CS 440/ECE 448 Lecture 12: Probability

Slides by Svetlana Lazebnik, 9/2016

Modified by Mark Hasegawa-Johnson, 2/2019

Outline

- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events
  - Joint, Marginal, and Conditional
  - Independent vs. Conditionally Independent events
- Classification Using Probabilities

Outline

- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events
  - Joint, Marginal, and Conditional
  - Independence and Conditional Independence
- Classification Using Probabilities

Motivation: Planning under uncertainty

- Recall: representation for planning
- **States** are specified as conjunctions of predicates
  - Start state: At(Me, UIUC) $\wedge$ TravelTime(35min,UIUC,CMI) $\wedge$ Now(12:45)
  - Goal state: At(Me, CMI, 15:30)
- **Actions** are described in terms of preconditions and effects:
  - Go(t, src, dst)
    - **Precond:** At(Me,src) $\wedge$ TravelTime(dt,src,dst) $\wedge$ Now($\leq$t)
    - **Effect:** At(Me, dst, t+dt)

Making decisions under uncertainty

- Suppose the agent believes the following:

  P(Go(deadline-25) gets me there on time) = 0.04

  P(Go(deadline-90) gets me there on time) = 0.70

  P(Go(deadline-120) gets me there on time) = 0.95

  P(Go(deadline-180) 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*:

$$\text{Prob}(\text{A succeeds}) \times \text{Utility}(\text{A succeeds}) + \text{Prob}(\text{A fails}) \times \text{Utility}(\text{A fails})$$

Making decisions under uncertainty

- More generally: the expected utility of an action is defined as:

$$\mathrm{E[Utility|Action]} = \sum_{outcomes} P(outcome|action) Utility(outcome)$$

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

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 an event that has never happened before?
    - What is the probability that the Earth will warm by 0.15 degrees this year?
- **Subjectivism**
  - Probabilities are degrees of belief
  - But then, how do we assign belief values to statements?
  - In practice: models. Represent an *unknown event* as a series of *better-known events*
- A theoretical problem with Subjectivism:

Why do "beliefs" need to follow the laws of probability?

The Rational Bettor Theorem

- Why should a rational agent hold beliefs that are consistent with axioms of probability?
  - For example, $P(A) + P(\neg A) = 1$

- Suppose an agent believes that $P(A)=0.7$, and $P(\neg A)=0.7$
- Offer the following bet: if A occurs, agent wins \$100. If A doesn't occur, agent loses \$105. Agent believes $P(A)>100/(100+105)$, so agent accepts the bet.
- Offer another bet: if ¬A occurs, agent wins \$100. If ¬A doesn't occur, agent loses \$105. Agent believes $P(\neg A)>100/(100+105)$, so agent accepts the bet. Oops…

- **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

Outline

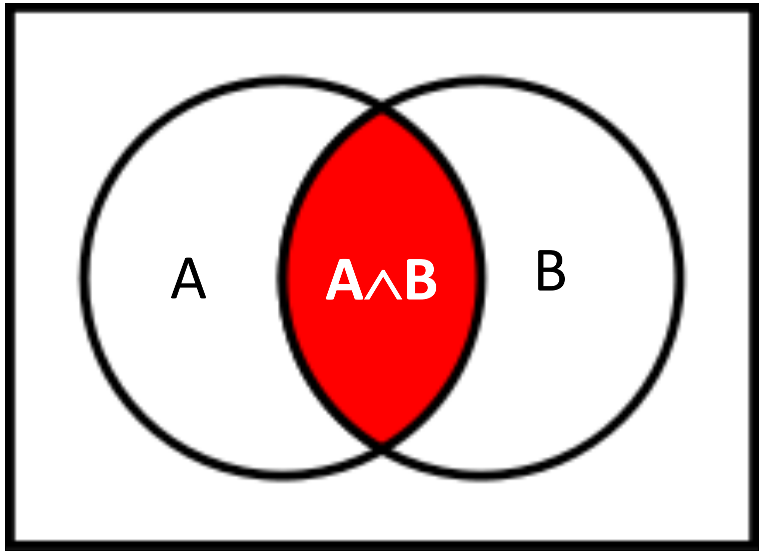
- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events
  - Joint, Marginal, and Conditional
  - Independence and Conditional Independence
- Classification Using Probabilities

Events

- Probabilistic statements are defined over *events*, or sets of world states
  - *A = "It is raining"*
  - *B = "The weather is either cloudy or snowy"*
  - *C = "I roll two dice, and the result is 11"*
  - *D = "My car is going between 30 and 50 miles per hour"*
- An EVENT is a SET of OUTCOMES
  - B = { outcomes : cloudy OR snowy }
  - C = { outcome tuples (d1,d2) such that d1+d2 = 11 }
- Notation: P(A) is the probability of the set of world states (outcomes) in which proposition A holds

Kolmogorov's axioms of probability

- For any propositions (events) A, B
  - $0 \leq P(A) \leq 1$
  - $P(True) = 1$ 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*

Outcomes = Atomic events

- ***OUTCOME or ATOMIC EVENT:*** is 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 outcomes:

  Outcome #1: $\neg Cavity \wedge \neg Toothache$
  Outcome #2: $\neg Cavity \wedge Toothache$
  Outcome #3: $Cavity \wedge \neg Toothache$
  Outcome #4: $Cavity \wedge Toothache$

Outline

- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events
  - Joint, Marginal, and Conditional
  - Independence and Conditional Independence
- Classification Using Probabilities

Joint probability distributions

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

Atomic event	**P**
¬Cavity ∧ ¬Toothache	0.8
¬Cavity ∧ Toothache	0.1
Cavity ∧ ¬Toothache	0.05
Cavity ∧ Toothache	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, each of which takes values from a domain of size D:
  - What is the size of the probability table?
  - Impossible to write out completely for all but the smallest distributions

Marginal distributions

- The marginal distribution of event X_k is just its probability, $P(X_k)$.
- If you're given the joint distribution, $P(X_1, X_2, \ldots, X_N)$, from it, how can you calculate $P(X_k)$?
- You calculate $P(X_k)$ from $P(X_1, X_2, \ldots, X_N)$ by **marginalizing**.

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 ∧ ¬Toothache	0.8
¬Cavity ∧ Toothache	0.1
Cavity ∧ ¬Toothache	0.05
Cavity ∧ Toothache	0.05

P(Cavity)	
¬Cavity	0.9
Cavity	0.1

P(Toothache)	
¬Toothache	0.85
Toochache	0.15

Conditional distributions

- The conditional probability of event X_k, given event X_j, is the probability that X_k has occurred if you already know that X_j has occurred.
- The conditional distribution is written $P(X_k | X_j)$.
- The probability that both X_j and X_k occurred was, originally, $P(X_j, X_k)$.
- But now you know that X_j has occurred. So all of the other events are no longer possible.
  - Other events: probability used to be $P(\neg X_j)$, but now their probability is 0.
  - Events in which X_j occurred: probability used to be $P(X_j)$, but now their probability is 1.
- So we need to renormalize: the probability that both X_j and X_k occurred, GIVEN that X_j has occurred, is $P(X_k | X_j)=P(X_j, X_k)/P(X_j)$.

Conditional Probability: renormalize (divide)

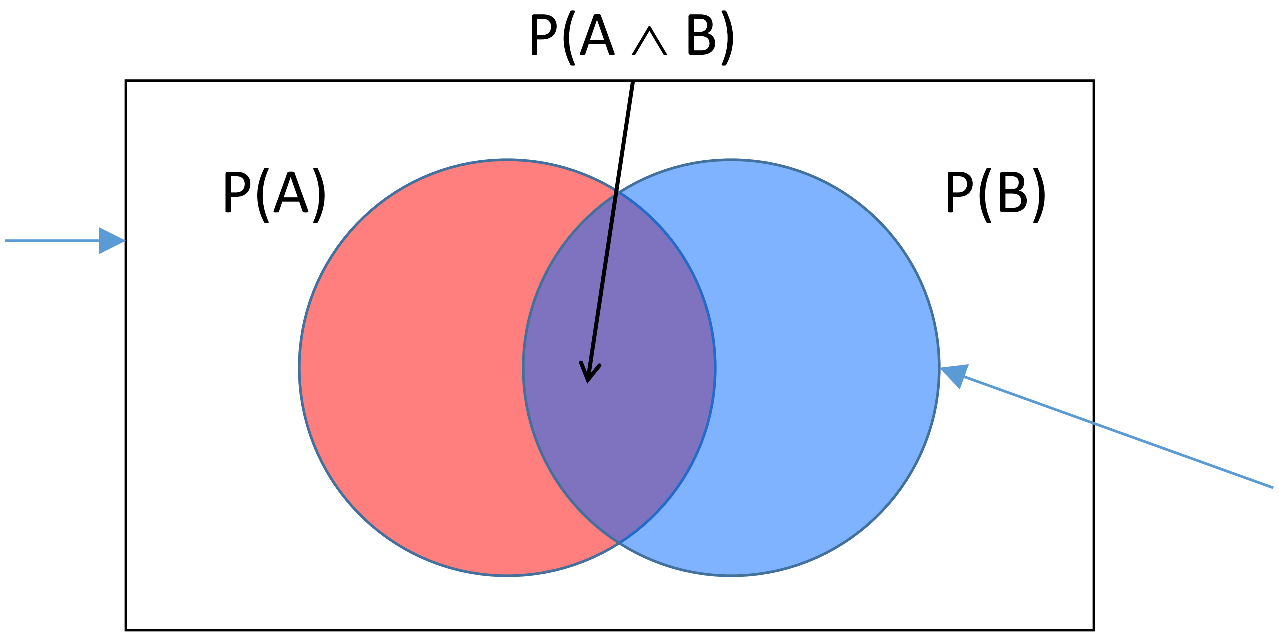
- Probability of cavity given toothache:

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

- For any two events A and B,

$$P(A\,|\,B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(A,B)}{P(B)}$$

The set of all possible events used to be this rectangle, so the whole rectangle used to have probability=1.



Now that we know B has occurred, the set of all possible events = the set of events in which B occurred. So we renormalize to make the area of this circle = 1.

Conditional probability

P(Cavity, Toothache)	
¬*Cavity* ∧ ¬*Toothache*	0.8
¬*Cavity* ∧ *Toothache*	0.1
Cavity ∧ ¬*Toothache*	0.05
Cavity ∧ *Toothache*	0.05

P(Cavity)	
¬*Cavity*	0.9
Cavity	0.1

P(Toothache)	
¬*Toothache*	0.85
Toochache	0.15

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

  p(*Cavity*|¬*Toothache*) = 0.05/0.85 = 1/17

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

  p(¬*Cavity*|*Toothache*) = 0.1/0.15 = 2/3

Conditional distributions

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

P(Cavity, Toothache)	
¬Cavity ∧ ¬Toothache	0.8
¬Cavity ∧ Toothache	0.1
Cavity ∧ ¬Toothache	0.05
Cavity ∧ Toothache	0.05

P(Cavity | Toothache)	
¬Cavity	0.667
Cavity	0.333

P(Cavity|¬Toothache)	
¬Cavity	0.941
Cavity	0.059

P(Toothache | Cavity)	
¬Toothache	0.5
Toochache	0.5

P(Toothache | ¬Cavity)	
¬Toothache	0.889
Toochache	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 ∧ ¬Toothache	0.8
¬Cavity ∧ Toothache	0.1
Cavity ∧ ¬Toothache	0.05
Cavity ∧ Toothache	0.05

Select

Toothache, Cavity = false	
¬Toothache	0.8
Toochache	0.1

Renormalize

P(Toothache | Cavity = false)	
¬Toothache	0.889
Toochache	0.111

Normalization trick

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

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

by marginalization

Product rule

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

- 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:

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

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

Outline

- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events, and Random Variables
  - Joint, Marginal, and Conditional
  - Independence and Conditional Independence
- Classification Using Probabilities

Independence ≠ Mutually Exclusive

- Two events A and B are *independent* if and only if
  $p(A \wedge B) = p(A, B) = p(A)\, p(B)$
  - 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! Quite the opposite! If you know A happened, then you know that B _didn't_ happen!!
    $p(A \vee B) = p(A) + p(B)$

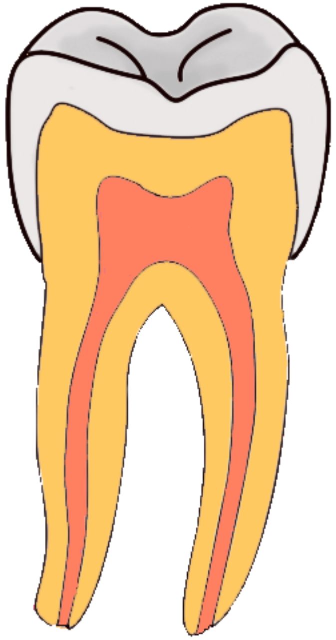
Independence ≠ Conditional Independence

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

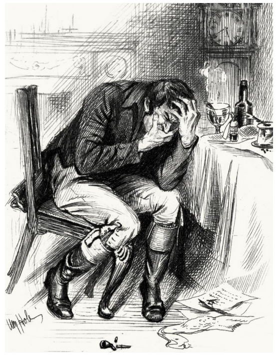
By William Brassey Hole(Died:1917)

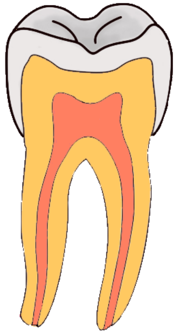


By Aduran, CC-SA 3.0

By Dozenist, CC-SA 3.0

These Events are not Independent





- If the patient has a toothache, then it's likely he has a cavity. Having a cavity makes it more likely that the probe will catch on something.

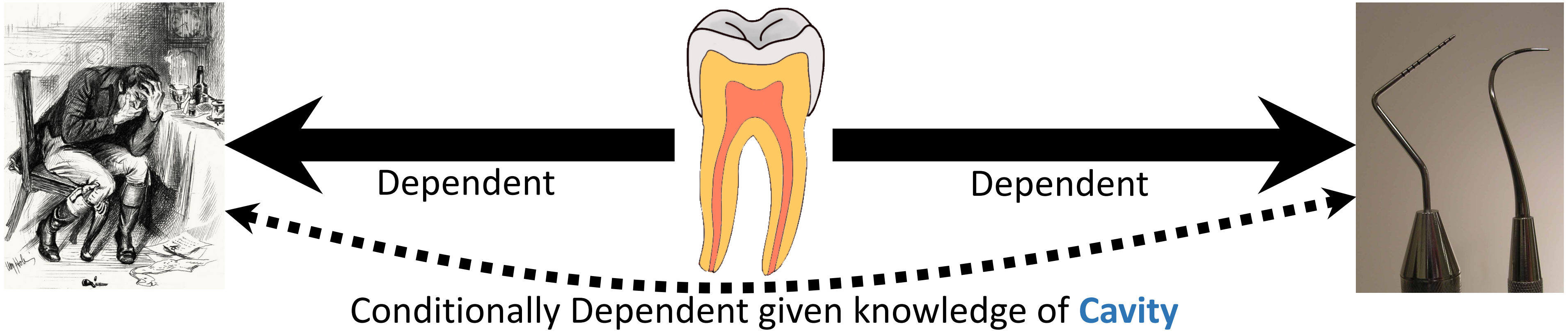
$$P(Catch|Toothache) > P(Catch)$$

- If the probe catches on something, then it's likely that the patient has a cavity. If he has a cavity, then he might also have a toothache.

$$P(Toothache|Catch) > P(Toothache)$$

- So Catch and Toothache are not independent

…but they are Conditionally Independent

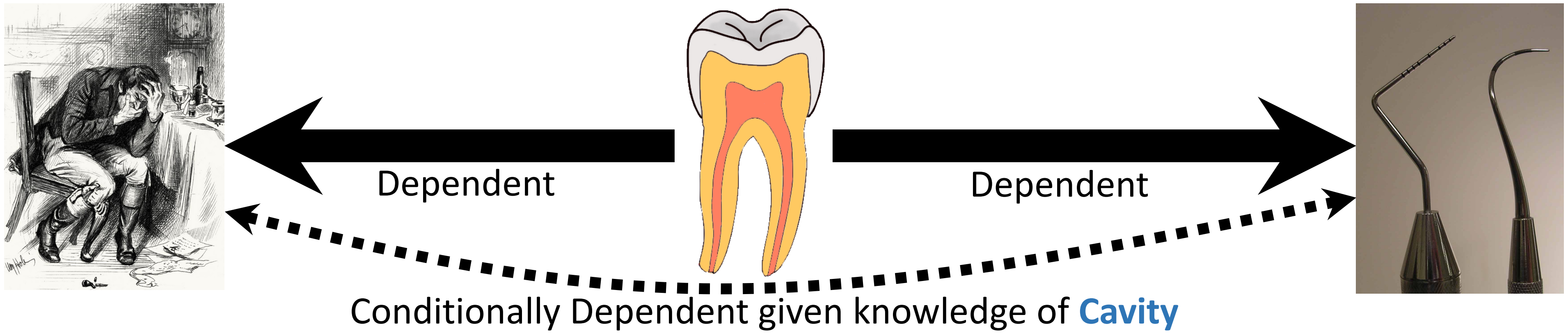


- Here are some reasons the probe might not catch, despite having a cavity:
  - The dentist might be really careless
  - The cavity might be really small
- Those reasons have nothing to do with the toothache!

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

- **Catch** and **Toothache** are **<u>conditionally independent</u>** given knowledge of **Cavity**

...but they are Conditionally Independent



These statements are all equivalent:

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

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

$$P(Toothache, Catch|Cavity) = P(Toothache|Cavity)\,P(Catch|Cavity)$$

...and they all mean that **Catch** and **Toothache** are **conditionally independent** given knowledge of **Cavity**

Outline

- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events
  - Joint, Marginal, and Conditional
  - Independent vs. Conditionally Independent events
- Classification Using Probabilities

Classification using probabilities

- Suppose you know that you have a toothache.
- Should you conclude that you have a cavity?
- Goal: make a decision that **minimizes your probability of error**.
- Equivalent: make a decision that **maximizes the probability of being correct**. This is called a MAP (maximum a posteriori) decision. You decide that you have a toothache if and only if

$$P(Cavity|Toothache) > P(\neg Cavity|Toothache)$$

Bayesian Decisions

- What if we don't know $P(Cavity|Toothache)$? Instead, we only know $P(Toothache|Cavity), P(Cavity)$, and $P(Toothache)$?
- Then we choose to believe we have a Cavity if and only if

$$P(Cavity|Toothache) > P(\neg Cavity|Toothache)$$

Which can be re-written as

$$\frac{P(Toothache|Cavity)P(Cavity)}{P(Toothache)} > \frac{P(Toothache|\neg Cavity)P(\neg Cavity)}{P(Toothache)}$$

Outline

- Motivation: Why use probability?
- Review of Key Concepts
  - Outcomes, Events
  - Joint, Marginal, and Conditional
  - Independent vs. Conditionally Independent events
- Classification Using Probabilities