Naïve Bayes & Expectation Maximization

CSE 573

Logistics

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

573 Schedule

Selected Topics

Supervised Learning

Reinforcement Learning

Planning

Logic-Based *Probabilistic*

Knowledge Representation & Inference

Search

Problem Spaces

Agency

Coming Soon

Selected Topics

Artificial Life

Crossword Puzzles

Intelligent Internet Systems

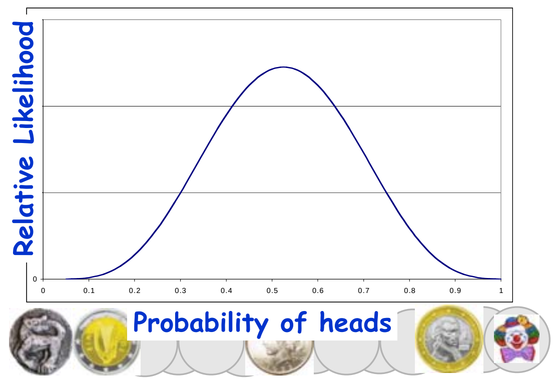
Topics

- Test & Mini Projects
- Review
- Naive Bayes
  - Maximum Likelihood Estimates
  - Working with Probabilities
- Expectation Maximization
- Challenge

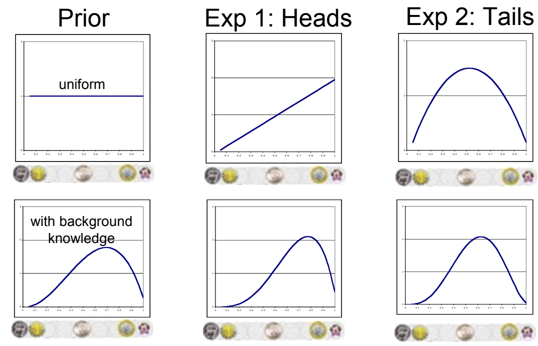
Estimation Models

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

Continuous Case

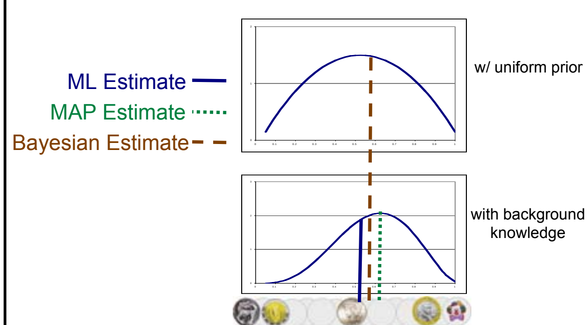


Continuous Case



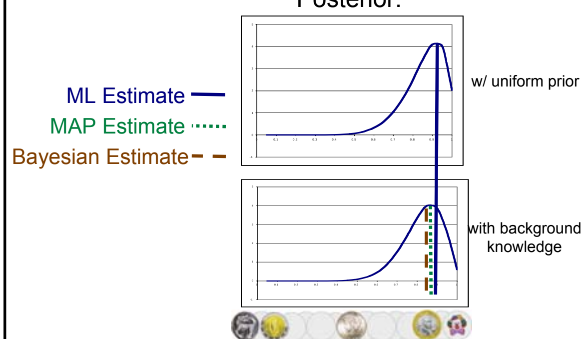
Continuous Case

Posterior after 2 experiments:



After 10 Experiments...

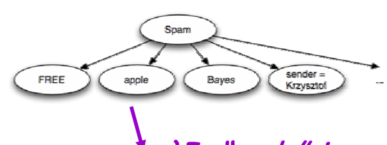
Posterior:



Topics

- Test & Mini Projects
- Review
- Naive Bayes
  - Maximum Likelihood Estimates
  - Working with Probabilities
- Expectation Maximization
- Challenge

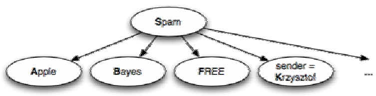
Naive Bayes



= 'Is "apple" in message?'

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

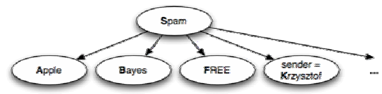
Naive Bayes



- All nodes are children of a single root node
- Why?

Expressive and accurate? **No** - why?
Easy to learn?

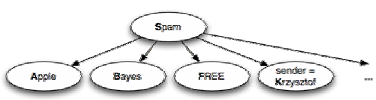
Naive Bayes



- All nodes are children of a single root node
- Why?

Expressive and accurate? **No**
Easy to learn? **Yes**

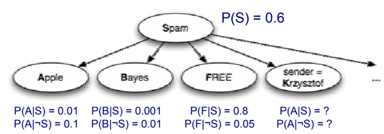
Naive Bayes



- All nodes are children of a single root node
- Why?

Expressive and accurate? **No**
Easy to learn? **Yes**
Useful? **Sometimes**

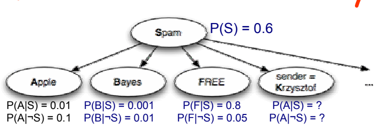
Inference In Naive Bayes



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

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

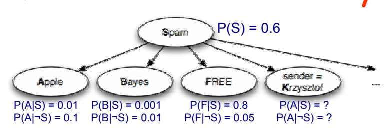
Inference In Naive Bayes



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

Independence to the rescue!

Inference In Naive Bayes

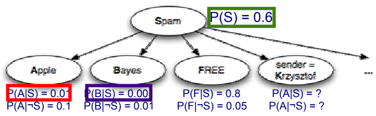


$$P(S|E) = \frac{P(A|S)P(\neg B|S)P(F|S)P(\neg K|S)P(\ldots|S)P(S)}{\boxed{P(E)}}$$
$$P(\neg S|E) = \frac{P(A|\neg S)P(\neg B|\neg S)P(F|\neg S)P(\neg K|\neg S)P(\ldots|\neg S)P(\neg S)}{\boxed{P(E)}}$$

Spam if P(S|E) > P(¬S|E)

But...

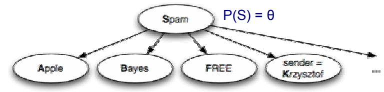
Inference In Naive Bayes



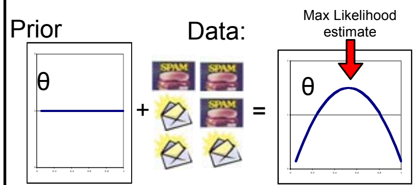
$$P(S|E) \propto \boxed{P(A|S)}\boxed{P(\neg B|S)}P(F|S)P(\neg K|S)P(\ldots|S)\boxed{P(S)}$$

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

Parameter Estimation Revisited



Can we calculate Maximum Likelihood estimate of θ easily?



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

Topics

- Test & Mini Projects
- Review
- Naive Bayes
  - Maximum Likelihood Estimates
  - Working with Probabilities
    - Smoothing
    - Computational Details
    - Continuous Quantities
- Expectation Maximization
- Challenge

Evidence is Easy?

$$P(X_i \mid S) = \frac{\#}{\# + \#}$$

- Or…. Are their problems?

Smooth with a Prior

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

p = prior probability
m = weight

Note that if m = 10, it means "I've seen 10 samples that make me believe $P(X_i \mid S) = p$"

Hence, m is referred to as the equivalent sample size

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

P(S | X)

- Easy to compute from data if X discrete

Instance	X	Spam?
1	T	F
2	T	F
3	F	T
4	T	T
5	T	F

- P(S | X) = $\frac{1}{4}$ *ignoring smoothing…*

P(S | X)

- What if X is real valued?

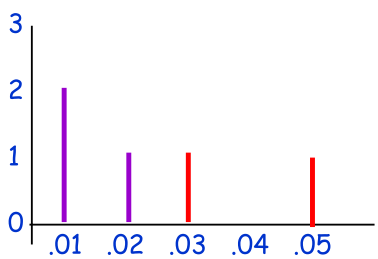
Instance	X	Spam?
1	~~0.01~~ <T	False
2	~~0.01~~ <T	False
3	~~0.02~~ <T	False
4	~~0.03~~ >T	True
5	~~0.05~~ >T	True

- What now?

Anything Else?

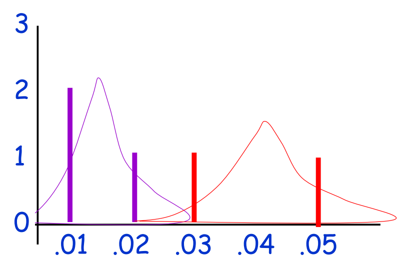
#	X	S?
1	0.01	F
2	0.01	F
3	0.02	F
4	0.03	T
5	0.05	T

P(S|.04)?



Fit Gaussians

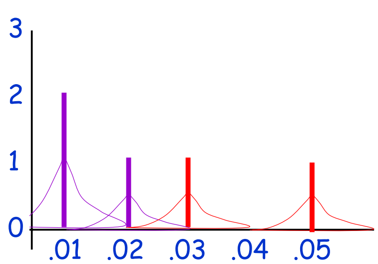
#	X	S?
1	0.01	F
2	0.01	F
3	0.02	F
4	0.03	T
5	0.05	T



Smooth with Gaussian then sum

"Kernel Density Estimation"

#	X	S?
1	0.01	F
2	0.01	F
3	0.02	F
4	0.03	T
5	0.05	T

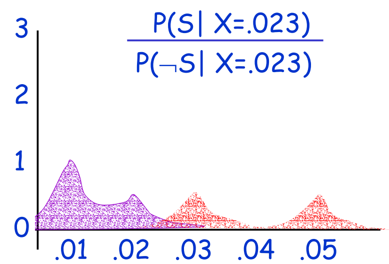


Spam?

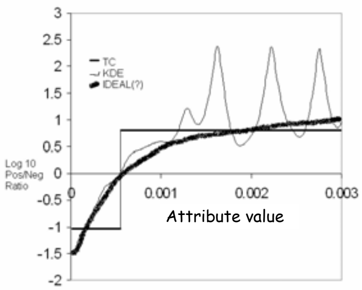
#	X	S?
1	0.01	F
2	0.01	F
3	0.02	F
4	0.03	T
5	0.05	T

$$\frac{P(S|\ X=.023)}{P(\neg S|\ X=.023)}$$

What's with the shape?



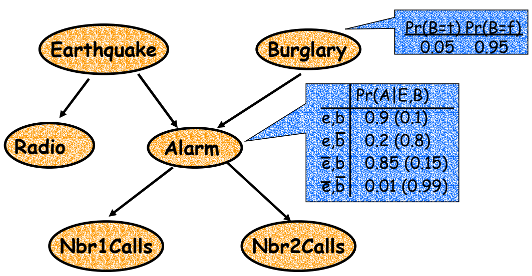
Analysis



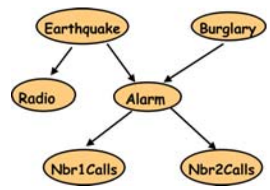
Topics

- Test & Mini Projects
- Review
- Naive Bayes
- Expectation Maximization
  - Review: Learning Bayesian Networks
    - Parameter Estimation
    - Structure Learning
  - Hidden Nodes
- Challenge

An Example Bayes Net



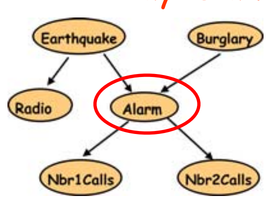
Parameter Estimation and Bayesian Networks



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

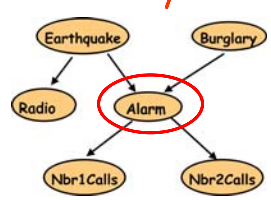


E	B
T	F
F	F
F	T
F	F
F	T
...	

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

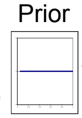


E	B
T	F
F	F
F	T
F	F
F	T
...	

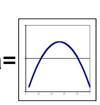
A
T
F
T
T
F

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

Prior



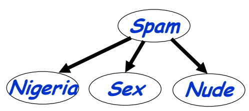
+ data=

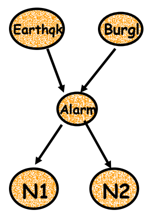


Now compute either MAP or Bayesian estimate

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????
- When scoring, add a penalty
  - $\propto$ model complexity

Problem !?!

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!

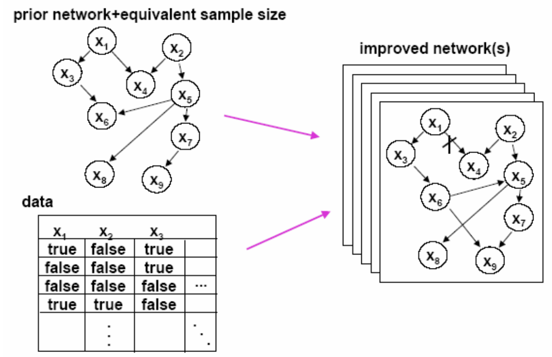
- Start with some network structure
- Try to make a change
  - (add or delete or reverse edge)
- See if the new network is any better

What should be the initial state?

Initial Network Structure?

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

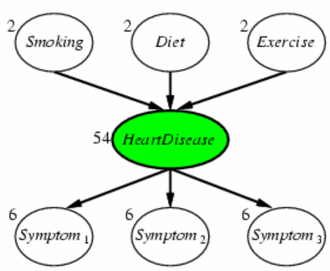
Learning BN Structure



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!

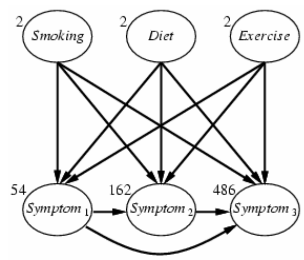
Hidden Variables



- But we can't observe the disease variable
- Can't we learn without it?

We *–could–*

- But we'd get a fully-connected network



- With 708 parameters (*vs.* 78)

  Much harder to learn!

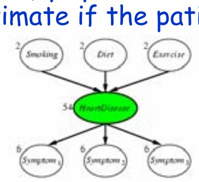
Chicken & Egg Problem

- If we knew that a training instance (patient) had the disease…

  It would be easy to learn P(symptom | disease)
  But we can't observe disease, so we don't.

- If we knew params, e.g. P(symptom | disease) then it'd be easy to estimate if the patient had the disease.

  But we don't know these parameters.



Expectation Maximization (EM)

(high-level version)

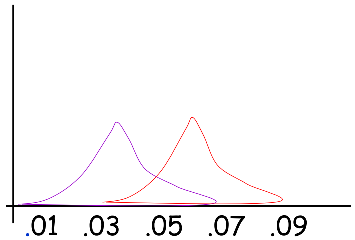
- Pretend we ***do*** know the parameters

  Initialize randomly

- **[E step]** Compute probability of instance having each possible value of the hidden variable
- **[M step]** Treating each instance as fractionally having ***both*** values compute the new parameter values
- Iterate until convergence!

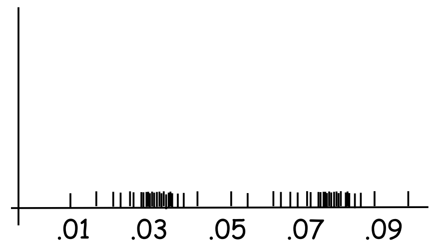
Simplest Version

- Mixture of two distributions

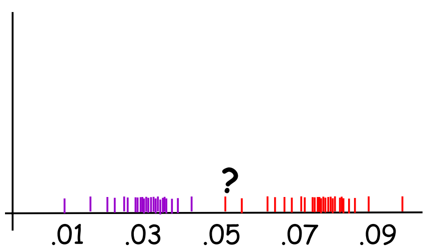


- Know: form of distribution & variance, % =5
- Just need ***mean*** of each distribution

Input Looks Like



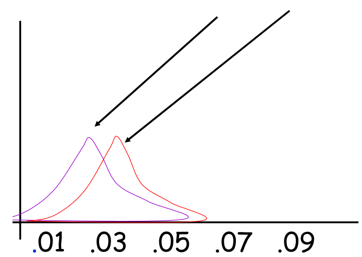
We Want to Predict



Expectation Maximization (EM)

- Pretend we ***do*** know the parameters

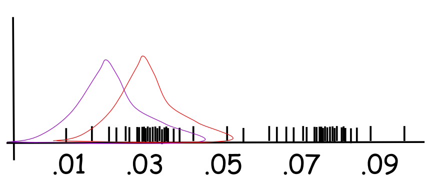
  Initialize randomly: set θ_1=?; θ_2=?



Expectation Maximization (EM)

- Pretend we ***do*** know the parameters

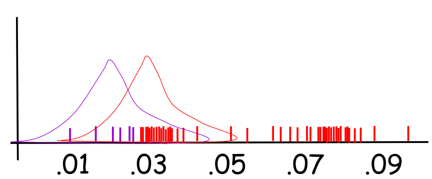
  Initialize randomly
- **[E step]** Compute probability of instance having each possible value of the hidden variable



Expectation Maximization (EM)

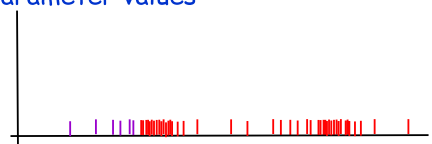
- Pretend we ***do*** know the parameters

  Initialize randomly
- **[E step]** Compute probability of instance having each possible value of the hidden variable



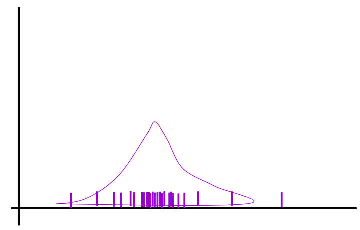
Expectation Maximization (EM)

- Pretend we ***do*** know the parameters
  - Initialize randomly
- **[E step]** Compute probability of instance having each possible value of the hidden variable
- **[M step]** Treating each instance as fractionally having ***both*** values compute the new parameter values



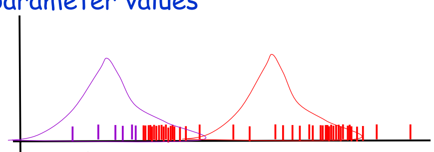
ML Mean of Single Gaussian

$U_{ml} = \text{argmin}_u \sum_i (x_i - u)^2$



Expectation Maximization (EM)

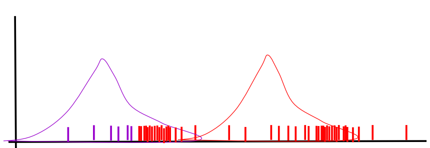
- Pretend we ***do*** know the parameters
  - Initialize randomly
- **[E step]** Compute probability of instance having each possible value of the hidden variable
- **[M step]** Treating each instance as fractionally having ***both*** values compute the new parameter values



Iterate

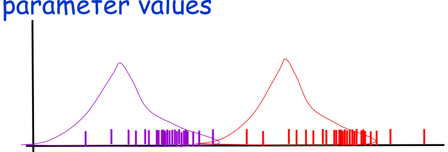
Expectation Maximization (EM)

-
- **[E step]** Compute probability of instance having each possible value of the hidden variable



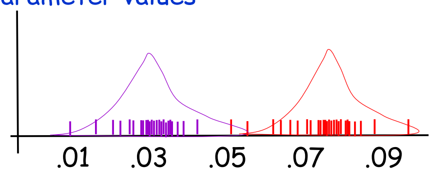
Expectation Maximization (EM)

-
- **[E step]** Compute probability of instance having each possible value of the hidden variable
- **[M step]** Treating each instance as fractionally having ***both*** values compute the new parameter values



Expectation Maximization (EM)

-
- **[E step]** Compute probability of instance having each possible value of the hidden variable
- **[M step]** Treating each instance as fractionally having ***both*** values compute the new parameter values



Until Convergence

- Problems
  - Need to assume form of distribution
  - Local Maxima
- But
  - It really works in practice!
  - Can easilly extend to multiple variables
    - E.g. Mean & Variance
    - Or much more complex models…

Crossword Puzzles