STAT-INF/COMB-EVID/NAÏVE BAYES

**DAY 1**

Slides on ML/MAP are about 45-50 minutes. I start with likelihood weighting example from bayes nets from previous lecture. students do a full example of weighting, generating samples. takes up 20 mins or so.

**DAY 2 : Lesson plan – Stat inf day 2 – combining evidence**

HAS SLIDES

Sally Clark

* English lawyer who was found guilty in 1999 of the murder of her two young sons.
* She had two babies, one in 1996, and one in 1998. Both died a few weeks after they were born.
* There is a phenomenon called sudden infant death syndrome (SIDS), where babies seemingly die for no reason in their first year of life. The cause is unknown, but people currently believe it might be due to the position in which the baby sleeps, or due to overheating, or suffocating due to blankets in the crib, or exposure to tobacco smoke.
* So when Clark’s first baby died, it was chalked up to SIDS.
* When the second baby died, they arrested her and put her on trial for the murder of both her sons, claiming that it was highly unlikely to have two deaths from SIDS in the same family.
* A pediatrician (Roy Meadow) testified that the probability of a single child dying from SIDS is 1/8543.   
  (WRITE ON BOARD). P(1 death via SIDS) = 1/8543.
* Now the problem is that two babies died, and Meadow assumed those events were conditionally independent. So he squared (1/8543) and came up with 1/73 million.
* That was the first mistake. In fact, the Royal Statistical Society (UK) later claimed that the 1/73 million number had “no statistical basis.”
* Second mistake the Prosecutor’s fallacy. The press reported the 1-in-73 million number as the probability that Clark was innocent (that is, the babies had died from SIDS).
* But that doesn't make any sense at all. This tells us **nothing** about the probability of Clark's innocence or guilt, because it doesn't consider any alternative hypotheses, and their probabilities. Specifically, the pediatrician gave us information about the probability of a child dying from SIDS. But they didn't give us any information about the probability of a child being murdered, which is probably (hopefully) even **smaller** than SIDS.
* Remember what we are comparing here are:
* P(H | D) for 2 hypotheses, SIDS, Murder  
    
  P(SIDS | D1, D2) vs P(M | D1, D2)
* H-MAP says these are proportional to  
    
  P(D1, D2|SIDS)P(SIDS) vs P(D1 , D2|M)P(M)

Unfortunately, the story does not have a happy ending. Clark was convicted in November 1999 and was sentenced to life in prison. The conviction was upheld on appeal in October 2000, but overturned in a second appeal in January 2003, when some additional microbiological evidence came to light that suggested the children's deaths were due to natural causes. So Clark was released from prison. Unfortunately, she never recovered from the effects of her conviction and imprisonment, and she became dependent on alcohol afterwards, eventually dying from acute alcohol intoxication in 2007.

So I bring up this story as a cautionary tale about being very careful what interpretations you assign to probabilities you're using, and what conclusions you draw from them.

Be very careful about when things are independent and when they're not. Now fortunately in this class we won't be dealing with any life or death situations, but when you're out in the real world, whether you're in the comp sci industry or any industry, (as you can see even the legal industry), you'll often have to deal with probabilities, so just be real sure that you know what you're doing.

DAY 3 – Naïve Bayes

* Run through slides.
* Do example:
* Do Email example from disc8 plan.
* Do SNL example.
* Slides are too short – we got through (s17) both the email example from disc8 and SNL. Things to add – doing this in your head, or email Bayesian spam poisoning.



features: luxury, ~brands, save.

argmax:

spam-> P(lux | spam)P(~brands|spam)P(save|spam)P(Spam) = .4 \* (1-.3)\*(.4)(.8)=.0896

notspam P(lux | ~spam)P(~brands|~spam)P(save|~spam)P(~Spam)

.01 \* (1-.8) \* .1 \* .2=.00004

So MAP picks SPAM!

**Spam example 1:**

Suppose I have 20 emails, that have been already classified into spam (15 emails) and non-spam (5 emails). Suppose I only care about the presence or absence of the words luxury, brands, and save.  
  
Suppose 6 of the spam emails contain "luxury," 3 of the spam emails contain "brands," and 7 of the spam emails contain "save."  
  
Suppose 1 of the non-spam emails contains "luxury," 2 of the non-spam emails contain "brands," and 2 of the non-spam emails contain "save."  
  
Suppose a new email arrives that contains the words "luxury" and "save" but not "brands." Should this be classified as spam or not spam?

work:

P(spam) = 15/20 [ recommend doing these calcs with fractions b/c we need to smooth later ]

P(not spam) = 5/20

P(lux | spam) = 6/15

P(brands | spam) = 3/15

P(save | spam) = 7/15

P(lux | notspam) = 1/5

P(brands | notspam) = 2/5

P(save | notspam) = 2/5   
  
HMAP = argmax[i] P(D|Hi)P(Hi)

(split into spam/not spam)

spam -> P(D | spam)P(spam)

D = luxury and save, but not brands

P(lux, save, ~brands | spam)P(spam)

P(lux | spam) P(save | spam) P(~brands | spam) P(spam)

(6/15) \* (7/15) \* (1 – 3/15) \* (15/20)

0.112 approx

notspam -> P(D | notspam)P(notspam)

D = luxury and save, but not brands

P(lux, save, ~brands | notspam)P(notspam)

P(lux | notspam) P(save | notspam) P(~brands | notspam) P(notspam)

(1/5) \* (2/5) \* (1 – 2/5) \* (5/20)

0.012 approx

So spam wins! bc .112 > .012

[Note these don't sum to 1]

**Smoothed versions**

P(lux | spam) = 6/15 -> 7/17

P(brands | spam) = 3/15 -> 4/17

P(save | spam) = 7/15 -> 8/17

P(lux | notspam) = 1/5 -> 2/7

P(brands | notspam) = 2/5 -> 3/7

P(save | notspam) = 2/5 -> 3/7

Talk about the (wrong) assumption that words are independent.

Naïve Bayes uses HMAP to predict. And HMAP uses argmax, so all we care about is which probability is higher,

P(D|spam)P(spam) versus P(D|~spam)P(~spam)

The **actual** probabilities don't matter (as much). If the spam/not spam calculation comes out 60/40 or 90/10, the email is still labeled as spam.

So the assumption is mathematically wrong, but it is useful in the real world because it tends to not mess up the math that much, or at least not enough to switch something from spam to not spam or vice versa.

Bayesian Poisoning

* Bayesian poisoning refers to a situation where a spammer will construct certain emails in a way such that over time, they can change the probabilities in a Bayesian spam filter in their favor.
* What do you think a spammer could do? There are (at least) two different techniques.
  + Send out "not-spam" emails (emails with a lot of legitimate text, designed to get through the spam filter), but include a small number of spam words. Over time, the spam words will become associated with not-spam. Then you can send spam emails with those words and they will get through (e.g., Viagra).
  + Send out "spam" emails but start including legitimate words in them. Then over time, the not-spam words become associated with spam, so the filter gets confused.