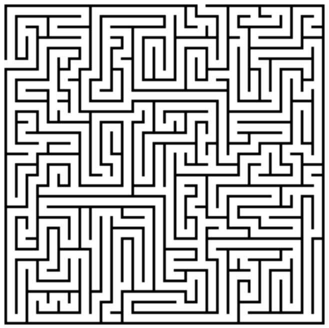
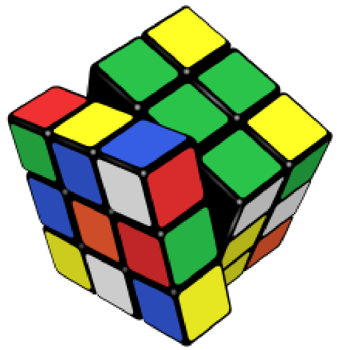
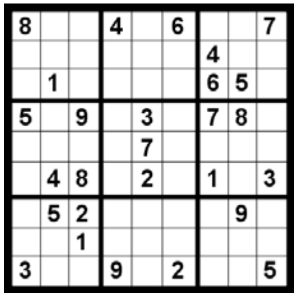
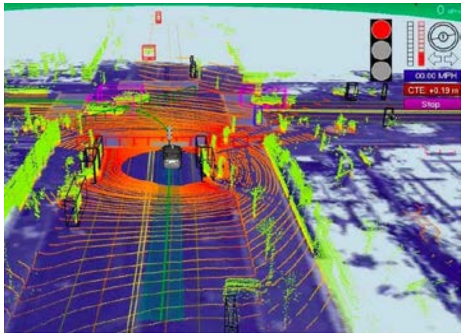
Where are we in CS 440?

- Now leaving: search, games, and planning


- Entering: probabilistic reasoning and machine learning




Probability: Review of main concepts

Motivation: Planning under uncertainty

- Recall: representation for planning
- **States** are specified as conjunctions of predicates
  - Start state: At(P1, CMI) $\wedge$ Plane(P1) $\wedge$ Airport(CMI) $\wedge$ Airport(ORD)
  - Goal state: At(P1, ORD)
- **Actions** are described in terms of preconditions and effects:
  - Fly(p, source, dest)
    - **Precond:** At(p, source) $\wedge$ Plane(p) $\wedge$ Airport(source) $\wedge$ Airport(dest)
    - **Effect:** $\neg$At(p, source) $\wedge$ At(p, dest)

Motivation: Planning under uncertainty

- Let action A_t = leave for airport t minutes before flight
  - Will A_t succeed, i.e., get me to the airport in time for the flight?
- Problems:
  - Partial observability (road state, other drivers' plans, etc.)
  - Noisy sensors (traffic reports)
  - Uncertainty in action outcomes (flat tire, etc.)
  - Complexity of modeling and predicting traffic
- Hence a purely logical approach either
  - Risks falsehood: “A_{25} will get me there on time,” or
  - Leads to conclusions that are too weak for decision making:
    - A_{25} will get me there on time if there's no accident on the bridge and it doesn't rain and my tires remain intact, etc., etc.
    - A_{1440} will get me there on time but I’ll have to stay overnight in the airport

Probability

Probabilistic assertions summarize effects of

- Laziness: reluctance to enumerate exceptions, qualifications, etc.
- Ignorance: lack of explicit theories, relevant facts, initial conditions, etc.
- Intrinsically random phenomena

Making decisions under uncertainty

- Suppose the agent believes the following:

  P(A_{25} gets me there on time) = 0.04

  …

  P(A_{120} gets me there on time) = 0.95

  P(A_{1440} gets me there on time) = 0.9999

- Which action should the agent choose?
  - Depends on preferences for missing flight vs. time spent waiting
  - Encapsulated by a *utility function*
- The agent should choose the action that maximizes the *expected utility*:

  P(A_t succeeds) * U(A_t succeeds) + P(A_t fails) * U(A_t fails)

- More generally: EU(A) = $\sum_{outcomes\ of\ A} P(outcome) U(outcome)$

- **Utility theory** is used to represent and infer preferences
- **Decision theory** = probability theory + utility theory

Monty Hall problem

- You're a contestant on a game show. You see three closed doors, and behind one of them is a prize. You choose one door, and the host opens one of the other doors and reveals that there is no prize behind it. Then he offers you a chance to switch to the remaining door. Should you take it?

http://en.wikipedia.org/wiki/Monty_Hall_problem

Monty Hall problem

- With probability 1/3, you picked the correct door, and with probability 2/3, picked the wrong door. If you picked the correct door and then you switch, you lose. If you picked the wrong door and then you switch, you win the prize.

- Expected utility of switching:

  EU(Switch) = (1/3) * 0 + (2/3) * Prize

- Expected utility of not switching:

  EU(Not switch) = (1/3) * Prize + (2/3) * 0

Where do probabilities come from?

- **Frequentism**
  - Probabilities are relative frequencies
  - For example, if we toss a coin many times, P(heads) is the proportion of the time the coin will come up heads
  - But what if we’re dealing with events that only happen once?
    - E.g., what is the probability that Team X will win the Superbowl this year?
    - “Reference class” problem
- **Subjectivism**
  - Probabilities are degrees of belief
  - But then, how do we assign belief values to statements?
  - What would constrain agents to hold consistent beliefs?

Probabilities and rationality

- Why should a rational agent hold beliefs that are consistent with axioms of probability?
  - For example, $P(A) + P(\neg A) = 1$
- If an agent has some degree of belief in proposition A, he/she should be able to decide whether or not to accept a bet for/against A (De Finetti, 1931):
  - If the agent believes that $P(A) = 0.4$, should he/she agree to bet \$4 that A will occur against \$6 that A will not occur?
- **Theorem:** An agent who holds beliefs inconsistent with axioms of probability can be convinced to accept a combination of bets that is guaranteed to lose them money

Random variables

- We describe the (uncertain) state of the world using ***random variables***
  - Denoted by capital letters
  - **R**: *Is it raining?*
  - **W**: *What's the weather?*
  - **D**: *What is the outcome of rolling two dice?*
  - **S**: *What is the speed of my car (in MPH)?*
- Just like variables in CSPs, random variables take on values in a *domain*
  - Domain values must be mutually exclusive and exhaustive
  - **R** in {True, False}
  - **W** in {Sunny, Cloudy, Rainy, Snow}
  - **D** in {(1,1), (1,2), … (6,6)}
  - **S** in [0, 200]

Events

- Probabilistic statements are defined over *events*, or sets of world states
  - *"It is raining"*
  - *"The weather is either cloudy or snowy"*
  - *"The sum of the two dice rolls is 11"*
  - *"My car is going between 30 and 50 miles per hour"*
- Events are described using propositions about random variables:
  - R = True
  - W = "Cloudy" $\vee$ W = "Snowy"
  - D $\in$ {(5,6), (6,5)}
  - $30 \leq S \leq 50$
- Notation: P(A) is the probability of the set of world states in which proposition A holds

Kolmogorov's axioms of probability

- For any propositions (events) A, B
  - $0 \le P(A) \le 1$
  - $P(True) = 1$ and $P(False) = 0$
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
    - Subtraction accounts for double-counting

- Based on these axioms, what is $P(\neg A)$?

- These axioms are sufficient to completely specify probability theory for *discrete* random variables
  - For continuous variables, need *density functions*

Atomic events

- ***Atomic event:*** a complete specification of the state of the world, or a complete assignment of domain values to all random variables
  - Atomic events are mutually exclusive and exhaustive

- E.g., if the world consists of only two Boolean variables *Cavity* and *Toothache*, then there are four distinct atomic events:

$$Cavity = false \wedge Toothache = false$$
$$Cavity = false \wedge Toothache = true$$
$$Cavity = true \wedge Toothache = false$$
$$Cavity = true \wedge Toothache = true$$

Joint probability distributions

- A ***joint distribution*** is an assignment of probabilities to every possible atomic event

Atomic event	**P**
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

  – Why does it follow from the axioms of probability that the probabilities of all possible atomic events must sum to 1?

Joint probability distributions

- Suppose we have a joint distribution of *n* random variables with domain sizes *d*
  - What is the size of the probability table?
  - Impossible to write out completely for all but the smallest distributions
- Notation:

  $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ refers to a single entry (atomic event) in the joint probability distribution table

  $P(X_1, X_2, \ldots, X_n)$ refers to the entire joint probability distribution table

Marginal probability distributions

- From the joint distribution P(X,Y) we can find the ***marginal distributions*** P(X) and P(Y)

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

P(Cavity)	
Cavity = false	?
Cavity = true	?

P(Toothache)	
Toothache = false	?
Toochache = true	?

Marginal probability distributions

- From the joint distribution P(X,Y) we can find the ***marginal distributions*** P(X) and P(Y)

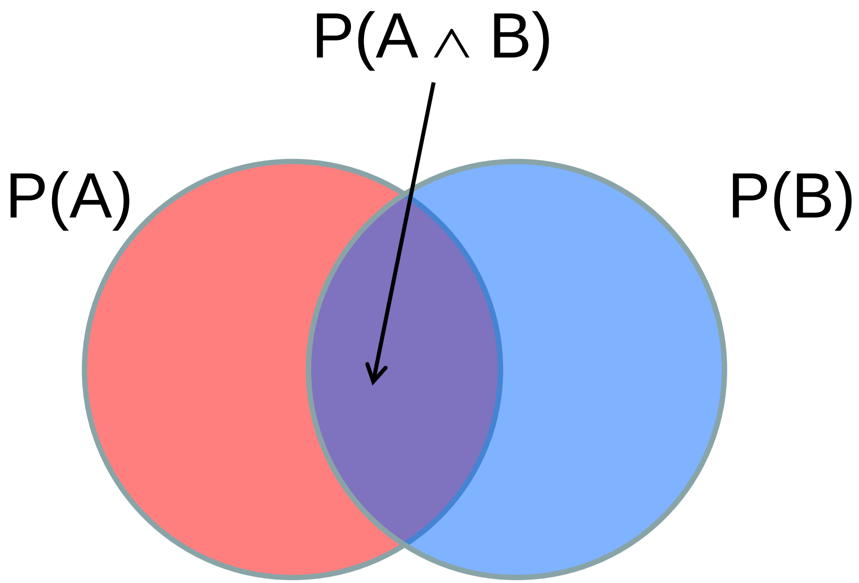
$$P(X = x) = P\big((X = x \wedge Y = y_1) \vee \ldots \vee (X = x \wedge Y = y_n)\big)$$

$$= P\big((x, y_1) \vee \ldots \vee (x, y_n)\big) = \sum_{i=1}^{n} P(x, y_i)$$

- General rule: to find P(X = x), sum the probabilities of all atomic events where X = x. This is called ***marginalization*** (we are *marginalizing out* all the variables except X)

Conditional probability

- Probability of cavity given toothache:
  P(*Cavity* = *true* | *Toothache* = *true*)
- For any two events A and B, $P(A \mid B) = \dfrac{P(A \wedge B)}{P(B)} = \dfrac{P(A, B)}{P(B)}$



Conditional probability

P(Cavity, Toothache)	
Cavity = *false* $\wedge$ *Toothache* = *false*	0.8
Cavity = *false* $\wedge$ *Toothache* = *true*	0.1
Cavity = *true* $\wedge$ *Toothache* = *false*	0.05
Cavity = *true* $\wedge$ *Toothache* = *true*	0.05

P(Cavity)	
Cavity = *false*	0.9
Cavity = *true*	0.1

P(Toothache)	
Toothache = *false*	0.85
Toothache = *true*	0.15

- What is P(*Cavity* = *true* | *Toothache* = *false*)?

  0.05 / 0.85 = 0.059

- What is P(*Cavity* = *false* | *Toothache* = *true*)?

  0.1 / 0.15 = 0.667

Conditional distributions

- A conditional distribution is a distribution over the values of one variable given fixed values of other variables

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

P(Cavity | Toothache = true)	
Cavity = false	0.667
Cavity = true	0.333

P(Cavity|Toothache = false)	
Cavity = false	0.941
Cavity = true	0.059

P(Toothache | Cavity = true)	
Toothache= false	0.5
Toothache = true	0.5

P(Toothache | Cavity = false)	
Toothache= false	0.889
Toothache = true	0.111

Normalization trick

- To get the whole conditional distribution P(X | Y = y) at once, select all entries in the joint distribution table matching Y = y and renormalize them to sum to one

P(Cavity, Toothache)	
Cavity = false $\wedge$ *Toothache = false*	0.8
Cavity = false $\wedge$ *Toothache = true*	0.1
Cavity = true $\wedge$ *Toothache = false*	0.05
Cavity = true $\wedge$ *Toothache = true*	0.05

Select

Toothache, Cavity = false	
Toothache= false	0.8
Toothache = true	0.1

Renormalize

P(Toothache | Cavity = false)	
Toothache= false	0.889
Toothache = true	0.111

Normalization trick

- To get the whole conditional distribution P(X | Y = y) at once, select all entries in the joint distribution table matching Y = y and renormalize them to sum to one
- Why does it work?

$$\frac{P(x,y)}{\sum_{x'} P(x',y)} = \frac{P(x,y)}{P(y)}$$ by marginalization

Product rule

- Definition of conditional probability: $P(A \mid B) = \dfrac{P(A, B)}{P(B)}$

- Sometimes we have the conditional probability and want to obtain the joint:

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

Product rule

- Definition of conditional probability: $P(A \mid B) = \dfrac{P(A,B)}{P(B)}$

- Sometimes we have the conditional probability and want to obtain the joint:

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

- The chain rule:

$$P(A_1, \ldots, A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \ldots P(A_n \mid A_1, \ldots, A_{n-1})$$
$$= \prod_{i=1}^{n} P(A_i \mid A_1, \ldots, A_{i-1})$$

The Birthday problem

- We have a set of *n* people. What is the probability that two of them share the same birthday?
- Easier to calculate the probability that *n* people *do not* share the same birthday

$$
\begin{aligned}
& P(B_1, \ldots B_n \text{ distinct}) \\
& = P(B_n \text{ distinct from } B_1, \ldots B_{n-1} \mid B_1, \ldots B_{n-1} \text{ distinct}) \\
& \quad P(B_1, \ldots B_{n-1} \text{ distinct}) \\
& = \prod_{i=1}^{n} P(B_i \text{ distinct from } B_1, \ldots B_{i-1} \mid B_1, \ldots B_{i-1} \text{ distinct})
\end{aligned}
$$

The Birthday problem

$$P(B_1, \ldots B_n \text{ distinct})$$

$$= \prod_{i=1}^{n} P(B_i \text{ distinct from } B_1, \ldots B_{i-1} \mid B_1, \ldots B_{i-1} \text{ distinct})$$

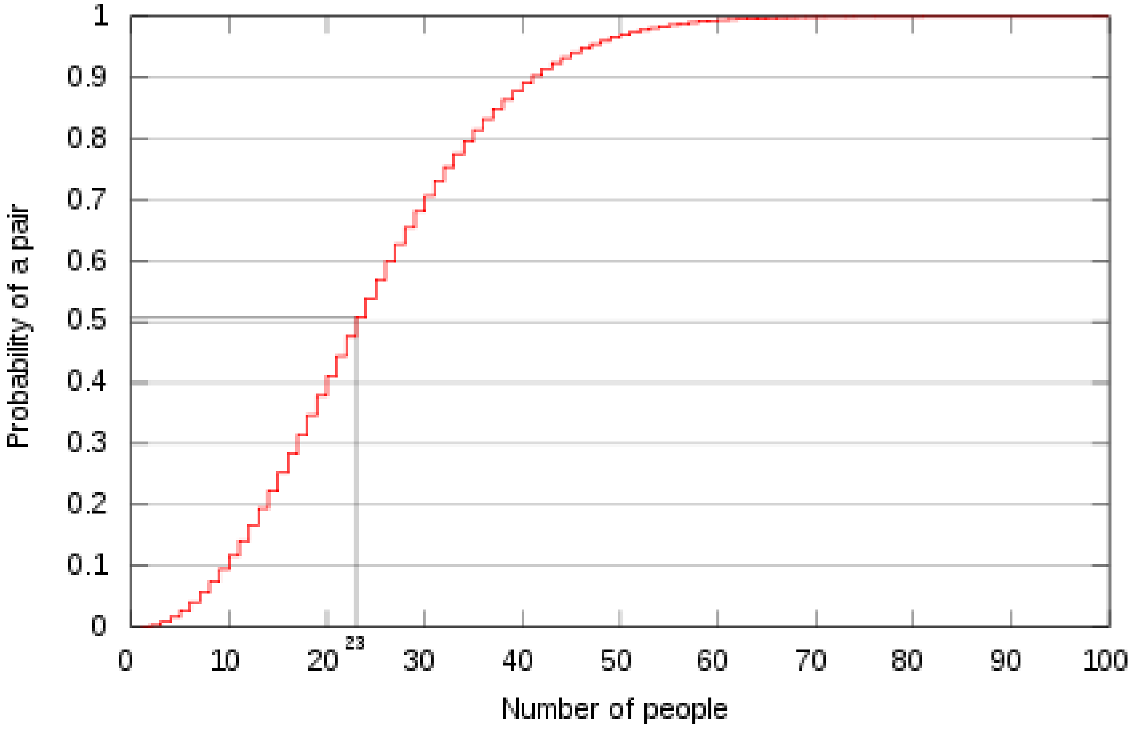
$$P(B_i \text{ distinct from } B_1, \ldots, B_{i-1} \mid B_1, \ldots, B_{i-1} \text{ distinct}) = \frac{365 - i + 1}{365}$$

$$P(B_1, \ldots, B_n \text{ distinct}) = \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{365 - n + 1}{365}$$

$$P(B_1, \ldots, B_n \text{ not distinct}) = 1 - \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{365 - n + 1}{365}$$

The Birthday problem

- For 23 people, the probability of sharing a birthday is above 0.5!



http://en.wikipedia.org/wiki/Birthday_problem

Independence

- Two events A and B are *independent* if and only if $P(A \wedge B) = P(A)\,P(B)$
  - In other words, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
  - This is an important simplifying assumption for modeling, e.g., *Toothache* and *Weather* can be assumed to be independent
- Are two *mutually exclusive* events independent?
  - No, but for mutually exclusive events we have $P(A \vee B) = P(A) + P(B)$
- **Conditional independence**: A and B are *conditionally independent* given C iff $P(A \wedge B \mid C) = P(A \mid C)\,P(B \mid C)$

Conditional independence: Example

- *Toothache*: boolean variable indicating whether the patient has a toothache
- *Cavity*: boolean variable indicating whether the patient has a cavity
- *Catch*: whether the dentist's probe catches in the cavity

- If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether he/she has a toothache

  P(*Catch* | *Toothache, Cavity*) = P(*Catch* | *Cavity*)
- Therefore, *Catch* is conditionally independent of *Toothache* given *Cavity*
- Likewise, *Toothache* is conditionally independent of *Catch* given *Cavity*

  P(*Toothache* | *Catch, Cavity*) = P(*Toothache* | *Cavity*)
- Equivalent statement:

  P(*Toothache, Catch* | *Cavity*) = P(*Toothache* | *Cavity*) P(*Catch* | *Cavity*)

Conditional independence: Example

- How many numbers do we need to represent the joint probability table P(*Toothache, Cavity, Catch*)?

  $2^3 - 1 = 7$ independent entries

- Write out the joint distribution using chain rule:

  P(*Toothache, Catch, Cavity*)

  = P(*Cavity*) P(*Catch* | *Cavity*) P(*Toothache* | *Catch, Cavity*)

  = P(*Cavity*) P(*Catch* | *Cavity*) P(*Toothache* | *Cavity*)

- How many numbers do we need to represent these distributions?

  1 + 2 + 2 = 5 independent numbers

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in *n* to linear in *n*