Bayes' law and independence

CS B553
Spring 2013

Announcements

- Readings and lecture notes online on OnCourse
  - Under the "Wiki" tab
- Assignment 1 online now
  - Due next Thursday
  - Work alone or in partnerships
- Office hours change
  - Today's office hours moved to 5:15pm-6:15pm (for today only)

Bayes' Law

- For two events *A* and *B*,

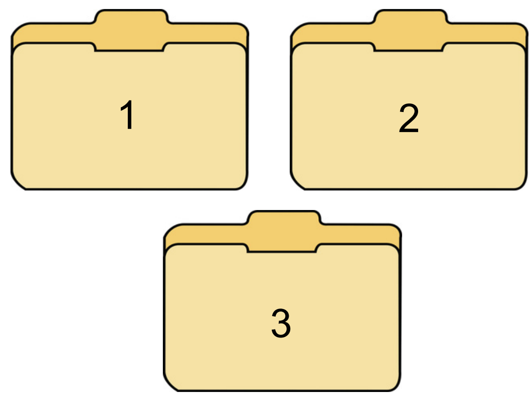
Likelihood

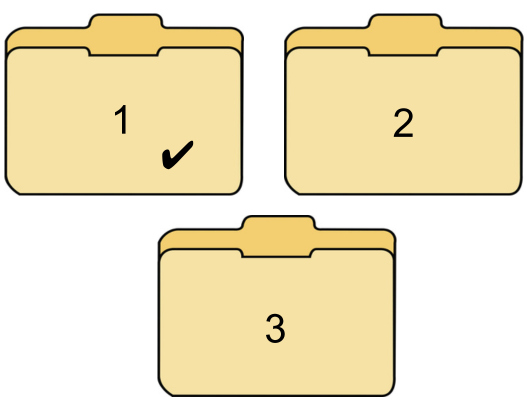
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

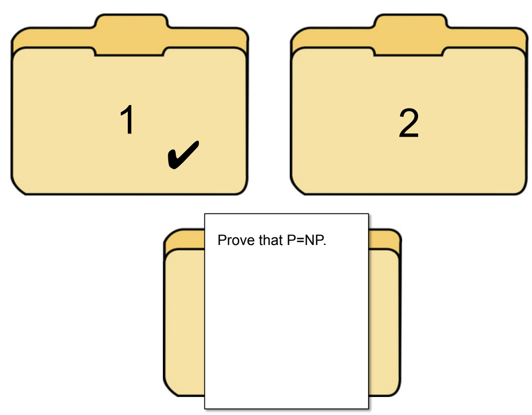
Priors

Posterior

- Useful when you want to know something about A, but all you can directly observe is B
  - This process is called *Bayesian inference*







Assumptions

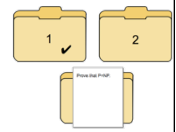
- Easy exam randomly placed in one of the 3 folders
- The teacher always reveals a hard exam
  - If the student chooses a hard exam, the teacher reveals the other hard exam
  - If the student chooses an easy exam, the teacher reveals one of the hard exams, chosen at random

Using Bayes' law…

Given that #1 was chosen by the student,

P(2 easy | 3 shown) = P(3 shown | 2 easy) P(2 easy) / P(3 shown)

P(2 easy) = ?
P(3 shown | 2 easy) = ?
P(3 shown) = ?

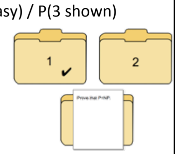


Using Bayes' law…

Given that #1 was chosen by the student,

P(2 easy | 3 shown) = P(3 shown | 2 easy) P(2 easy) / P(3 shown)

P(2 easy) = 1/3
P(3 shown | 2 easy) = ?
P(3 shown) = ?

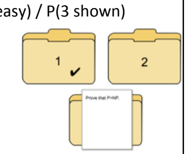


Using Bayes' law…

Given that #1 was chosen by the student,

P(2 easy | 3 shown) = P(3 shown | 2 easy) P(2 easy) / P(3 shown)

P(2 easy) = 1/3
P(3 shown | 2 easy) = 1
P(3 shown) = ?

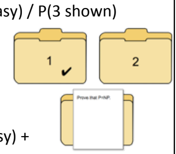


Using Bayes' law…

Given that #1 was chosen by the student,

P(2 easy | 3 shown) = P(3 shown | 2 easy) P(2 easy) / P(3 shown)

P(2 easy) = 1/3
P(3 shown | 2 easy) = 1
P(3 shown) = P(1 easy) P(3 shown | 1 easy) +
P(2 easy) P(3 shown | 2 easy) +
P(3 easy) P(3 shown | 3 easy)

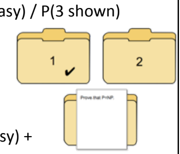


Using Bayes' law…

Given that #1 was chosen by the student,

P(2 easy | 3 shown) = P(3 shown | 2 easy) P(2 easy) / P(3 shown)

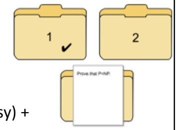
P(2 easy) = 1/3
P(3 shown | 2 easy) = 1
P(3 shown) = P(1 easy) P(3 shown | 1 easy) +
P(2 easy) P(3 shown | 2 easy) +
P(3 easy) P(3 shown | 3 easy)
= (1/3) (1/2) + (1/3) (1) + (1/3) (0) = 1/2



Using Bayes' law…

Given that #1 was chosen by the student,

P(2 easy | 3 shown) = P(3 shown | 2 easy) P(2 easy) / P(3 shown)



P(2 easy) = 1/3

P(3 shown | 2 easy) = 1

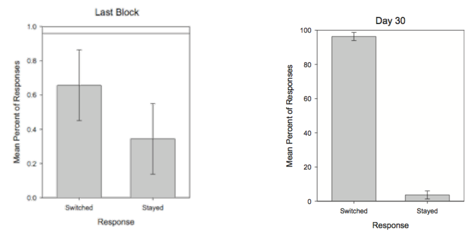
P(3 shown) = P(1 easy) P(3 shown | 1 easy) +
P(2 easy) P(3 shown | 2 easy) +
P(3 easy) P(3 shown | 3 easy)
= (1/3) (1/2) + (1/3) (1) + (1/3) (0) = 1/2

P(2 easy | 3 shown) = (1)(1/3) / (1/2) = 2/3

Monty Hall Problem

Behavior of undergrads after 200 trials:

Behavior of pigeons:



Herbranson & Schroeder 2010

Back to AI…

- In AI we often want to predict an unknown answer given known answers to past problems
  - E.g., Given current weather observations, will it rain later?
- Whether it will rain (R) may depend on hundreds or thousands of observations, $V_1, V_2, \dots V_{1000}$
  - Temperatures across U.S., moisture in atmosphere, etc…
- Given enough days of data, we could estimate a joint probability function $P(R, V_1, V_2, \dots, V_{1000})$
  - Then problem would be solved!
  - How many days of data would you need?

A huge problem

- Say all variables of $(R, V_1, V_2, \dots, V_{1000})$ are binary
  - Need at least 2^{1000} days of data just to observe all possible combinations of the variables
  - Need to observe multiple days for each combination of variables to estimate conditional probability robustly
  - Simply impossible from a computational, representational, or intuitive point of view
- This seemed fatal for the first ~30 years of AI research
  - Graphical models are a framework for avoiding this problem by making assumptions about the structure of a model

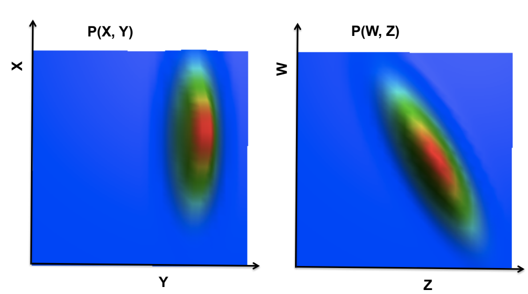
Trivial example

- Suppose you try to predict the weather by flipping 1000 coins each day
  - Here we again need to model $P(R, V_1, V_2, \dots, V_{1000})$
- Clearly all of these variables are independent, so the joint probability distribution can be factored as,

$$P(R, V_1, V_2, ..., V_{1000}) = P(R) \prod_{1}^{1000} P(V_i)$$

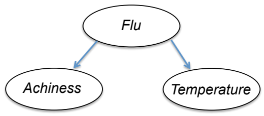
  - How many parameters does this model have?

Independent vs correlated joint distribution examples



Another example

- Say we want to decide whether someone has the flu (F) based on their temperature (T) and achiness (A)
- A, T, and F are clearly **not** independent
- But a weaker assumption of conditional independence may be appropriate, $A \perp T | F$
  - Says that A and T are independent *for a given value of F*
  - We can represent this assumption with a *Bayesian network*:

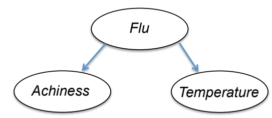


Another example

- Now we can factor P(A,T,F) as:

$$P(A,T,F) = P(A|F)P(T|F)P(F)$$

- To decide whether someone has the flu given observed symptoms, we'll want to compute P(F | A, T)
  - How to compute this?



Back to the weather…

- We want to compute probability of rain (R) given observed variables V_1, V_2, … V_{1000}. Using Bayes' law,

$$P(R|V_1, V_2, ..., V_{1000}) = \frac{P(V_1, V_2, ..., V_{1000}|R)P(R)}{P(V_1, V_2, ..., V_{1000})}$$

  - Now, assuming that V_1 … V_{1000} are conditionally independent given R:

$$P(V_1, V_2, ..., V_{1000}|R) = \prod_{1}^{1000} P(V_i|R)$$

  - Under this assumption, what is $P(V_1, V_2, \dots V_{1000})$?
  - How many parameters do we need to estimate in this factored model?

Naïve Bayes model

- Assuming conditional independence among observed variables is called *naïve Bayes*
  - Class label *C* we want to infer
  - Set of observable variables X1, X2, … Xn
  - Assume that observable variables are independent conditioned on the class label *C*
  - Estimate prior distribution P(C) and conditional distributions P(X1|C), …, P(Xn | C) from training data
  - Use Bayes' Law to calculate P(C | X1 … Xn)

Bayes' Law: An example

- You're a juror in a murder case
  - You need to decide between guilt (G) and innocence ($\bar{G}$)
  - You have heard some evidence (E)
  - Bayesian approach: Compute *P(G|E),* and vote to convict if

$$P(G|E) > \tau$$

  where τ is a threshold

  - Using Bayes' law,

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

Bayes' Law: An example

- Say you have to decide before hearing any evidence
  - what is the *prior* probability, *P(G)*?

- How to estimate *P(G)*?
  - Based on population constraints
    - 1 person in Bloomington (~20,000 people) did it
    - *P(G)* ≈ 1/20000 = 0.00005
  - Based on historical data
    - U.S. murder conviction rate: 0.06/1000 [BJS96]
    - *P(G)* ≈ 0.00006

Bayesian inference example

- Eyewitness testimony (T) identifies the suspect
  - Now we want to compute *P(G|T)*, $P(G \mid T) = \frac{P(T \mid G)P(G)}{P(T)}$
  - P(G) =
  - P(T|G) =
  - P(T) =

Bayesian inference example (2)

- Now you hear that the murderer had a red car, and that the suspect owns a red car (R)
  - We want to compute P(G|R,T),

$$P(G \mid T,R) = \frac{P(T,R \mid G)P(G)}{P(T,R)}$$

  - Assuming that T and R are independent conditioned on G,

$$\underbrace{P(G \mid T,R)}_{\text{The } posterior \text{ probability}} = \frac{P(R \mid G)P(T \mid G)P(G)}{P(R)P(T)} = \underbrace{\left(\frac{P(R \mid G)}{P(R)}\right)}_{\text{New evidence}} \underbrace{P(G \mid T)}_{\text{Probability given all prior knowledge}}$$

Bayesian inference example (3)

- Given the testimony (T) and red car evidence (R),

$$P(G \mid T,R) = \frac{P(R \mid G)P(G \mid T)}{P(R)}$$

  - P(G|T) =
  - P(R|G) = P(R|$\bar{G}$) ≈

$$\begin{aligned} P(R) &= P(R \mid G)P(G) + P(R \mid \bar{G})P(\bar{G}) \\ &= (1)(0.00005) + (0.13)(1 - 0.00005) \\ &\approx 0.13004 \end{aligned}$$

$$P(G \mid T,R) = \frac{(1)(0.0000859)}{0.13} \approx 0.0006608$$

Bayesian inference example (4)

- Now suppose a partial fingerprint (F) matches the suspect

$$P(G \mid T,R,F) = \frac{P(F \mid G)P(G \mid T,R)}{P(F)}$$

  - P(F|G) = 1, P(F|$\bar{G}$)=0.001
  - P(G|T,R) = 0.0006608

$$\begin{aligned} P(F) &= P(F \mid G)P(G) + P(F \mid \bar{G})P(\bar{G}) \\ &= (1)(0.00005) + (0.001)(0.99995) \\ &\approx 0.00105 \end{aligned}$$

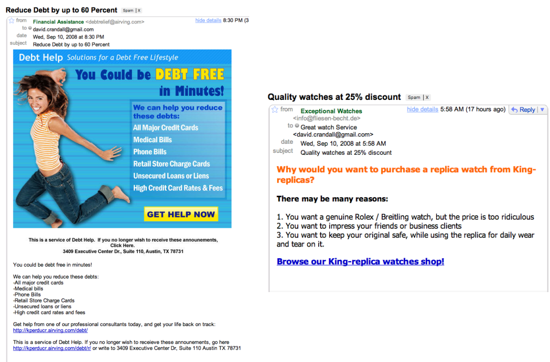
$$P(G \mid T,R,F) = \frac{(1)(0.0006608)}{0.00105} \approx 0.6209$$

Naïve bayes: Pros and cons

- **Pro:** Notice that we avoided making hard classification decisions until all evidence had been considered
  - Explicitly modeled uncertainty
- **Pro:** Easy to estimate model parameters from training data or human intuition
  - Avoids brittleness of early AI systems
- **Con:** Strong conditional independence assumptions
  - More complex systems can't be modeled

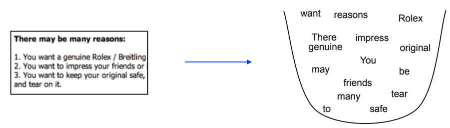
Spam classification

- Spam = junk e-mail
- A big problem! [Commtouch07]
  - ~96% of all email traffic on the Internet
  - ~150 billion junk emails per day
  - >2 petabytes (= 2,000 terabytes = 2,000,000 gigabytes) daily
  - Spreads malware, worms, phishing schemes, etc.
- Possible solutions
  - Block e-mails from blacklisted users and servers
  - Accept e-mails only from whitelisted addresses
  - Cost-based solutions (e.g. micropayments)
  - Filtering rules (ignore mail with "debt", "viagra", "stock")
  - Content-based statistical filtering



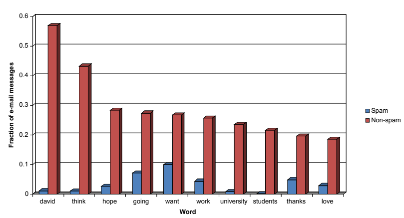
Modeling a document

– Represent a document as an unordered collection of words (a *bag of words* model)



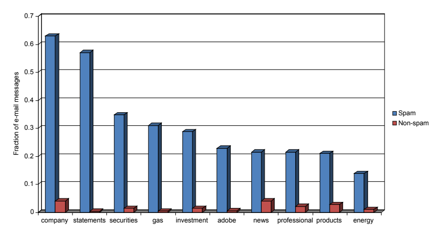
Statistical motivation

- Spam and (my) non-spam are statistically very different



Statistical motivation

- Spam and (my) non-spam are statistically very different



Bayesian spam filtering

- Suppose we get an email containing the word "debt"
  – What is the probability it is spam (S), P(S|debt)?

$$P(S \mid \text{debt}) = \frac{P(\text{debt} \mid S)P(S)}{P(\text{debt})}$$

  – P(debt | S) = 0.309 , P(debt | S) = 0.00447
  – P(S) = 0.5
  – P(debt) = 0.157

$$P(S \mid \text{debt}) = \frac{P(\text{debt} \mid S)P(S)}{P(\text{debt})} \approx 0.986$$

0.986/0.014 ≈ 70:1 odds that message is spam

$$P(\bar{S} \mid \text{debt}) = 1 - P(S \mid \text{debt}) \approx 0.014$$

More examples

– Assuming a uniform prior, *P(S)*=0.5

Word	P(word|spam)	P(word|not spam)	P(word)	P(spam|word)
debt	0.309	0.00447	0.157	0.986
news	0.215	0.0395	0.127	0.845
investment	0.288	0.0137	0.151	0.955
david	0.012	0.575	0.294	0.020
want	0.101	0.268	0.185	0.274
thanks	0.0491	0.196	0.123	0.200

Computed from Bayes' Law

Bayesian spam filtering

- A new email has the words "debt" and "price"
  - What is the probability it is spam (S), P(S|debt, price)?

$$P(S \mid \text{debt, price}) = \frac{P(\text{debt, price} \mid S)P(S)}{P(\text{debt, price})}$$

  - If we assume that the occurrence of the words "debt" and "price" are independent events given S,

$$P(S \mid \text{debt, price}) = \frac{P(\text{debt} \mid S)P(\text{price} \mid S)P(S)}{P(\text{debt})\, P(\text{price})}$$

Bayesian spam filtering

- Generalize to an arbitrary number of words,

$$P(S \mid W_1, W_2, W_3, \dots, W_n) = \frac{P(W_1 \mid S)P(W_2 \mid S) \dots P(W_n \mid S)P(S)}{P(W_1)P(W_2) \dots P(W_n)}$$

which is equivalent to,

$$P(S \mid \bigcap_{i=1}^{n} W_i) = P(S) \prod_{i=1}^{n} \frac{P(W_i \mid S)}{P(W_i)}$$

- For example,

$$P(S \mid \text{debt, free, credit}) = P(S)\left(\frac{P(\text{debt} \mid S)}{P(\text{debt})}\right)\left(\frac{P(\text{free} \mid S)}{P(\text{free})}\right)\left(\frac{P(\text{credit} \mid S)}{P(\text{credit})}\right)$$

A practical spam filter [Graham02]

- Break a message into *tokens* of words, numbers, etc.
- Look for the 15 "most interesting words"
  - I.e. words for which P(S|W) is farthest from 0.5
  - Then compute $P(S|W_1, W_2, \dots, W_{15})$

```
Dear Sir or Madam:

Please reply to
Receiver: China Enterprise Management Co., Ltd. (CMC)
E-mail: unido@chinatop.net

As one technical organization supported by China Investment and
Technical Promotion Office of United Nation Industry Development
Organization (UNIDO), we cooperate closely with the relevant Chinese
Quality Supervision and Standardization Information Organization. We
provide the most valuable consulting services to help you to open
Chinese market within the shortest time:

1. Consulting Service on Mandatory National Standards of The People's
Republic of China.

2. Consulting Service on Inspection and Quarantine Standards of The
People's Republic of China.

3. Consulting Service for Permission to Enter Chinese Market

We are very sorry to disturb you!

More information, please check our World Wide Web:
http://www.chinatop.net

Sincerely yours
```

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

$P(S|W_1, W_2, \dots, W_{15})=0.9$

A true negative

```
Hi,

Do you have any examples online of that continuation style
web programming that you describe?

For example, you mention that you needed the user to go
to a color picker screen and then return to the
same spot. I'm interested on what was required to
achieve that. Did you have to use real continuations
to achieve that?


Dru Nelson
San Carlos, California
```

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

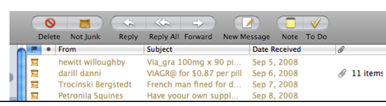
A false negative

```
Dear Paul Graham,

As a person involved in website development you recognize that finding a
good web host offering reasonably priced quality services can be quite
difficult. We believe that offering a combination of quality services,
timely and responsive customer support and good prices is where our company
excels.

HOSTEX ( http://www.hostex.com ) is a flexible company providing quality and
cost effective web hosting solutions with focus on small and medium businesses.

Recently we introduced new service plans offering flexibility, performance and value:

Our plans are:

Basic    - $6.95 / month
         - 40 Mb of diskspace
         - 10 POP3 mailboxes, unlimited aliases, autoresponders
         - unlimited ftp accounts, URL protection, shell access and more
         - no setup fee

Advanced - $14.95 / month
         - 80 Mb of diskspace
         - 15 POP3 mailboxes, unlimited aliases, autoresponders
         - unlimited ftp accounts, URL protection, shell access
         - CGI/SSI scripting (Perl, Python, Tcl)
         - PHP4 scripting
         - Frontpage 2002 extensions
         - Graphical web statistics
         - various optional services and more
         - no setup fee

Professional     - customizable, make-your-own plan
                 - select only those services that you need

All plans include full 30-day money back guarantee and there is no setup fee.
```

perl	0.01
python	0.01
tcl	0.01
scripting	0.01
morris	0.01
graham	0.01491078
guarantee	0.9762507
cgi	0.9734398
paul	0.027040077
quite	0.030676773
pop3	0.042199217
various	0.06080265
prices	0.9359873
managed	0.06451222
difficult	0.071706355

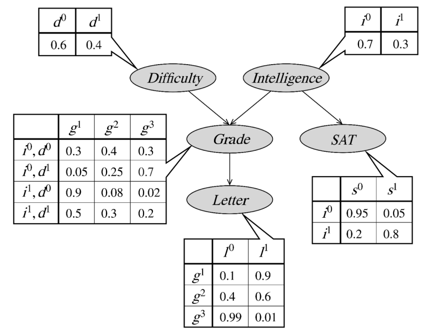
Learning

- The advantage of a Bayesian classifier is that it can learn optimal values for its parameters
  - Given a set of training data
  - No need for hand-crafted rules. More accurate, less work.
  - But a good set of training data is critical
- The classifier can be continue to learn with time
  - User corrects the classifier's errors, classifier adjusts probabilities accordingly



Implementation issues

- What do we do about words that were not seen during training?
- How do we handle very small numbers?
- Do we need the denominator?
- What is the consequence of the naïve Bayes assumption?



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01