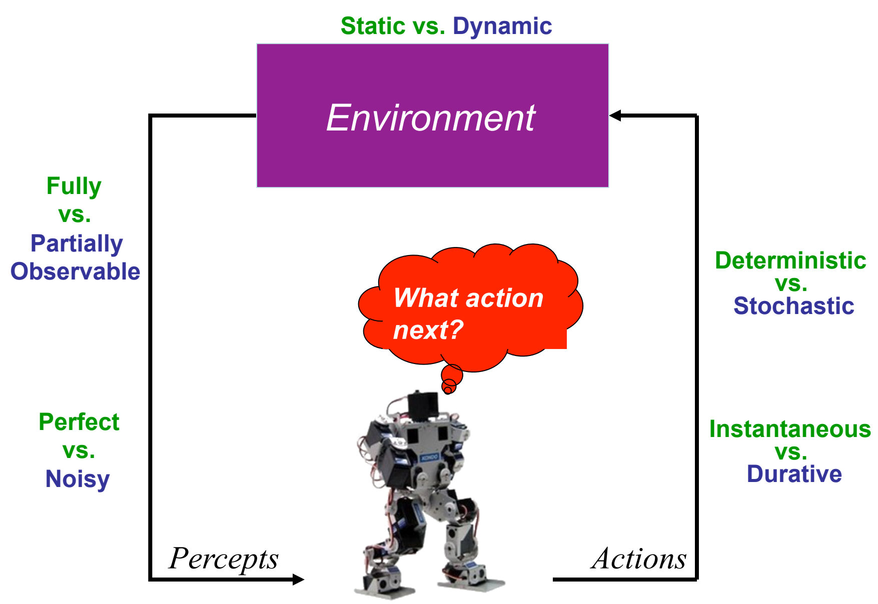
CSE 473: Artificial Intelligence
Fall 2014

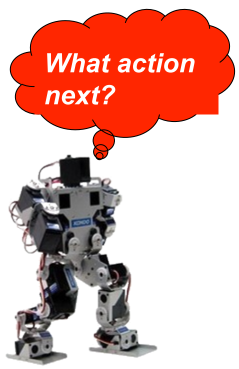
Bayesian Networks - Learning

Dan Weld

Slides adapted from Jack Breese, Dan Klein, Daphne Koller, Stuart Russell, Andrew Moore & Luke Zettlemoyer

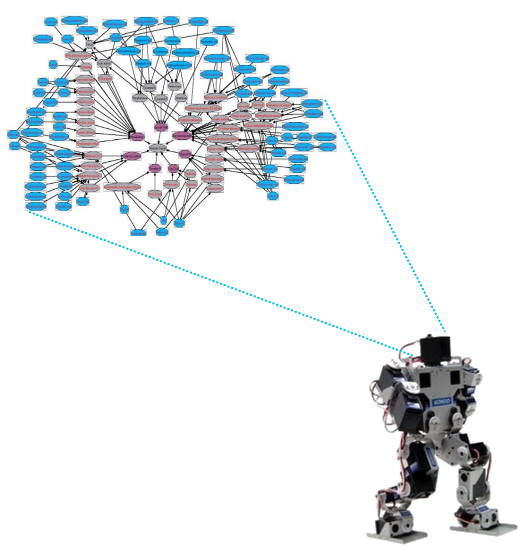


Algorithms



- Blind search
- Heuristic search
- Mini-max & Expectimax
- MDPs
- Reinforcement learning
- State estimation
- Variable Elimination

Knowledge Representation

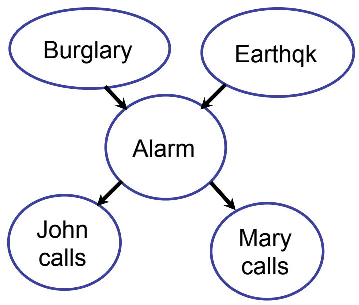


- Problem spaces
- Constraint networks
- HMMs
- Bayesian networks
- First-order logic
- Markov logic networks
- …

Example: Alarm Network

Only 10 params

B	P(B)
+b	0.001
←b	0.999



E	P(E)
+e	0.002
←e	0.998

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	←a	0.05
+b	←e	+a	0.94
+b	←e	←a	0.06
←b	+e	+a	0.29
←b	+e	←a	0.71
←b	←e	+a	0.001
←b	←e	←a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	←j	0.1
←a	+j	0.05
←a	←j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	←m	0.3
←a	+m	0.01
←a	←m	0.99

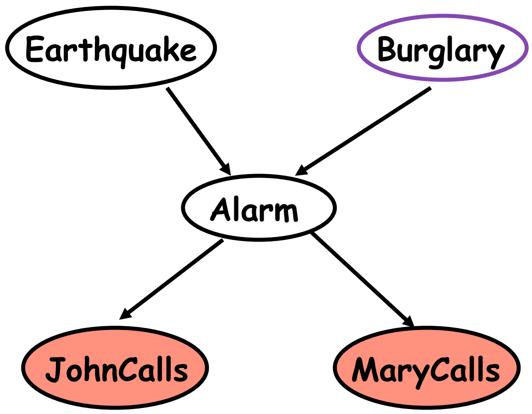
Probabilities in BNs

- Bayes' nets implicitly encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$$

- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain independence assumptions
  - Compare to the exact decomposition according to the chain rule!

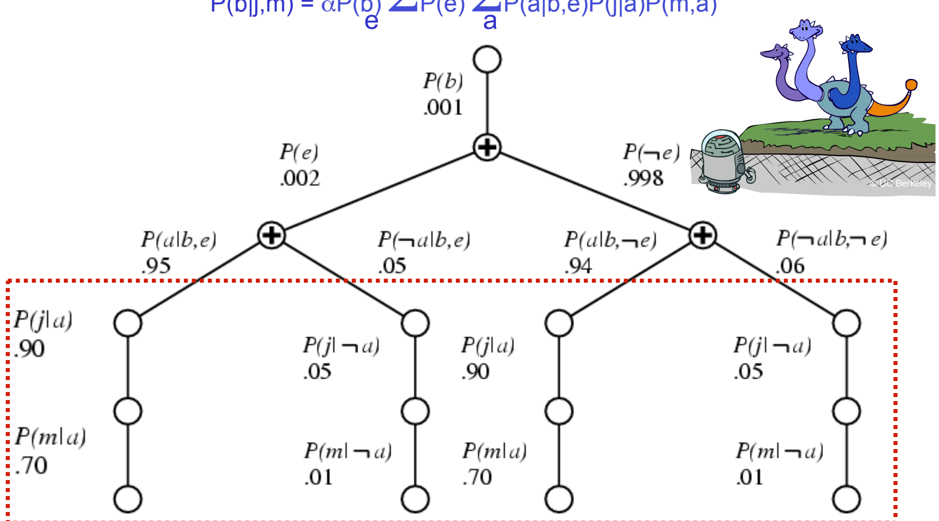
P(B | J=true, M=true)



$$P(b|j,m) = \alpha \sum_{e,a} P(b,j,m,e,a)$$

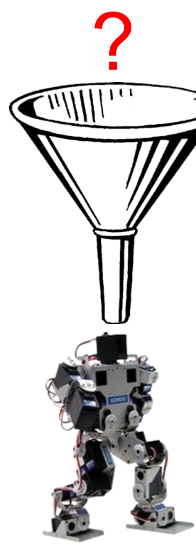
Variable Elimination

$$P(b|j,m) = \alpha P(b) \sum_{e} P(e) \sum_{a} P(a|b,e)P(j|a)P(m,a)$$



Repeated computations ➔ Dynamic Programming

Learning



What is Machine Learning ?

Machine Learning

Study of algorithms that

- improve their performance
- at some task
- with experience

Exponential Growth in Data

Supremacy of Machine Learning

- Machine learning is preferred approach to
  - Speech recognition, Natural language processing
  - Web search – result ranking
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Sensor networks
  - …
- This trend is accelerating
  - Improved machine learning algorithms
  - Improved data capture, networking, faster computers
  - Software too complex to write by hand
  - New sensors / IO devices
  - Demand for self-customization to user, environment

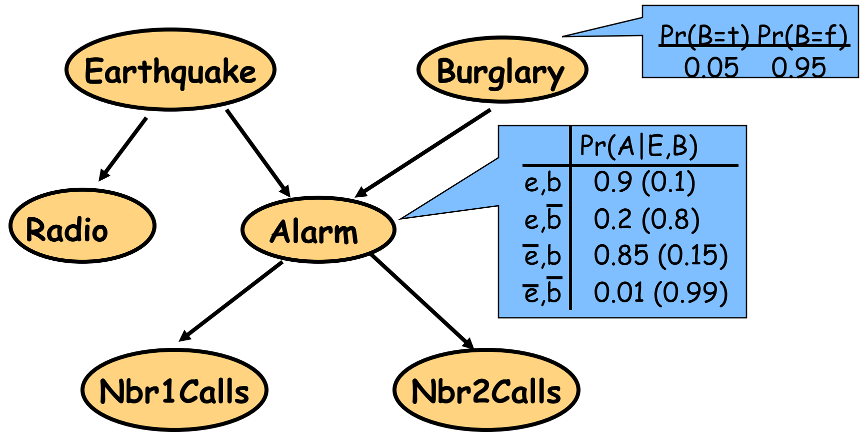
Space of ML Problems

Type of Supervision
(eg, Experience, Feedback)

What is Being Learned?

	Labeled Examples	Reward	Nothing
Discrete Function	Classification		Clustering
Continuous Function	Regression		
Policy	Apprenticeship Learning	Reinforcement Learning	

The Origin of Bayes Nets

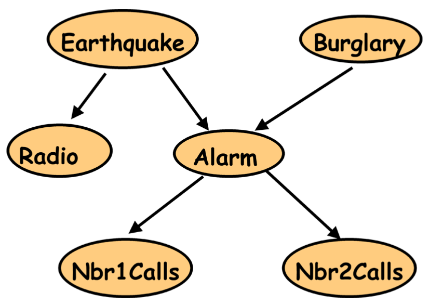


 20

Learning Topics

- Learning Parameters for a Bayesian Network
  - Fully observable
    - Maximum Likelihood (ML)
    - Maximum A Posteriori (MAP)
    - Bayesian
  - Hidden variables (EM algorithm)
- Learning Structure of Bayesian Networks

Parameter Estimation and Bayesian Networks

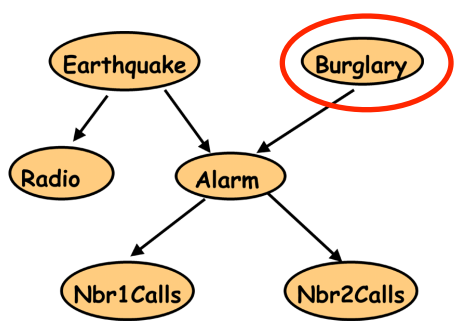


E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
...					

We have:

- Bayes Net structure and observations
- We need: Bayes Net parameters

Parameter Estimation and Bayesian Networks

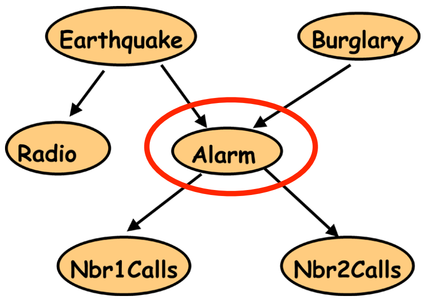


B
F
F
T
F
T

P(B) = ? = 0.4

P(¬B) = 1- P(B) = 0.6

Parameter Estimation and Bayesian Networks

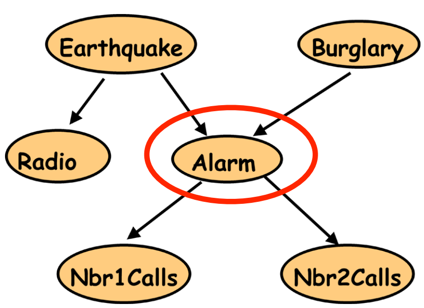


E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

P(A|E,B) = ?
P(A|E,¬B) = ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = ?

Parameter Estimation and Bayesian Networks



E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

P(A|E,B) = ?
P(A|E,¬B) = ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = 0.5

Parameter Estimation and Bayesian Networks

Coin Flip

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Prior: Probability of a hypothesis before we make any observations

Coin Flip

C_1	C_2	C_3
		
$P(H\|C_1) = 0.1$	$P(H\|C_2) = 0.5$	$P(H\|C_3) = 0.9$

Which coin will I use?

$P(C_1) = 1/3$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

Uniform Prior: All hypothesis are equally likely before we make any observations

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} \qquad P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$$

C_1	C_2	C_3
		
$P(H\|C_1)=0.1$	$P(H\|C_2) = 0.5$	$P(H\|C_3) = 0.9$
$P(C_1)=1/3$	$P(C_2) = 1/3$	$P(C_3) = 1/3$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.6$

Posterior: Probability of a hypothesis given data

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 1/3$

Using Prior Knowledge

- Should we always use a ***Uniform Prior*** ?
- Background knowledge:

Heads => we have to buy Dan chocolate

Dan ***likes*** chocolate…

=> Dan is more likely to use a coin biased in his favor

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Using Prior Knowledge

We can encode it in the prior:

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

C_1

C_2

C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = ?$ $P(C_2|H) = ?$ $P(C_3|H) = ?$

$$P(C_1|H) = \alpha P(H|C_1)P(C_1)$$

C_1

C_2

C_3

$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

$P(C_1) = 0.05$ $P(C_2) = 0.25$ $P(C_3) = 0.70$

Experiment 1: Heads

Which coin did I use?

$P(C_1|H) = 0.006$ $P(C_2|H) = 0.165$ $P(C_3|H) = 0.829$

Compare with ML posterior after Exp 1:

$P(C_1|H) = 0.066$ $P(C_2|H) = 0.333$ $P(C_3|H) = 0.600$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 0.05$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 0.25$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = ?$ $P(C_2|HT) = ?$ $P(C_3|HT) = ?$

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 0.05$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 0.25$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Experiment 2: Tails

Which coin did I use?

P(C_1|HT) = 0.035   P(C_2|HT) = 0.481   P(C_3|HT) = 0.485

$$P(C_1|HT) = \alpha P(HT|C_1)P(C_1) = \alpha P(H|C_1)P(T|C_1)P(C_1)$$

C_1	C_2	C_3
		
P(H|C_1) = 0.1	P(H|C_2) = 0.5	P(H|C_3) = 0.9
P(C_1) = 0.05	P(C_2) = 0.25	P(C_3) = 0.70

Experiment 2: Tails

Which coin did I use?

P(C_1|HT) = 0.035   P(C_2|HT)=0.481   P(C_3|HT) = 0.485

C_1	C_2	C_3
		
P(H|C_1) = 0.1	P(H|C_2) = 0.5	P(H|C_3) = 0.9
P(C_1) = 0.05	P(C_2) = 0.25	P(C_3) = 0.70

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin:

C_3 

Best estimate for P(H)

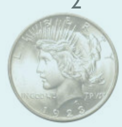
$P(H|C_3) = 0.9$

C_1

$P(H|C_1) = 0.1$

$P(C_1) = 0.05$

C_2



$P(H|C_2) = 0.5$

$P(C_2) = 0.25$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Your Estimate?

Maximum A Posteriori (MAP) Estimate:
The best hypothesis that fits observed data
assuming a non-uniform prior

Most likely coin:

C_3 

Best estimate for P(H)

$P(H|C_3) = 0.9$

C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Did We Do The Right Thing?

$P(C_1|HT)=$0.035 $P(C_2|HT)=$0.481 $P(C_3|HT)=$0.485

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Did We Do The Right Thing?

$P(C_1|HT)=$0.035 $P(C_2|HT)=$0.481 $P(C_3|HT)=$0.485

C_2 and C_3 are almost equally likely

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

A Better Estimate

Recall: $P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i)$ = 0.680

$P(C_1|HT)=0.035$ $P(C_2|HT)=0.481$ $P(C_3|HT)=0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error, given data assuming an arbitrary prior

$$P(H) = \sum_{i=1}^{3} P(H|C_i)P(C_i) = 0.680$$

$P(C_1|HT)=0.035$ $P(C_2|HT)=0.481$ $P(C_3|HT)=0.485$

C_1

$P(H|C_1) = 0.1$

C_2

$P(H|C_2) = 0.5$

C_3

$P(H|C_3) = 0.9$

Comparison
After more experiments: HTHHHHHHHH

ML (Maximum Likelihood):
- P(H) = 0.5
- after 10 experiments: P(H) = 0.9

MAP (Maximum A Posteriori):
- P(H) = 0.9
- after 10 experiments: P(H) = 0.9

Bayesian:
- P(H) = 0.68
- after 10 experiments: P(H) = 0.9

Summary

Easy to compute

	Prior	Hypothesis
Maximum Likelihood Estimate	Uniform	The most likely
Maximum A Posteriori Estimate	Any	The most likely
Bayesian Estimate	Any	Weighted combination

Still easy to compute
Incorporates prior knowledge

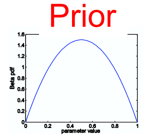
Minimizes error
Great when data is scarce
Potentially much harder to compute

Bayesian Learning

Use Bayes rule:

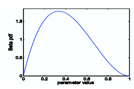
Data Likelihood

Prior



$$P(Y \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid Y)\, P(Y)}{P(\mathbf{X})}$$

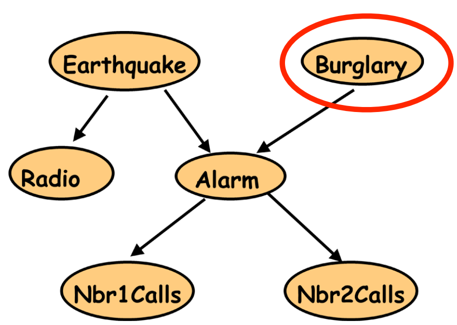
Posterior



Normalization

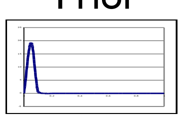
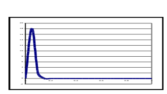
Or equivalently: $P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$

Parameter Estimation and Bayesian Networks



B
F
F
T
F
T

Prior

P(B) =  + data = 

Now compute either MAP or Bayesian estimate

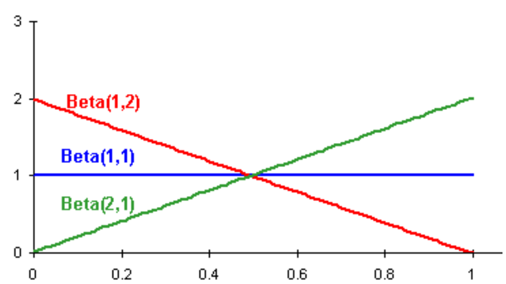
What Prior to Use?

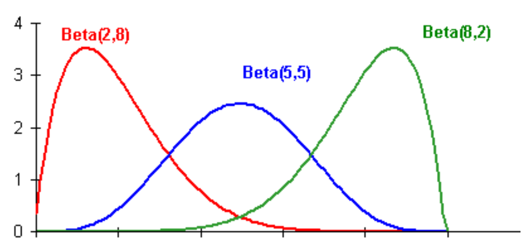
- Prev, you ***knew***: it was one of only three coins

  - Now more complicated…
- The following are two common priors
- Binary variable Beta
  - Posterior distribution is binomial
  - Easy to compute posterior
- Discrete variable Dirichlet
  - Posterior distribution is multinomial
  - Easy to compute posterior

54

Beta Distribution





Beta Distribution

- Example: Flip coin with B*eta* distribution as prior over p [prob(heads)]
  1. Parameterized by two positive numbers: a, b
  2. Mode of distribution (E[p]) is *a/(a+b)*
  3. Specify our prior belief for *p* = *a/(a+b)*
  4. Specify confidence in this belief with high initial values for *a* and *b*
- Updating our prior belief based on data
  - incrementing *a* for every *heads* outcome
  - incrementing *b* for every tails outcome

One Prior: Beta Distribution

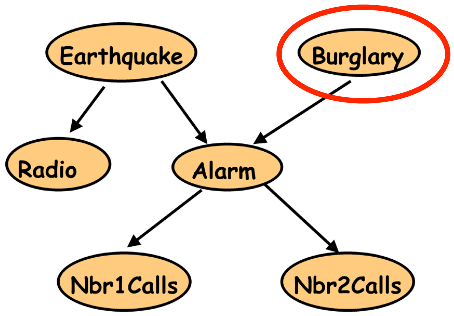
$$\beta_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

$$0 \leq x \leq 1 \text{ and } a, b > 0$$

$$\text{Here } \Gamma(y) = \int_0^\infty x^{y-1} e^{-x} dx$$

For any positive integer *y*, Γ(*y*) = (*y*-*1*)!

Parameter Estimation and Bayesian Networks



B
F
F
T
F
T

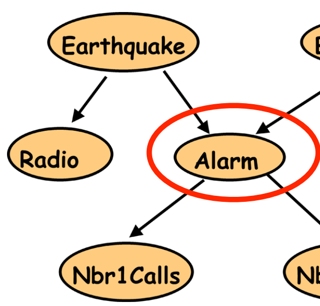
Prior

P(B|data) = ? Beta(1,4) "+ data" = (3,7)

B	¬B
.3	.7

Prior P(B)= 1/(1+4) = 20% with equivalent sample size 5

Parameter Estimation and Bayesian Networks

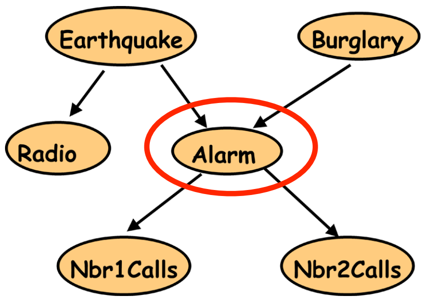


E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

P(A|E,B) = ?
P(A|E,¬B) = ?
P(A|¬E,B) = ?
P(A|¬E,¬B) = ?

Parameter Estimation and Bayesian Networks

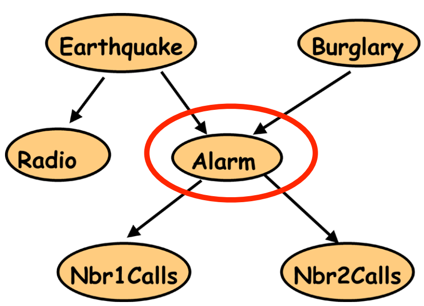


E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

P(A|E,B) = ? Prior
P(A|E,¬B) = ?
P(A|¬E,B) = ? Beta(2,3)
P(A|¬E,¬B) = ?

Parameter Estimation and Bayesian Networks



E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

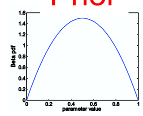
P(A|E,B) = ? Prior
P(A|E,¬B) = ?
P(A|¬E,B) = ? Beta(2,3) + data= (3,4)
P(A|¬E,¬B) = ?

Bayesian Learning

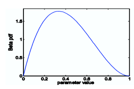
Use Bayes rule:

Data Likelihood

Prior



Posterior



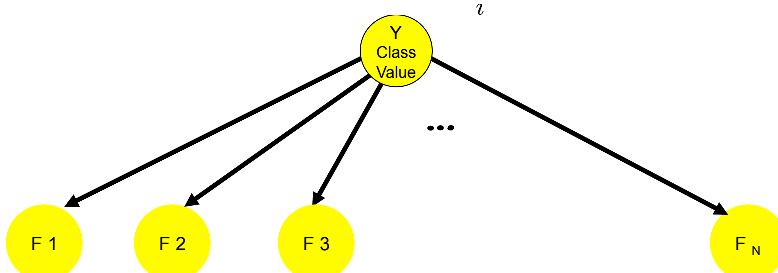
$$P(Y \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid Y)\, P(Y)}{P(\mathbf{X})}$$

Normalization

Or equivalently: $P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$

Naïve Bayes

$$P(\mathsf{Y}, \mathsf{F}_1 \ldots \mathsf{F}_n) = P(\mathsf{Y}) \prod_i P(\mathsf{F}_i \mid \mathsf{Y})$$



Assume that features are conditionally independent given class variable
Works well in practice
But forces probabilities towards 0 and 1

Naïve Bayes

- Naïve Bayes assumption:
  - Features are independent given class:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

  - More generally:

$$P(X_1 ... X_n|Y) = \prod_i P(X_i|Y)$$

- How many parameters now?
  - Suppose **X** is composed of *n* binary features

NB with Bag of Words for text classification

- Learning phase:
  - Prior P(Y)
    - Count how many documents from each topic (prior)
  - $P(X_i|Y)$
    - For each of m topics, count how many times you saw word X_i in documents of this topic (+ k for prior)
    - Divide by number of times you saw the word (+ k×m)
- Test phase:
  - For each document
    - Use naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Probabilities: Important Detail!

- $P(spam \mid X_1 \ldots X_n) = \prod_i P(spam \mid X_i)$

Any more potential problems here?

- We are multiplying lots of small numbers
  Danger of underflow!
  - 0.5^{57} = 7 E -18
- Solution? Use logs and add!
  - $p_1 * p_2 = e^{\log(p1)+\log(p2)}$
  - Always keep in log form