Team 44: Jared Fernandez, Isaac Lee, Murphy Angelo
EECS 349: Machine Learning
Professor Downey
May 17, 2017

Project Status Report

Introduction

We use machine learning to explore data on American colleges and predict post-graduate earnings. There are hundreds of different features that make each college unique, including: location, admission rates, tuition, demographics, average student loans, and average standardized test scores. Through machine learning, we discover which of these features truly matter and predict the average post-graduate earnings for a given college.

Data Analysis

Describe the data set you have utilized to date. What types of attributes are there, how many attributes, how many examples, and how have you partitioned the data for the purpose of development/training/validation/testing.

The data was obtained from the College Scorecard Dataset provided by the Department of Education. The target class was determined by grouping each instance into buckets based on the school's median income ten years after graduation. The classes were grouped into eight buckets of $15,000 increments (Ex. $0-$15k, $15k-$30k, etc.). This set includes demographic data, average standardized test scores, admission rates, tuition, etc. There are 7703 total examples given in the dataset. However, when adjusting for missing data points we end up with approximately 4350 examples.

Preliminary Results

We first calculated the ZeroR accuracy by determining the maximal class (i.e. the most common income tier). We then tested the accuracy of a decision tree algorithm on our dataset, using the DecisionTreeClassifier function of the scipy package. The accuracy was tested using 10-fold cross validation.

Our initial tests used the scipy python package's DecisionTreeClassifier function

ZeroR Accuracy: 0.42 Tree Accuracy: 0.67 (+/- 0.05)

Future Plans

In the future, we plan to test the accuracy of various machine learning techniques in order to optimize our solution. These techniques include but may not be limited to nearest neighbor with various distance measures, neural networks, and naive bayes.

We also plan to make a website that allows people to enter in their own data for a given institution. This allows people to interact with our model and forecast how their own attributes impact their predicted post-graduate earnings.

We would appreciate any suggestions in regards to what machine learning techniques we should try in order to optimize our solution! :)