CSE 473: Artificial Intelligence
Spring 2012

Bayesian Networks - Learning

Dan Weld

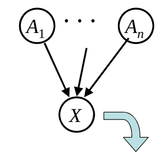
Slides adapted from Jack Breese, Dan Klein, Daphne Koller, Stuart Russell, Andrew Moore & Luke Zettlemoyer

Bayes' Net Semantics

Formally:

- A set of nodes, one per variable X
- A directed, acyclic graph
- A CPT for each node
  - CPT = "Conditional Probability Table"
  - Collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \ldots a_n)$$



$P(X|A_1 \ldots A_n)$

A Bayes net = Topology (graph) + Local Conditional Probabilities

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:

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

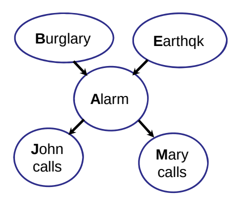
- This lets us reconstruct any entry of the full joint
- Not every BN can represent every joint distribution
  - The topology enforces certain *independence* assumptions
  - Compare to the exact decomposition according to the chain rule!

Example: Alarm Network

Only 10 params

B	P(B)
+b	0.001
¬b	0.999

E	P(E)
+e	0.002
¬e	0.998

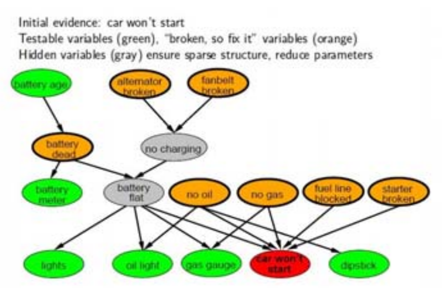


B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	¬a	0.05
+b	¬e	+a	0.94
+b	¬e	¬a	0.06
¬b	+e	+a	0.29
¬b	+e	¬a	0.71
¬b	¬e	+a	0.001
¬b	¬e	¬a	0.999

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

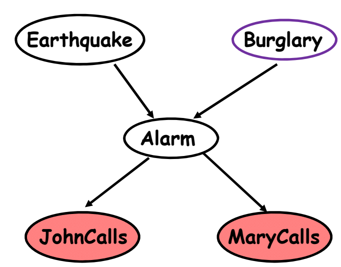
A	M	P(M|A)
+a	+m	0.7
+a	¬m	0.3
¬a	+m	0.01
¬a	¬m	0.99

Example: Car Diagnosis



© D. Weld and D. Fox 5

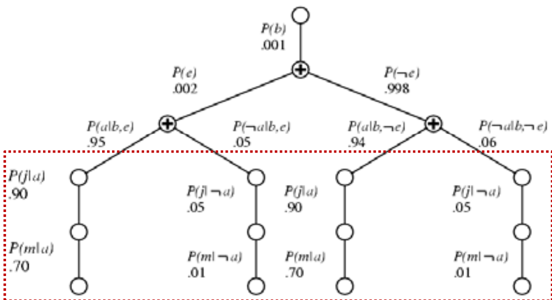
P(B | J=true, M=true)



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

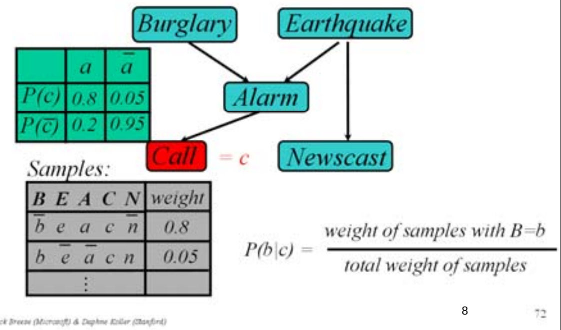
Variable Elimination

P(b|j,m) = αP(b) $\sum_e$ P(e) $\sum_a$ P(a|b,e)P(j|a)P(m,a)



Repeated computations ➔ Dynamic Programming

Likelihood weighting P(B|C)



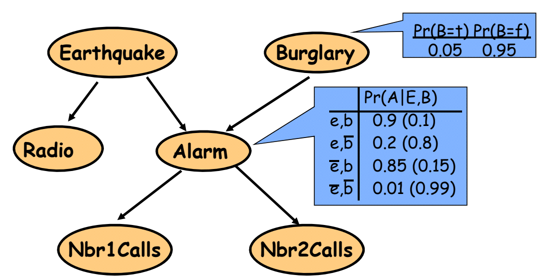
8

MCMC with Gibbs Sampling

- Fix the values of observed variables
- Set the values of all non-observed variables randomly
- Perform a random walk through the space of complete variable assignments. On each move:
  1. Pick a variable X
  2. Calculate Pr(X=true | Markov blanket)
  3. Set X to true with that probability
- Repeat many times. Frequency with which any variable X is true is it's posterior probability.
- Converges to true posterior when frequencies stop changing significantly
  - stable distribution, mixing

9

The Origin of Bayes Nets

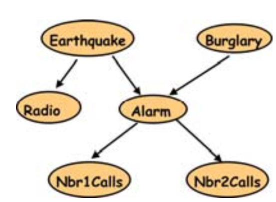


 10

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

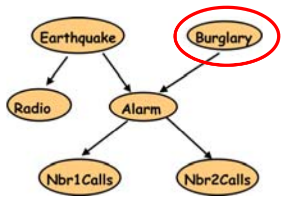
Parameter Estimation and Bayesian Networks



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

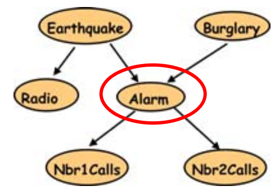


B
F
F
T
F
T

P(B) = ? = 0.4

P(¬B) = 1- P(B) = 0.6

Parameter Estimation and Bayesian Networks



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

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

Parameter Estimation and Bayesian Networks

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?

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

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

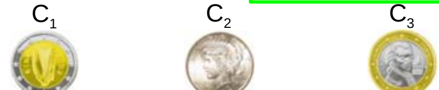
Experiment 1: Heads

Which coin <u>did</u> 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

$P(H|C_3) = 0.9$
$P(C_3) = 1/3$

Terminology

- **Prior**:
  - Probability of a hypothesis before we see any data
- **Uniform Prior**:
  - A prior that makes all hypothesis equally likely
- **Posterior**:
  - Probability of a hypothesis after we saw some data
- **Likelihood**:
  - Probability of data given hypothesis

Experiment 2: Tails

Now, 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

$P(H|C_1) = 0.1$
$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$
$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$
$P(C_3) = 1/3$

Experiment 2: Tails

Now, Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

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

C_1

$P(H|C_1) = 0.1$
$P(C_1) = 1/3$

C_2

$P(H|C_2) = 0.5$
$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$
$P(C_3) = 1/3$

Experiment 2: Tails

Which coin did I use?

$P(C_1|HT) = 0.21$ $P(C_2|HT) = 0.58$ $P(C_3|HT) = 0.21$

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

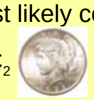
$P(H|C_3) = 0.9$
$P(C_3) = 1/3$

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: Best estimate for P(H)

C_2



$P(H|C_2) = 0.5$

C_1

$P(H|C_1) = 0.1$
$P(C_1) = 1/3$

C_2

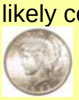
$P(H|C_2) = 0.5$
$P(C_2) = 1/3$

C_3

$P(H|C_3) = 0.9$
$P(C_3) = 1/3$

Your Estimate?

Maximum Likelihood Estimate: The best hypothesis that fits observed data assuming uniform prior

Most likely coin: C_2 

Best estimate for P(H): $P(H|C_2) = 0.5$

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

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 C_2 C_3

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

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 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?

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

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

C_1 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?

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

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

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	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?

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

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

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 C_2 C_3

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

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 C_2 C_3

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

Comparison

After more experiments: HTHHHHHHHHH

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

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

Maximum Likelihood Estimate: Easy to compute

Maximum A Posteriori Estimate: Still easy to compute; Incorporates prior knowledge

Bayesian Estimate: Minimizes error; Great when data is scarce; Potentially much harder to compute

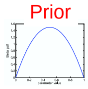
Bayesian Learning

Use Bayes rule:

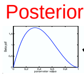
$$P(Y \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid Y)\, P(Y)}{P(\mathbf{X})}$$

Data Likelihood: $P(\mathbf{X} \mid Y)$

Prior: $P(Y)$



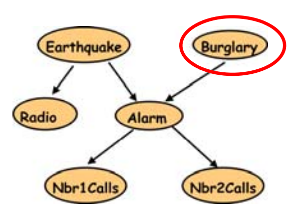
Posterior: $P(Y \mid \mathbf{X})$



Normalization: $P(\mathbf{X})$

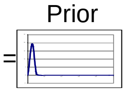
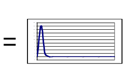
Or equivalently: $P(Y \mid \mathbf{X}) \propto P(\mathbf{X} \mid Y)\, P(Y)$

Parameter Estimation and Bayesian Networks



B
F
F
T
F
T

Prior

P(B) =  + data = 

Now compute either MAP or Bayesian estimate

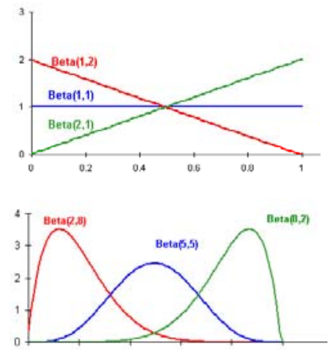
What Prior to Use?

- 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

 43

Beta Distribution



Beta Distribution

- Example: Flip coin with Beta 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

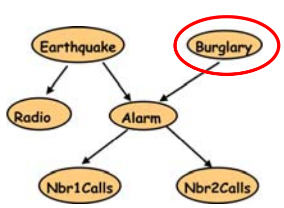
$$\beta_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},$$

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

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

For any positive integer y, $\Gamma(y) = (y-1)!$

Parameter Estimation and Bayesian Networks



B
F
F
T
F
T

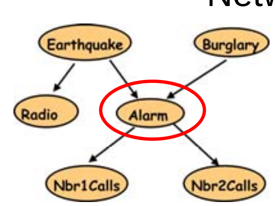
Prior

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

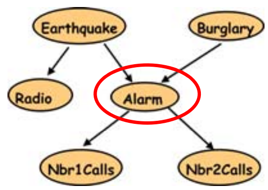


E	B
T	F
F	F
F	T
F	F
F	T
…	

A
T
F
T
T
F

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

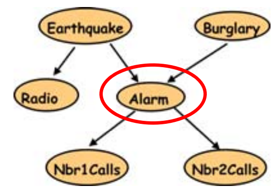
Parameter Estimation and Bayesian Networks



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

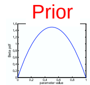
P(A|E,B) = ?   Prior
P(A|E,¬B) = ?
P(A|¬E,B) = Beta(2,3) + data= (3,4)
P(A|¬E,¬B) = ?

Bayesian Learning

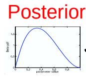
Use Bayes rule:

Data Likelihood

Prior



Posterior



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

Normalization

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