Artificial Intelligence

CPSC 481

Stochastic Methods for Reasoning in Uncertain Situations

Part A

Lecture Overview

- Concepts of reasoning in uncertain situations
- Review of probability theory
- Probabilistic reasoning
  - Probabilistic inference using Bayesian theorem

Uncertainty

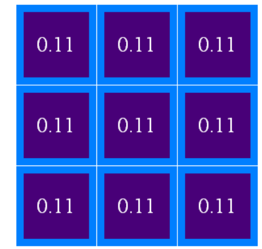
- General situation:

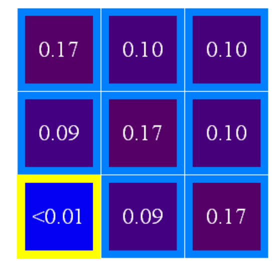
  - **Observed variables (evidence/fact)**: Agent knows **certain** things about the state of the world (e.g., sensor readings or symptoms or weather now)

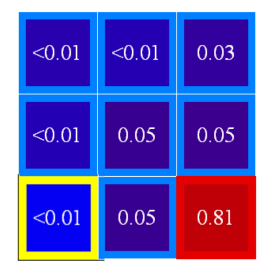
  - **Unobserved variables**: Agent needs to **reason** about **other** aspects (e.g. where an object is or what disease is present or weather tomorrow)

  - **Model**: Agent knows **something** about how the known variables **relate** to the unknown variables

- Probabilistic reasoning gives us a framework for managing our beliefs and knowledge

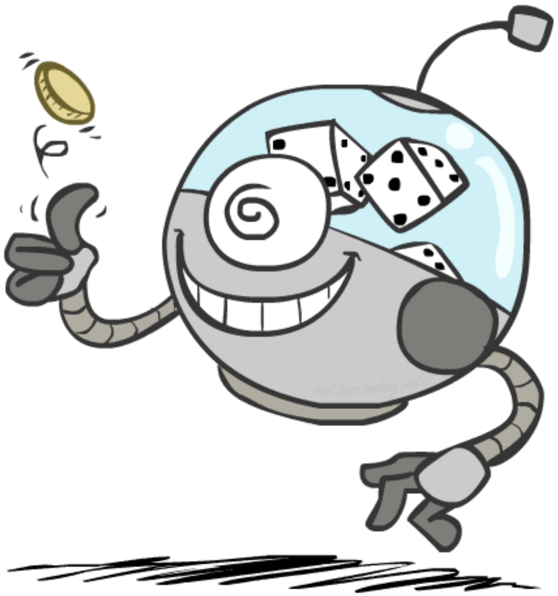






Random Variables

- A random variable is **some aspect of the world** about which we (may) have **uncertainty**
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost (Pacman project)?

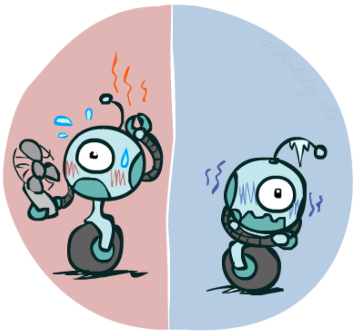


- We denote random variables with **capital letters**

- Random variables have **domains**
  - R in {true, false}   (often write as {+r, -r})
  - T in {hot, cold}
  - D in $[0, \infty)$
  - L in possible locations, maybe {(0,0), (0,1), …}

Probability Distributions

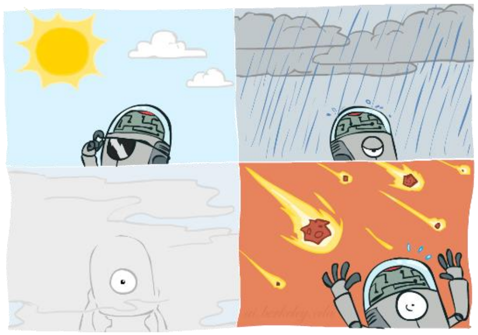
- Associate a probability with each value
  - Temperature:



$$P(T)$$

T	P
hot	0.5
cold	0.5

  - Weather:



$$P(W)$$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\ldots$$

OK *if* all domain entries are unique

- A **distribution** is a **TABLE** of probabilities of values

- A **probability** (lower case value) is a **single number**

$$P(W = rain) = 0.1$$

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint Distributions

- A *joint distribution* over **a set of random variables**: $X_1, X_2, \ldots X_n$ specifies a real number for each *outcome*:

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

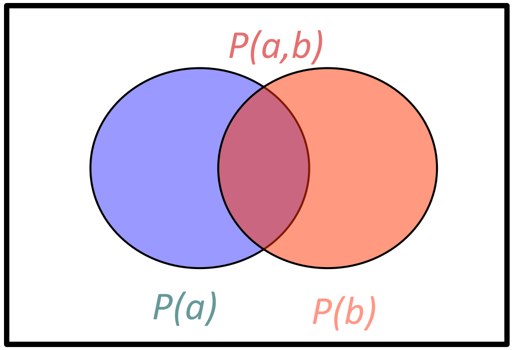
  - Must obey: $P(x_1, x_2, \ldots x_n) \geq 0$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

- Size of distribution if n variables with domain sizes d?
  - For all but the smallest distributions, impractical to write out!



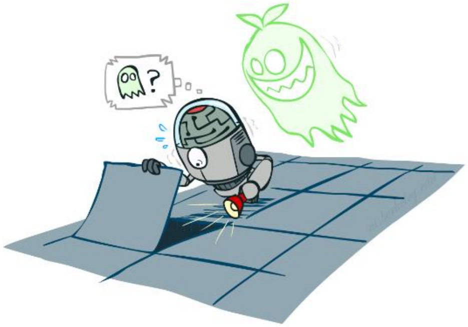
Probabilistic Models

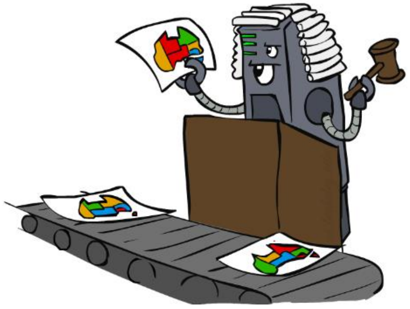
- A probabilistic model is a **joint distribution** over **a set of random variables**

- Probabilistic models:
  - (Random) **variables** with **domains**
  - **Joint distributions**: say whether outcomes are likely
  - ***Normalized**:* sum to 1.0
  - Ideally: only certain variables directly interact

Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3





Events and Sample Space

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?

- Typically, the events we care about are *partial outcomes*, like P(T=hot)

- The set of **all possible outcomes** of an event E is the sample space S for event E
  - The sample space for it's hot = {it's hot and sun, it's hot and rain}

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Student Participation: Events

- P(+x, +y) ?

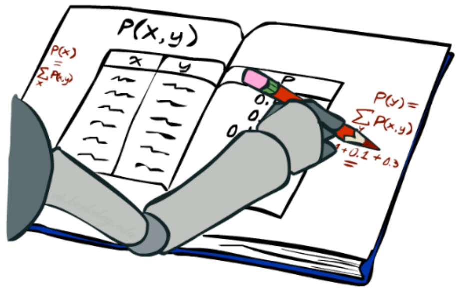
- P(+x) ?

- P(-y OR +x) ?

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal Distributions



- Marginal distributions are **sub-tables** which **eliminate variables**
- Marginalization (summing out): Combine collapsed rows by adding

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(s) = \sum_t P(t, s)$$

$P(W)$

W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Student Participation : Marginal Distributions

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$$P(x) = \sum_y P(x, y)$$

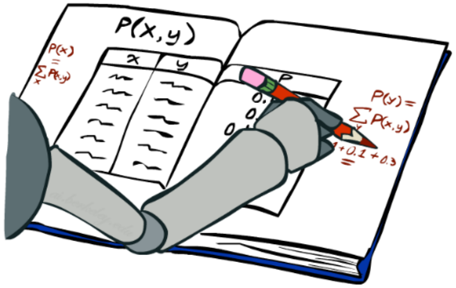
$$P(y) = \sum_x P(x, y)$$

$P(X)$

X	P
+x	
-x	

$P(Y)$

Y	P
+y	
-y	



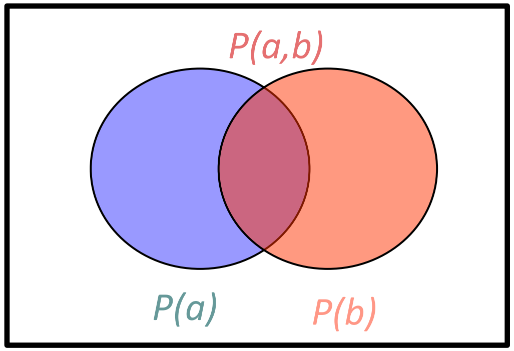
Conditional Probabilities

- a measure of the **probability** of an **event occurring** given that **another event** has (by assumption, presumption, assertion or evidence) **occurred**
- P(a | b), b is the event occurred, a is the event occurring
- P(hot | sun) = ?
- P(Sick | Cough) = ?
- P(Cough | Sick) = ?

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$P(T,W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$

$$= 0.2 + 0.3 \quad = 0.5$$

Student Participation: Conditional Probabilities

$P(X,Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- P(+x | +y) ?
- P(-x | +y) ?
- P(-y | +x) ?

Conditional Distributions

- Conditional distributions are probability **distributions** over **some variables** given fixed values of others

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Joint Distribution

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

How to get conditional distributions from join distributions and vice versa?

Normalization Trick

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W = s | T = c) = \frac{P(W = s, T = c)}{P(T = c)}$$
$$= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$
$$= \frac{0.2}{0.2 + 0.3} = 0.4$$

$$P(W = r | T = c) = \frac{P(W = r, T = c)}{P(T = c)}$$
$$= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)}$$
$$= \frac{0.3}{0.2 + 0.3} = 0.6$$

$$P(W | T = c)$$

W	P
sun	0.4
rain	0.6

Normalization Trick

$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

SELECT the joint probabilities matching the **evidence**

$P(c, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to **one**)

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

- Why does this work? Sum of selection is P(evidence)! (P(T=c), here)

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)}$$

Student Participation: Normalization Trick

- P(X | Y=-y) ?

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

SELECT the joint probabilities matching the evidence

NORMALIZE the selection (make it sum to one)

Exercise Questions for Probability

- **Scenario**:
  - We toss a **fair coin three successive** times. What is the probability of seeing **more heads** than tails coming up when the **first toss is a head**?
- Q1: What is the sample space?
- Q2: What are the events and the probabilities of each event?
- Q3: What are the random variables?
- Q4: What is the probability distribution for "number of heads"?
- Q5: What is the conditional probability and how can we compute it?

Inference using Conditional Probability

Scenario:

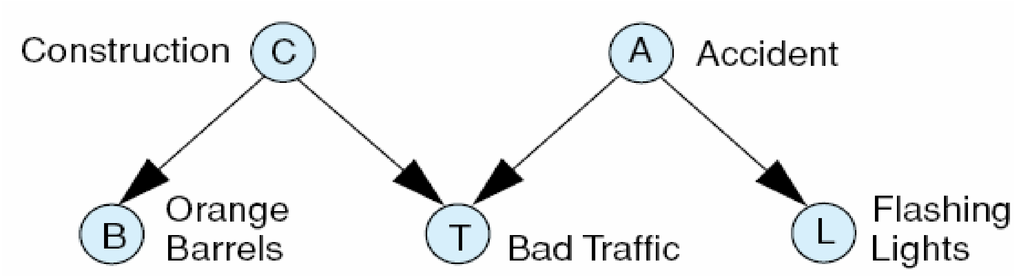
Suppose you are **driving** the interstate highway system and realize you are gradually **slowing down** because of increased **traffic congestion**.

You begin to search for **possible explanations** of the slowdown. Could it be **road construction**? Has there been an **accident**? Perhaps there are other possible explanations.

After a few minutes you come across **orange barrels** you determine that the best explanation is **road construction**.

Similarly, if you would have seen **flashing lights** in the distance ahead, such as from a police vehicle or an ambulance, the best explanation given this evidence would be a traffic **accident**.

Inference using Conditional Probability



The traffic problem of Bayesian representation with potential explanations

C	T	p
t	t	.3
t	f	.2
f	t	.1
f	f	.4

C is true = .5

T is true = .4

build a **joint probability** distribution for the road construction and bad traffic relationship

Question1: What is the probability of road construction, p(C=t)?

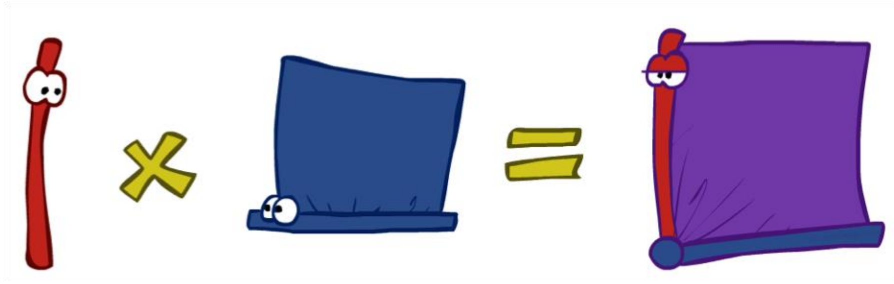
Question2: What is the probability of road construction given the fact that we have bad traffic, p(C=t|T=t)?

Question3: If we see the presence of orange barrels, will the probability of be increased?

The Product Rule

- Sometimes have conditional distributions but want the joint

$$P(y)P(x|y) = P(x,y) \qquad \Longleftrightarrow \qquad P(x|y) = \frac{P(x,y)}{P(y)}$$



The Product Rule

$$P(y)P(x|y) = P(x,y)$$

- Example:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\Longleftrightarrow$

$P(D,W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

The Chain Rule

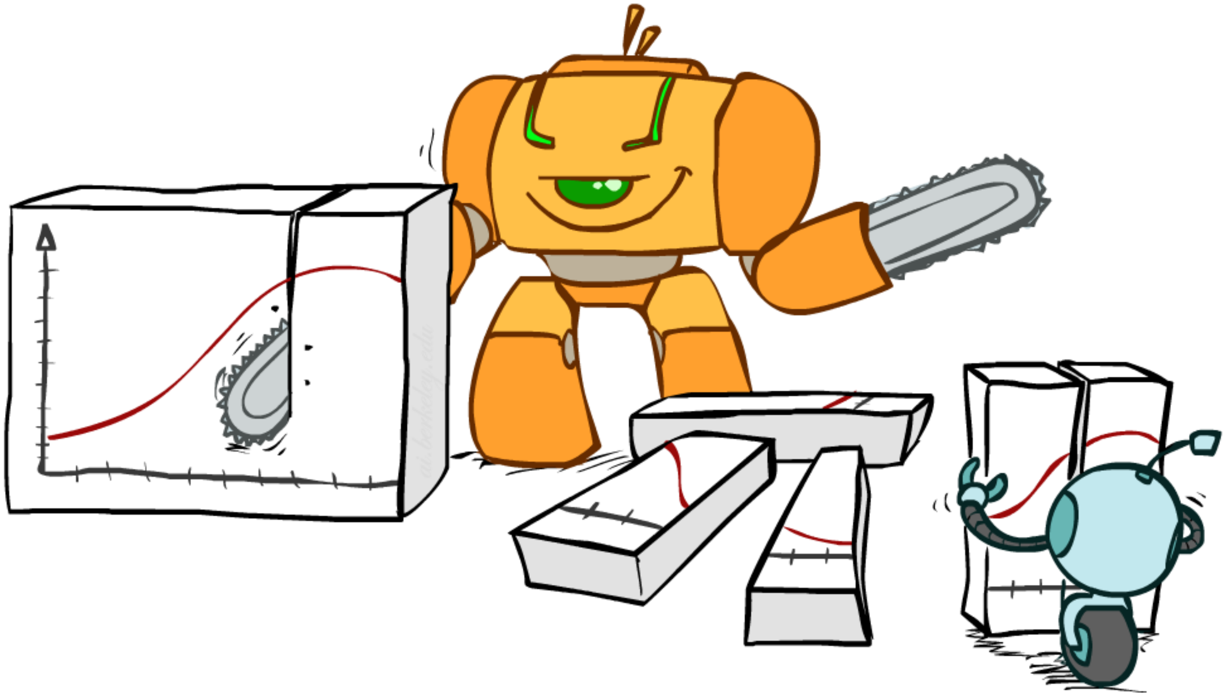
- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \ldots x_n) = \prod_i P(x_i|x_1 \ldots x_{i-1})$$

- Why is this always true?

Bayes Rule



Bayes' Rule

- Two ways to factor a joint distribution over two variables: **p(x, y) = p(y, x)**

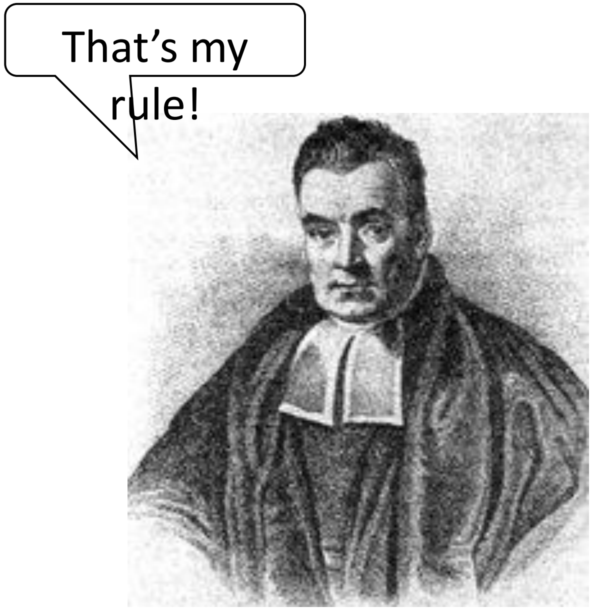
$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

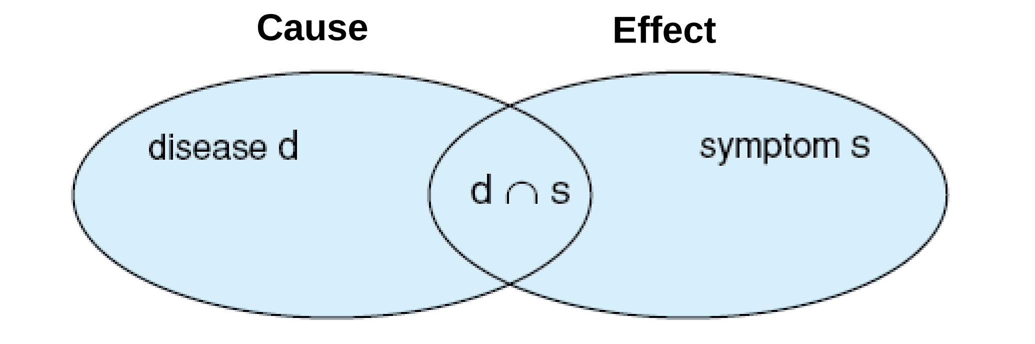
$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- **Why is this at all helpful?**
  - Lets us build one conditional from its reverse
  - **Often one conditional is tricky but the other one is simple**
  - Foundation of many systems

- In the running for most important AI equation!



Bayes' Rule from Conditional Probability



Medical diagnosis system: study the relationship between disease and symptom

$$p(d|s) = p(d \cap s) \, / \, p(s).$$

$$p(s|d) = p(s \cap d) \, / \, p(d).$$

$$p(s \cap d) = p(s|d) \, p(d).$$

$$p(d|s) = \frac{p(s|d)p(d)}{p(s)}$$

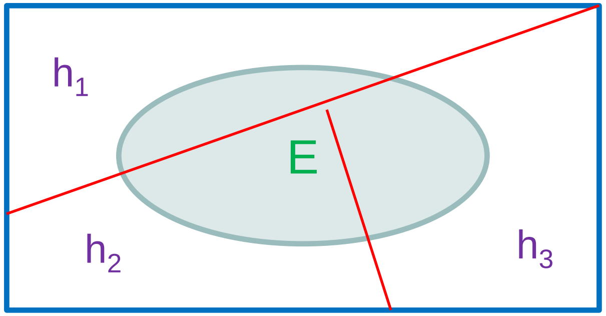
Bayes' rule

Interpretation of Bayesian rule:
p(d|s) means that given symptom **s**, the probability of the disease **d**, or the probability of the disease **d** to cause the symptom **s**.

In many cases, it is difficult to compute it. So we can change the calculations of p(d|s) as a function of p(s|d),
$p(d|s) = \frac{p(s|d)p(d)}{p(s)}$ because when we know what the **disease** (or cause) is, telling those **symptoms** (or effect) of the disease, is much **easier** than figuring out a disease based on symptoms.

Total Probability Theorem

- **Assume** that the entire sample space and event E within it are **partitioned** by **the set of disjoint** (discrete) **hypotheses** h_i (union of h_i is entire sample space). For example, see the Venn diagram below for a sample space with three hypotheses and E.



- $E = (h_1 \cap E) \cup (h_2 \cap E) \cup \ldots \cup (h_n \cap E)$
- *By total probability theorem:* $\mathbf{p(E)} = \sum_i p(E|h_i)p(h_i)$

- We can derive $p(E) = p(E|h_1)p(h_1)+p(E|h_2)p(h_2)+\ldots+p(E|h_n)p(h_n)$ from $p(E)=p((E\cap h_1)\cup(E\cap h_2)\cup\ldots\cup(E\cap h_n))=p(E\cap h_1)+p(E\cap h_2)\cup\ldots\cup p(E\cap h_n) - p(E \cap h_1 \cap Eh_2 \cap\ldots \cap h_n) = p(E\cap h_1)+p(E\cap h_2)\cup\ldots\cup p(E\cap h_n)$ since h_i are disjoints and the set of hypotheses h_i partition E so $h_i \cap E=\{\}$.

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

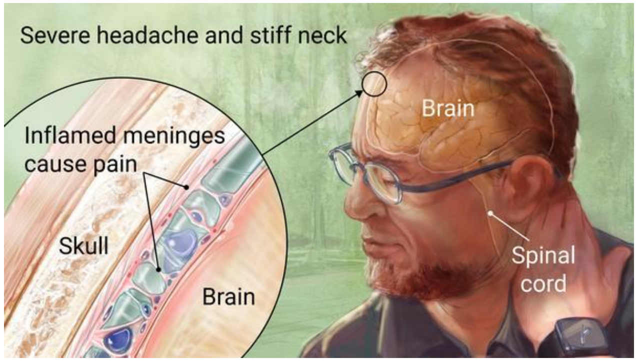
$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{Example givens}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.999}$$

  - Note: posterior probability of meningitis still **very small**
  - Note: you should still get stiff necks checked out! Why?



Student Participation: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is P(W | dry) ?

General Form of Bayes' Theorem

We assume the set of hypotheses H partition the evidence set E.

$$p(H|e) = \frac{p(e|H)p(H)}{p(e)} \quad \text{where } p(e) = \sum p(e|Hi)p(Hi)$$

Total probability theorem

Physical meanings of these probabilities:

p(H): How probable was our hypothesis before observing the evidence?

Likelihood, p(e|H): How probable is the evidence, given that our hypothesis is true?

Marginal, p(e): How probable is the new evidence under all possible hypotheses?

Conditional, p(H|e): How probable is our hypothesis, given the observed evidence?
Not directly computable

Probabilistic inference

- **Maximum Likelihood Hypothesis**
  - **Arg $\max(h_i)p(E|h_i)p(h_i)$**
- From $p(h_i|e) = p(e|h_i)p(h_i)/p(e)$, if we want to get the maximum value over all h_i of $p(e|h_i)p(h_i)$, then we can drop $p(e)$ for easier computation and choose h_i that gives maximum likelihood.
- Choose a hypothesis/class whose probability is the highest: P(+m|+s) = ?, P(-m|+s) = ?

Example using Bayes' Theorem

- **Scenario**: Assume only three car dealers you will go since they all sell a1 model of car. Probability that you will go to each dealer is $p(d1)=0.2$, $p(d2)=0.4$, $p(d3)=0.4$. Once you are at a dealer, the probability to purchase a particular model of car a1 at d1,d2,and d3 is 0.2, 0.4, and 0.3, respectively.

- **Question**: You purchased an a1 model of car. What is the probability that you purchased it at the dealer d2?

- **Answer**: Given that you purchased an **a1** model of car, you bought it from **d2** out of three possible dealers. Basically you need to calculate **p(d2|a1)** based on the evidences you are likely to go to each dealer.

$$p(d2|a1) = [p(a1|d2)p(d2)] / [p(a1|d1)p(d1)+p(a1|d2)p(d2)+p(a1|d3)p(d3)]$$

$$= [(0.4)(0.4)]/[(0.2)(0.2)+(0.4)(0.4)+(0.4)(0.3) = 0.16/0.32 = 0.5$$

Independence

- Two variables are ***independent***: P(x | y)= P(x)
- in a joint distribution:

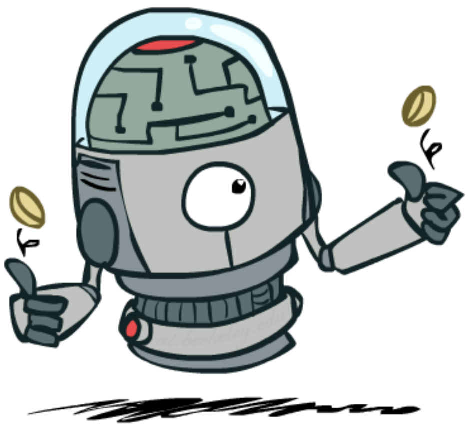
$$P(x, y) = P(x|y)P(y)$$

$$P(X, Y) = P(X)P(Y)$$

$$\forall x, y \, P(x, y) = P(x)P(y)$$

$$X \perp\!\!\!\perp Y$$

  - Says the joint distribution *factors* into a product of two simple ones
  - Usually variables aren't independent!



- Can use independence as a *modeling assumption*
  - Independence can be a simplifying assumption
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity}?

Example: Independence

- N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

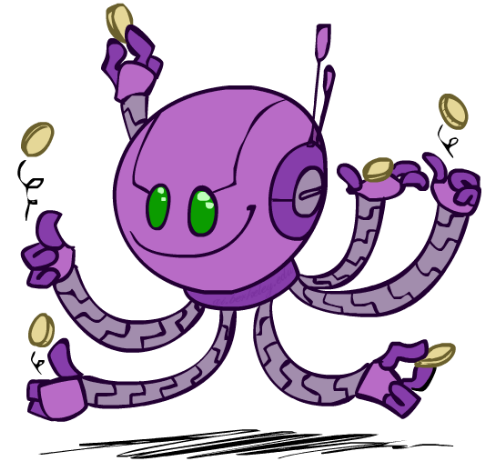
$P(X_2)$

H	0.5
T	0.5

$\ldots$

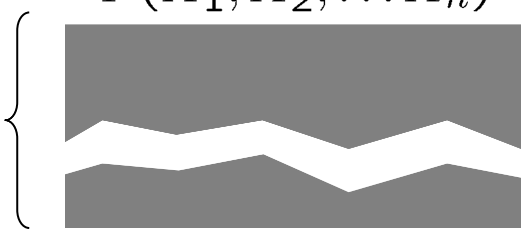
$P(X_n)$

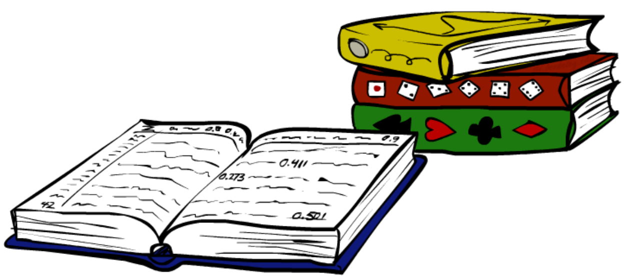
H	0.5
T	0.5



$P(X_1, X_2, \ldots X_n)$

2^n





Naïve Bayes Approach

- If we assume each feature F_i is independent of other features F_j, $i \neq j$,
- Based on chain rule:

  $P(h_i,F_1,\ldots,F_n) = P(h_i)P(F_1|h_i)P(F_2|h_i,F_1)\ldots P(F_n|h_i,F_1,\ldots,F_{n-1}) = P(h_i)P(F_1|h_i)P(F_2|h_i)\ldots P(F_n|h_i)$
- Given a hypothesis, the pieces of evidence are independent. Then based on product rule and independence:

$$P(h_i|F_1, F_2, \ldots, F_n) = P(h_i, F_1, F_2, \ldots, F_n) / P(F_1, F_2, \ldots, F_n) =$$

$$\frac{P(h_i)P(F_1|h_i)\ldots P(Fn|hi)}{P(F_1,\ldots,Fn)} = \frac{P(h_i)}{P(F_1,\ldots,Fn)}\prod_{j=1}^{n} p(Fj|hi)$$

- Naïve Bayes assume $\mathrm{p(E|h_j)} \approx \prod_{i=1}^{n} p(e_i|h_j)$

Naïve Bayes' classifier

- Naïve Bayes' classifier can be defined:
- $\text{Argmax}(\text{C}_\text{j}) \prod_{i=1}^{n} p(e_i|\text{C}_\text{j})p(\text{C}_\text{j})$
- **Justification**:
  - For attributes of a fruit, Apple with features, F = (shape, color, size), each of these attributes contribute **independently** to the probability that the fruit is an apple.
  - Many situations this assumption works reasonably well, e.g., text document classification, e.g., SPAM filtering.

Naïve Bayes' classifier Example

Training dataset: apples, grapes, and some other fruits

Features: size, color, shape

Hypotheses: apple, not an apple

$P(h_i|F_1, F_2, \ldots, F_n)$: given the features, what is the probability of h_i is true?

The training results are the probabilities: $p(e_i|\mathrm{C_j})$, $p(\mathrm{C_j})$, the probability of apple is red, the probability of "not an apple" is red

For testing: the classifier will return a class based on maximum likelihood: the probability of "an apple" is 0.8, the probability of "not an apple" is 0.2, then classified as apple

Application of Naïve Bayes:
Spoken Language Understanding

DEFINITION

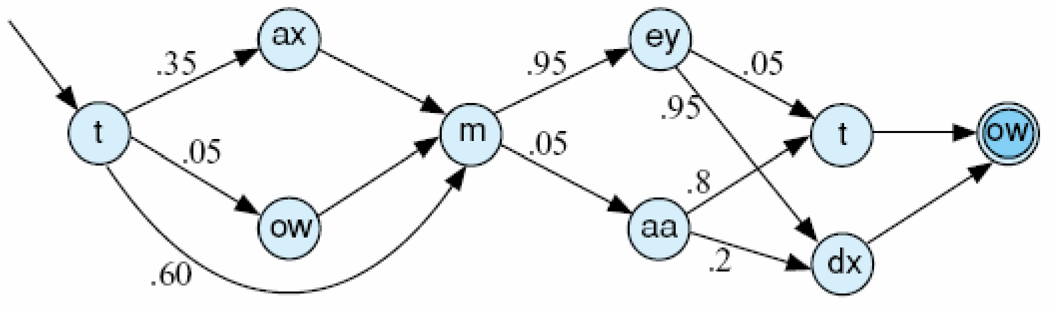
PROBABILISTIC FINITE STATE MACHINE

A *probabilistic finite state machine* is a finite state machine where the next state function is a probability distribution over the full set of states of the machine.

PROBABILISTIC FINITE STATE ACCEPTOR

A *probabilistic finite state machine* is an *acceptor*, when one or more states are indicated as the *start* states and one or more as the *accept* states.

A probabilistic finite state acceptor for the pronunciation of "**tomato**" adapted from Jurafsky and Martin (2009).



Speech Recognition

A phoneme recognition algorithm has identified the phone **ni** (as in "**knee**") that occurs just after the recognized word (phone) **I**, and we want to associate **ni** with either a word or the first part of a word.

$$p(\text{word} \mid [\text{ni}]) \propto p([\text{ni}] \mid \text{word}) \times p(\text{word})$$

This simplified conditional probability formula can be used to calculate the probabilities.

word	frequency	probability p(word)
knee	61	.000024
the	114834	.046
neat	338	.00013
need	1417	.00056
new	2625	.001

The ***ni*** words with their frequencies and probabilities from the Brown (~1M words from written text such as newspaper, books, academic writings collected at Brown Univ.) and Switchboard (1.4M words from phone conversations) corpora of 2.5M words.

+Which word seem to be the first choice for matching **ni**?

Speech Recognition

The results of calculation for the ***ni*** phone/word probabilities from the Brown and Switchboard corpora.

word	p([ni] | word)	p(word)	p([ni] | word) x p(word)
new	0.36	0.001	0.00036
neat	0.52	0.00013	0.000068
need	0.11	0.00056	0.000062
knee	1.0	0.000024	0.000024
the	0.0	0.046	0.0

Q1: Why **p(ni | the)** is impossible?
Q2: What is the most likely word for decoding "**ni**"?

Naïve Bayes Evaluation

- Advantages of Naïve Bayes Approach:
  - Efficient classification (Training data is only needed to estimate mean and variance.)
  - Not sensitive to irrelevant features
  - Handle both real and discrete data including streaming data
- Disadvantage:
  - Many other situations violate this assumption.
  - **Solution**: Consider the relationships between attributes.

References

- George Fluger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th edition, **Chapters 5, 9, and 13,** Addison Wesley, 2009.
- Russel and Norvig, Artificial Intelligence: A Modern Approach, 3rd edition, Prentice Hall, 2010.
- Some of slides are revised based on Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley