Did you make the Wright Choice for your College?

Matt Piekenbrock (U00625376),
Jace Robinson (U00667313),
Ning Xie (U00833572)

October 19, 2016

Introduction

It is well known that U.S. university students often graduate with copious amounts of financial debt. Despite the U.S. Department of Educations Mission Statement of "...promote[ing] student achievement and preparation for global competitiveness by fostering educational excellence and ensuring equal access." [1], college tuition rates are often far above an affordable threshold of being considered practical, which often limits many aspiring students' enrollment choices to educational institutions that seem more affordable. Even worse, students who do graduate often find themselves facing repayment obligations that far outweigh their potential income earnings. These notions collectively spark a number of questions, centered around the same idea: Is college generally worth the massive financial debt that prerequisites attendance and, if so, which colleges are offer truly cost-effective educations, given their tuition cost?

Goals

The average person in the U.S. may perhaps agree that students who graduate from the "elite" colleges (i.e. Stanford, MIT, Carnegie Mellon, UC Berkeley, etc.) tend to earn more than graduates from more middle-ground universities, however, the in-depth relationships between future income and choice of university are not at all well known, nor are they intrinsically deterministic.

The recent manifestation of openly accessible, public-domain data sets provides an opportunity never before possible to know more about the underlying relationships regarding the true quality of college education. Using the College Scorecard data set collected by the United States Department of Education [4], we intend to provide a detailed 'bang-for-buck" analysis into the underlying reasons regarding the relative success–or failure–graduates from particular universities of interest are facing.

Data Set

The college scorecard data set [4] is provided to the public by the U.S. Department of Education. This data set contains over a thousand features on thousands of universities since 1996. Looking closer at the data for the year 2014-15, it contains 1743 features on 7703 colleges. Some examples of colleges included are Wright State Main-Campus, Ohio State, University of Phoenix, and Harvard. The over 1700 features describe various interesting details about each college such as admission rates, ACT/SAT requirements, tuition costs, student ethnicity proportions, average student family's income levels, retention rates, debt levels, post graduation income levels, and many, many more. Some multi-class features of interest include the region (Great Lakes, New England, etc...), the locale of institution (city, suburb, town, rural), or control of institution (public, private nonprofit, private for-profit). An additional notable classification is the *Carnegie Classification*, which is used to classify universities into broad groups such as large, medium, or small in population size, or for defining an undergraduate profile such as majority full-time or part-time students with low or high transfer rates. From the rich set of features this data set provides, this research project seeks to answer some exciting questions relating to the quantitative differences in colleges.

Objective

In an effort to achieve the global goal of defining the best college for the price, several smaller questions can be asked and answered within the scope of this project.

1. What are the universities that provide the maximum value (to be defined) to its students?
2. Examine trends about Wright State, Ohio State, and University of Dayton over time. What are the projected employment rates, post graduate income levels, and debt levels in the next 10 years (extrapolation on regression)?
3. Predict post graduate income by college based features such as family income, graduation rates, ethnicity proportions, acceptance rates, population size, debt levels, and tuition levels (regression problem). What set of features provide the best prediction accuracy?
4. Explore additional regression problems for predicting tuition costs or debt levels.
5. Examine the relationships between the region of a college and various input features such as income levels, tuition costs, graduation rates, etc... (classification problem). Are there significant differences in these features in differing regions?

6. Explore additional similar multi-class classification problems for Carnegie classifications and control of institution.

The field of Machine Learning is vast; the book "Machine Learning: A Probabilistic Perspective" [3] covers 28 major categories of the machine learning field, discretized into distinct chapters, each of which presents numerous models, methods, and statistical estimation techniques that could all potentially be used *in some way* to address the above goals. Although some techniques may offer advantages for specific goals compared to others (mentioned below), the correct algorithms or "tools" to do a generalizable cost-analysis of aforementioned scorecard data are, relatively, unknown to the authors. Despite this, and following the curriculum of the class, we've chosen to include Bayesian equivalents to several of the machine learning models presented throughout the semester, including:

1. Linear Regression
2. Logistic Regression
3. Bayesian Linear Regression
4. Bayesian Logistic Regression
5. Support Vector Machine (SVM)
6. Softmax Regression
7. Fuzzy Logic Techniques (time-permitting)

The reasons for a Bayesian approach are manyfold; primarily, we believe inference-based techniques using probabilistic models should have a systematic framework for incorporating *a priori* knowledge into the calculation of the model itself, and that the probabilities estimated by such models should be built using the full degree of belief we have about the underlying data set. From a practical point of view, Bayesian statistics have been on the rise in popularity [2] in the past decade, and an exploration into Bayesian methods of probabilistic inference seems to be both valuable and effective to understanding more about the field of Machine Learning itself.

A support vector machine allows for a more complex decision boundary than the logistic regression. In cases where the data is linearly separable, it is expected the two algorithms will achieve similar performance. In cases where the separation of data is more complex, a support vector machine has the advantage of being able represent non-linear boundaries in high dimension space using a *kernel trick*.

A limitation of logistic regression and SVM is the requirement of a binary output variable. In a multi-class setting, these problem can be transformed into a collection of one-versus-all binary problems. An alternative to changing the problem is to use an algorithm such as Softmax Regression which allows for a

categorical decision. The performance differences of this approach versus the former will be explored.

In many classification settings, a binary or categorical yes or no response does not always capture the complete complexity of human reasoning. Often there are non-distinct boundaries representing a smoother transition between yes or no. Frequently people describe colleges with different levels of severity such as a *great* college, *good* college, *okay* college, or *bad* college. Through the use of *Fuzzy Logic*, we are able to capture this nuanced transition from yes and no. The exploration of fuzzy logic techniques will be a *time-permitting* optional investigation by the authors.

To determine the relative performance of each of the algorithms on during this investigation, the following questions will be explored:

1. Does the incorporation of prior knowledge in the Bayesian algorithms improve regression and classification accuracy? Use performance metrics such a mean-squared error, Precision, Recall, F1, and more.

2. Does an algorithm such as the SVM lead to improved classification accuracy? Use performance metrics such a Precision, Recall, F1, and more.

3. Does the use of Softmax regression against the one-versus-all algorithms lead to improved classification accuracy? Does one-versus-all *miss* meaningful information?

4. Can Fuzzy Logic allow for a more insightful classification of the dataset as opposed to multi-class classification using logistic regression and SVM (time-permitting)?

In the completion of this report, the authors and future readers are expected to have a stronger understanding of the differing characteristics of colleges around the nation. Along with the dataset analysis, readers will be exposed to how a diverse set of algorithms perform on similar problems, providing practical knowledge on the strengths and weaknesses of each of the techniques.

Approach

Evaluation

Conclusion

References

[1] Overview and mission statement — u.s. department of education. http://www2.ed.gov/about/landing.jhtml. (Accessed on 10/18/2016).

[2] D. Ashby. Bayesian statistics in medicine: a 25 year review. *Statistics in medicine*, 25(21):3589–3631, 2006.

[3] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[4] U. D. of Education. College scorecard data, 2016.

Appendices

A Milestones

Task	Projected Completion Date	Assigned Member	Completion Date
Implement Bayesian Linear Regression on sample dataset	10-27-16	Matt	
Implement Bayesian Logistic Regression on sample dataset	10-27-16	Jace	
Implement SVM on sample dataset	10-27-16	Ning	
Analyze College Scorecard dataset with Linear Regression	11-3-16	Matt	
Analyze College Scorecard dataset with Logistic Regression	11-3-16	Jace	
Implement SoftMax on sample dataset	11-3-16	Ning	
Analyze College Scorecard dataset with Bayesian Linear Regression	11-10-16	Matt	
Analyze College Scorecard dataset with Bayesian Logistic Regression	11-10-16	Jace	
Analyze College Scorecard dataset with SVM and Softmax	11-10-16	Ning	
Create nearly completed regression analysis report	11-17-16	Matt	
Create nearly completed logistic regression analysis report	11-17-16	Jace	
Create nearly completed SVM and Softmax analysis report	11-17-16	Ning	
Merge analysis into single report	11-24-16	ALL	