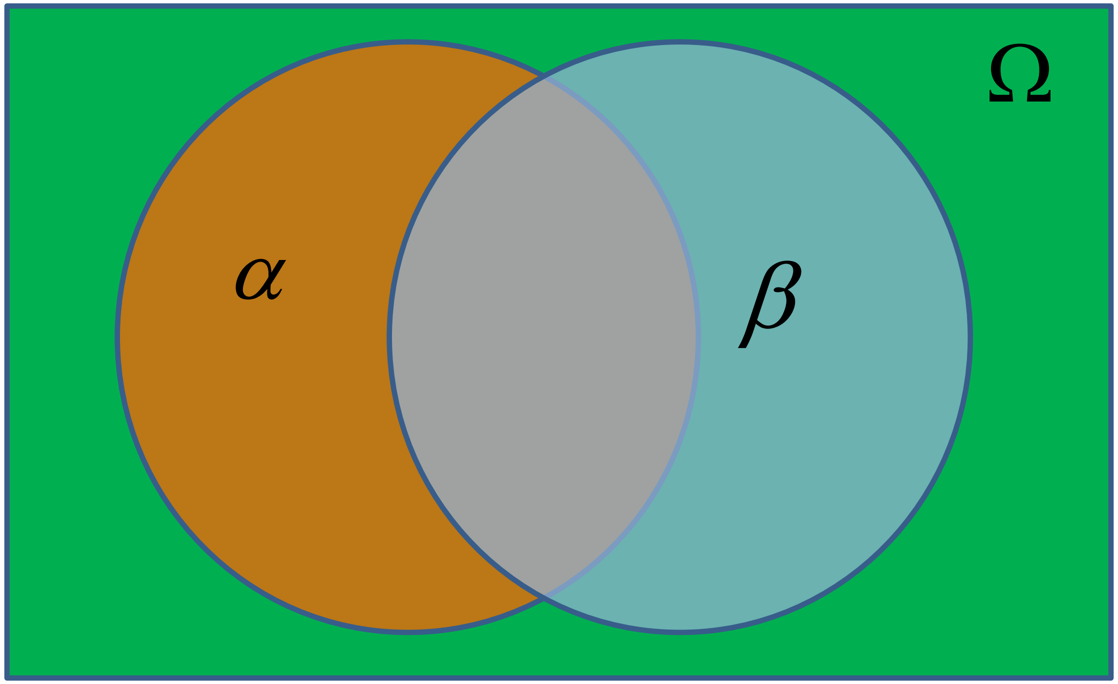
Instead of $a \wedge \neg a$,
$P(a) + P(\neg a) = 1$

Events

- Event space Ω
  - E.g. for dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$

- Set of measurable events $S \subseteq 2^{\Omega}$
  - E.g.,
    α = event we roll an even number = $\{2, 4, 6\} \in S$
  - S must:
    - Contain the empty event $\emptyset$ and the trivial event Ω
    - Be closed under union & complement
      - $\alpha, \beta \in S \rightarrow \alpha \cup \beta \in S$ and $\alpha \in S \rightarrow \Omega - \alpha \in S$

Probability Distributions



Can visualize probability as fraction of area

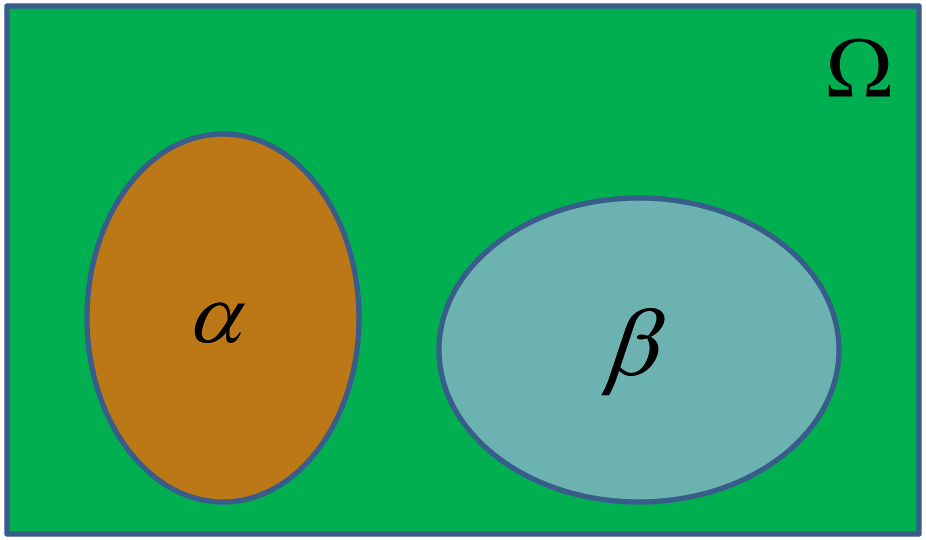
Probability Distributions

- A probability distribution P over (Ω, S) is a mapping from S to real values such that:

$$P(\alpha) \geq 0$$

$$P(\Omega) = 1$$

$$\alpha, \beta \in S \wedge \alpha \cap \beta = \varnothing \rightarrow P(\alpha \cup \beta) = P(\alpha) + P(\beta)$$



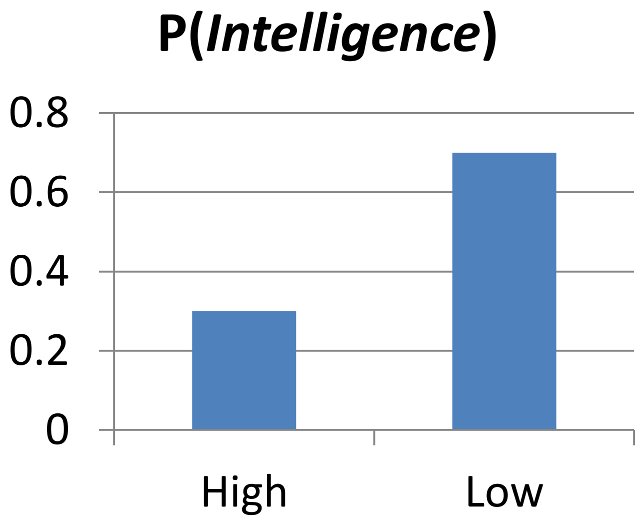
Probability: Interpretations & Motivation

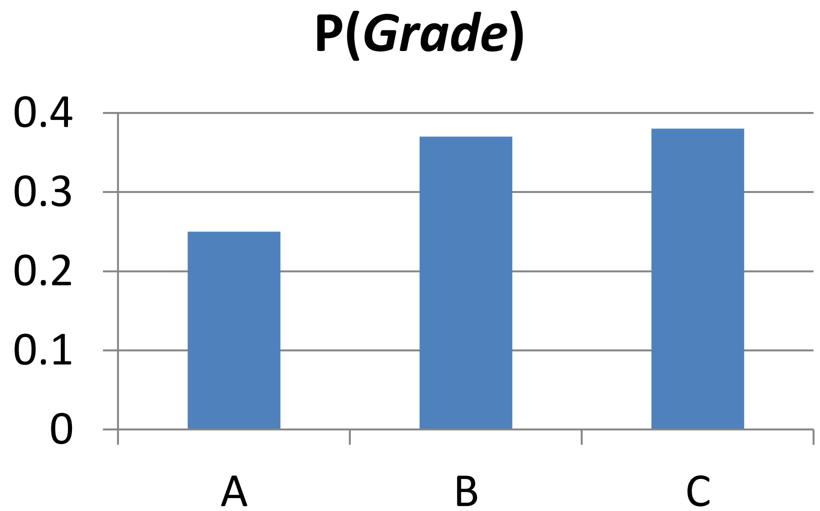
- Interpretations
  - Frequentist
  - Bayesian/subjective
- Why use probability for subjective beliefs?
  - Beliefs that violate the axioms can lead to bad decisions *regardless* of the outcome [de Finetti, 1931]
  - Example: P(A) = 0.6, P(not A) = 0.8 ?
  - Example: P(A) > P(B) and P(B) > P(A) ?

Random Variables

- A random variable is a function from Ω to a value
  - A short-hand for referring to *attributes* of events.
- E.g., your grade in this course
  - Let Ω = set of possible scores on hmwks and test
  - Cumbersome to have separate events GradeA, GradeB, GradeC
  - So instead define a random variable *Grade*
    - Deterministic function from Ω to {A, B, C}

Distributions

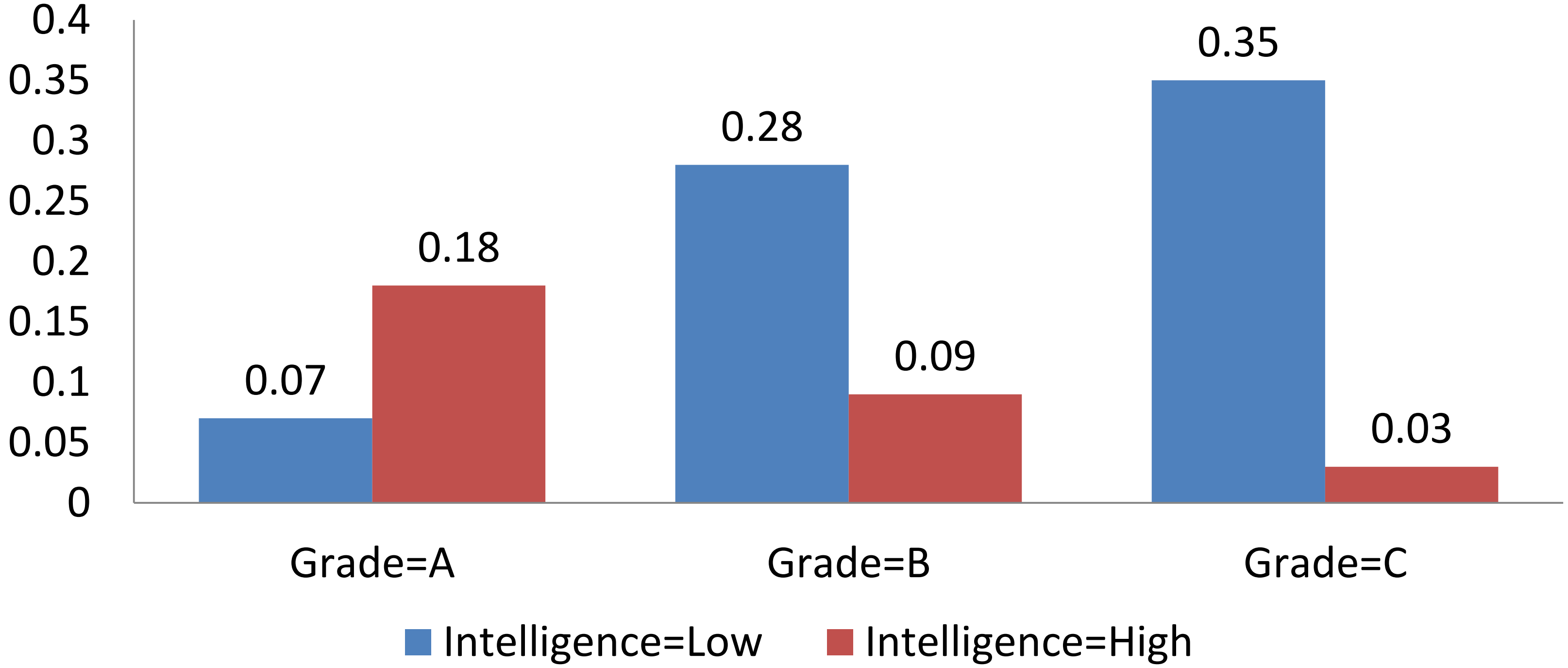




- Called “marginal” because they apply to only one r.v.

Joint Distribution

P(*Intelligence, Grade*)



Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

Joint Distribution specified with 2*3 – 1 = 5 values

Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

P(Grade = A, Intelligence = Low)? 0.07

Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

P(Grade = A)?   0.07 + 0.18 = 0.25

Joint Distribution

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

$P(\text{Grade} = A \vee \text{Intelligence} = \text{High})$?

$0.07 + 0.18 + 0.09 + 0.03 = 0.37$

=> Given the joint distribution, we can compute probabilities for any proposition by summing events.

Conditional Probability

- P(*Grade* = A | *Intelligence* = High) = 0.6
  - the probability of getting an A given **only** *Intelligence =* High, and nothing else.
    - If we know *Motivation* = High or *OtherInterests* = Many, the probability of an A changes even given high *Intelligence*
- Formal Definition:
  - $P(\alpha \mid \beta) = P(\alpha, \beta) / P(\beta)$
    - When $P(\beta) > 0$

Conditional Probability

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

P(*Grade* = A | *Intelligence* = High) ?

P(*Grade* = A, *Intelligence* = High) = 0.18

P(*Intelligence* = High) = 0.18+0.09+0.03 = 0.30

=> P(*Grade* = A | *Intelligence* = High) = 0.18/0.30 = **0.6**

Conditional Probability

		Intelligence	
		Low	High
Grade	A	0.07	0.18
	B	0.28	0.09
	C	0.35	0.03

P(*Intelligence* | *Grade* = A)?

Intelligence	
Low	High
0.28	0.72

Conditional Probability

		Intelligence	
		Low	High
Grade	A	0.28	0.72
	B	0.76	0.24
	C	0.92	0.08

P(*Intelligence* | *Grade*)?

Actually three separate distributions, one for each *Grade* value
(has three independent parameters total)

Chain Rule

$$P(X_1 = x_1, \ldots, X_n = x_n) =$$

$$\prod_{i=1}^{n} P(X_i = x_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1)$$

- E.g., P(*Grade*=B, *Int.* = High)
  = P(*Grade*=B | *Int.*= High)P(*Int.* = High)
- Can be used for distributions…
  - P(*A*, *B*) = P(*A* | *B*)P(*B*)

Queries

- Given subsets of random variables ***Y*** and ***E***, and assignments ***e*** to ***E***
  - Find P(***Y*** | ***E*** = ***e***)
- Answering queries = **inference**
  - The whole point of probabilistic models, more or less
  - P(*Disease* | *Symptoms*)
  - P(*StockMarketCrash* | *RecentPriceActivity*)
  - P(*CodingRegion* | *DNASequence*)
  - P(*PlayTennis* | *Weather*)
  - …(the other key task is **learning**)

Answering Queries: Summing Out

		Intelligence = Low		Intelligence=High	
		Time=Lots	Time=Little	Time=Lots	Time=Little
Grade	A	0.05	0.02	0.15	0.03
	B	0.14	0.14	0.05	0.0
	C	0.10	0.25	0.01	0.02

$$P(Grade \mid Time = \text{Lots})?$$

$$\sum_{v \in Val(Intelligence)} P(Grade, Intelligence = v \mid Time = Lots)$$

Answering Queries: Solved?

- Given the joint distribution, we can answer any query by summing
- …but, joint distribution of 500 Boolean variables has 2^500 -1 parameters (about 10^150)
- For non-trivial problems (~25 boolean r.v.s or more), using the joint distribution requires
  - Way too much **computation** to compute the sum
  - Way too many **observations** to learn the parameters
  - Way too much **space** to store the joint distribution

Conditional Independence (1 of 3)

- Independence
  - P(*A*, *B*) = P(*A*)*P(*B*), denoted $A \perp B$
  - E.g. consecutive dice rolls
    - Gambler's fallacy
  - Rare in (real) applications

Conditional Independence (2 of 3)

- Conditional Independence
  - P(*A*, *B* | *C*) = P(*A* | *C*) P (*B* | *C*), denoted (*A* ⊥ *B* | *C*)
  - Much more common
  - E.g.,
    (*GetIntoNU* ⊥ *GetIntoStanford* | *Application*),
    but **NOT** (*GetIntoNU*⊥ *GetIntoStanford*)

Conditional Independence (3 of 3)

- How does Conditional Independence save the day?

  P(*NU, Stanford, App*) =
  P(*NU*|*Stanford, App*)*P(*Stanford* |*App*)*P(*App*)

  Now, (***A*** ⊥ ***B*** | ***C***) means P(***A*** | ***B***, ***C***) = P(***A*** | ***C***)

  So since (*NU* ⊥ *Stanford* | *App*), we have
  P(*NU, Stanford, App*) =
  P(*NU* | *App*)*P(*Stanford* |*App*)*P(*App*)

  Say *App* ∈ {Good, Bad} and *School* ∈ {Yes, No, Wait}

  All we need is 4+4+1=**9** numbers
  (vs. 3*3*2-1=**17** for the full joint)

- Full joint has size **exponential** in # of r.v.s
  Conditional independence eliminates this!

Bayes' Rule

- $P(A \mid B) = P(B \mid A)\, P(A) \,/\, P(B)$
- Example:

  P(symptom| disease) = 0.95, P(symptom| $\neg$disease) = 0.05

  P(disease = 0.0001)

  P(disease | symptom)

  $$= \frac{P(\text{symptom} \mid \text{disease}) * P(\text{disease})}{P(\text{symptom})}$$

  $$= \frac{0.95*0.0001}{0.95*0.0001 + 0.05*0.9999} = \mathbf{0.002}$$

What have we learned?

- Probability – a calculus for dealing with uncertainty
  - Built from small set of axioms (ignore at your peril)
- Joint Distribution P(A, B, C, …)
  - Specifies probability of all combinations of r.v.s
  - Intractable to compute exhaustively for non-trivial problems
- Conditional Probability P(A | B)
  - Specifies probability of A given B
- Conditional Independence
  - Can radically reduce number of variable combinations we must assign unique probabilities to.
- Bayes' Rule