Yoni Brande

PHSX 815 HW12

Picking Universities to Apply To

Given data on all the universities in the country, we would like to be able to rank them by some measure of quality. The IPEDS dataset provided has 144 separate parameters that we can use for ranking, although in reality not all of these will be useful. Several of them are organizational (ID number, name, year of data), and some of them are not likely to be relevant to every applicant. To try and provide a broadly useful metric of utility for a student, we chose to focus on 6 parameters: cost, size, admission rate, 4-yr graduation rate, endowment per student, and diversity. These are somewhat intentionally vague, to highlight the effects of possible bias or parameter covariances, but they each correlate to columns in the dataset: "Total price for in-state students living on campus 2013-14", "Estimated enrollment, total", "Percent admitted - total", "Graduation rate - Bachelor degree within 4 years, total", "Endowment per FTE", and "Percent of undergraduate enrollment that are White". For “diversity” I took 100 – the percentage of white undergraduates to determine the non-whiteness of a student body. This does not distinguish between the other racial categories outlined in the dataset, which makes this an imperfect metric. Each parameter is normalized to that column’s max value, converted to a percentile, and added in a weighted sum: 1.5*(cost) + 1*size + 0.75*admit + 1*gradrate + 1.25*endow + 1.5*diversity. We would like to select diverse yet somewhat selective schools with high spending per student, and a reasonable 4-yr graduation rate.

Some trends can be noted: 4yr graduation rate is directly correlated with total cost and also with endowment per student. Total cost and endowment per student are also directly correlated with each other. There seems to be a negative trend between admissions percentage and endowment per student. Some parameters don’t correlate, such as percentage of white undergraduates and endowment per student. Although we don’t separate public and private (or junior/senior colleges and universities), the correlations we do see seem to all relate to spending, possibly as a proxy for spending on/by students, or public educational spending. Understanding these relationships may help tweak our weighting scheme, or to select different parameters.

Our model is fairly simple, with three layers of fully connected neurons, and L2 regularization to account for the covariances in some of our parameters. The model was trained with 100 epochs, a batch size of 20, and a 20% validation split. When running the model on the IPEDS data with a 30% test fraction, we find our validation loss and training loss follow very similar trends, and that our model has test data accuracy around 25%, predicts a reasonable spread of quality labels, but is peaked around 5, and skewed somewhat lower. In general, we predict few very bad or very good schools, which in this case may be a benefit, as a smaller pool of high-quality schools simplifies our college application process.

We might note that some of the parameters we chose to use are very highly skewed, like endowment and total enrollment. There are many smaller schools with smaller endowments, and highly weighting these parameters might overlook very good schools. Likewise, it would be easy to overselect for a ludicrously high-endowment school, even if it’s expensive or small. We also have no information about available academic programs, possibly the biggest real-life factor in school choice for students who already have an idea of what they want to study. In a perfect world, the university data would have reasonable distributions for all parameters of interest, with no covariances between parameters, and we

would also have more than about 1500 data points, and be able to make predictions about university quality without the confounding effects. As it is, if we had another, unclassified sample of universities, our model would likely replicate the results found for the IPEDS database and not suddenly come up with a drastically different quality distribution.