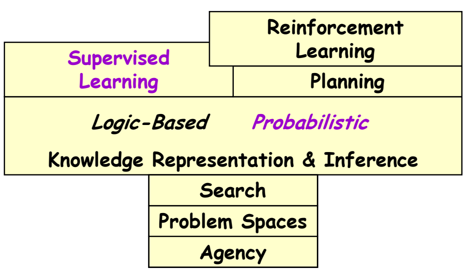
Statistical Learning

CSE 573

Logistics

- Team Meetings
- Midterm
  - Open book, notes
  - Studying
    - See AIMA exercises

573 Topics



Topics

- Parameter Estimation:
  - Maximum Likelihood (ML)
  - Maximum A Posteriori (MAP)
  - Bayesian
  - Continuous case
- Learning Parameters for a Bayesian Network
- Naive Bayes
  - Maximum Likelihood estimates
  - Priors
- Learning Structure of 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 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 C_2 C_3

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

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

Terminology

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

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

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

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

Experiment 2: Tails

Which coin did I use?

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

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

C_1 C_2 C_3

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

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

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 C_2 C_3

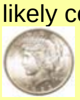
$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

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

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: C_2



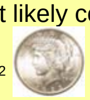
Best estimate for P(H): $P(H|C_2) = 0.5$

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

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

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 <u>did</u> 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 <u>did</u> 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 <u>did</u> I use?

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

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

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

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

Experiment 2: Tails

Which coin <u>did</u> I use?

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

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

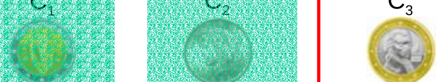
$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

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

Experiment 2: Tails

Which coin <u>did</u> I use?

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



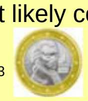
$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

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

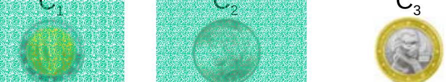
Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: C_3



Best estimate for P(H): $P(H|C_3) = 0.9$



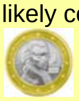
$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

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

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?

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

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

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

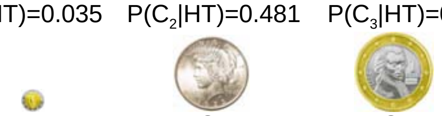
$P(H|C_1) = 0.1$ $P(H|C_2) = 0.5$ $P(H|C_3) = 0.9$

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error, given data and (generally) assuming a ***non-uniform 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: HTH^8

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

Comparison

ML (Maximum Likelihood):
- Easy to compute

MAP (Maximum A Posteriori):
- Still easy to compute
- Incorporates prior knowledge

Bayesian:
- Minimizes error => great when data is scarce
- Potentially much harder to compute

Summary For Now

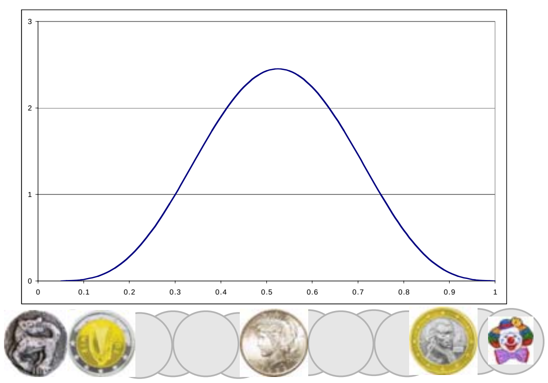
	Prior	Hypothesis
Maximum Likelihood Estimate	Uniform	The most likely
Maximum A Posteriori Estimate	Any	The most likely
Bayesian Estimate	Any	Weighted combination

Continuous Case

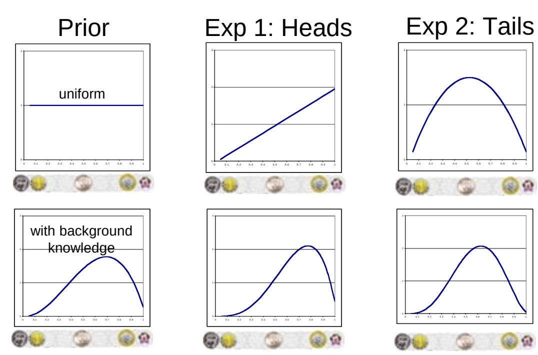
- In the previous example,
  we chose from a discrete set of three coins
- In general,
  we have to pick from a continuous distribution of biased coins

Continuous Case

Continuous Case

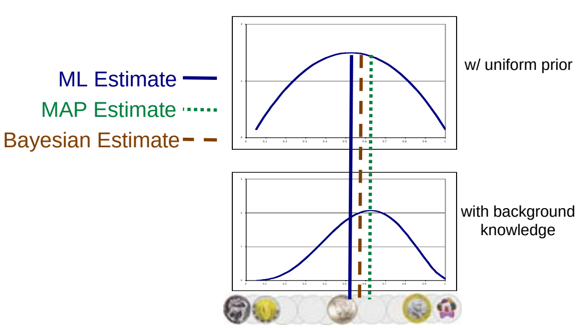


Continuous Case



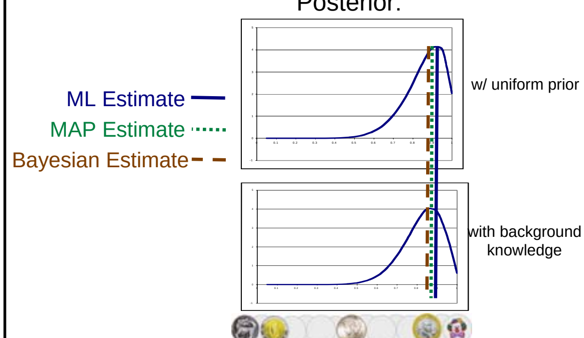
Continuous Case

Posterior after 2 experiments:



After 10 Experiments...

Posterior:



After 100 Experiments...

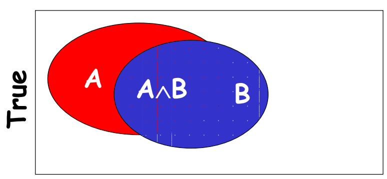
Topics

- Parameter Estimation:
  - Maximum Likelihood (ML)
  - Maximum A Posteriori (MAP)
  - Bayesian
  - Continuous case
- Learning Parameters for a Bayesian Network
- Naive Bayes
  - Maximum Likelihood estimates
  - Priors
- Learning Structure of Bayesian Networks

Review: Conditional Probability

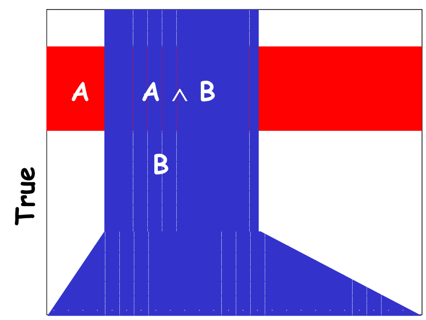
- P(*A* | *B*) is the probability of *A* given *B*
- Assumes that *B* is the only info known.
- Defined by:

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



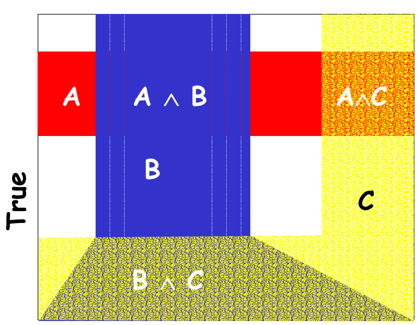
Conditional Independence

A&B *not* independent, since P(A|B) < P(A)



Conditional Independence

But: A&B are *made* independent by ¬C



P(A|¬C) = P(A|B,¬C)

Bayes Rule

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

Simple proof from def of conditional probability:

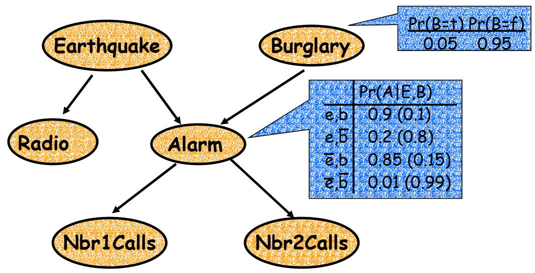
$$P(H \mid E) = \frac{P(H \wedge E)}{P(E)} \quad \text{(Def. cond. prob.)}$$

$$P(E \mid H) = \frac{P(H \wedge E)}{P(H)} \quad \text{(Def. cond. prob.)}$$

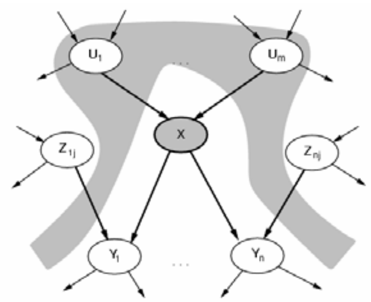
$$P(H \wedge E) = P(E \mid H)P(H) \quad \text{(Mult by P(H) in line 1)}$$

QED: $$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)} \quad \text{(Substitute \#3 in \#2)}$$

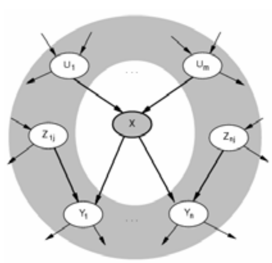
An Example Bayes Net



Given Parents, X is Independent of Non-Descendants

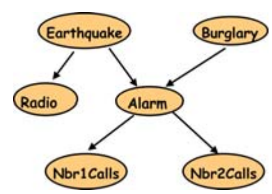


Given Markov Blanket, X is Independent of All Other Nodes



MB(X) = Par(X) ∪ Childs(X) ∪ Par(Childs(X))

Parameter Estimation and Bayesian Networks



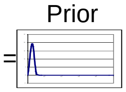
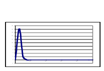
E	B	R	A	J	M
T	F	T	T	F	T
F	F	F	F	F	T
F	T	F	T	T	T
F	F	F	T	T	T
F	T	F	F	F	F
...					

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

Parameter Estimation and Bayesian Networks

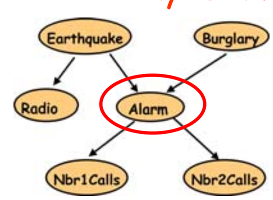
B
F
F
T
F
T

Prior

P(B) =  + data = 

Now compute either MAP or Bayesian estimate

Parameter Estimation and Bayesian Networks

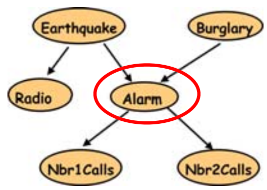


E	B
T	F
F	F
F	T
F	F
F	T
...	

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

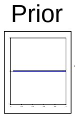


E	B
T	F
F	F
F	T
F	F
F	T
...	

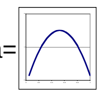
A
T
F
T
T
F

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

Prior



+ data=



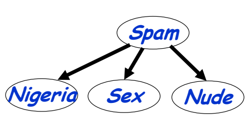
Now compute either MAP or Bayesian estimate

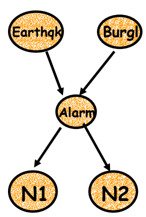
Topics

- Parameter Estimation:
  - Maximum Likelihood (ML)
  - Maximum A Posteriori (MAP)
  - Bayesian
  - Continuous case
- Learning Parameters for a Bayesian Network
- ~~Naive Bayes~~
  - ~~Maximum Likelihood estimates~~
  - ~~Priors~~
- Learning Structure of Bayesian Networks

Recap

- Given a BN structure (with discrete or continuous variables), we can learn the parameters of the conditional prop tables.





What if we *don't* know structure?

Learning The Structure of Bayesian Networks

- Search thru the space…
  - of possible network structures!
  - (for now, assume we observe all variables)
- For each structure, learn parameters
- Pick the one that fits observed data best
  - Caveat – won't we end up fully connected????

Problem !?!?

- When scoring, add a penalty
  - ∝ model complexity

Learning The Structure of Bayesian Networks

- Search thru the space
- For each structure, learn parameters
- Pick the one that fits observed data best
- Problem?
  - Exponential number of networks!
  - And we need to learn parameters for each!
  - Exhaustive search out of the question!
- So what now?

Learning The Structure of Bayesian Networks

Local search!

Start with some network structure

Try to make a change

(add or delete or reverse edge)

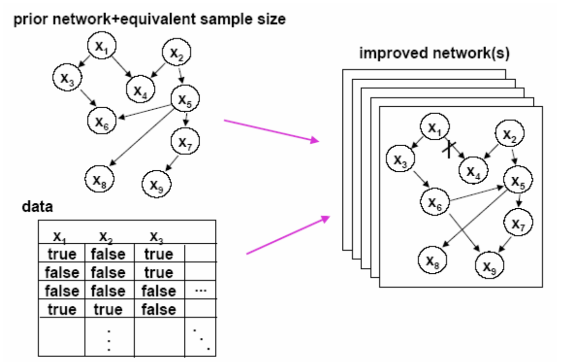
See if the new network is any better

What should be the initial state?

Initial Network Structure?

- Uniform prior over random networks?
- Network which reflects expert knowledge?

Learning BN Structure



X_1	X_2	X_3	
true	false	true	
false	false	true	
false	false	false	...
true	true	false	
	⋮		⋱

 57

The Big Picture

- We described how to do MAP (and ML) learning of a Bayes net (including structure)
- How would Bayesian learning (of BNs) differ?
  - •Find all possible networks
  - •Calculate their posteriors
  - •When doing inference, return weighed combination of predictions from all networks!