2005

An Information Systems Teaching Case: Bayesian Probability Applied to Spam eMail Filters

Samuel S. Conn
Regis University, sconn@regis.edu

Daniel Likarish
Regis University, dlikaris@regis.edu

An Information Systems Teaching Case: Bayesian Probability Applied to Spam E-mail Filters

Samuel S. Conn
Regis University School for Professional Studies
MSCIT Program
sconn@regis.edu

Daniel Likarish
Regis University School for Professional Studies
MSCIT Program
dlikaris@regis.edu

ABSTRACT
Information Systems professionals can participate in the strategic planning and policy development of the business organization by applying sound techniques for rational decision making. Decision Support Systems often utilize inferential techniques to provide analysis and knowledge creation for business and its information systems. One common method of reasoning under uncertainty is the application of the Bayesian probability model. This teaching case can be used in an Information Systems program to teach one method of inferential reasoning as applied to policy and business rules for spam email filters.

Keywords: Bayesian probability, inferential reasoning, spam email filters, teaching case

INTRODUCTION

Inference techniques are widely used in Information Systems with respect to Decision Support Systems, intelligent systems, and various types of reasoning engines (Gelman, 2003). Knowledge bases allow the acquisition and organization of knowledge, but without some type of reasoning strategy no inferences can be drawn from the knowledge base. As decisions are made or inferred from a base of knowledge, the degree of confidence in the conclusion can be expressed as a probability (Sedlmeier, 2001). This has conceptual value in Information Systems. In the case of email, one acknowledged problem is spam. Spam email creates problems for businesses trying to limit the amount of non-business traffic coming in to corporate email systems, and personal productivity is often hindered by unwanted spam emails in the inbox (Tschabitscher, 2005). The application of an inference technique using Bayesian probability modeling for controlling spam email offers a teaching case for Information Systems. This paper presents a teaching case for the use of a Bayesian approach to email filtering, and thus presents several important outcomes in the study of Information Systems. One important outcome is the use of inferential techniques with information and knowledge in decision support. Another important outcome is the synthesis of the Bayesian approach with a common Information System problem in email systems. And finally, an important outcome of this teaching case is a demonstration of the role of subjective probabilities in Information Systems.

Bayesian filtering is based on a mathematical theory which promotes that by assigning spam probability numbers to individual trigger words within a message, the message can be classified as spam or not spam (Jak.com, 2005). The theory also notes that the Bayesian filter will learn from its experience and more intelligently assign spam probability values to new and existing words (tokens) based on the historic message content of the email (GFI, 2005). The filter will then aggregate all individual values. If the aggregate reaches some pre-determined threshold value, the email is classified as spam (Spam-Protection, 2005). The Achilles heel of the spammers is their message. The real advantage of the Bayesian approach is that you know what you are measuring (Giarrantano, 1998). The Bayesian approach assigns an actual probability and considers all the evidence in the email, both good and bad.

BAYES' THEOREM AS A MECHANISM FOR COMBINING NEW AND EXISTENT EVIDENCE

An English mathematician, Thomas Bayes, was a reverend who lived from 1702 to 1761 and discovered an important probabilistic relationship using conditional probabilities. He created a theorem that uses conditional probabilities to alter computations so that new and relevant information can be accommodated (Bernardo, 2000). The expression of Bayes' theorem uses certain notational form for applying the theorem to any given situation (Neapolitan, 2003). The general notational form used in this teaching case is:

NOTATION	SEMANTICS
¬A	Complement of A (Not A)
P(B|A)	Probability of event B, given event A
P(B|¬A)	Probability of event B, given Not A

These notations can be used to express Bayes' theorem and apply it to the given teaching case herein. Bayes' theorem notes that if we have some event A and we then calculate the probability of A by itself (i.e. P(A)), then we have an unconditional probability. However, if we have a new event B that can affect A, we know the probability of B if A occurs (i.e. P(B|A)), as well as the probability of B occurring given not A (i.e. P(B|¬A)). This new event can result in many potential ways. So the value of Bayes' theorem is that we can use this new event information to reformulate our calculation of the probability of event A occurring given the new event B. Thus the calculation for the probability of A is more realistic by including the information from event B (Winkler, 2003). Bayes' theorem is expressed as:

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

To use Bayes Theorem, events A and B must be identified. The probabilities can then be calculated. The a priori information is defined before the application of new information. Posterior probabilities can be calculated using the new information (Base, 2003).

THE PROBLEM CASE

The ABC Technology Company is experiencing a problem with spam emails. The company is a medium size company with 200 employees who rely on email to conduct their daily operations. The employees email inboxes routinely fill up with spam email and this causes serious delays on the email server, regular disk capacity warnings to employees for exceeding disk quotas, and a general nuisance in parsing emails in their email client. After conferencing with the IS department, ABC Technology Company's management has decided to take action. The management has decided on the following plan.

The business organization needs to implement spam filtering. It plans to do this by creating software in the front end access networks of its infrastructure. The basic strategy of the organization is to create software that can be a predictive indicator for the value of email traffic coming into the organization. Following solid reasoning under uncertainty techniques, the IS department decides that spam filtering software using a Bayesian probability model will work best. The software will be required to rate incoming email messages as high-risk (H), medium-risk (M), or low-risk (L). A rating of H would indicate that the email is highly likely to be spam email and should be filtered. A rating of M would indicate that the email is somewhat likely to be a spam email and should either be filtered or not. And a rating of L would indicate that the email is not very likely to be spam email and should not be filtered out. Historical data in the organization says that 40% of the messages received in the organization are spam emails. The estimated cost of blocking a message to the organization that is actually not a spam is $0.10 and the estimated cost of allowing a spam email to pass is $0.12. If an email is not a spam email, then there is no cost associated with blocking it or allowing it to pass through the email system. When the software was implemented, the test results with email messages that were not spam emails showed that the software rated 20% of the messages that are actually not spams as high-risk (H), 30% of the messages that are actually not spams as medium-risk (M), and the remaining 50% of the messages that are not spams were rated as low-risk (L). The information systems manager must determine a security policy to block or allow (to pass) an email based on the rating it receives.

APPLICATION IN DECISION THEORY: THE INITIAL PROBLEM

In order to tune the software, the probability that a particular email is rated as H, M, or L for being spam must be computed. Therefore, three computations are required to know the conditional probabilities that:

- a message rated as high-risk is actually a spam email
- a message rated as medium-risk is actually a spam email
- a message rated as low-risk is actually a spam email

APPLICATION TO RATIONAL DECISION MAKING

ABC Technology Company's management is trying to decide whether of not to block or allow email based on the rating as determined by the filtering software. They ask the information systems manager to calculate the respective costs associated with blocking or allowing email based on the ratings in order to determine an overall strategy for the organization. The management considers that they are risk neutral in their decision making, so the cost information will provide objective criteria by which to make a decision. So the essential proofs that must be considered are whether the decision should be to block or allow a message when:

- the message is rated by the software as high-risk.
- the message is rated by the software as medium-risk.
- the message is rated by the software as low-risk.

ANALYSIS OF ORGANIZATIONAL COST

ABC Technology Company's management has asked the information systems security manager to determine the cost of implementing the decisions made in the above section. Specifically, management would like to know two things:

(1) If the organization follows the strategy determined by the information systems security manager, what is the expected cost of the strategy (per message) to the organization?

(2) Everything else remaining the same, at least how high should the cost of allowing a spam to get through be for a risk neutral rational decision maker to Block a message when the software rates it as Medium-Risk?

TEACHING CASE SOLUTION

The first step is to identify and state the basic events associated with the problem. Noting that we can have the event of email being either spam (S) or not spam (¬S), there can also be the event that an email can be blocked (B) or email can be allowed (A) to pass through. Thus the basic events would be stated as:

S = event that an email is SPAM
¬S = event that an email is NOT SPAM (HAM)
B = event that an email is BLOCKED
A = event that an email is ALLOWED

The next step is to express all known probabilities associated to the basic events. There is a priori knowledge and unconditional probabilities that are evident from the known data. We know that the probability of email being spam is 40%, so P[S] = 0.4, and thus we know that the probability of an email not being spam is 60%, or P[¬S] = 0.6. From this we can infer the following:

P[H|S] = .6	P[H|¬S] = .2
P[M|S] =.3	P[M|¬S]= .3
P[L|S] =.1	P[L|¬S] = .5

Now Bayesian probability can be used to calculate the conditional probabilities for the questions of whether or not to block or allow certain emails rated as H, M, or L risk. The posterior calculation to answer the question of whether or not a message rated as H (high-risk) is actually a spam email is:

$$P[S|H] = (P[H|S])\,(P[S]) \,/\, (P[H|S])\,(P[S]) + (P[H|\neg S])\,(P[\neg S])$$

$$= \frac{.6 \times .4}{(.6 \times .4) + (.2 \times .6)}$$

$$= .24 \,/\, .36 = .67 = 67\%$$

The posterior calculation to answer the question of whether or not a message rated as M (medium-risk) is actually a spam email is:

$$P[S|M] = (P[M|S])\,(P[S]) / (P[M|S])\,(P[S]) + (P[M|\neg S])\,(P[\neg S])$$

$$= \frac{.3 \times .4}{(.3 \times .4) + (.3 \times .6)}$$

$$= .12 / .30 = .40 = 40\%$$

And the posterior calculation to answer the question of whether or not a message rated as L (low-risk) is actually a spam email is:

$$P[S|L] = (P[L|S])\,(P[S]) / (P[L|S])\,(P[S]) + (P[L|\neg S])\,(P[\neg S])$$

$$= \frac{.1 \times .4}{(.1 \times .4) + (.5 \times .6)}$$

$$= .04 / .34 = .1176 = 11.76\%$$

Now that the conditional probabilities are known, the given costs can be used to compute the expected costs (value) or EC. ECs can be calculated as the sum of (cost x probability) over all possible outcomes. So the information systems security manager could recommend to management whether to block or allow the email when it is rated by the filtering software as high-risk based on the following:

- Looking only at email rated H, 67% is (S) spam (free to block, .12 cents to allow) and 33% is (N) non-spam (.10 cents to block, free to allow).
- The EC of blocking email that is rated H is 67% * 0 + 33% * 10, which equals 3.3 cents per email.
- Similarly, the EC of allowing email rated H is 67% * 12 + 33% * 0, which equals 8.04 cents per email.
- The recommendation should be to block email rated H because it is cheaper to do so.

For the email that is rated M, the information systems security manager could make a recommendation based on the following:

- Looking only at email rated M, 40% is S (spam) and 60% is N (non-spam).
- The EC of blocking email that is rated M is 40% * 0 + 60% * 10 cents, which equals 6 cents.
- Similarly, the EC of allowing email rated M is 40% * 12 + 60% * 0, which equals 4.8 cents.
- The decision should be to allow email rated M because it is cheaper to do so.

And for email that is rated L, the information systems security manager could make a recommendation based on the following:

- Looking only at email rated L, 11.76% is S (spam) and 88.24% is N (non-spam).
- The EC of blocking email that is rated L is 11.76% * 0 + 88.24% * 10, which equals 8.82 cents.
- The EC of allowing email that is rated L is 11.76% * 12 + 88.24% * 0, which equals 1.41 cents.
- The decision should be to allow email rated L because it is cheaper to do so.

The organization's management would consider the information systems security manager's recommendation, and then ask for an analysis of organizational cost. In the previous three calculations the attention was restricted to a subset of all the emails based on the filtering software's rating. Now given the strategy of blocking email rated H, and allowing email rated M and L, the EC of this strategy for all email can be calculated. Hence the following is true:

- If the email is rated H, it will be blocked and the expected cost is 3.3 cents.
- If the email is rated M, it will be allowed and the expected cost is 4.8 cents.
- If the email is rated L, it will be allowed and the expected cost is 1.4 cents.

And to calculate the EC, the percentages of each outcome that was previously calculated can be used. These are stated as:

- 36% (or 24% + 12%) of all email is rated H
- 30% of all email is rated M
- 34% of all email is rated L

Thus the EC for the strategy would be stated as:

$$36\% * 3.3 + 30\% * 4.8 + 34\% * 1.4\% = 3.1 \text{ cents per email}$$

The final question asked of the information systems security manager by the organization's management is how high the cost of allowing a spam email to get through should be to make the decision to block an email when the filtering software rates it as M. Since nothing changes except for the cost (C) of allowing spam email, the EC of blocking email rated M is still \$0.06. The EC cost of allowing email is .4 x C + .6 x 0. So the information systems security manager recommends blocking email rated M if the EC of allowing email is more than \$0.06. This will happen if .4 x C > 6, or restated C > 6/.4 which equals \$0.15.

THE TEACHING CASE AS A PEDAGOGICAL METHOD

Information Systems education frequently uses case-based pedagogies because it is proven to be effective. Case-based learning is a core feature of education programs in many parts of the world (Jackson, 2003). This teaching case can be used as an individual assignment or as a group assignment in an Information Systems course offering instruction in reasoning under uncertainty, decision support, or email system engineering using Bayesian filter design. Case-based learning is also adaptable to various learning theories. According to Crittenden (2005), case-based education is a viable part of social learning theory. And the case method of instruction has a lengthy history as an effective form of active learning. Kock, Aiken and Sandas (2004) point out the increasing need for more realistic courses in Information Systems teaching the use of complex and domain-specific IT applications. The use of case-based learning is also consistent with the current trend toward interdisciplinary courses between business and information systems. Case-based teaching has proven successful in IT related disciplines (e.g. Rosson, Carroll and Rodi (2004)). Also, case-based teaching has extreme pedagogical flexibility as seen in Helps and Renshaw's (2004) learn-expand-teach model that uses teaching cases. Polak (1999) used case-based learning early on to teach group decision support systems. Teaching cases should be used to support and enhance a variety of pedagogical methods and instructional styles in information systems education.

SUMMARY AND CONCLUSIONS

Bayesian probability methods are commonly used in spam email filters. This teaching case for an Information Systems curriculum applies a business case where Bayesian probabilities are used to help managers with rational decision making. This example underscores one way in which Information Systems professionals can provide valuable information to an organization's management and participate in strategic decisions regarding business rules. The problem case is presented first to lay the foundation for application of Bayes probability. Next the application to decision theory and rational decision making is framed by stating the computations and proofs that must be made. And finally, the central problem of the teaching case is presented by focusing on an analysis of organizational cost. Management poses two central questions with respect to email filtering policy that can be answered through application of Bayes probability model. Students should work to demonstrate a working knowledge of Bayesian probability to create a policy that meets the business rules of the organization. Following this the teaching case solution is presented. The solution first states the a priori knowledge given in the teaching case in terms of the unconditional probabilities. It then applies a Bayesian probability calculation to solve for the unknowns.

REFERENCES

1. Andrew Gelman, J. B. C., Hal S. Stern, Donald B. Rubin. (2003). *Bayesian Data Analysis, Second Edition*. Boca Raton: Chapman & Hall/CRC.

2. Base, C. H., & Brase, C. P. (2003). *Understandable Statistics*. Boston: Houghton Mifflin Company.

3. Bernardo, J. M., & Smith, A. F. M. (2000). *Bayesian Theory*. Indianapolis, IN: John Wiley & Sons.

4. Crittenden, W. F. (2005). A social learning theory of cross-functional case education. *Journal of Business Research, 58*, 960-966.

5. GFI. (2005). *Why Bayesian Filtering is the Most Effective Anti-spam Technology*. Retrieved January 18, 2005, from the World Wide Web: http://www.gfi.com/whitepapers/why-bayesian-filtering.pdf

6. Giarrantano, J., & Riley, G. (1998). *Expert Systems Principles and Programming*. Boston: Internation Thomson Publishing Company.

7. Jackson, J. (2003). Case-based learning and reticence in a bilingual context: perceptions of business students in Hong Kong. *System, 31*, 457-469.

8. Jak.com. (2005). *Bayesian Anti Spam Filters - Glue & Duct tape for an End-of-Life Technology*. Retrieved January 10, 2005, from the World Wide Web: http://www.jak.com/Bayesian-Anti-Spam-Filters.htm

9. Kock, N., Aiken, R., & Sandas, C. (2004). Isolated versus integrated case studies: A comparison in the contect of teaching complex and domain-specific IT applications. *Computers and Education, Article in Press and available online at Science Direct*.

10. Neapolitan, R. E. (2003). *Learning Bayesian Networks*. Indianapolis, IN: Prentice Hall.

11. Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian Reasoning in Less That Two Hours. *Journal of Experimental Psychology: General, 130*(3), 380-400.

12. Spam-Protection.com. (2005). *Bayesian Email*. Retrieved January 20, 2005, from the World Wide Web: http://www.spam-protection.com/bayesian_email_filters.htm

13. Tschabitscher, H. (2005). *What You Need to Know About Bayesian Spam Filtering*. Retrieved January 15, 2005, from the World Wide Web: http://email.about.com/cs/bayesianfilters/a/bayesian_filter.htm

14. Winkler, R. L. (2003). *An Introduction to Bayesian Inference and Decision, Second Edition*: Probablilistic Pub.