11/21/2020

IU Course Recommender

Jared Winslow and Noah Litwiller

Problem space:

The primary objective of this project is to match a student with a set of courses they will be most likely to succeed in. Rather than recommend classes to students based on interest, the main purpose of the recommender is to match the difficulty level of courses with the current ability of the student. A major challenge in solving this problem is to find a way to accurately determine which courses a student will succeed in. In the interest of simplifying this problem, we have chosen to only use data from a single semester, and as such no long term trends are used to solve this problem. To accomplish this goal we will use regression combined with a metric to gauge course difficulty and student compatibility. This compatibility score evaluates the difference between the attributes of a student and a course. For example, a student taking a class outside of their major will often find that course to be more difficult so this should decrease their overall compatibility with that course. Other factors including GPA, grade distribution, course subject area, course level, and student year are also taken into account. Our course data for this project is taken from Indiana University's grade distribution database, and we generated a set of students arbitrarily for testing. A major challenge of this problem is to find a generalizable way to solve it, as no two students are alike. For this reason, as seen in the above compatibility score description, we used attributes that have less variation between students of different majors.

Techniques Implemented:

For the course recommendation portion of our project we implemented our own classification solution which is a nearest neighbor classifier tailored to this specific problem. First our classifier takes a student input and creates a compatibility score with each course in our dataset. As mentioned in the previous section, this score attempts to quantify differences between the average GPA for the course and the student's GPA, the course subject and the student's major, and the course level and the student's year in school. We consider a smaller score to be better than a higher score because it means the student and the course are "closer" to each other. After creating scores with each course, we then loop through all of the courses and pull out the best 5 courses for the student. We chose 5 courses because, with the average course giving 3 credit hours, this gives the student a typical 15 hour course load. However, it is simple to increase or decrease the number of outputted courses in our solution as we are just looping through a list.

In addition to creating a course recommender we also wanted to see what insights we could gather from the grade distribution dataset. To do this we utilized three types of regression models, the first of which is linear regression. The purpose of linear regression is to model the relationship between one or more explanatory variables and a response variable. To accomplish this each explanatory variable is assigned a weight. As the linear regression model is trained these weights converge, and a line formed out of the sum of the products of the weights and their

corresponding variables is created. Then, when a new set of explanatory variable values is given, the linear regression function can output a predicted value of the response variable.

The second regressive model we used is polynomial regression. The training process for this model is very similar to that of a linear regression with the one difference being that polynomial terms can be provided as explanatory variables. These polynomial terms are usually one of the explanatory variables that may be used in linear regression but raised to a power. This type of regression is useful because it is a model for a wide variety of relationships between attributes, not just those that are completely linear.

Lastly, we used stochastic gradient descent regression. Stochastic gradient descent begins with an n-dimensional gradient formed by the input attributes. Similar to linear regression, each input attribute has an associated weight. At each iteration, the algorithm updates these weights and in doing so traverse towards the locally lowest point in the gradient. When the lowest nearby point of the gradient is reached, the weights have converged and the model has been generated. The prediction process for new inputs is identical to linear regression where each input value is multiplied by its corresponding weight to produce the predicted output value.

Comparison to Human Thought Process:

The most clear similarity to the human thought process in our algorithm is in how we determine the compatibility score between a student and a course. If a student were to look at a course themselves they would see the course subject and see if it was the same as their major. They would also be able to look at the course level and see if it would be a good fit for them based on their year in school. The student could also draw conclusions from the course average GPA compared to their own. One major difference here is that a student can make a much more precise calculation based on dozens of other things they might know about themselves. Perhaps they excel in a particular subject that is not their major, or maybe there is a course subject that is very similar to their major subject. Not to mention the fact that students can pick off interest, and we don't have enough data to infer students' interests. Due to the fact that our recommender takes the exact difference of only a few attributes it will not be able to make such a fine tuned prediction.

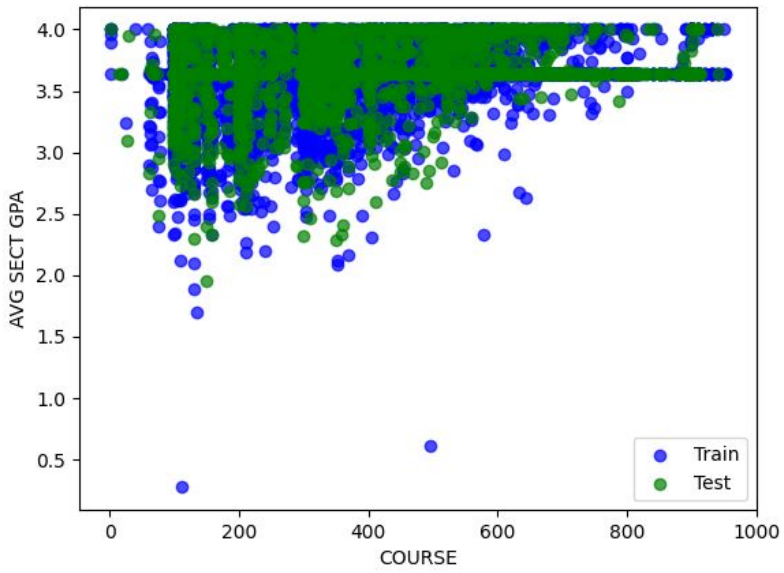
Empirical Analysis and Insights:

During exploratory analysis, we investigated the relationship between numerous variables. Our primary focus was the extent to which average class GPA, our primary dependent variable, varied with respect to different independent variables. The variables we analyzed are as follows:

==

Session (standard/first eight weeks/second eight weeks/nonstandard)
Department
Course Number (0-499 for undergraduate/500-999 for graduate)
Length of Course Title in Characters
Instructor Name
Average Student Cumulative GPA (the average GPA of students enrolled in the class)
Percent Majors
A%
B%
C%
D%
F%
W%
==

Many of these were not related to average class GPA, at least so far as we could tell. However, we did discover multiple interesting relationships. The first example is a relationship between average class GPA and course number. For context, courses ranging 0-99 are often remedial, courses ranging 100-499 are undergraduate level, and courses ranging 500-999 are graduate level. Below is a graph showing the relationship.

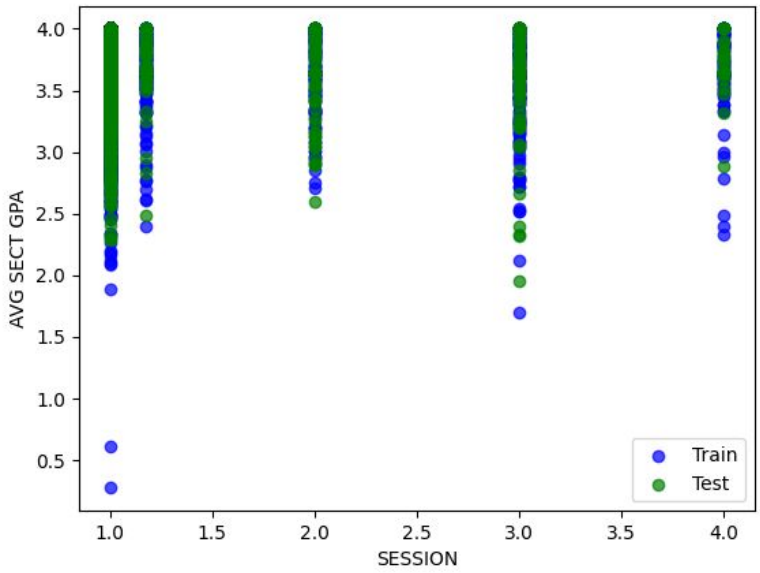


There is a clear positive relationship between course number and average class GPA. There also seems to be a difference in GPA corresponding to the three groups mentioned above. It is also important to note that there are less data points for remedial and graduate courses because there are fewer of those courses. Since this is observational data, we can't determine the causal direction of the relationship. However, we have two hypotheses for this occurrence: 1) students who take higher level classes are better at the classes they take, be it from prior education or the

ability to get into the higher class in the first place or 2) teachers are more relaxed in their grading at higher course levels.

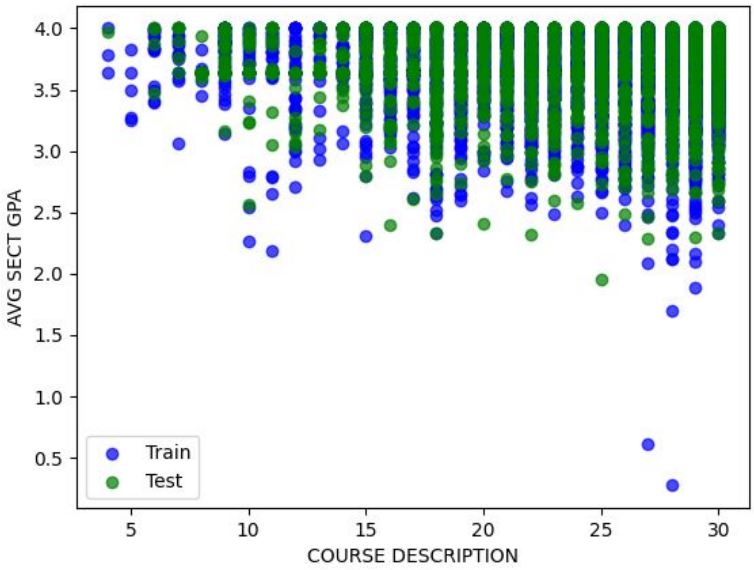
If we look closely, there seem to be two outliers in terms of average class GPA score. We checked which courses these were:
ENGR-E 111 Software Systems Engineering
Average Class GPA: 0.275
INFO-I 450/495 Design and Development of an Information System
Average Class GPA 0.616

Our next example is a relationship between average class GPA and session, a categorical variable. Regarding session membership, 1.0 denotes standard session, 2.0 denotes first eight weeks, 3.0 denotes second eight weeks, and 4.0 denotes nonstandard session.



If we treat the standard session as the baseline, we can look at the other sessions and compare their results. Just from viewing the graph, we can see that the first eight weeks session has a smaller variance. On the other hand, the second eight weeks session has a larger range and has classes with lower scores than any other sessions (disregarding the outliers). This stands out because we would not expect a larger range from a smaller sample size (as compared to the standard session). One hypothesis for this is that students become tired after working for the first eight weeks and are not able to perform as well during the second eight weeks. Again, we see the same two outliers as before.

Our last example is a relationship between average class GPA and length of the course title in characters. Note: this title is not necessarily the same title that can be found in bulletins and acts more like an abbreviated course description. The graph is below:



This was surely the most unexpected of the trends we found. There seems to be a clear negative relationship between the length of the course title in characters and the average GPA of the class. One hypothesis for this result is that as courses become more difficult, they are likewise more specific and granular. This move from breadth towards depth could be reflected in the course titles, where the titles end up more specific, and hence longer. The outliers agree with this relationship.

Other outliers we checked:
CHEM-C 342 Organic Chemistry Lectures 2
65% W
MATH-M 118 Finite Mathematics
63% W
MATH-M 119 Brief Survey of Calculus 1
28% F

It is not surprising to find that the courses that are outliers in terms of withdrawals and failures are the required math courses for IU (and organic chemistry), but the magnitude of the withdrawal percentage is startling.

Outside code:
For this project we made use of four outside libraries. We used Pandas to read the data into python from the CSV file it was stored in. Next, we used Numpy to manipulate the data and convert it into a form we could run regression on. Lastly, we utilized multiple packages from SK-Learn. We used preprocessing, replace_values, and model_selection to prepare our data for regression; we used linear_model to run linear, polynomial regression, and stochastic gradient

descent regression. Finally, MatPlotLib was used for data visualization, both to help find trends in the data and to present our conclusions.

Limitations:

As defined above, our course recommender does not recommend based on student interest. This is a big limitation because students pick classes not only based on their perceived chance of success in the course but also on whether a class sounds interesting or not; in fact, this is probably how most classes are chosen, aside from required classes of course. Overall, this is not our goal, and given the data, we are only about to recommend based on a student's predicted success in classes.

Our course recommendation method does not consider trends over time, but, instead, only uses data from the spring 2020 semester. This may limit the accuracy as courses could be more or less difficult based on the semester they appear in, or if a course has a different professor in another semester. This is especially true given our choice of spring 2020; compared to previous years, spring 2020 classes went online mid semester because of the Covid19 pandemic. In this way, we are more confident in predictions for the time period of the spring 2020 semester rather than for all semesters at IU.

Another limitation of our course recommender is that it uses a small number of attributes to create the compatibility metric between a student and a course. There are many more attributes besides GPA, subject, and level that factor into a student's potential success in a course. We are limited to the data of the grade distribution. If we wanted to investigate other variables, it would be nice to have data on: expected hours of work per week, type of assignments, number of group assignments, number of presentations, temperament of professor, and prerequisites to name a few.

Additionally, the quantifiable attributes of a student will never be a totally accurate indicator of their potential success in a course, especially when trends over time are not taken into account. A student may have experience prior to college in a specific non-major course area or they might have a busier schedule during a semester leading to less time to focus on schoolwork.