CSc 421 Assignment 3

Neil Burroughs
University of Victoria
British Columbia, Canada
inb@uvic.ca

April 20, 2006

1 Probabilty Theory

As a matter of completeness I will outline the probabilty of winning for this problem. From the table we can see that the player has three possible first choices, two of which are goats. The host must choose a goat. If the players first choice is a goat then switches the player will always win. Since there are two doors with goats behind them the player has a probability of winning of $P(W|S) = \frac{2}{3}$.

Scenario	First Choice	Host Choice	Switch Choice	Result
1	Car	Goat x	Goat y	Lose
2	Goat 1	Goat 2	Car	Win
3	Goat 2	Goat 1	Car	Win

1.1 Variation I

Given that the player won a car, what is the probability that they switched doors where the decision to switch doors was decided by the toss of a fair coin?

Note that $P(SW)$ is the prob. of switching, $P(DS)$ is the prob. that the player didn't switch, and $P(W)$ is the probability that the player won a car.

We know that $P(SW) = P(DS) = 0.5$ because a fair coin toss decides. Also the probability of winning given that the player switches is $P(W|SW) = \frac{2}{3}$ and the probability of winning if the player didn't switch is $P(W|DS) = 1 - P(W|SW) = \frac{1}{3}$. From this we can calculate the probability that the player switched doors given thay they won $P(SW|W)$ using Bayes Theorem:

$$\begin{aligned}
P(SW|W) &= \frac{P(W|SW)P(SW)}{P(W|SW)P(SW) + P(W|DS)P(DS)} \\
&= \frac{\frac{2}{3}\frac{1}{2}}{\frac{2}{3}\frac{1}{2} + \frac{1}{3}\frac{1}{2}} = \frac{\frac{1}{3}}{\frac{1}{3} + \frac{1}{6}} = \frac{\frac{1}{3}}{\frac{1}{2}} \\
&= \frac{2}{3}
\end{aligned}$$

1.2 Variation II

Over two games, what is the probability that the player wins two goats given that on the first game the player didn't switch and on the second game the player did switch? Once again the decision to switch is based on a fair coin toss.

We note that $P(W|DS) = \frac{1}{3}$ but that winning a goat is not winning as such so $P(G|DS) = 1 - P(W|DS) = \frac{2}{3}$. Also note that the $P(W|SW) = \frac{2}{3}$ and its inverse $P(W|SW) = 1 - P(G|SW) = \frac{1}{3}$. Since these are two different events we can calculate the result as:

$$P(G|DS)P(G|SW) = \frac{2}{3}\frac{1}{3} = \frac{2}{9}$$

The probability of winning two cars under the same conditions as the previous problem is calculated in the same way:

$$P(W|DS)P(W|SW) = \frac{1}{3}\frac{2}{3} = \frac{2}{9}$$

2 Email categorization

2.1 Constructing the model

The training data should be divided into the predefined categories. For a given category we read the group of training documents and count the occurrences of each unique word. For example, the word 'java' may appear 15 times in a the training data for the Computer Science Friends category while the word 'enlarged' may appear 13 times in the Spam category. These word counts are divided by the number of words read for the particular category to generate a probability of occurrence for each particular word. For example, the Spam category may have had 1000 words total so the word 'enlargement' would have a probability of $P(enlargement|Spam) = 0.013$. These probabilties are stored in a table along with their associated word for each category. In this way, training is performed for each category.

2.2 Categorizing a new document

We are provided with a document and want to know what category it belongs in. The first step is to scan the document and count the unique words. For each word and its count, divide the word by the count to get its probability in the email documnet. For example, word=Viagra, occurrences=3, word count=45:

$$P(word = Viagra|Email) = \frac{3}{45}$$

Next we calculate the probability that the email is in a given email category. For each unique word in the email we multiply its probability in the email by the probability in the category. For k words in the email we sum each of the k products, or:

$$P(Email|Category) = \sum_{i=0..k} P(w_i \in email)P(w_i \in category)$$

We can calculate the probability of the email being in each of the categories available. Each of these values is compared and the greatest probability value indicates which category the email belongs in.

2.3 Implementing the better email categorizer

The email categorizer was implemented in Python. Training data was created from my own personal email which proved to be a problem since I receive very little spam. Instead I consider jokes sent by friends as spam, other email from friends as friendly, and colloquium email from the CSc department as what it is. This training data most certainly affected performance.

There is roughly the same amount of words in both the spam and friends categories with much less in the CSc Colloquium category. Results are less than perfect with 50 percent of spam being caught, but all friends emails categorized correctly.

2.4 Practical issues solved/unsolved

Some words are not useful in categorizing documents. Words such as 'and' and 'the' are not indicative of anything in particular; at least not to any significant degree. To solve this problem we can build a list of words deemed not significant and use that list to filter them out of the statistics. However this is not a simple task. Words may be used in different contexts so that a word deemed insignificant in one category of email may be very significant in another. So what is needed is a list of insignificant words for *each* category. This implementation though only uses a single list for all categories due to the expected time a full analysis of each category would require.

When dealing with email it must be determined which portion of the email is important for statistical analysis. An email is divided into a header and a body. The header may hold incriminating information such as location of sender. Email coming from countries in Eastern Europe or parts of Africa may be more likely to be spam as opposed to email coming from Canada. This is especially true if you don't know anyone in Africa. This implementation only considers words located in the body of the message.

Generating statistics in the body of the message has some problems. Not every email contains nice paragraphs of text. Instead there is much punctuation, there is even large portions of HTML, and still others have words cut off during wrapping/encoding of messages by the sender. This implementation eliminates some punctuation but leaves in other pieces. For example, commas are replaced by spaces, but colons are left since some of the example email was found to have terms like From: and Where: as fields. These could be significant indicators. Parsing HTML is beyond the scope of this assignment since it is anything but an easy task. The training data and test emails were filtered of HTML before testing. Word splitting is tricky. Some words are cut with an '=' sign added. It is a simple matter of rejoining those words. In clean cuts with no indication there is little possibility of recovering the words without some analysis such as comparing join segments to the database while looking for a maximum score. ie if the score of the two separated words is less when looking up in the database than it is after joining them then join. This is a problem at the very beginning when no words are in the database.

It appears that spam email is now including paragraphs of words that are very likely to appear in normal email so as to increase the score of the spam away from categorizing it as spam. This is very devious and it indicates that spammers are very unethical people. I am not sure how to correct this problem but it may be possible to split a single email into parts and classify those sections. If certain portions are very likely to be spam then the entire email could be spam.

3 Baysian Networks

The network and probability tables are included. Probability estimates were made by assigning values through reasonable guesses. Since this network is closely related to my personal experience (though I can't remember the last time I had lox and eggs for breakfast) real values could be found simply by documenting situations and choices made over a given period of time. This period of time would have to be long in order to get values for things like Hot Water (the hot water tank rarely breaks).

4 Exact Inference in Baysian Networks

4.1 Query 1

$P(Energy = avg|ToBed = early, Dreams = nightmare, MidnightSnack = pizza, \quad AlarmVolume = soft, Rested, Water = cold, Shower, Breakfast = cereal, Coffee)$

Known Probabilities:

$P_1(ToBed = early) = 0.3$
$P_2(MidnightSnack = pizza|ToBed = early) = 0.05$
$P_3(Dreams = nightmare|MidnightSnack = pizza) = 0.4$
$P_4(AlarmVolume = soft) = 0.6$
$P_5(Water = cold) = 0.1$
$P_6(Breakfast = cereal) = 0.8$

Conditional Probabilities:

$P_C(Coffee) = 0.8$
$P_{\overline{C}}(\neg Coffee) = 0.2$
$P_S(Shower|Water = cold) = 0.2$
$P_{\overline{S}}(\neg Shower|Water = cold) = 0.8$
$P_R(Rested|ToBed = early, Dreams = nightmare, AlarmVolume = soft) = 0.2$
$P_{\overline{R}}(\neg Rested|ToBed = early, Dreams = nightmare, AlarmVolume = soft) = 0.8$
$P_{Esc}(Energy = avg|Shower = true, Coffee = true, Breakfast = cereal) = 0.35$
$P_{Es\overline{c}}(Energy = avg|Shower = true, Coffee = false, Breakfast = cereal) = 0.45$
$P_{E\overline{s}c}(Energy = avg|Shower = false, Coffee = true, Breakfast = cereal) = 0.4$
$P_{E\overline{sc}}(Energy = avg|Shower = false, Coffee = false, Breakfast = cereal) = 0.5$

$$\begin{aligned}
P &= P_1P_2P_3P_4P_5P_6P_CP_SP_RP_{Esc} + P_1P_2P_3P_4P_5P_6P_CP_SP_{\overline{R}}P_{Esc} \\
&+ P_1P_2P_3P_4P_5P_6P_CP_{\overline{S}}P_RP_{E\overline{s}c} + P_1P_2P_3P_4P_5P_6P_CP_{\overline{S}}P_{\overline{R}}P_{E\overline{s}c} \\
&+ P_1P_2P_3P_4P_5P_6P_{\overline{C}}P_SP_RP_{Es\overline{c}} + P_1P_2P_3P_4P_5P_6P_{\overline{C}}P_SP_{\overline{R}}P_{Es\overline{c}} \\
&+ P_1P_2P_3P_4P_5P_6P_{\overline{C}}P_{\overline{S}}P_RP_{E\overline{sc}} + P_1P_2P_3P_4P_5P_6P_{\overline{C}}P_{\overline{S}}P_{\overline{R}}P_{E\overline{sc}} \\
&= 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.8 \times 0.2 \times 0.2 \times 0.35 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.8 \times 0.2 \times 0.8 \times 0.35 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.8 \times 0.8 \times 0.2 \times 0.45 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.45 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.2 \times 0.2 \times 0.2 \times 0.4 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.2 \times 0.2 \times 0.8 \times 0.4 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.2 \times 0.8 \times 0.2 \times 0.5 \\
&+ 0.3 \times 0.05 \times 0.4 \times 0.6 \times 0.1 \times 0.8 \times 0.2 \times 0.8 \times 0.8 \times 0.5 \\
&= 0.0000032 + 0.0000129 + 0.0000166 + 0.0000664 + 0.0000009 + 0.0000037 + 0.0000046 + 0.0000184 \\
&= 0.0001267
\end{aligned}$$

4.2 Query 2

$P(Energy = avg|ToBed = early, MidnightSnack = none, Dreams, AlarmVolume = soft, \quad Rested, Water = hot, Shower = true, Breakfast = cereal, Coffee = true)$

Known Probabilities:
$P_1(ToBed = early) = 0.3$
$P_2(MidnightSnack = none|ToBed = early) = 0.8$
$P_3(AlarmVolume = soft) = 0.6$
$P_4(Water = hot) = 0.8$
$P_5(Shower = true|Water = hot) = 0.95$
$P_6(Breakfast = cereal) = 0.8$
$P_7(Coffee) = 0.8$
$P_8(Energy = avg|Shower = true, Coffee = true, Breakfast = cereal) = 0.35$

Conditional Probabilities:
$P_{D=n}(Dreams = nightmare|MidnightSnack = none) = 0.2$
$P_{D=s}(Dreams = sweet|MidnightSnack = none) = 0.5$
$P_{D=w}(Dreams = weird|MidnightSnack = none) = 0.3$
$P_{Rn}(Rested|ToBed = early, Dreams = nightmare, AlarmVolume = soft) = 0.2$
$P_{\overline{Rn}}(Rested|ToBed = early, Dreams = nightmare, AlarmVolume = soft) = 0.8$
$P_{Rs}(Rested|ToBed = early, Dreams = sweet, AlarmVolume = soft) = 0.85$
$P_{\overline{Rs}}(Rested|ToBed = early, Dreams = sweet, AlarmVolume = soft) = 0.15$
$P_{Rw}(Rested|ToBed = early, Dreams = weird, AlarmVolume = soft) = 0.5$
$P_{\overline{Rw}}(Rested|ToBed = early, Dreams = weird, AlarmVolume = soft) = 0.5$

$$\begin{aligned}
P &= P_1P_2P_3P_4P_5P_6P_7P_8P_{D=n}P_{Rn} + P_1P_2P_3P_4P_5P_6P_7P_8P_{D=n}P_{\overline{Rn}} \\
&+ P_1P_2P_3P_4P_5P_6P_7P_8P_{D=s}P_{Rs} + P_1P_2P_3P_4P_5P_6P_7P_8P_{D=s}P_{\overline{Rn}} \\
&+ P_1P_2P_3P_4P_5P_6P_7P_8P_{D=w}P_{Rw} + P_1P_2P_3P_4P_5P_6P_7P_8P_{D=w}P_{\overline{Rn}} \\
&= 0.3 \times 0.8 \times 0.6 \times 0.8 \times 0.95 \times 0.8 \times 0.8 \times 0.35 \times 0.2 \times 0.2 \\
&+ 0.3 \times 0.8 \times 0.6 \times 0.8 \times 0.95 \times 0.8 \times 0.8 \times 0.35 \times 0.2 \times 0.8 \\
&+ 0.3 \times 0.8 \times 0.6 \times 0.8 \times 0.95 \times 0.8 \times 0.8 \times 0.35 \times 0.5 \times 0.85 \\
&+ 0.3 \times 0.8 \times 0.6 \times 0.8 \times 0.95 \times 0.8 \times 0.8 \times 0.35 \times 0.5 \times 0.15
\end{aligned}$$

$$\begin{aligned}
&+ && 0.3 \times 0.8 \times 0.6 \times 0.8 \times 0.95 \times 0.8 \times 0.8 \times 0.35 \times 0.3 \times 0.5 \\
&+ && 0.3 \times 0.8 \times 0.6 \times 0.8 \times 0.95 \times 0.8 \times 0.8 \times 0.35 \times 0.3 \times 0.5 \\
&= && 0.0009806 + 0.0392235 + 0.0104187 + 0.0018386 + 0.0036772 + 0.0036772 \\
&= && 0.0598158
\end{aligned}$$

4.3 Query 3

$P(Energy = avg|ToBed = usual, MidnightSnack = icecream, Dreams = sweet, AlarmVolume = soft, Rested = true, Water = hot, Shower = true, Breakfast, Coffee)$

Known Probabilities:
$P_1(ToBed = usual) = 0.6$
$P_2(MidnightSnack = icecream|ToBed = usual) = 0.1$
$P_3(Dreams = sweet|MidnightSnack = icecream) = 0.5$
$P_4(AlarmVolume = soft) = 0.6$
$P_5(Water = hot) = 0.8$
$P_6(Shower = true|Water = hot) = 0.95$
$P_7(Rested = true|Dreams = sweet, ToBed = usual, AlarmVolume = soft) = 0.75$
Conditional Probabilities:
$P_{B=c}(Breakfast = cereal) = 0.8$
$P_{B=l}(Breakfast = loxandeggs) = 0.1$
$P_{B=y}(Breakfast = yogourt) = 0.4$
$P_C(Coffee = true) = 0.8$ $P_{\overline{C}}(Coffee = true) = 0.2$ $P_{Ec,b=c}(Energy = avg|Shower = true, Coffee = true, Breakfast = cereal) = 0.35$
$P_{Ec,b=l}(Energy = avg|Shower = true, Coffee = true, Breakfast = loxandeggs) = 0.45$
$P_{Ec,b=y}(Energy = avg|Shower = true, Coffee = true, Breakfast = yogourt) = 0.45$
$P_{E\overline{c},b=c}(Energy = avg|Shower = true, Coffee = false, Breakfast = cereal) = 0.35$
$P_{E\overline{c},b=l}(Energy = avg|Shower = true, Coffee = false, Breakfast = loxandeggs) = 0.45$
$P_{E\overline{c},b=y}(Energy = avg|Shower = true, Coffee = false, Breakfast = yogourt) = 0.45$

Variables eliminated:
$P_{1-7} = P_1P_2P_3P_4P_5P_6P_7 = 0.01026$ $P_{1-7,B=c} = P_{1-7}P_{B=c} = 0.008208$ $P_{1-7,B=c} = P_{1-7}P_{B=l} = 0.001026$ $P_{1-7,B=c} = P_{1-7}P_{B=y} = 0.004104$

$$\begin{aligned}
P &= && P_{1-7}P_{B=c}P_CP_{Ec,b=c} + P_{1-7}P_{B=c}P_{\overline{C}}P_{E\overline{c},b=c} \\
&+ && P_{1-7}P_{B=l}P_CP_{Ec,b=l} + P_{1-7}P_{B=l}P_{\overline{C}}P_{E\overline{c},b=l} \\
&+ && P_{1-7}P_{B=y}P_CP_{Ec,b=y} + P_{1-7}P_{B=y}P_{\overline{C}}P_{E\overline{c},b=y} \\
&= && P_{1-7,B=c}P_CP_{Ec,b=c} + P_{1-7,B=c}P_{\overline{C}}P_{E\overline{c},b=c} \\
&+ && P_{1-7,B=l}P_CP_{Ec,b=l} + P_{1-7,B=l}P_{\overline{C}}P_{E\overline{c},b=l} \\
&+ && P_{1-7,B=y}P_CP_{Ec,b=y} + P_{1-7,B=y}P_{\overline{C}}P_{E\overline{c},b=y} \\
&= && 0.008208 \times 0.8 \times 0.35 + 0.008208 \times 0.2 \times 0.35 \\
&+ && 0.001026 \times 0.8 \times 0.45 + 0.001026 \times 0.2 \times 0.45 \\
&+ && 0.004104 \times 0.8 \times 0.45 + 0.004104 \times 0.2 \times 0.45 \\
&= && 0.0022982 + 0.0005746 + 0.0003694 + 0.0000923 + 0.0014774 + 0.0003694 \\
&= && 0.0051813
\end{aligned}$$

4.4 Query 4

$P(Energy = avg|ToBed = usual, MidnightSnack = none, Dreams = sweet, AlarmVolume, \quad Rested = true, Water = cold, Shower = true, Breakfast = LoxandEggs, Coffee)$

Known Probabilities:
$P_1(ToBed = usual) = 0.6$
$P_2(MidnightSnack = none|ToBed = usual) = 0.7$
$P_3(Dreams = sweet|MidnightSnack = none) = 0.5$
$P_4(Water = cold) = 0.1$
$P_5(Shower = true|Water = cold) = 0.2$
$P_6(Breakfast = loxandeggs) = 0.1$

Conditional Probabilities:
$P_{R,A=s}(Rested = true|Dreams = sweet, ToBed = usual, AlarmVolume = soft) = 0.75$ $P_{R,A=l}(Rested = true|Dreams = sweet, ToBed = usual, AlarmVolume = loud) = 0.74$ $P_C(Coffee = true) = 0.8$ $P_{\overline{C}}(Coffee = false) = 0.2$ $P_{Ec}(Energy = avg|Shower = true, Coffee = true, Breakfast = loxandeggs) = 0.45$
$P_{E\overline{c}}(Energy = avg|Shower = true, Coffee = false, Breakfast = loxandeggs) = 0.35$

Variables eliminated:
$P_{1-6} = P_1P_2P_3P_4P_5P_6 = 0.00042$ $P_{1-6,R,A=s} = P_{1-6}P_{R,A=s} = 0.000315$ $P_{1-6,R,A=l} = P_{1-6}P_{R,A=l} = 0.0003108$ $P_{C,Ec} = P_CP_{Ec} = 0.36$ $P_{\overline{C},Ec} = P_{\overline{C}}P_{Ec} = 0.09$ $P_{C,E\overline{c}} = P_CP_{E\overline{c}} = 0.28$ $P_{\overline{C},E\overline{c}} = P_{\overline{C}}P_{E\overline{c}} = 0.07$

$$\begin{aligned}
P &= P_{1-6}P_{R,A=s}P_CP_{Ec} + P_{1-6}P_{R,A=s}P_CP_{E\overline{c}} \\
&+ P_{1-6}P_{R,A=s}P_{\overline{C}}P_{Ec} + P_{1-6}P_{R,A=s}P_{\overline{C}}P_{E\overline{c}} \\
&+ P_{1-6}P_{R,A=l}P_CP_{Ec} + P_{1-6}P_{R,A=l}P_CP_{E\overline{c}} \\
&+ P_{1-6}P_{R,A=l}P_{\overline{C}}P_{Ec} + P_{1-6}P_{R,A=l}P_{\overline{C}}P_{E\overline{c}} \\
&= P_{1-6,R,A=s}P_CP_{Ec} + P_{1-6,R,A=s}P_CP_{E\overline{c}} \\
&+ P_{1-6,R,A=s}P_{\overline{C}}P_{Ec} + P_{1-6,R,A=s}P_{\overline{C}}P_{E\overline{c}} \\
&+ P_{1-6,R,A=l}P_CP_{Ec} + P_{1-6,R,A=l}P_CP_{E\overline{c}} \\
&+ P_{1-6,R,A=l}P_{\overline{C}}P_{Ec} + P_{1-6,R,A=l}P_{\overline{C}}P_{E\overline{c}} \\
&= P_{1-6,R,A=s}P_{C,Ec} + P_{1-6,R,A=s}P_{C,E\overline{c}} \\
&+ P_{1-6,R,A=s}P_{\overline{C},Ec} + P_{1-6,R,A=s}P_{\overline{C},E\overline{c}} \\
&+ P_{1-6,R,A=l}P_{C,Ec} + P_{1-6,R,A=l}P_{C,E\overline{c}} \\
&+ P_{1-6,R,A=l}P_{\overline{C},Ec} + P_{1-6,R,A=l}P_{\overline{C},E\overline{c}} \\
&= 0.000315 \times 0.36 + 0.000315 \times 0.28 \\
&+ 0.000315 \times 0.09 + 0.000315 \times 0.07 \\
&+ 0.0003108 \times 0.36 + 0.0003108 \times 0.28 \\
&= 0.0003108 \times 0.09 + 0.0003108 \times 0.07 \\
&= 0.0005006
\end{aligned}$$

5 Section 6

The code for calculating π can be seen in figure 1. One instance of results output by this code is in table 1.

darts	10	100	1000	10000
π	3.2	3.24	3.12	3.1424

Table 1: Results on calculating π for a number of random darts

To calculate the cummulative distribution of $X \in x_1, \ldots, x_j$ we sum the probabilities $P(x_1) + \ldots + P(x_j)$ for each $j \in 1 \ldots k$. To calculate in $O(k)$ time we add the current sample to the sum of all previous samples or $x_j + \sum_{j-1}^{i=1} x_i$. This is made possible by keeping the current sum handy at each step of the distribution calculation.

```
import random

def go(num_points):
    total=num_points
    inside=0
    while num_points>0:
        x = random.random()
        y = random.random()
        if (x*x+y*y) <= 1:
            inside = inside + 1
        num_points = num_points - 1
    print ’inside:’ , inside , ’ inside/4:’,4*(float(inside)/float(total))
```

Figure 1: Python function for estimating π

To generate a single sample $P(X = x_i)$ from the cumulative distribution simply take the cumulative distribution probability at x_i and subtract the probability at x_{i-1} where i is not the first position. If $i = 1$ or is the first position then it is simply the probability at x_1.

The cumulative distribution function $C(x)$ is can be defined as $C(x_i) = \sum_{j=1}^{i} P(x_j)$. Then the probability $P(X = x_i) = C(x_i) - C(x_{i-1})$ can be calculated in less than $O(k)$ time provided we can maintain a table with the cummulative distribution sums at each step rather than calculating the entire distribution for each step.

Therefore to generate N samples of X where $N >> k$ we simply use the cummulative distribution table as explained above and do one subtraction to calculate each sample in constant time.