Statistical Learning

Today

- 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?

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

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

C_1 C_2 C_3



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

$\boxed{P(C_1) = 1/3}$ $P(C_2) = 1/3$ $P(C_3) = 1/3$

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



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

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

Experiment 2: Tails

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

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

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

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

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

C_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

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_2

$P(H|C_2) = 0.5$

$P(C_2) = 1/3$

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 Uniform Prior?
- Background knowledge:
  - Heads => you go first in Abalone against TA
  - TAs are nice people
  - => TA is more likely to use a coin biased in your 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?

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

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_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Your Estimate?

What is the probability of heads after two experiments?

Most likely coin: C_3 

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

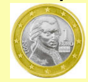
C_3

$P(H|C_3) = 0.9$

$P(C_3) = 0.70$

Your Estimate?

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

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?

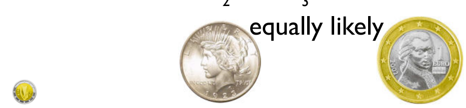
$P(C_1|HT) = 0.035$ $P(C_2|HT) = 0.481$ $P(C_3|HT) = 0.485$

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

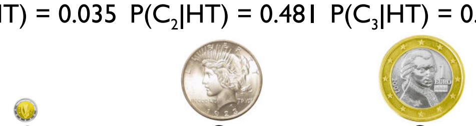


C_1	C_2	C_3
$P(H|C_1) = 0.1$	$P(H|C_2) = 0.5$	$P(H|C_3) = 0.9$

A Better Estimate

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

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



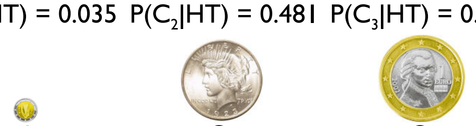
C_1	C_2	C_3
$P(H|C_1) = 0.1$	$P(H|C_2) = 0.5$	$P(H|C_3) = 0.9$

Bayesian Estimate

Bayesian Estimate: Minimizes prediction error, given data and (generally) assuming a <u>non-uniform prior</u>

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

- ML (Maximum Likelihood):
  P(H) = 0.5
- MAP (Maximum A Posteriori):
  P(H) = 0.9
- Bayesian:
  P(H) = 0.68

Comparison

- ML (Maximum Likelihood):
  P(H) = 0.5
  after 10 experiments (HTH^8): P(H) = 0.9
- MAP (Maximum A Posteriori):
  P(H) = 0.9
  after 10 experiments (HTH^8): P(H) = 0.9
- Bayesian:
  P(H) = 0.68
  after 10 experiments (HTH^8): P(H) = 0.9

Comparison

- ML (Maximum Likelihood):
- MAP (Maximum A Posteriori):
- Bayesian:
  - Minimizes error => great when data is scarce
  - Potentially much harder to compute

Comparison

- ML (Maximum Likelihood):
- MAP (Maximum A Posteriori):
  - Still easy to compute
  - Incorporates prior knowledge
- Bayesian:
  - Minimizes error => great when data is scarce
  - Potentially much harder to compute

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**:
- **Uniform Prior**:
- **Posterior**:
- **Likelihood**:

Summary For Now

- **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

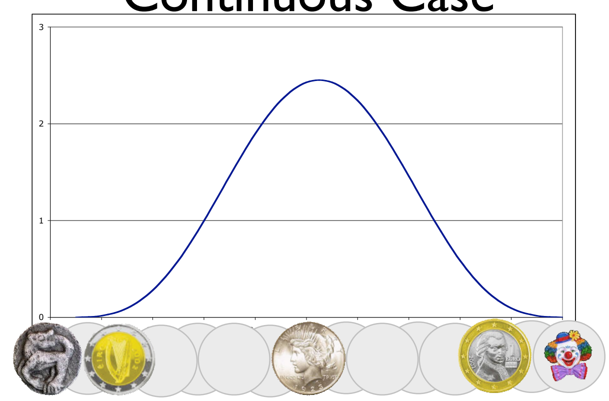
	Prior	Hypothesis	
Maximum Likelihood Estimate	Uniform	The most likely	Point
Maximum A Posteriori Estimate	Any	The most likely	Point
Bayesian Estimate	Any	Weighted combination	Average

Continuous Case

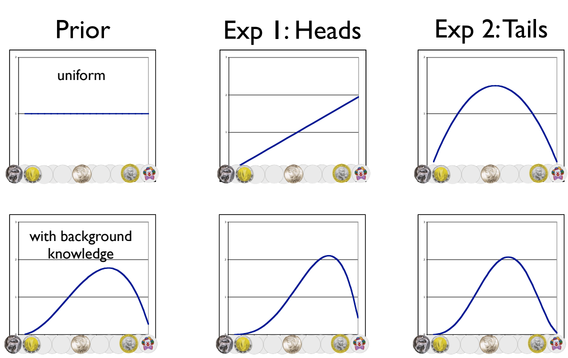
- In the previous example, we chose from a discrete set of three coins
- In general, we have to pick from a continuous distribution of biased coins

Continuous Case

Continuous Case

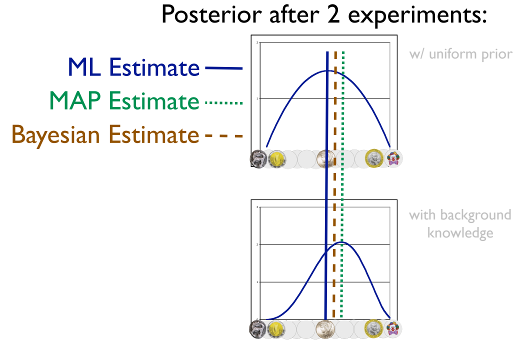


Continuous Case

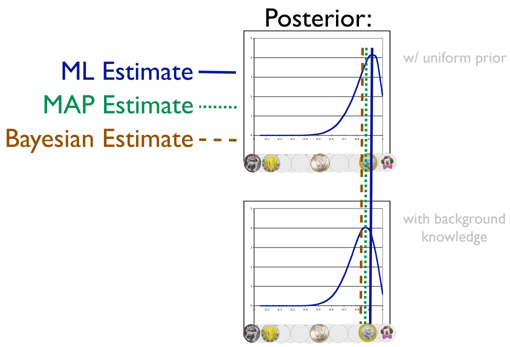


Continuous Case

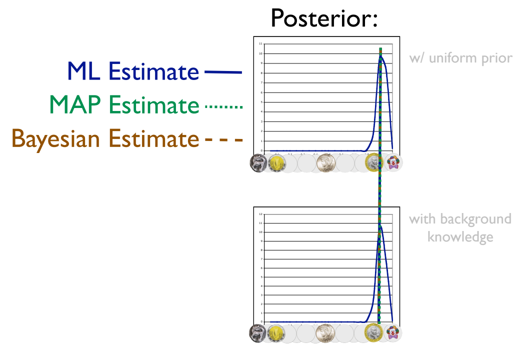
Posterior after 2 experiments:



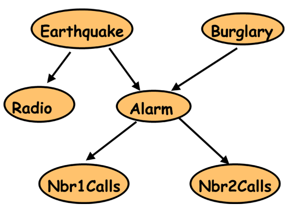
After 10 Experiments...



After 100 Experiments...



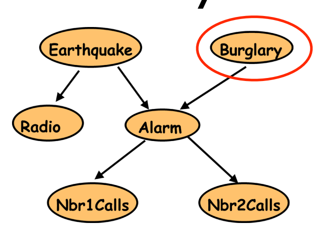
Parameter Estimation and Bayesian Networks



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
...					

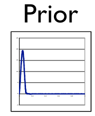
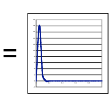
We have:
- Bayes Net structure and observations
- We need: Bayes Net parameters

Parameter Estimation and Bayesian Networks

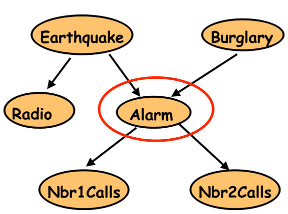


B
F
F
T
F
T

Prior

P(B) = ?  + data =  Now compute either MAP or Bayesian estimate

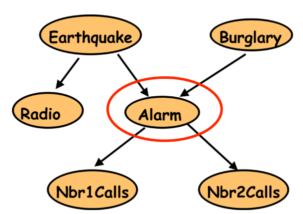
Parameter Estimation and Bayesian Networks



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

A
T
F
T
T
F

Parameter Estimation and Bayesian Networks

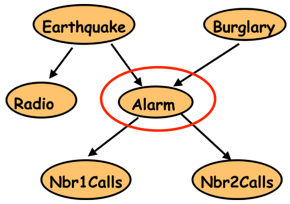


E	B
T	F
F	F
F	T
F	F
F	T
...	

A
T
F
T
T
F

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

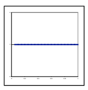
Parameter Estimation and Bayesian Networks



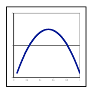
E	B	A
~~T~~	~~F~~	~~T~~
~~F~~	~~F~~	~~F~~
F	T	T
~~F~~	~~F~~	~~T~~
F	T	F
...		

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

Prior

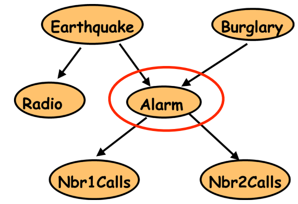


+ data =



Now compute either MAP or Bayesian estimate

Parameter Estimation and Bayesian Networks

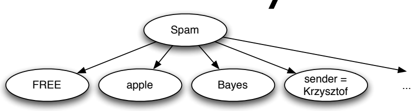


E	B	A
T	F	T
F	F	F
F	T	T
F	F	T
F	T	F
...		

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

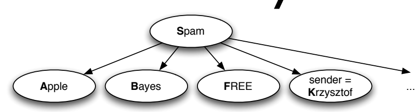
You know the drill...

Naive Bayes



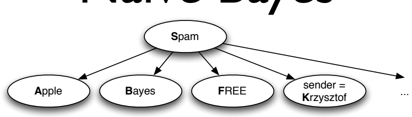
- A Bayes Net where all nodes are children of a single root node
- Why?
  - Expressive and accurate?
  - Easy to learn?

Naive Bayes



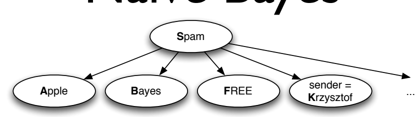
- A Bayes Net where all nodes are children of a single root node
- Why?
  - Expressive and accurate? **No** - why?
  - Easy to learn?

Naive Bayes



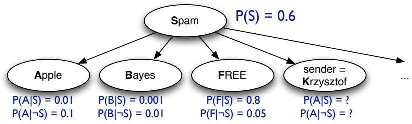
- A Bayes Net where all nodes are children of a single root node
- Why?
  - Expressive and accurate? **No**
  - Easy to learn? **Yes**

Naive Bayes



- A Bayes Net where all nodes are children of a single root node
- Why?
  - Expressive and accurate? **No**
  - Easy to learn? **Yes**
  - Useful? **Sometimes**

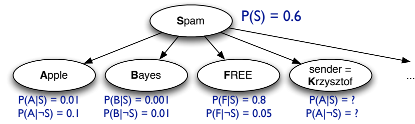
Inference In Naive Bayes



- Goal, given evidence (words in an email) decide if an email is spam

$$E = \{A, \neg B, F, \neg K, \ldots\}$$

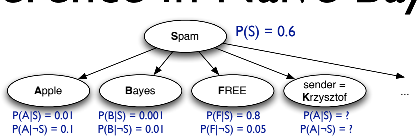
Inference In Naive Bayes



$$\begin{aligned} P(S|E) &= \frac{P(E|S)P(S)}{P(E)} \\ &= \frac{P(A, \neg B, F, \neg K, \ldots|S)P(S)}{P(A, \neg B, F, \neg K, \ldots)} \\ &= \frac{P(A|S)P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S)}{P(A)P(\neg B)P(F)P(\neg K)P(\ldots)} \end{aligned}$$

Independence to the rescue!

Inference In Naive Bayes



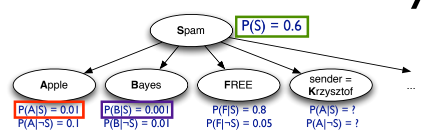
$$P(S|E) = \frac{P(A|S)P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S)}{\boxed{P(A)P(\neg B)P(F)P(\neg K)P(\ldots)}}$$

$$P(\neg S|E) = \frac{P(A|\neg S)P(\neg B|\neg S)P(F|\neg S)P(\neg K|\neg S)P(\ldots|\neg S)P(\neg S)}{\boxed{P(A)P(\neg B)P(F)P(\neg K)P(\ldots)}}$$

Spam if P(S|E) > P(¬S|E)

But...

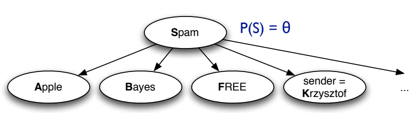
Inference In Naive Bayes



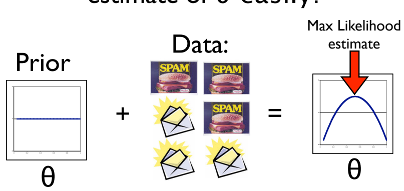
$$P(S|E) \propto \boxed{P(A|S)}\boxed{P(\neg B|S)}P(F|S)P(\neg K|S)P(\ldots|S)\boxed{P(S)}$$

$$P(\neg S|E) \propto P(A|\neg S)P(\neg B|\neg S)P(F|\neg S)P(\neg K|\neg S)P(\ldots|\neg S)P(\neg S)$$

Parameter Estimation Revisited

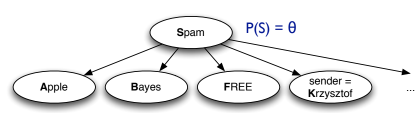


- Can we calculate Maximum Likelihood estimate of θ easily?



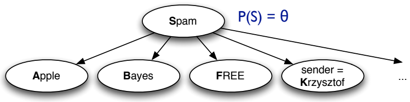
Looking for the maximum of a function:
- find the derivative
- set it to zero

Parameter Estimation Revisited



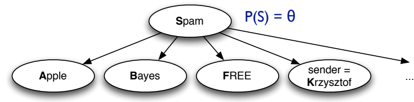
- What function are we maximizing? P(data|hypothesis)

Parameter Estimation Revisited



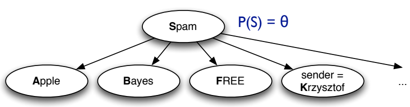
- What function are we maximizing? P(data|hypothesis)
- hypothesis = h_θ (one for each value of θ)

Parameter Estimation Revisited



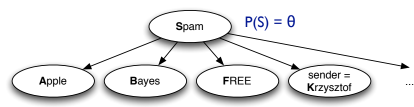
- What function are we maximizing? P(data|hypothesis)
- hypothesis = h_θ (one for each value of θ)
- $P(data|h_\theta)$ = P($|h_\theta$)P($|h_\theta$)P($|h_\theta$)P($|h_\theta$)

Parameter Estimation Revisited



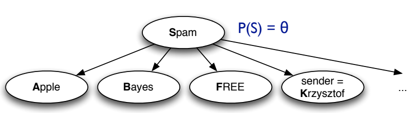
- What function are we maximizing? P(data|hypothesis)
- hypothesis = h_θ (one for each value of θ)
- $P(data|h_\theta)$ = P($|h_\theta$)P($|h_\theta$)P($|h_\theta$)P($|h_\theta$)
  = $\theta \quad (1-\theta) \quad (1-\theta) \quad \theta$

Parameter Estimation Revisited



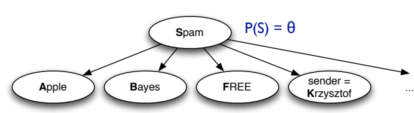
- What function are we maximizing? P(data|hypothesis)
- hypothesis = h_θ (one for each value of θ)
- $P(data|h_\theta)$ = P($|h_\theta$)P($|h_\theta$)P($|h_\theta$)P($|h_\theta$)
  = $\theta \quad (1-\theta) \quad (1-\theta) \quad \theta$
  = $\theta^{\#}$$(1-\theta)^{\#}$

Parameter Estimation Revisited



- To find θ that maximizes $\theta^{\#}$$(1-\theta)^{\#}$ we take a derivative of the function and set it to 0. And we get:

Parameter Estimation Revisited



- To find θ that maximizes $\theta^{\#}$$(1-\theta)^{\#}$ we take a derivative of the function and set it to 0. And we get:
- P(S) = θ = #  / (#  + # )
- You knew it already, right?

Problems With Small Samples

- What happens if in your training data apples are not mentioned in any spam message?
- P(A|S) = 0
- Why is it bad?

$P(S|E) \propto$ **0** $P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S)$ **= 0**

Smoothing

- Smoothing is used when samples are small
- Add-one smoothing is the simplest smoothing method: just add 1 to every count!

Priors!

- Recall that P(S) = $\frac{\#}{\# + \#}$

Priors!

- Recall that P(S) = $\frac{\#}{\# + \#}$
- If we have a slight hunch that P(S) ≈ *p*

P(S) = $\frac{\# \; + p}{\# + \# + 1}$

Priors!

- Recall that P(S) = $\frac{\#}{\# + \#}$
- If we have a slight hunch that P(S) ≈ *p*

P(S) = $\frac{\# \; + p}{\# + \# + 1}$

- If we have a **big** hunch that P(S) ≈ *p*

$\frac{\# \; + mp}{\# + \# + m}$

where *m* can be any number > 0

Priors!

P(S) = $\frac{\# \; + mp}{\# + \# + m}$

- Note that if *m* = 10 in the above, it is like saying "I have seen 10 samples that make me believe that P(S) = *p*"
- Hence, m is referred to as the equivalent sample size

Priors!

P(S) = (#  + mp) / (#  + #  + m)

- Where should p come from?
- No prior knowledge => p=0.5
- If you build a personalized spam filter, you can use p = P(S) from some body else's filter!

Inference in Naive Bayes Revisited

- Recall that

$$P(S|E) \propto P(A|S)P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S)$$

Is there any potential for trouble here?

Inference in Naive Bayes Revisited

- Recall that

$$P(S|E) \propto P(A|S)P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S)$$

- We are multiplying lots of small numbers together => danger of underflow!
- Solution? Use logs!

Inference in Naive Bayes Revisited

$$log(P(S|E)) \propto log(P(A|S)P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S))$$

$$\propto log(P(A|S)) + log(P(\neg B|S)) + log(P(F|S)) + log(P(\neg K|S)) + log(P(\ldots|S)) + log(P(S)))$$

- Now we add "regular" numbers -- little danger of over- or underflow errors!

Learning The Structure of Bayesian Networks

- General idea: look at all possible network structures and pick one that fits observed data best
- Impossibly slow: exponential number of networks, and for each we have to learn parameters, too!
- What do we do if searching the space exhaustively is too expensive?

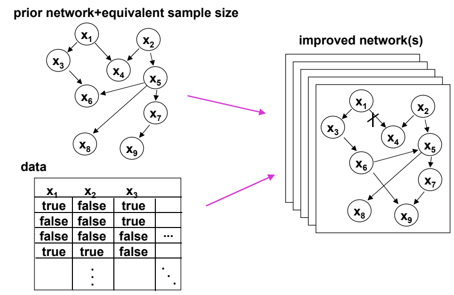
Learning The Structure of Bayesian Networks

- Local search!
  - Start with some network structure
  - Try to make a change (add, delete or reverse node)
  - See if the new network is any better

Learning The Structure of Bayesian Networks

- What network structure should we start with?
  - Random with uniform prior?
  - Networks that reflects our (or experts') knowledge of the field?

Learning The Structure of Bayesian Networks



x_1	x_2	x_3	
true	false	true	
false	false	true	
false	false	false	...
true	true	false	
	⋮		⋱

Learning The Structure of Bayesian Networks

- We have just described how to get an ML or MAP estimate of the structure of a Bayes Net
- What would the Bayes estimate look like?
  - Find all possible networks
  - Calculate their posteriors
  - When doing inference: result weighed combination of all networks!