CS / Philo 372

Week 13

Uncertainty
EM Algorithm

Basic Probability

- Conditional – or Posterior - Probability
  - P(A|B) == the probability that A will occur given that B has occurred
  - P(A|B) = P(A and B both occur) / P(B)
    - For example: in a sentence from the “Wizard of Oz” what is the probability of the next word being “witch” given that the previous word was “wicked”
      - P(witch | wicked)
- Prior Probability
  - The probability of something given no information
  - P(“the”) = 0.07 in English text

Bayes Rule

- Bayes Theorem
  - P(A|B) = (P(B|A)/P(B)) P(A)
    - P(B|A)/P(B) is called the Likelihood of A given B
- Example
  - 2 bowls of cookies
    - Bowl A: 30 Chocolate chip, 10 plain
    - Bowl B: 20 Chocolate chip, 20 plain
  - Suppose randomly pick a bowl and then randomly select a cookie from that bowl. When you do so, you get a chocolate chip
  - What is the probability that you picked from bowl A?
    - P(A | cc)
  - Know: P(cc|A) = 0.75, P(cc|B) = 0.5, P(cc)=0.675, P(A) = 0.5
  - P(A|cc) = P(cc|A)*P(A)/P(cc) = 0.75*0.5/0.675 = 0.6

Bayes and Medicine

- Suppose: a disease
  - 1% of the population has the disease
  - 3% of all people tested will test positive
  - 99% of people with disease will test positive
- What is the false positive rate?
- What is the probability that you have the disease if you have a positive test?

Bayes and Medicine

- Suppose: a disease
  - 1% of the population has the disease
  - 3% of all people tested will test positive
  - 99% of people with disease will test positive
  - what is the probability that you have the disease if you have a positive test?
    - P(+) = 0.01 or P(-)=0.99
    - P(tp) = 0.03 or P(tn)=0.97
    - P(tp|+) = 0.99
    - P(tp)=P(-)*P(tp|-)+P(+)*P(tp|+)
      0.03=0.99*P(tp|-)+0.01*0.99
      0.0211/0.99 = 0.0213 =P(tp|-)
  - P(+ | tp) = P(tp | +) P(+) / P(tp)
    - = 0.99*0.01/0.03=0.33

Bayes and Documents

- Want to compute
  - P(is relevant to Q | document D)
    = (P(D|rel) * P(rel)) / P(D)
- So, by Bayes would need:
  - P(D | relevant to Q), P(relevant to Q), P(D)
  - P(is relevant to Q) = probability of picking a relevant document from among all documents. (This is the same for all documents)
  - P(D) = P(t1) * ... for each word in D
    - *This we know, but it drops out as a constant term because it is the same for all documents*
    - *Good thing, it is essentially 0*

Naïve Bayes Classifiers

- Idea – base classification decisions on a Bayes model
  - P(C | F1,F2, ...,Fn) = P(C)*P(F1,...Fn | C)/P(F1..Fn)
  - Note that
    - P(C) is not dependent on data
    - P(F1..Fn) is not dependent on data or classification
    - so only care about P(F1..Fn|C)
  - P(F1..Fn|C) = P(F1|C)*...*P(Fn|C)
- P(C) and P(F|C) can be estimated from training data
- So why is this naïve?

NB Example

No.	Credit History	Debt	Income	Loan Approved?
1	Bad	High	Low	No
2	Unknown	High	Middle	No
3	Unknown	Low	Middle	Yes
4	Unknown	Low	Upper	Yes
5	Unknown	Low	Upper	Yes
6	Bad	Low	Low	No
7	Good	Low	Middle	Yes
8	Good	High	Upper	Yes
9	Good	Low	Low	Yes
10	Bad	High	Upper	No

- P(no)=0.4
- P(chb|no)=1.0
- P(chu)|no)=0.25
- P(chg|no)=0.0
- P(dh|no)=0.75
- P(dl|no)=.1666
- P(IL|no)=.666
- P(Im|no)=.333
- P(Ih|no)=.333

NP Example contd

- So, what does the NB Classifier do with
  - CH=Unknown, Dept=low, Income=low
- P(No|data)/P(yes|data)
  - [0.25/0.75] * [0.166/0.866] * [0.666/0.333]
  - 0.333 * 0.2 * 2
  - .1333
  - since this is less than 1.0 say YES to loan

Application to Spam Filtering

- Most adaptive spam filters are based on naïve bayes classifiers
  - Originally suggested by Paul Graham (2002)
- Main Idea: rather than computing product and taking a threshold compute ratio:
  - P(spam|Doc) / P(not spam | Doc)
    - if > 1 then spam
  - P(spam|Doc) = P(spam)*P(W1|spam)*...*P(Wn|spam)
    - where W1..Wn are the words in a document
  - Usually do things with logs to avoid floating point problems

Graham's Examples

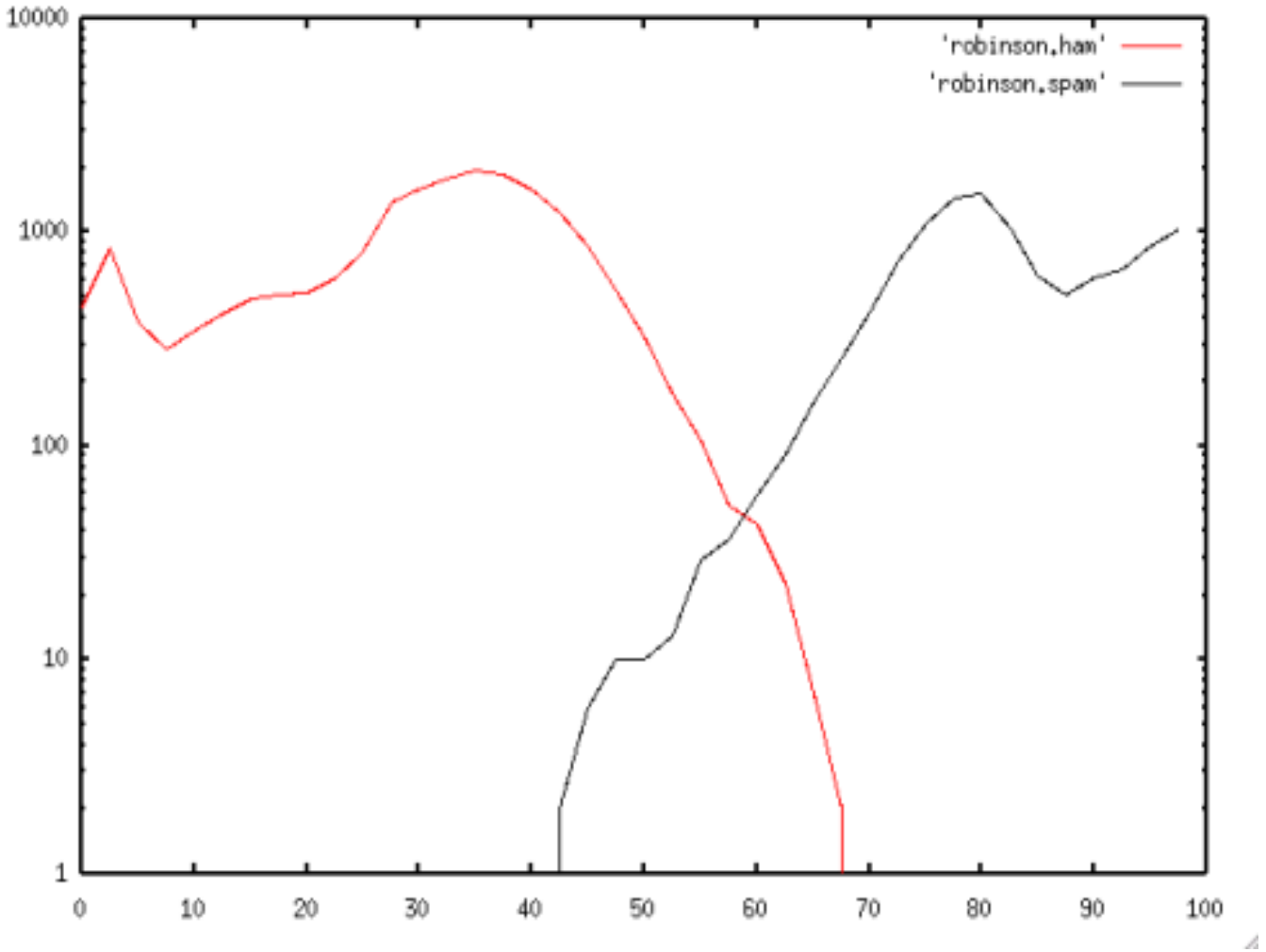
- Words & their conditional probabilities

Spam:
madam 0.99
promotion 0.99
republic 0.99
shortest 0.047225013
mandatory 0.047225013
standardization 0.07347802
sorry 0.08221981
supported 0.09019077
people's 0.09019077
enter 0.9075001
quality 0.8921298
organization 0.12454646
investment 0.8568143
very 0.14758544
valuable 0.82347786

Not Spam:
continuation 0.01
describe 0.01
continuations 0.01
example 0.033600237
programming 0.05214485
i'm 0.055427782
examples 0.07972858
color 0.9189189
localhost 0.09883721
hi 0.116539136
california 0.84421706
same 0.15981844
spot 0.1654587
us-ascii 0.16804294
what 0.19212411

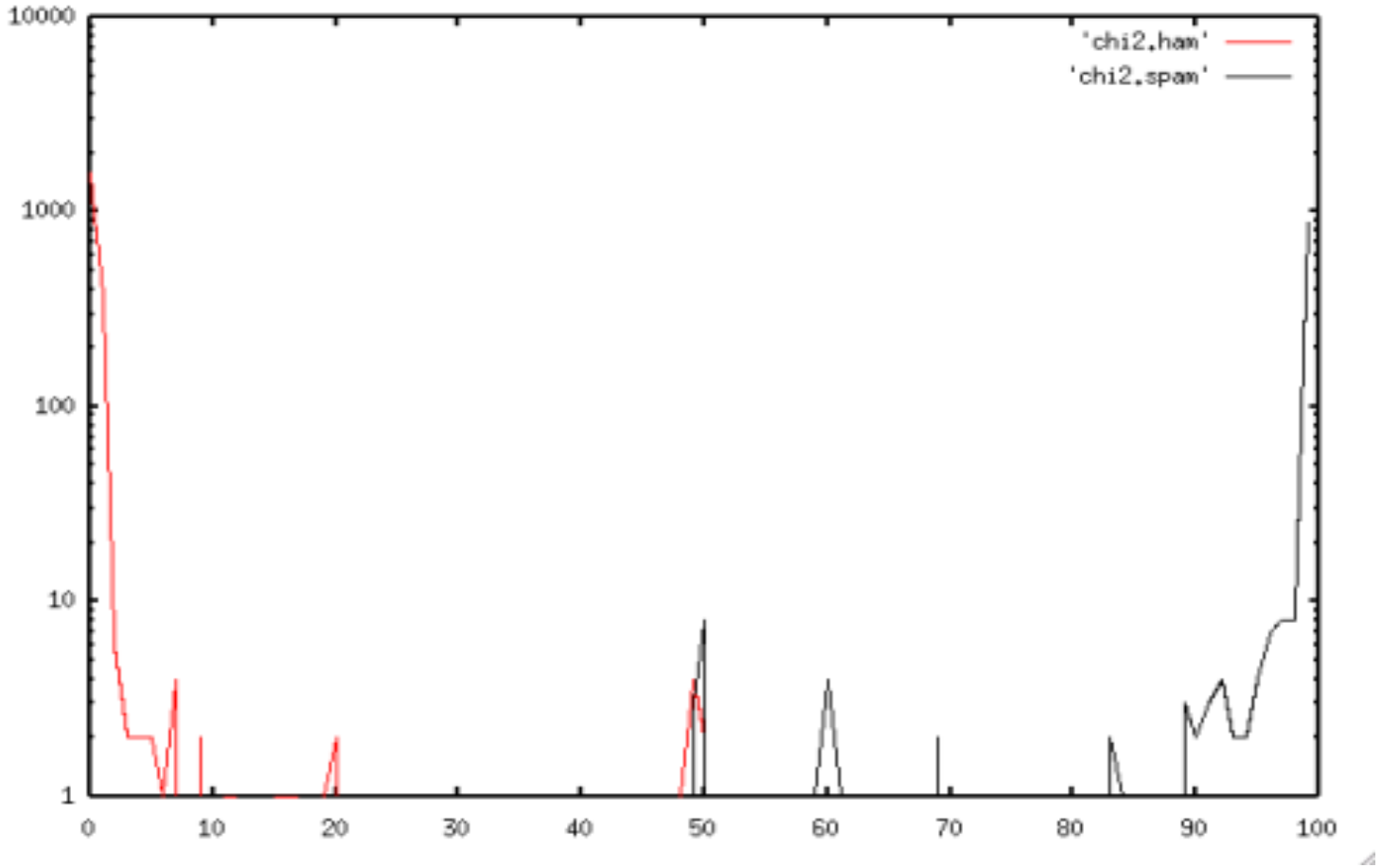
Against of corpus of about 4000 spam and 4000 non-spam

Data on spam filtering



- Xaxis=score
- Yaxis=histogram

A Better Bayes Spam Filter



- Xaxis=score
- Yaxis=histogram
- It looks better, but I worry ..... why?

Problems with NB for spam filtering

- Biggest problem is that spammers know about NB spam filters
  - So they write spam to avoid them
  - How?
    - adaptive spam filter so cannot be not filtered for everyone
- Question, for spam filtering where do you want the threshold?
  - Why?

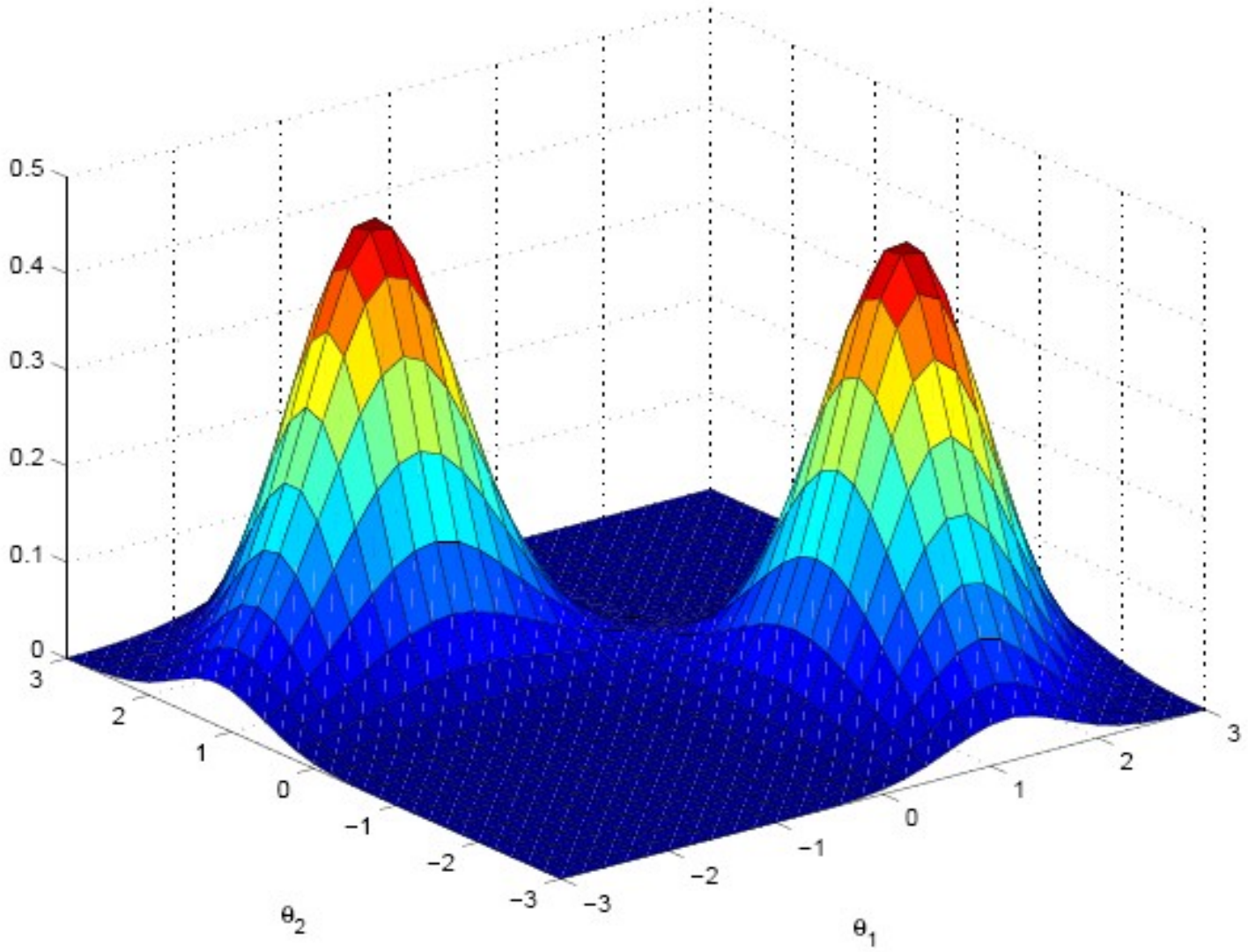
Handling Missing Data

- In decision trees
  - During Construction
    - Do you explicitly have a value "unknown"? If not, then what?
  - During classification
    - traverse all possible trees then weight?
- In NB classifiers
  - Just use the prior on the class?

More Handling Missing Data

- Problem: just admitting that the data is missing and trying to reason around it (as decision trees) does not really solve the problem
- Better approach would be to have the algorithm automatically fill in missing data
  - this is essentially what decision trees do during the classification task, but they only fill in 1 particular missing element,
    - which one?

EM algorithm



EM Algorithm

- If SVM is not most interesting thing in past 20 years of AI then EM is

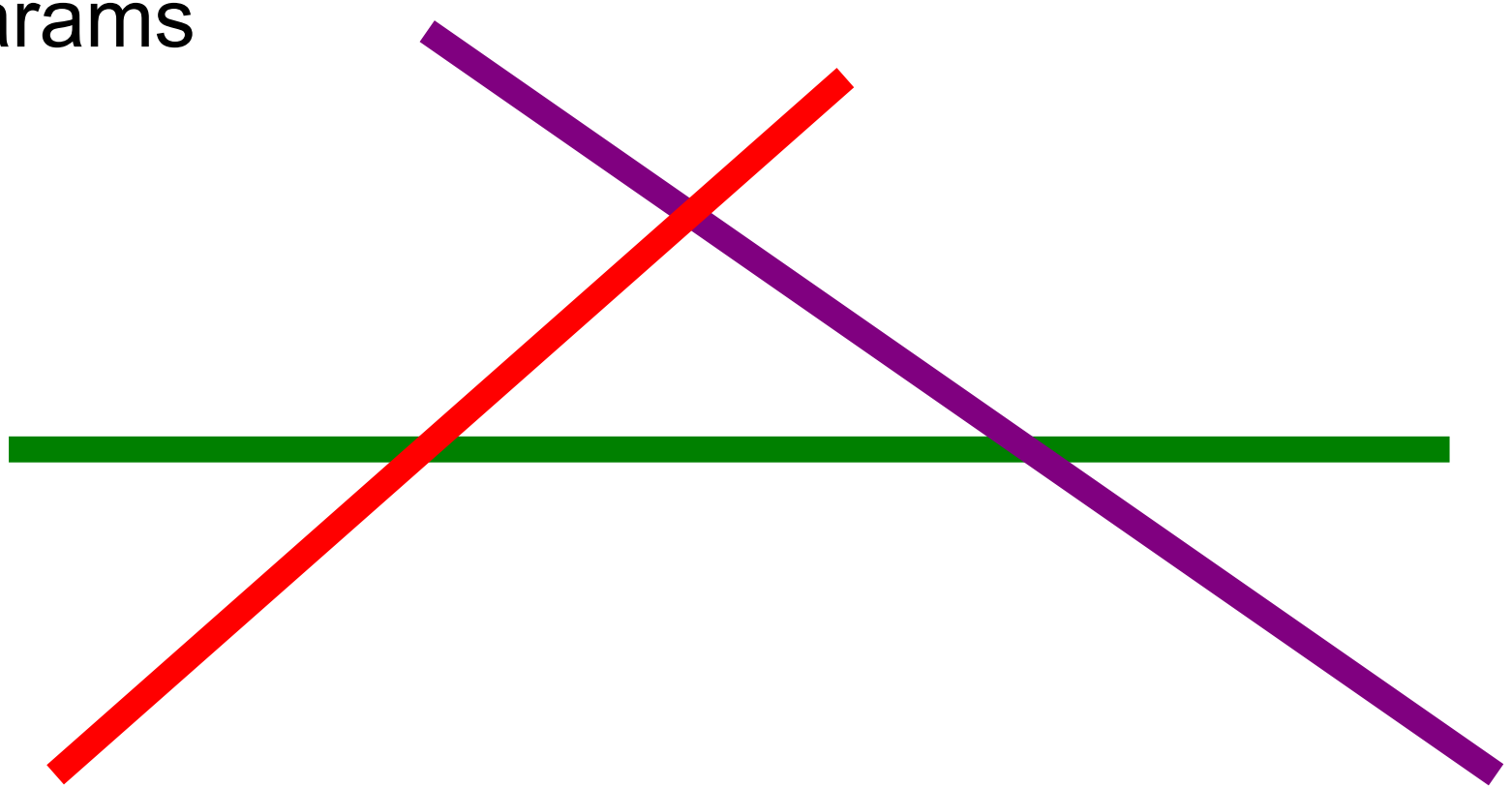
- Iteratively apply
  - E Step -- Expectation Step
    - compute the expected values of the unknown variables
  - M Step – Maximization Step
    - recompute the distributions of all the variable

K-means Clustering Analogy

- Initialize: Pick N starting points
  - set variances & covariances to 1
- Put remaining points with closest of N
  - use Mahalanobis distance
- compute centroid of each of the N groups
  - For each centroid recompute variances and covariances
- Call the centroids the starting points
- Goto "Put"

EM

- Now, rather than just taking the centroids as the values of the missing data, use the data in the cluster to estimate the values of the missing params



EM Issues

- As in K-means clusters may end up covering a single point
  - stats do not work as a result
- The math is really hairy
  - "gaussian mixture models" etc
  - The math is even worse if do not make gaussian assumption
    - the lines on the previous page – NO
    - Mahalanobis distance – NO
- EM can (almost) be seen as mathematical justification for Caruana's MCL