Vir Desai
720329873
Collaborators: None

HW3

1. Bayes Nets (written questions)

A

P(J|G) = P(J|G, I) and P(M|G, B, I) = P(M|G, B, I, J)

The network asserts these (2 and 3). Three is considering a Markov blanket of M.

B

P(b, I, ~m, g, j) = P(b)*P(i|b, ~m)*P(~m)*P(g|b, i, ~m)*P(j|g) = 0.9*0.5*0.9*0.8*0.9 = 0.2916

This is the probability that someone broke an election law, and was indicted, and their prosecutor was not politically motivated, and the person was found guilty, and they were ultimately put in jail.

C

The probability that someone goes to jail, given that he or she broke the law and the prosecutor was not politically motivated [P (j | b, ~m)] is equal to 0.36. The work to show this is presenting in a picture as Figure 3 at the bottom in the Figures section. I decided to multiply all the probabilities of being indicted by the probabilities of being found guilty. This produces a probability chart shown in the bottom left of Figure 3. Now I can assume that the only values that matter are when B is true and M is false so I eliminated the other possibilities on the chart. This leaves 1 with probability 0.4 and 1 with probability of 0 for the probability of being found guilty. Therefore I eliminated the line where probability is 0 and was left with the probability of being found guilty as 0.4. Now I used the chart in the bottom right of Figure 3 to calculate the probability of the person going to jail. Therefore I multiplied the probability of someone being found guilty (0.4) by the 0.9 and being found not guilty (0.6) by 0.0 and found that the probability for going to jail is 0.36.

D

I would use the process of likelihood weighting to sample a joint assignment to the variables of this Bayes Net conditioned on the fact that there was a guilty verdict. Because of a guilty verdict only being able to occur if a person was indicted, indictment becomes a fixed evidence variable to its observed value of true and gives it weight for those values. This allows samples to only reflect the state of the world suggested by the evidence.

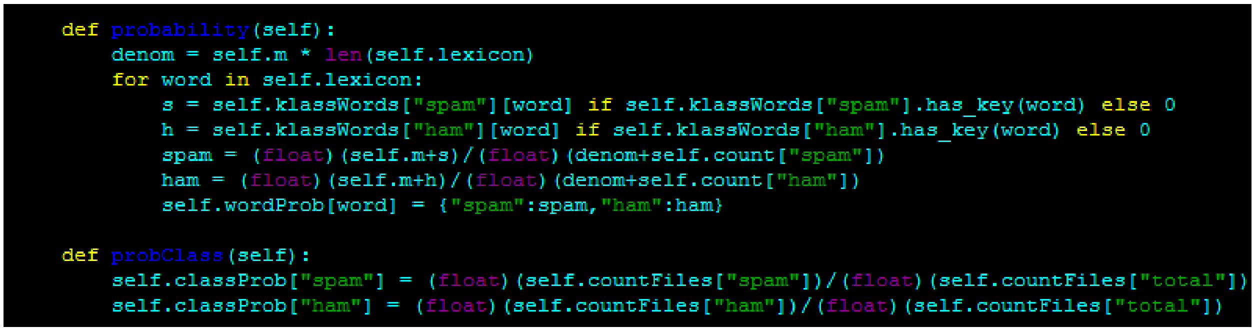
2. Naïve Bayes Spam Classification

- Description of your Naïve Bayes classifier implementation
  - Figure 1 in the Figures section at the bottom of this document shows the code I used to implement my Naïve Bayes classifier. I enumerated my lexicon in that class as a dictionary. This dictionary attributed words to numerical counter values of how many times the word had been seen in either spam or ham testing. I also have dictionaries in the class for the amount of times the words have been used in each email type, spam and ham, which are updated when my program parses through the documents. I start my probability function by creating the denominator value that needed to be added to the probability function based on the requirements provided. I used the length of the

lexicon, which is equivalent to the number of different words seen in all the training documents. For each word in the lexicon I went through and created a dictionary with that as a key and the value for that word key was a dictionary. That dictionary consisted of the two email types as keys and counts for them per class as the values. The s and h variables shown are the counts of the lexicon word if it exists in spam or ham. Then I calculated the spam and ham probability for each class using the formula provided and simply just implemented. This all was added to the dictionary for the words showing their probability of occurring in spam and probability of occurring in ham.

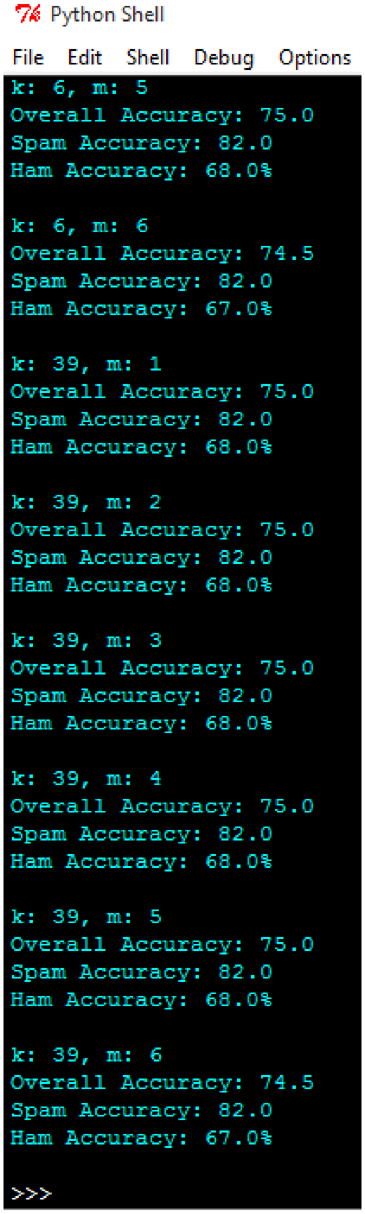
- Description of your results -- including overall accuracy, spam accuracy, ham accuracy, and descriptions of your experiments to set values for k and m. Also include discussion of whether MAP and ML classification would produce different results for the provided data.
  - The highest overall accuracy I am able to produce with my configuration and code is 76%. This occurs with my spam accuracy usually being above 80% and ham accuracy in the upper 60%. To be more precise, I am able to reproduce getting my 76% max overall accuracy when my m is anything from 1 to 5 and my k is anything from 2 to 40, inclusively. For m values between 1 and 4, inclusive, I am able to achieve 76% overall accuracy with k values between 0 and 40. My program's Spam accuracy with this max overall accuracy is generally 83-84% and my Ham accuracy with this max overall accuracy is generally 68%. An example of this is shown at the bottom section of pictures in Figure 2. I figured out my k and m value range and how the accuracies changed in correspondence to them by running a loop of executions of my code with randomly chosen k and m values and narrowing it down from there. In the code that can be executed on the server, a loop will run for 10 randomly chosen k values from 1 to 41 for all m values from 1 to 6.
  - It would not make any difference to use maximum likelihood classification for the provided data instead of maximum a posteriori because we are using the same number of spam and ham documents. Therefore the P(class), probability of spam or ham occurring, is the same for both classes, spam and ham.
- Include some example emails showing where your algorithm did well and where it failed (and why you think this behavior occurred).
  - My algorithm did not do well on emails such as 1097.2000-05-19.farmer.ham.txt, 4731.2001-07-09.farmer.ham.txt, 3338.2004-12-29.GP.spam.txt, and 4993.2005-08-13.GP.spam.txt (all shown at the bottom in Figures). The first two of these emails are ham messages marked as spam and the second two are spam messages marked as ham. This happens because generally the spam messages are trained to encounter messages which are not words/are misspelled or are non-alphanumeric characters such as '/' and ':' whereas ham messages are trained to encounter proper words which are not spelled incorrectly as well as few characters such as '/' and ':'. The messages which are ham registered as spam contain lots of single non-alphanumeric characters and those which are spam registered as ham have mostly properly spelled words and low usage of single non-alphanumeric characters.

Figures:



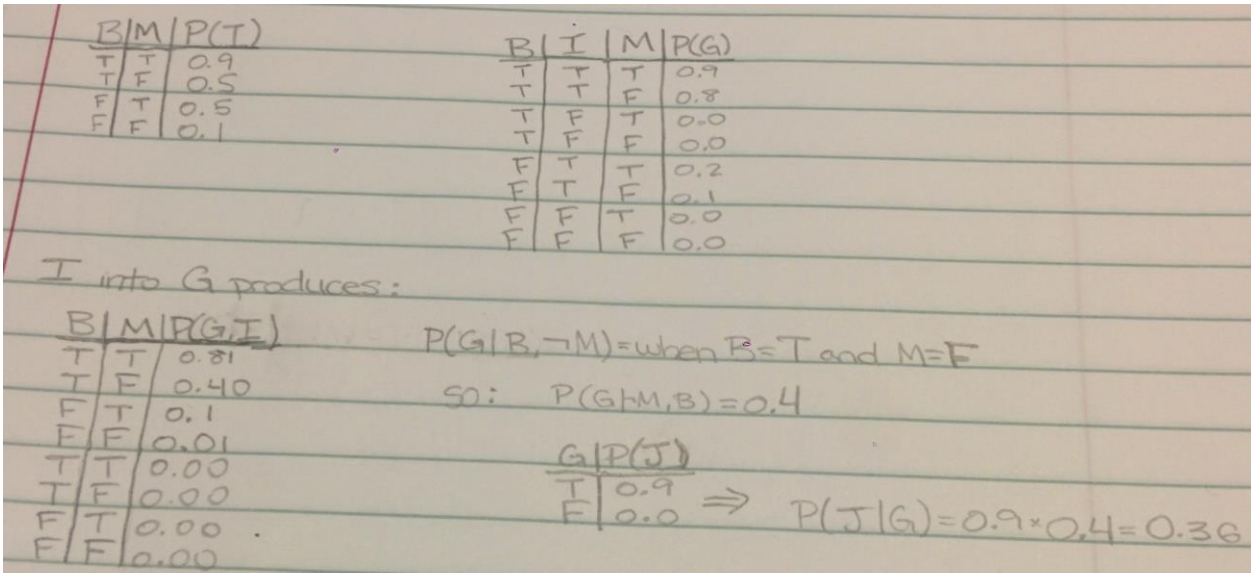
```
    def probability(self):
        denom = self.m * len(self.lexicon)
        for word in self.lexicon:
            s = self.klassWords["spam"][word] if self.klassWords["spam"].has_key(word) else 0
            h = self.klassWords["ham"][word] if self.klassWords["ham"].has_key(word) else 0
            spam = (float)(self.m+s)/(float)(denom+self.count["spam"])
            ham = (float)(self.m+h)/(float)(denom+self.count["ham"])
            self.wordProb[word] = {"spam":spam,"ham":ham}

    def probClass(self):
        self.classProb["spam"] = (float)(self.countFiles["spam"])/(float)(self.countFiles["total"])
        self.classProb["ham"] = (float)(self.countFiles["ham"])/(float)(self.countFiles["total"])
```

Figure 1: Top portion of Training implementation for creating the lexicon and figuring out probabilities or words in lexicon



```
k: 6, m: 5
Overall Accuracy: 75.0
Spam Accuracy: 82.0
Ham Accuracy: 68.0%

k: 6, m: 6
Overall Accuracy: 74.5
Spam Accuracy: 82.0
Ham Accuracy: 67.0%

k: 39, m: 1
Overall Accuracy: 75.0
Spam Accuracy: 82.0
Ham Accuracy: 68.0%

k: 39, m: 2
Overall Accuracy: 75.0
Spam Accuracy: 82.0
Ham Accuracy: 68.0%

k: 39, m: 3
Overall Accuracy: 75.0
Spam Accuracy: 82.0
Ham Accuracy: 68.0%

k: 39, m: 4
Overall Accuracy: 75.0
Spam Accuracy: 82.0
Ham Accuracy: 68.0%

k: 39, m: 5
Overall Accuracy: 75.0
Spam Accuracy: 82.0
Ham Accuracy: 68.0%

k: 39, m: 6
Overall Accuracy: 74.5
Spam Accuracy: 82.0
Ham Accuracy: 67.0%

>>>
```

Figure 2: Example of some of the accuracies produced by my code with different k and m values shown



B	M	P(T)
T	T	0.9
T	F	0.5
F	T	0.5
F	F	0.1

B	I	M	P(G)
T	T	T	0.9
T	T	F	0.8
T	F	T	0.0
T	F	F	0.0
F	T	T	0.2
F	T	F	0.1
F	F	T	0.0
F	F	F	0.0

I into G produces:

B	M	P(G,I)
T	T	0.81
T	F	0.40
F	T	0.1
F	F	0.01
T	T	0.00
T	F	0.00
F	T	0.00
F	F	0.00

$P(G|B,\neg M)$ = when B=T and M=F

so: $P(G|\neg M,B)=0.4$

G	P(J)
T	0.9
F	0.0

$\Rightarrow$ $P(J|G)=0.9\times0.4=0.36$

Figure 3: Work for variable elimination problem of the first section

1097.2000-05-19.farmer.ham.txt

```
Subject: customer meeting invitation
good afternoon all ! !
i have attached the invitation to our 8 th annual enron customer
meeting to
be held at the don ce sar in st . pete ' s beach , fl .
please feel free to fax or email to your customers as you see
necessary ,
but please use the rsvp forms for the trip they will be
attending ( july or august ) . the ' invitation ' attachement is
generic and
should be used for both rsvp forms .
the original invitation will be going out in the mail this evening .
with
a requested return date of june 8 , 2000 .
if you have any questions or need an assistance , please call me at
x 33278 .
thank you ,
heather choate
```

4731.2001-07-09.farmer.ham.txt

```
Subject: your order with amazon . com ( # 102 - 6820014 - 8227326 )
thanks for ordering from amazon . com ! your purchase information
appears below .
to see the latest information about your order , or to cancel or
modify a pending order , just click the " your account " link in the
top
right corner of any page on our web site or visit :
http : / / www . amazon . com / your - account
did you know you can view and edit your orders online ?
access " your account " ( http : / / www . amazon . com / your -
account ) , and you
```

can :
* track order status
* combine orders
* change payment options
* edit shipping address
* cancel unshipped items
* change gift messaging
* do much more . . .
if you ordered several items to be delivered to the same address , we
might send them in separate boxes to ensure quicker service . but
don ' t worry : you won ' t be charged any extra shipping fees .
thanks again for shopping with us .
- - amazon . com customer service
your purchase reads as follows :
e - mail address : dfarmer @ enron . com
billing address : daren farmer
2747 meadowtree
spring , tx 77388
united states
telephone : 281 - 288 - 8251
subtotal : $ 19 . 99
shipping @ $ 19 . 99 each
usually available in 24 hours
did you know that we have amazon . com gift certificates available ?
you can order a gift certificate in any dollar amount from $ 5
to $ 5 , 000 . we ' ll deliver it via e - mail or physical mail - - so
it ' s a
perfect last - minute gift . for more details on ordering gift
certificates , please visit the following url :
http : / / www . amazon . com / gift - certificates /
you can make changes to any unshipped orders in your account . just
click the " your account " link in the top right corner of any page on
our web site or visit :
http : / / www . amazon . com / your - account
if you ' ve explored the links on that page but still have a question
,
please visit our online help department .
http : / / www . amazon . com / help
thanks again for shopping at amazon . com !
amazon . com
earth ' s biggest selection
http : / / www . amazon . com

3338.2004-12-29.GP.spam.txt

Subject: how have you been , , obtain all of your meds here . . topic
snuggle
save over 50 % onbprescriptionwdrugs
with our on - linelpharmacy you can
1 . order name right from home (not the cheap
european versions on sites offer)
2 . get it shipped same day to your door step

3 . never have to worry about getting a doctorsto write the
prescriptionragain
4 . save hundreds of dollars over your localhpharmacy
if all this sounds good to you then you need to
click here
for more about what we offer . we carry everything from vicodin ,
valum ,
xanax , and viagra . so go to our site to see how much we can save you
today .

4993.2005-08-13.GP.spam.txt

Subject: my site links to your site now
hello my name is anthony lewis . i am seeking out possible link
partners that our
visitors would be interesting in visiting . i ' ve found your website
to be a very good fit
for our visitors . i have already gone ahead and added your link to
our website at :
i am contacting you to see if it is ok to have done so . also , i
would like to ask if
you mind linking back to us ? if so , please use the linking details
below and send
me the location of our link on your website .
here is our linking details :
title : acai superfood of the amazon
description : learn how you can build a free to join network and never
be penalized on volume .
url : http : / / www . big - money - busniess . com
we ' ve got several pr 6 and 7 websites , so we expect this site to
become atleast a
pr 5 within 1 month and will eventually become a 6 or 7 in 2 - 3
months .
i hope this can be a way for us to benefit our visitors with excellent
content . hope
to hear from you soon .
anthony lewis
www . big - money - busniess . com