Bayesian Inferencing

Dr Pradipta Biswas, *PhD (Cantab)*
Assistant Professor
Indian Institute of Science
https://cambum.net/index.htm

Why Uncertainty

- Theoretical Ignorance : Imperfect Domain Knowledge
- Practical Ignorance : Imprecise Case Data
- Laziness : Incomplete Data

Problems with Exact Method

- Exact methods are not known
- They are impractical to employ
    - Lack of data
    - Problems with collecting data
    - Difficulties with processing data

Inexact Methods

Inexact methods give approximate answers i.e. answers with variable precision. It may be possible to reject the first choice and go for another choice

So we go for modeling the uncertainty to get efficient and effective Inexact Methods

Basic Probabilistic Model

- Random Variables
  - Boolean
  - Discrete
  - Continuous
- Conditional Probability
- Expected everybody is familiar

Think: You are tossing a fair coin. The first three outcomes are head (incidentally). According to conditional probability of the prior events, what will be the outcome of all the next tosses?

Conditional Probability

Product Rule

$P(a \wedge b) = P(a \mid b)\, P(b)$

$P(a \wedge b) = P(b \mid a)\, P(a)$

$P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

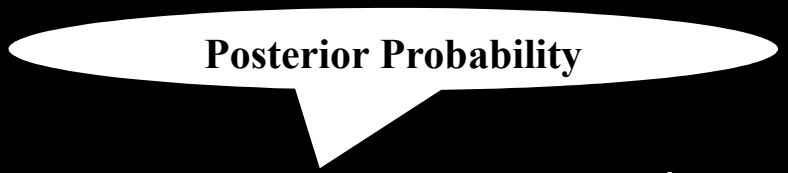
$$P(a \mid b) = \frac{P(b \mid a)\, P(a)}{P(b)}$$

Bayes' rule

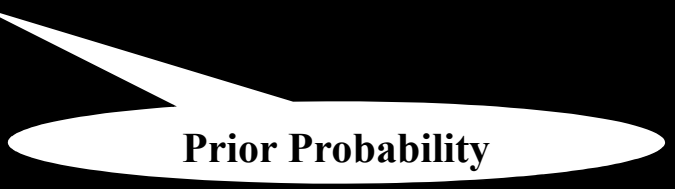
- Relates prior knowledge (prior to present experiment) for taking present inference

- **Intuition:** Before we observe the data, the parameters are described by a *prior* which is typically very broad. Once we observed the data, we can make use of Bayes' formula to find *posterior.* Since now we can utilize some more facts in the form of evidences, the *posterior* is narrower than *prior.*

Bayes' Rule



$$P(Y|X)=P(X|Y)*P(Y) / P(X)$$



Y=Cause

X=Evidence or Symptoms

- Bayes' rule with more than one evidence

$$P(Y|X,e)= P(X|Y,e)P(Y|e) / P(X|e)$$

Problems with Bayes' Rule

- When there are many causes and evidences the calculation becomes exponential

p(x1|y1^y2^y3…^yk) =

p (y1^y2^y3…^yk|x)*p(x1) / p(y1^y2^y3…^yk)

p(y1^y2^y3…^yk)=p(y1|y2^y3…^yk) *p(y2|y3…^yk)*…*p(yk)

-If there are m causes and n evidences, we needs (mn^k)+m+(n^k) probabilities

Independent Evidence Assumption

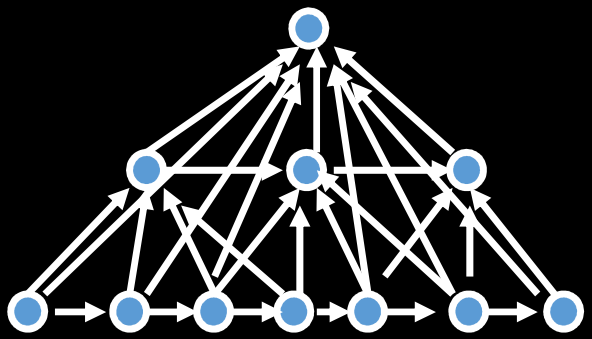
p(y1|y2^y3) = p(y1|y2) i.e. y1 does not depend on y3

Belief networks

- **A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distribution.**
- **Syntax:**
  - a set of nodes, one per random variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents
    $\mu(\boldsymbol{X}_i | \text{Parents}(\boldsymbol{X}_i))$
- **Conditional distributions are represented by conditional probability tables (CPT)**

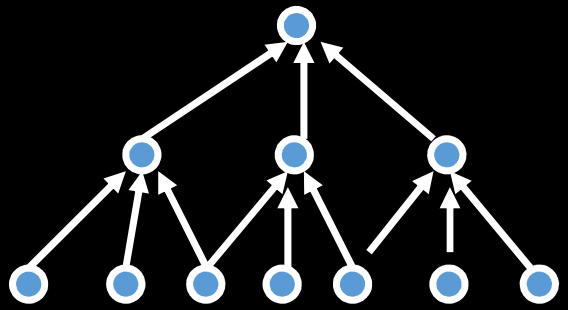
The importance of independency statements

n binary nodes,
fully connected



2^n -1 independent numbers

n binary nodes
each node max. 3 parents



less than $2^3 \cdot n$
independent numbers

Example

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

1% of people have a certain genetic defect.

90% of tests for the gene detect the defect (true positives).

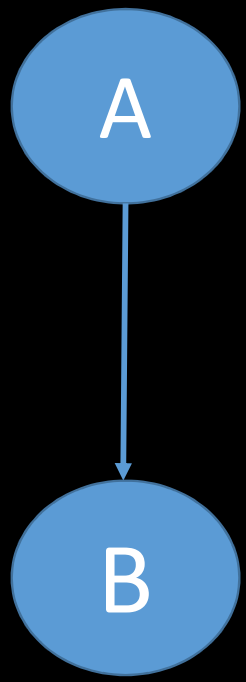
9.6% of the tests are false positives.

If a person gets a positive test result, **what are the odds they actually have the genetic defect?**

The first step into solving Bayes' theorem problems is to assign letters to events:

A = chance of having the faulty gene. That was given in the question as 1%. That also means the probability of *not* having the gene (~A) is 99%.

B = A positive test result.



Example

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

A = chance of having the faulty gene. That was given in the question as 1%. That also means the probability of *not* having the gene (~A) is 99%.

B = A positive test result.

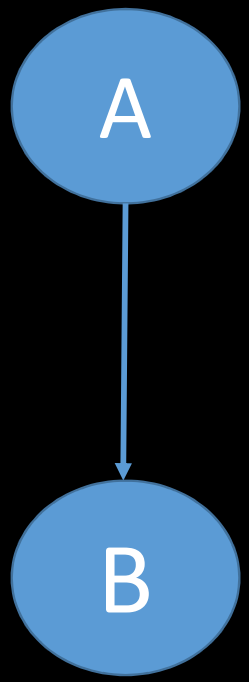
P(A|B) = Probability of having the gene given a positive test result.

P(B|A) = Chance of a positive test result given that the person actually has the gene. That was given in the question as 90%.

p(B|~A) = Chance of a positive test if the person *doesn't* have the gene. That was given in the question as 9.6%

Now we have all of the information we need to put into the equation:
P(A|B) = (.9 * .01) / (.9 * .01 + .096 * .99) = 0.0865 (8.65%).

The probability of having the faulty gene on the test is 8.65%.

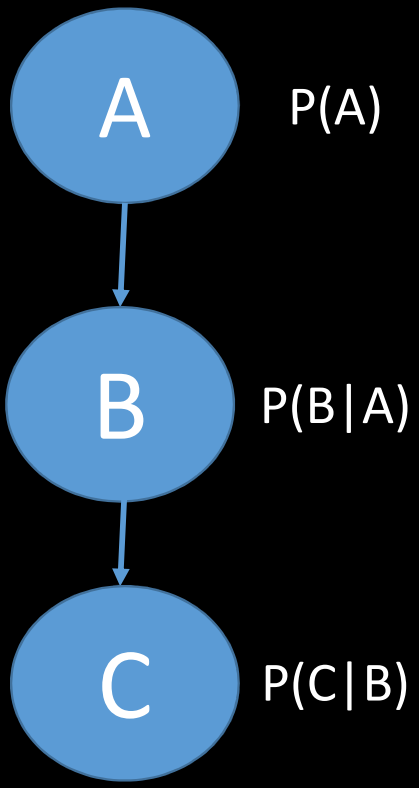


Example

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)}$$

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

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



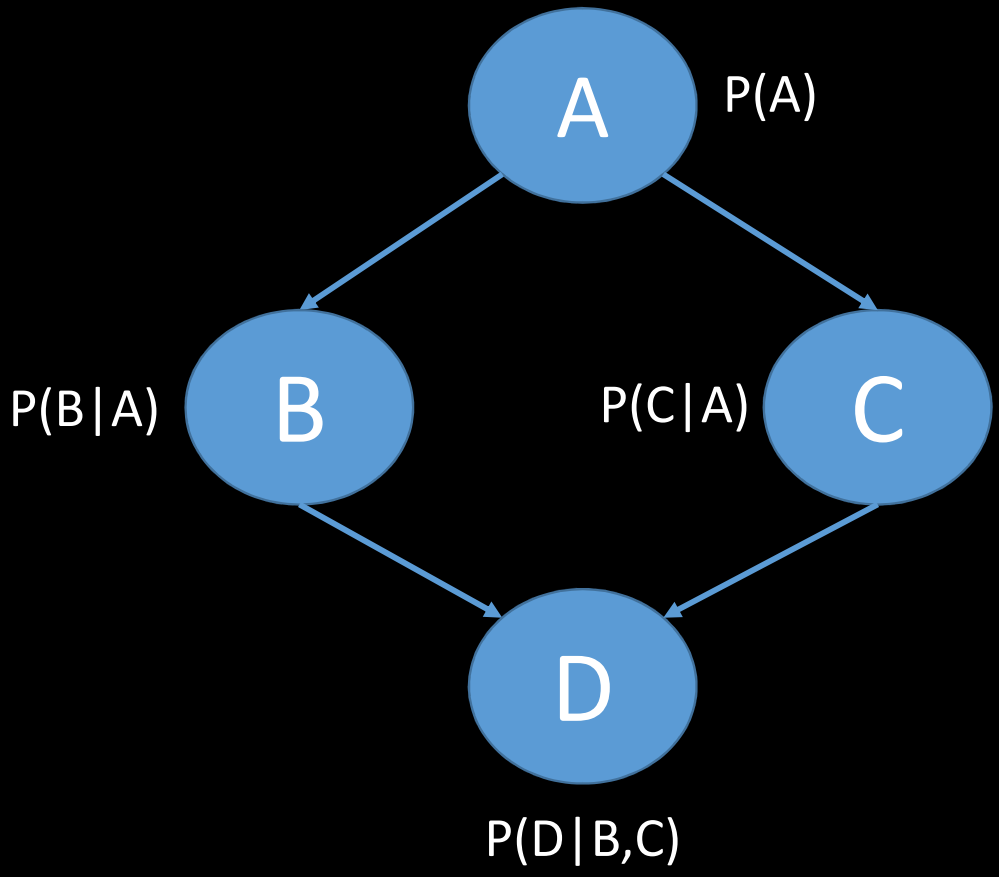
Bayesian Network

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)}$$

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

$$= \frac{\mathbf{P(D|B,C)}P(B,C|A)P(A)}{P(D)}$$

$$= \frac{P(D|B,C)\mathbf{P(B|A)P(C|A)}P(A)}{P(D)}$$



Bayesian networks

- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

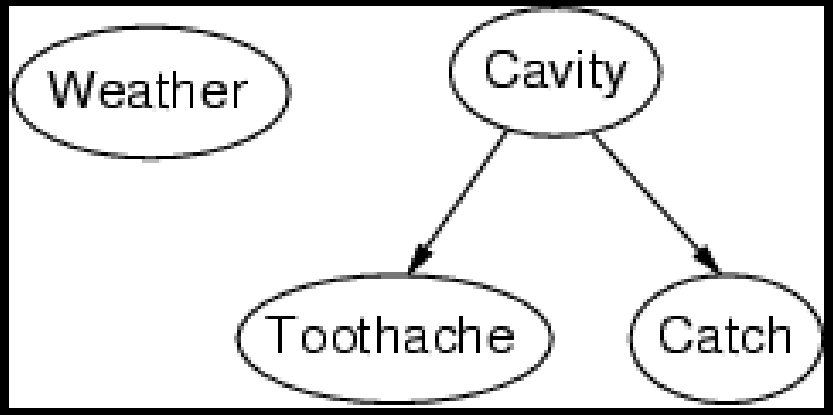
- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example

- Topology of network encodes conditional independence assertions:



- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

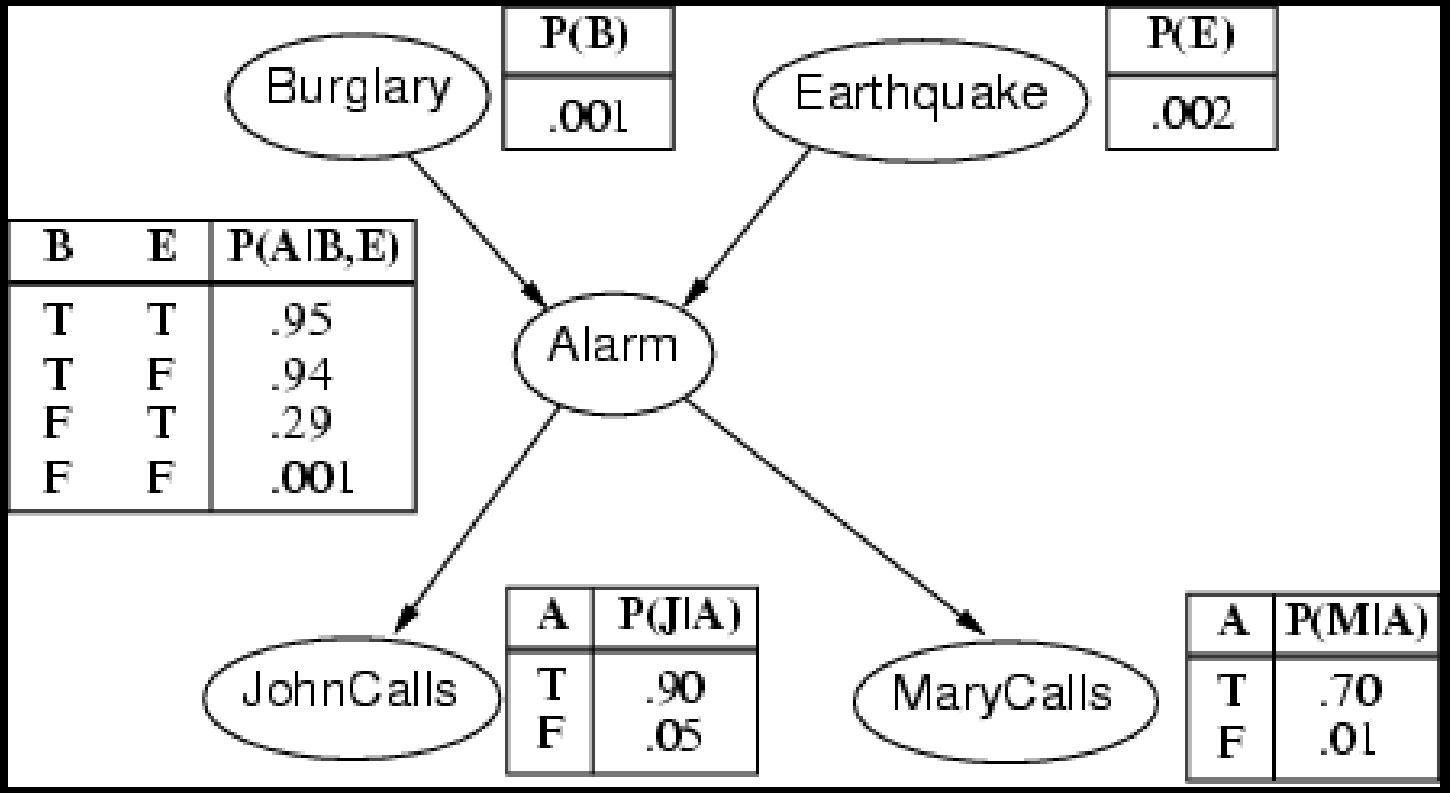
Example

- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes. Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



Example

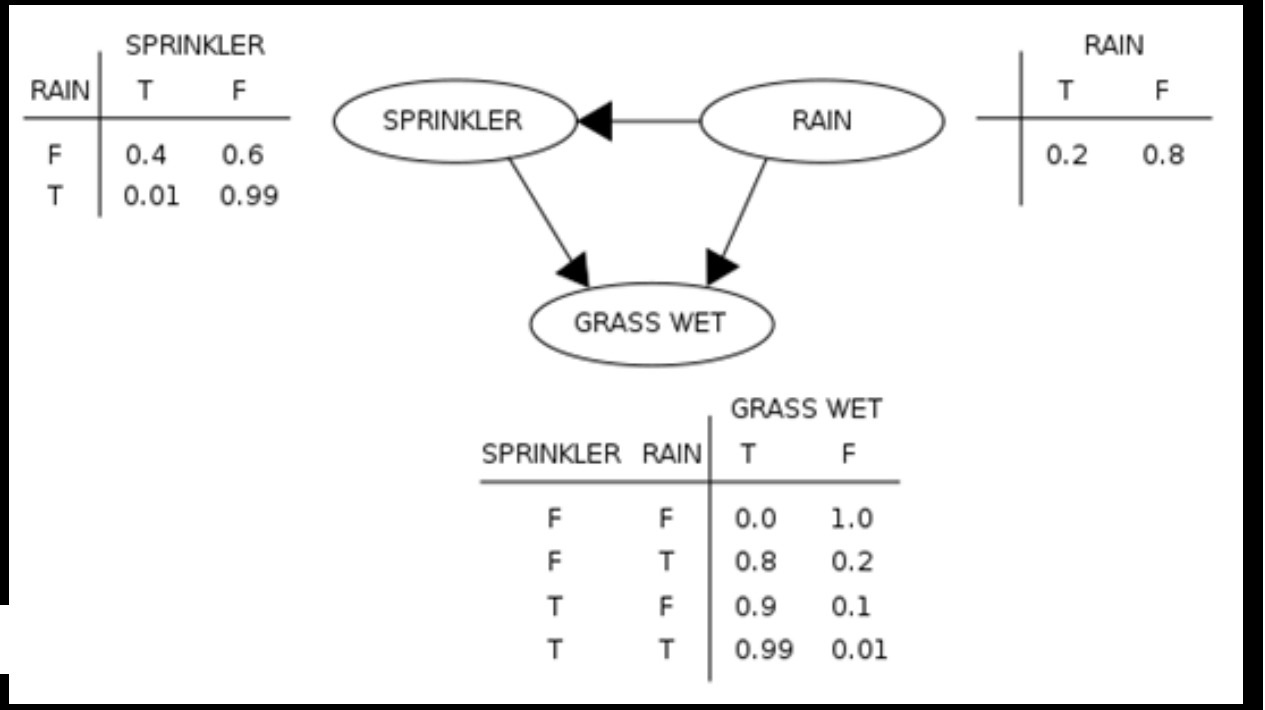
What is the probability of it is raining given the grass is wet?

$$P(R|G) = \frac{P(G|R)P(R)}{P(G)}$$

$$= \frac{P(G|R,S)P(S|R)P(R)}{P(G)}$$

$$\Pr(G,S,R) = \Pr(G|S,R)\Pr(S|R)\Pr(R)$$

$$\sum_{S,R\in\{T,F\}} \Pr(G=T,S,R)$$



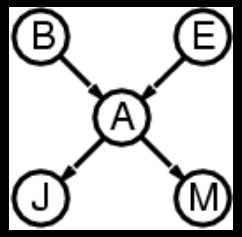
RAIN	SPRINKLER T	SPRINKLER F
F	0.4	0.6
T	0.01	0.99

RAIN T	RAIN F
0.2	0.8

SPRINKLER	RAIN	GRASS WET T	GRASS WET F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



- Each row requires one number p for $X_i = true$
  (the number for $X_i = false$ is just $1-p$)
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers
- I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution
- For burglary net, 1 + 1 + 4 + 2 + 2 = 10 numbers (vs. $2^5-1 = 31$)

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \ldots, X_n$
- 2. For $i = 1$ to n
  - add X_i to the network
  -
  - select parents from $X_1, \ldots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots X_{i-1})$$

This choice of parents guarantees:

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

(chain rule)

$$= \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

(by construction)

Summary of Bayesian Network

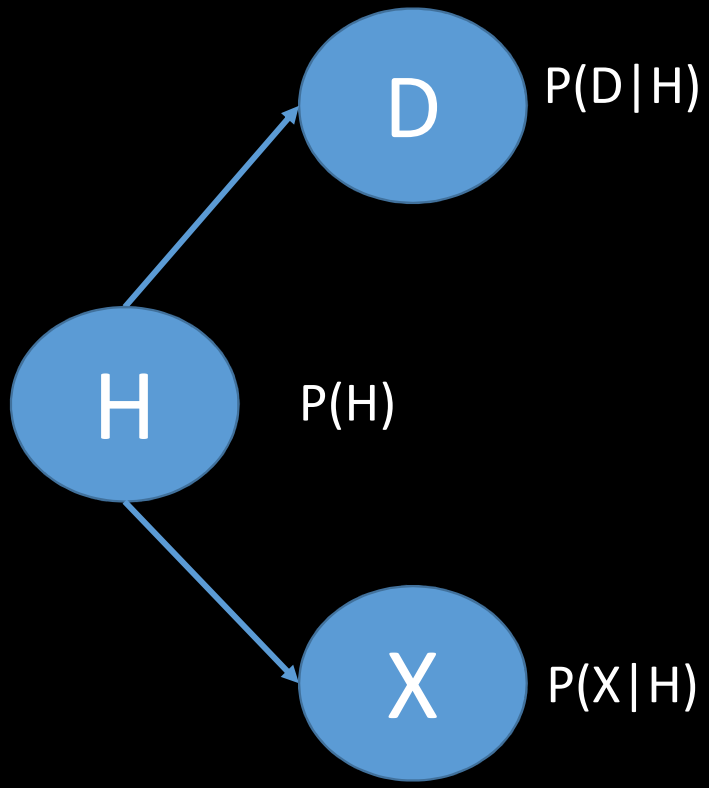
- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs = compact representation of joint distribution
- Generally easy for domain experts to construct

Bayesian Inferencing

$$P(X|D) = P(X|D,H)P(H|D)$$

$$= \mathbf{P(X|H)}P(H|D)$$

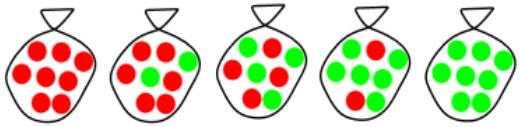
$$= P(X|H)\frac{P(D|H)P(H)}{P(D)}$$



$P(h_i|D)$ for j=1

$P(h_i)$	$P(d_1|h_i)$	$P(d_1|h_i).P(h_i)$	$P(h_i|d_1)$	$P(X|h_i)$
0.1	0	0	0	0
0.2	0.25	0.05	0.1	0.25
0.4	0.5	0.20	0.4	0.5
0.2	0.75	0.15	0.3	0.75
0.1	1	0.10	0.2	1
		∑ = 0.5		

Suppose there are five kinds of bags of candies:
10% are h_1: 100% cherry candies
20% are h_2: 75% cherry candies + 25% lime candies
40% are h_3: 50% cherry candies + 50% lime candies
20% are h_4: 25% cherry candies + 75% lime candies
10% are h_5: 100% lime candies



Then we observe candies drawn from some bag: ● ● ● ● ● ● ● ● ● ●

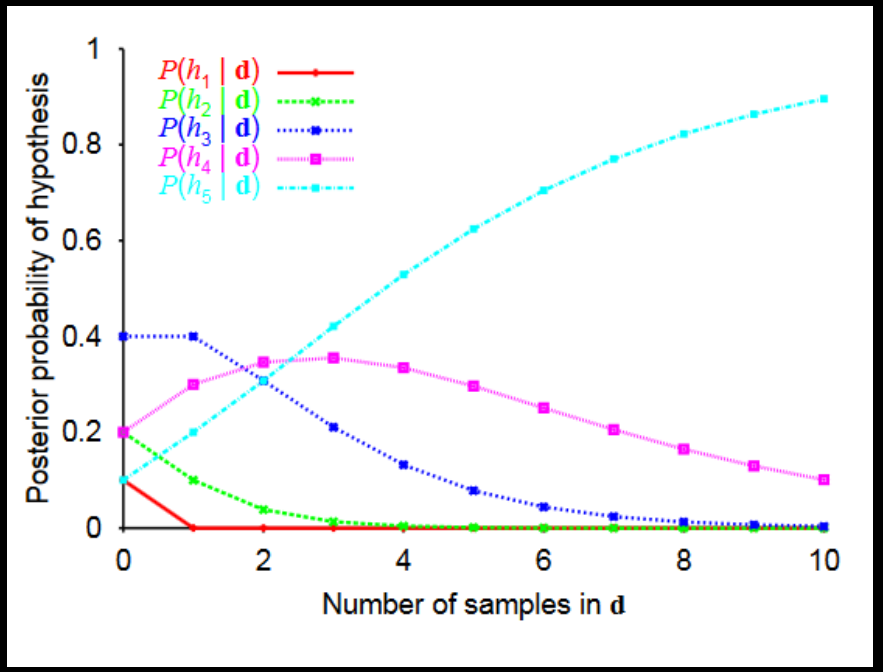
What kind of bag is it? What flavour will the next candy be?

$\sum_{i=1}^{n} P(X|hi).\ P(h_i|d_1) = 0 + 0.025 + 0.2 + 0.225 + 0.2$

$= 0.65$

$P(h_i|D)$ for j=2

$P(h_i)$	$P(d_{2,1}|h_i)$	$P(d_{2,1}|h_i).P(h_i)$	$P(h_i|d_{2,1})$	$P(X|h_i)$
0.1	0	0	0	0
0.2	0.25*0.25=0.0625	0.0125	0.038	0.25
0.4	0.5*0.5=0.25	0.1	0.308	0.5
0.2	0.75*0.75=0.5625	0.1125	0.346	0.75
0.1	1	0.1	0.308	1
		∑ = 0.325		



$\sum_{i=1}^{n} P(X|hi).\ P(h_i|d_{2,1})$ = 0 + 0.01 + 0.154 + 0.26 + 0.308

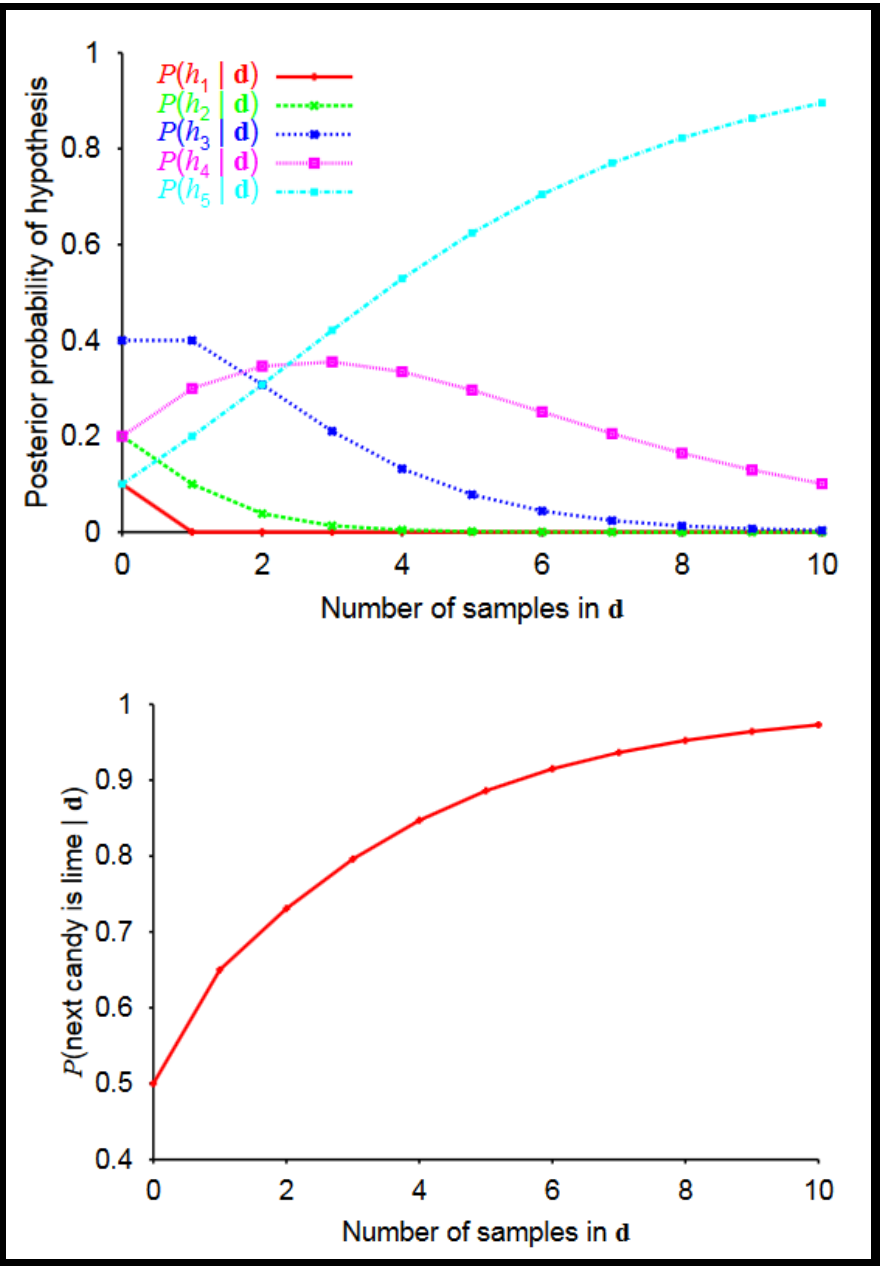
= 0.732

$P(h_i|D)$ for j=3

$P(h_i)$	$P(d_{3,21}|h_i)$	$P(d_{3,21}|h_i).P(h_i)$	$P(h_i|d_{3,21})$	$P(X|h_i)$
0.1	0	0	0	0
0.2	0.0156	0.0031	0.0131	0.25
0.4	0.125	0.05	0.210	0.5
0.2	0.422	0.0844	0.355	0.75
0.1	1	0.1	0.421	1
		∑ = 0.2375		

$\sum_{i=1}^{n} P(X|hi).\ P(h_i|d_{3,21}) = 0 + 0.003275 + 0.105 + 0.2662 + 0.421$

$= 0.795525$



MAP Learning

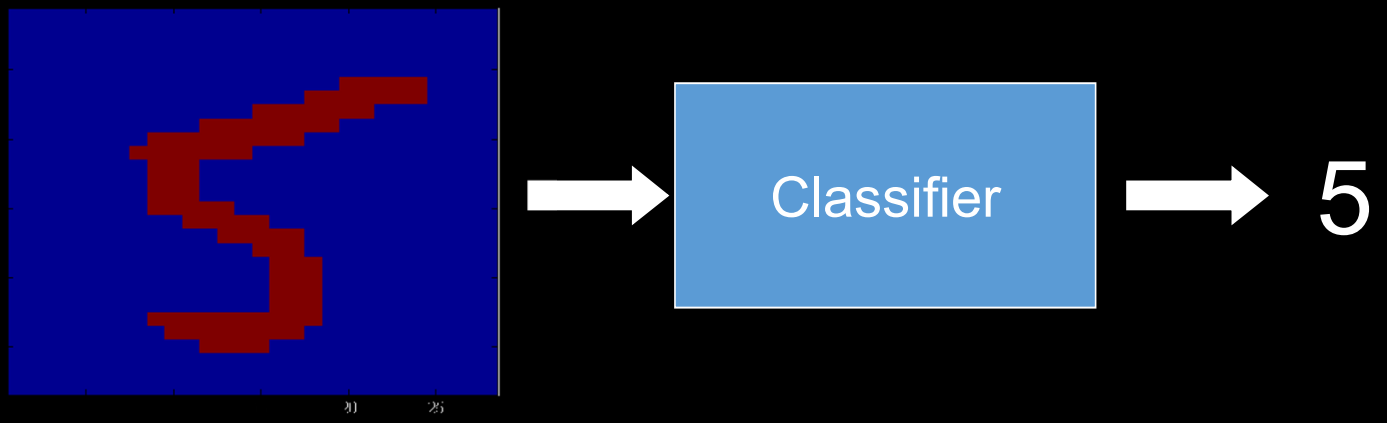
- Summing over the hypothesis space is often intractable
- (e.g., 18,446,744,073,709,551,616 Boolean functions of 6 attributes)
- Maximum a posteriori (MAP) learning: choose h MAP maximizing $P(h_i|d)$
- I.e., maximize $P(d|h_i)P(h_i)$ or $\log P(d|h_i) + \log P(h_i)$
- Log terms can be viewed as (negative of)
  - bits to encode data given hypothesis + bits to encode hypothesis
  - This is the basic idea of minimum description length (MDL) learning

Bayesian Classification

- Problem statement:
  - Given features $X_1, X_2, \ldots, X_n$
  - Predict a label Y

Another Application

- **Digit Recognition**



- **$X_1,\ldots,X_n \in \{0,1\}$ (Black vs. White pixels)**
- **$Y \in \{5,6\}$ (predict whether a digit is a 5 or a 6)**

The Bayes Classifier

- A good strategy is to predict:

$$\arg\max_{Y} P(Y|X_1, \ldots, X_n)$$

  - (for example: what is the probability that the image represents a 5 given its pixels?)

- So … How do we compute that?

The Bayes Classifier

- Use Bayes Rule!

Likelihood Prior

$$P(Y|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y)P(Y)}{P(X_1,\ldots,X_n)}$$

Normalization Constant

- Why did this help? Well, we think that we might be able to specify how features are "generated" by the class label

The Bayes Classifier

- Let's expand this for our digit recognition task:

$$P(Y=5|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=5)P(Y=5)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

$$P(Y=6|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y=6)P(Y=6)}{P(X_1,\ldots,X_n|Y=5)P(Y=5)+P(X_1,\ldots,X_n|Y=6)P(Y=6)}$$

- To classify, we have to compute these two probabilities and predict based on which one is greater

Model Parameters

- The problem with explicitly modeling $P(X_1,\ldots,X_n|Y)$ is that there are usually way too many parameters:
  - We'll run out of space
  - We'll run out of time
  - And we'll need tons of training data (which is usually not available)

The Naïve Bayes Model

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \ldots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

Why is this useful?

- # of parameters for modeling $P(X_1,\ldots,X_n|Y)$:
  - $2(2^n-1)$
- # of parameters for modeling $P(X_1|Y),\ldots,P(X_n|Y)$
  - $2n$

Take Away Points

- Bayesian Inferencing
- Bayesian Network
- Inferencing using Bayesian Network
- Introduction to Statistical Learning
- Naïve Bayes' Classifier