Machine Learning of the College Admissions Process
TJHSST Senior Research Project
Computer Systems Lab 2009-2010

Sam Rush
Period 4

January 24, 2010

Abstract

The goal of this project is to analyze the various biases that exist in the college admissions system by attempting to predict college decisions. This project will attempt to reduce college admissions to pure numbers, excluding data that is inaccessible such as essays and teacher recommendations. Past user-submitted data from the 2007[3], 2008[2], and 2009[1] *Senior Destinations* websites will be used to train an artificial neural network in a process known as machine learning. Then, factors such as the gender bias and the race bias will not only be proven to exist but will be quantifiable.

Keywords: college admissions, machine learning, neural network

1 Introduction

The college application process has become a hypercompetitive environment in which students embark on a four year process of padding their reśumeć to look impressive to an admissions officer. College admissions is often publicized as a wholistic process in which admissions officers look at everything without "weighting" certain aspects of your application such as GPA. Therefore students look to excel in all areas instead of taking the most efficient path, which is not immediately obvious. So, how do we determine what's *really* important to a college? In this paper I attempt to answer that question.

2 Background

2.1 Machine Learning

Machine learning is the programming technique in which a program's behavior is altered according to pre-existing data. A computer's ability to learn is its ability to recognize patterns among data sets and apply those patterns to other data. This is essentially data interpolation. In this project, supervised machine learning is used to translate an N-dimensional input vector into a single scalar output.

2.2 Senior Destinations

At TJHSST, it has become a tradition for one person in each class to create something called the Senior Destinations website. This site enables those in each class to submit their information (such as GPA, SAT scores, AP scores, etc.) along with where they applied and what happened to each application. Data still exists from the Senior Destinations sites of 2007[3], 2008[2], and 2009[1] and will be used to train the neural network. I should note that while a significant portion of the senior class does participate in this each year, the data is somewhat skewed toward the higher achieving portion of the class, since they are more likely to be enthusiastic about college and the admissions process.

3 Development

The project will consist of two parts: the 2010 Senior Destinations website and a College Analysis website. The first site will be coded from scratch in order to be cleaner and more complete than the previous years' sites. The College Analysis website will deal with the machine learning and analysis of college admissions.

3.1 Languages

3.1.1 PHP

PHP: **H**ypertext **P**reprocessor is the main language of this project. The output consists of the standard web elements: Hyper Text Markup Language (HTML), JavaScript, and Cascading Style Sheets (CSS). The websites will run on an Ubuntu Linux machine running the Apache HTTP Server.

3.1.2 MySQL

MySQL is a Structured Query Language that is the storage engine for this project due to its integration into PHP.

3.2 College Analysis Website

To make the College Analysis Website, data will first need to be imported from the Senior Destinations sites and missing data will need to be filled in. For example, the classes of 2007, 2008, and 2009 do not have gender and race data readily available. This data will be manually input by me using pictures from TJHSST yearbooks.

Another discrepancy between the data sets is the GPA. The class of 2010's GPAs are calculated in a different way from the other classes due to a system called FAIRGRADE[4]. Luckily, at TJ, one's FAIRGRADE GPA can be fairly easily predicted from their pre-FAIRGRADE GPA. To come up with a transformation between the two, I took the currently submitted 60 GPAs from the class of 2010 and 60 evenly distributed (by class rank) GPAs from the classes of 2007 and 2008. Then, I plotted the pre-FAIRGRADE GPAs on the X-axis and the FAIRGRADE GPAs on the Y-axis and took the quintic of best fit. This process works if you make the fair assumption that the distribution of GPAs is constant from class to class.

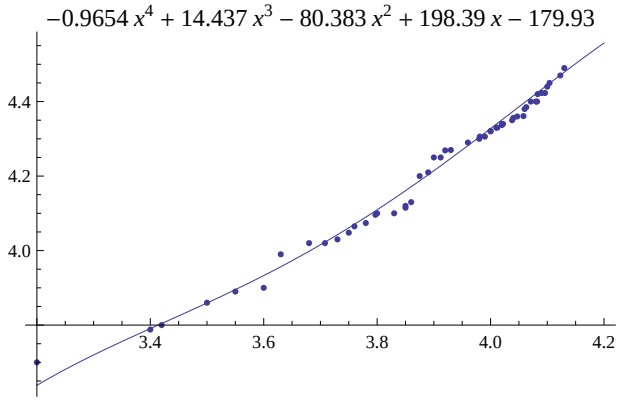


Figure 1: FAIRGRADE GPAS plotted versus Non-FAIRGRADE GPAs, with the quartic of best fit ($R^2 = .9903$) as the interpolation function.

3.2.1 Neural Network

Currently, I am modeling both factors (GPA and SAT) as sigmoids, or $f(x) = \frac{1}{a+e^{b(c-x)}}$. This roughly means that there's a cutoff (which could be steep or lax) GPA and a cutoff SAT for each college. The neural network trains three nodes to the optimal values of a, b, and c for each factor. In further development, I will instead break the range of factors up into four segments and use cubic splines in each segment to model the data. This will be accurate in most situations.

3.3 Testing

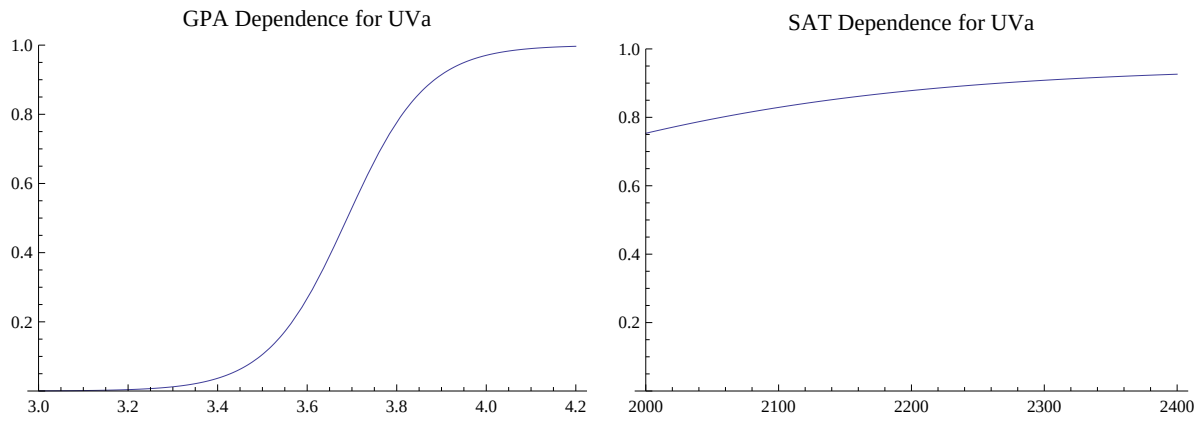
After the neural networks have been trained by the past admissions' data, the TJHSST class of 2010's application data will be run through the program and the computer will output its predictions of each result. In April, the users of the Destinations site will return to input their admissions results. At that time, the predictions will be compared with the actual results for accuracy. Then, the neural networks will be retrained with the inclusion of the 2010 data. With four years of data, I can then begin to investigate biases for each individual colleges. The gender weight, G, is the node in the neural network that will quantify the gender bias of the system. That is, if the weight is positive, males are more likely to be accepted than females and vice versa. Similarly, there will be weights for each race which will quantify those biases as well.

4 Expected Results

Due to the limited data from Senior Destinations sites, I do not expect this to be a stellar indicator for college admissions. I expect that the computer will be able to predict an applicant's fate 60% of the time. I expect that the gender weight will come out to be negative, indicating that you are more likely to get in if you are female. Additionally, I predict that the race weights for white and asian students will be negative, except for Michigan and California schools, where Affirmative Action is banned.

5 Preliminary Results

I have tested the program for the University of Virginia. For the GPA multiplier, the neural network gives $f(x) = \frac{1}{1.00026+e^{9.28027(3.68945-x)}}$. For the SAT multiplier, the neural network gives $f(x) = \frac{1}{1.05321+e^{.00583(1778-x)}}$. These two functions are plotted below.



To come up with an approximate probabiliity of admission, we can multiply these two functions. This is shown in the graph below.

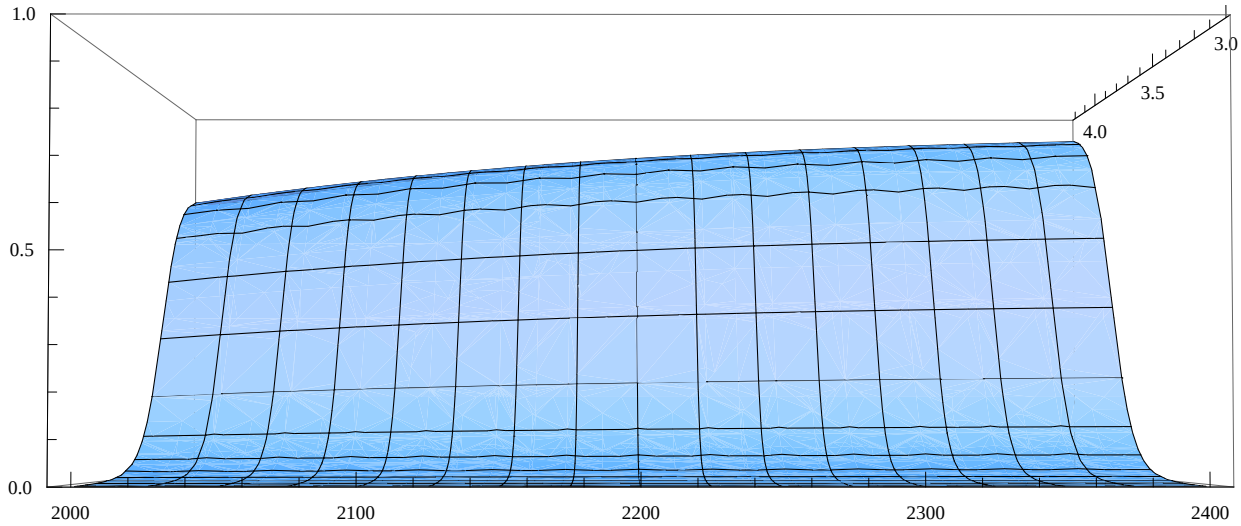


Figure 2: GPA vs. SAT vs. Acceptance Probability at UVa

6 Final Results

There are no final results as of this time.

References

[1] Thiagarajan, Arvind. "TJHSST Class of 2009 Senior Destinations" <http://www.kavitech.com/EduInfo/Destinations/Destinations.html>

[2] Chen, Jeff. "TJHSST Class of 2008 Senior Destinations" <http://www.tjhsst.edu/ jchen2/college>

[3] Wang, Jonathan and Zeng, Will. "TJHSST Class of 2007 Senior Destinations" <http://www.tjhsst.edu/ pwang/college/base.php

[4] "Fairfax County School Board Votes to Change Grading Scale." Fairfax County Public Schools, 1/23/2009. <http://commweb.fcps.edu/newsreleases/newsrelease.cfm?newsid=1058>

[5] Chang, Lin "Applying Data Mining to Predict College Admissions Yield: A Case Study" New Directions for Institutional Research, n131 p53-68 Fall 2006