TJHSST Computer Systems Lab Senior Research Project Human Cognitive Emulation 2006-2007

Lee Rumpf

June 12, 2007

Abstract

Attempting to recreate accurate human responces to stimuli is something that man has been working on since the dawn of computers. While doing so would require a lifetime of research and work, bits and pieces can be attempted by individuals. Using a survery format, this experiment hopes to produce a unique responce to a stimuli based on information gained about the user. While alone, the ramification of this lab can perhaps draw broad conclusions about groups of people and how they respond, combined with other techniques of emulating human thought patterns, computers can become closer and closer to accurately representing a real human.

1 Introduction

This project looks to accomplish two things: the stated goal of accurately predicting a response to a stimuli (in this case a simple question), and also to discover trends within the student body. In order to accomplish the first a survey was created. The idea is that if enough data is gathered from a survey which can group people together, i.e Sally and Sue both are female, seniors, took higher math, participated in sports, dedicated a lot of time to community service, etcetera... and both felt well prepared for college life, then if Sarah comes along with those same traits, she is very likely to feel

well prepared for college life as well. In order to take all of the data in and organize it, a program was written. This program, written in LISP, has its most important piece in what is known as ID3. ID3 is a method of tree sorting developed in 1975 that uses entropy to weigh the importance of each trait (Time in Community Service or Sex of Student) vs. the outcome (Preparedness for College Life). The reason it is important to sort the data in that manner is that when an outside user comes along, they need but follow the tree from trunk (most important factor) to branch (least important) in order to arrive at their predicted leaf (the response to the stimuli). Without it the outside user would have to slog through all the data to find the student or students with whom they share the most traits.

The second goal of this project is to gain an understanding of trends concerning the seniors at TJ and their 1) likelyhood to continue in the sciences at college 2) feelings concerning how prepared they will be for college life and 3) feelings concerning how prepared they will be for college academics. The nature of the program lent itself to this goal so no additional programming is needed. Besides looking at the raw data, the ID3 program shows the most important factor to least important factor when it comes to students and the previously mentioned three questions.

2 Background

The idea of using computers to emulate humans is and has been a hot topic in computer science. When this project began, I contacted Dr. Ann Speed at Sandia National Labs per recommendation of my father. Dr. Speed is a psychologist who works closely with their software developers to study the impact of computer aid in the battlefield and in the private sector. Their goal is to supplement the human brain with computer processing, more specifically to take up the slack that factors such as fatigue and stress cause the human mind.

The use of ID3 has been widespread since its introduction in 1975 by J. Ross Quinlan at the University of Sydney. Its power comes from its ability to sort traits by importance. To do so it uses a statistical property called information gain. Information gain measures how well a given trait (Time spent in Sys Lab, Age, etc..) separates the data according to the responses to the larger questions. To even begin to define information gain, a measure of the impurity or purity of a collection of data must be stated, that measure

is known as entropy. More specifically, in a collection of data S,

$Entropy(S) = -(pp)(log_2(pp)) - (np)(log_2(np))$ where pp is proportion of positive examples and pn is the proportion of negative ones. This equation works only for boolean (true/false) responses. If more than two outcomes are possible (not prepared, prepared, well prepared) then

$Entropy(S) = sum(k = 1tos)of - (pk)(log_2(pk))$ where pk is the proportion of S belonging to class k. All entropy calculations are in base 2 because entropy is a measure of the expected encoding length measured in bits. See appendix 1 for a graph of entropy relative to a boolean classification, as a proportion, pp, of positive examples between 0 and 1.

Now that Entropy has been calculated, we return to the idea of information gain. Information gain is the expected reduction in entropy caused by partitioning the data according to a trait. Again more specifically, Gain(S,A) of an trait A, relative to a set of data S, is defined as

$Gain(S, A) = Entropy(S) - (sumfromValues(A)of(|Sv|/|S|) * Entropy(Sv)$ where Values(A) is the set of all possible values for trait A (i.e Time spent in SysLab can be 0, ¡4, 5, etc..) and Sv is the subset of S for which trait A has the value v.

With a ranking of Gains for all the different traits, the work is now done for ID3 and the traits are sorted by importance.

3 Structure

LISP is not an easy language to comprehend or read, so for ID3 I used one provided in the book Machine Learning (Mitchel 1997). The main work done in the program was organizing over 200 data points in a number of different arrangements. Using VIM I was able to more easily do this. There was also the need to preset the creation of the tree in the body of the program, so that it would not have to be done every time CLISP was called to run the program.

The survey was written in HTML by Josiah Boning. I came up with the traits and values, but with little or no knowledge of web-based programming, I asked an Intranet Administrator to aid me in my project. Over 260 responses were received from the Intranet survey but only about 240 were usable as some students didn't compete the survey. The survey questions and values were:

1.How much time in a week do you spend in/around the Systems Lab?

None (0 hours) low (< 2 hrs) avg (around 4 hours) high (> 5 hours)
2. How much time in a week do you spend in/around Sports?
None (0 hours) low (< 4 hrs) avg (around 7 hours) high (> 10 hours)
3. How much time in a week do you spend in/around Drama?
None (0 hours) low (< 4 hrs) avg (around 7 hours) high (> 10 hours)
4. How much time in a week do you spend in/around VideoTech?
None (0 hours) low (< 2 hrs) avg (around 4 hours) high (>5 hours)
5. How much time do you spend doing Community Service in a year?
None (0 hours) low (< 12 hrs) avg (around 36 hours) high (> 48 hours)
6. How much time do you spend on your Homework per week?
None (0 hours) low (< 5 hrs) avg (around 10 hours) high (> 15 hours)
7. How much time do you spend playing games per week?
None (0 hours) low (< 2 hrs) avg (around 6 hours) high (> 10 hours)
8. How much time do you spend watching tv per week?
None (0 hours) low (< 2 hrs) avg (around 6 hours) high (> 10 hours)
9. Are you male or female?
male/female
10. What is your race?
white/asian/wasian/other
11. Did you take AP Chem?
yes/no
12. Did you take AP Physics?
yes/no
13. Did you take AP Bio?
yes/no
14. Did you take Multivar/Linear Algebra?
yes/no
3 important questions:

After attending Tj, do you plan on continuing with physical science/ math as your focus?
yes/no
How prepared do you feel for college academics?
well prepared / prepared / not prepared
How prepared do you feel for college life?
well prepared / prepared / not prepared

4 Results and Conclusion

The result of this project was a success. It does in fact accurately predict what a random senior would say when confronted with one of the three qualifying questions, based on their responses to the preceding 14. The more interesting results are for the secondary goal of this project, to discover trends concerning seniors and their feelings and plans for college. Because of the ID3 program, it was determined that the most important factor (or largest Information Gain) for deciding whether or not a senior will continue with science or math in college was how much time they spent in the Systems Lab. This makes sense because those who spent the most time in the Sys lab are almost definitely going to continue with a tech focus while those who spend no time have pretty good odds of not doing so. It was also determined that the most important factor concerning how prepared seniors feel for college academics is whether or not they took a higher than required math class. With a higher math background many students may feel more comfortable engaging in the rigors of college classrooms. Lastly, the program determined that time spent playing sports has the most impact on how prepared students feel for college life. This result was not as easy to analyze, but after much thought and discussion, it could be that participating in a sport leads directly to human interaction, and experience with human interaction would lead a student to be more comfortable when it comes to college life, a place where human interaction is required. Students who, say, spend the majority of their time in the computer lab are not forced to interact socially, and therefore may not feel as comfortable going into an environment where such interaction is necessary.

5 Discussion

Every reader can take what they will from the general results that were found from this project. For example a software firm may determine that it is more important to get their workers to participate in a team sport that before simply because it would seam that sports impact the social wellbeing of their workers. Additionally, the recent trend at TJ for more and more students to finish their senior year with multi-variable and linear algebra is a good sign, it may mean that graduates will be more confident going into their freshman year of college. The results could go to middle schools to

encourage an earlier start in high-school mathematics. As for the main goal of the project, there is little application beyond a "wow" factor that a program exists which knows an answer before it is said. It was a good experiment in evolutionary programming and a success as far as accomplishing its stated goals.

6 Bibliography

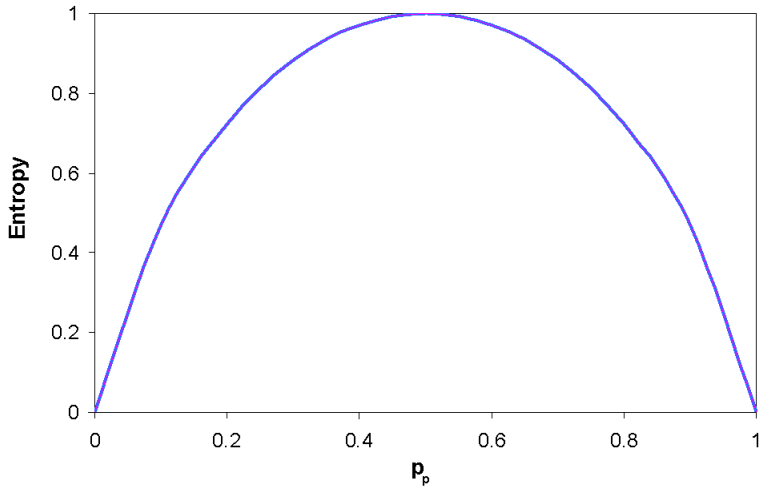
Forsythe, C., Xavier, P. (2002). Human emulation: Progress toward realistic synthetic human agents. Proceedings of the 11th Conference on Computer-Generated Forces and Behavior Representation, Orlando, FL.257-266. http://www.sandia.gov/cog.s

Project Title: Extensible Knowledge-Based Agents for Simulation An essential step in developing agent-based simulations for any application involves the representation of knowledge for the application domain. This project was undertaken with the objective to expand existing capabilities for human modeling and simulation to facilitate their application to new domain problems. Specifically, this project has emphasized the development of techniques for knowledge elicitation and modeling to support creation of individualized models of naturalistic decision making processes.

Forsythe, C. (2001). Toward a human emulator: A comprehensive representation of human cognition. Presentation at ATEDS/SA, March 13-15, San Diego CA. http://www.sandia.gov/cog.systems/documents/ForsytheATEDS.pdf

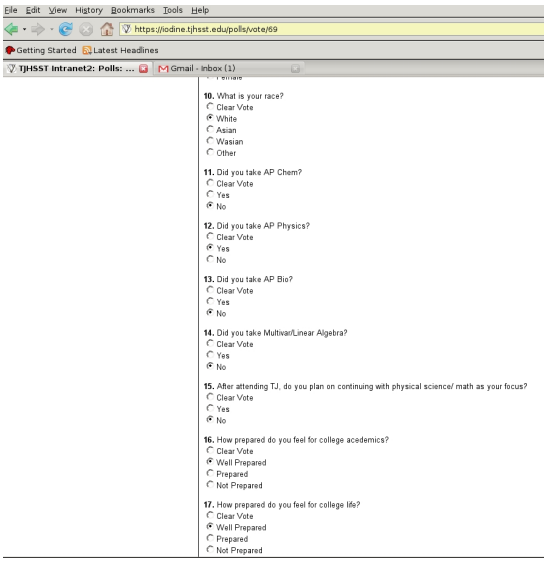
Mitchell, Tom M. (1997) Machine Learning. McGraw-Hill, Inc.

7 appendix 1



Entropy

8 appendix 2



Online Survey