**GA Data Science Course**

**Draft Paper**

**Fannie Tseng**

**December 22, 2015**

**Introduction**

This paper presents analyses on a dataset of student participation in MOOCs (Massive Open Online Courses) in 2012-13. The data represents student activity in the first year of Harvard X and MITx courses. The analyses in the paper were focused on answering the following research questions:

* Can unsupervised learning be used to identify alternate outcome groups, based on limited class interaction variables?
* Can supervised learning be used to predict class outcomes?

**Dataset Description**

The dataset, which has been de-identified and made publicly available by HarvardX-MITx, is at the level of one row per-person, per-course, with 641,138 person-course observations. The data includes administrative variables that describe student interactions with the course, including grades and course outcomes, and a limited set of demographic variables (gender, age, country name, highest level of education completed\_)

Course outcomes were pre-defined and coded in the dataset. The outcome definitions are presented below:

|  |  |
| --- | --- |
| Course Outcomes | |
| Registered | Signed up for course |
| Viewed | Accessed the ‘Courseware’ tab (the home of the videos, problem sets, and exams) within the edX platform for the course. |
| Explored | Accessed at least half of the chapters in the courseware |
| Certified | Earned a certificate. Certificates are based on course grades, and depending on the course, the cutoff for a certificate varies from 50% - 80%. |

The following courses were offered in 2012-13 and are included in the MOOC dataset:

|  |  |  |
| --- | --- | --- |
| HarvardX | CB22X | The Ancient Greek Hero |
|
| HarvardX | CS50x | Introduction to Computer Science I |
|
| HarvardX | ER22x | Justice |
|
| HarvardX | PH207x | Health in Numbers: Quantitative Methods in Clinical & Public Health Research |
|
| HarvardX | PH278x | Human Health and Global Environmental Change |
|
| MITx | 14.73x | The Challenges of Global Poverty |
|
| MITx | 2.01x | Elements of Structures |
|
| MITx | 3.091x | Introduction to Solid State Chemistry |
|
| MITx | 6.002x | Circuits and Electronics |
|
| MITx | 6.00x | Introduction to Computer Science and Programming |
|
| MITx | 7.00x | Introduction to Biology – The Secret of Life |
|
| MITx | 8.02x | Electricity and Magnetism |
|
| MITx | 8.MReV | Mechanics Review |
|

**Clustering Analyses**

For this paper, I analyzed student participation in two courses, “Justice” (hereafter called Justice), and “Introduction to Computer Science I”. These courses were chosen because course materials stated that the courses were introductory and that they did not have any prerequisites. Presumably anyone who is proficient in the English language would be able to learn from, if not complete, these courses.

In order to answer the first research question: “Can unsupervised learning be used to identify alternate outcome groups, based on limited class interaction variables?”, I utilized k-means clustering to perform cluster analyses. In order to employ k-means clustering, k, the number of clusters, is an input variable and needs to be determined prior to running the analyses. Given k, k-means groups the dataset into the k clusters that minimize the sum of the squared distance of each point in the cluster to its center.

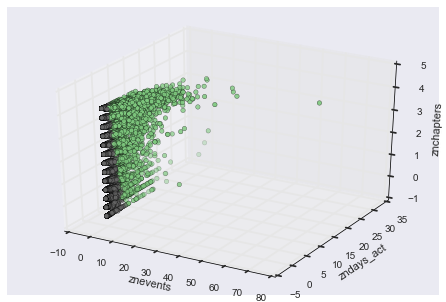
The k-means clustering was chosen in order to detect any patterns in course participation that might not align with the pre-defined outcomes discussed above. While the course interaction variables are fairly limited, they do provide some insight as to the students’ exposure with the course materials. The course interaction variables are presented below:

|  |  |
| --- | --- |
| Course interaction variables | |
| nevents | number of interactions with the course |
| ndays\_act | number of unique days student interacted with course |
| nplay\_video | number of play video events within the course |
| nchapters | number of chapters (within the Courseware) with which the student interacted |
| nforum\_posts | number of forum posts to the Discussion Forum |
| grade | final grade in the course, ranges from 0 to 1 |

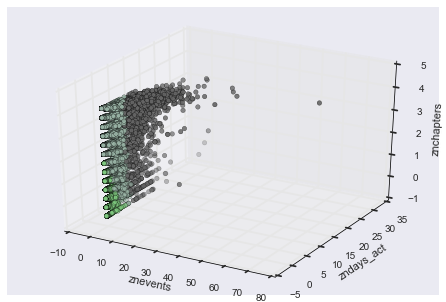
The goal of the clustering analyses was to identify different ways students interacted with the course materials. For the clustering analyses, I used: nevents, ndays\_act, nplay\_video, nchapters and nforum\_posts. Unfortunately, too few students earned grades in any of the classes to make this variable useful for these analyses.

As discussed above, for k-means clustering, one has to pre-determine the number of clusters. Therefore, I ran the model with different numbers of clusters for each course, visualized the results and ran additional tests. For the cluster visualizations, three of the input variables were plotted and then color-coded according to the k-means cluster each observation was assigned to. The input variables were also converted to z-scores for scaling purposes. Below are the results for CS50:

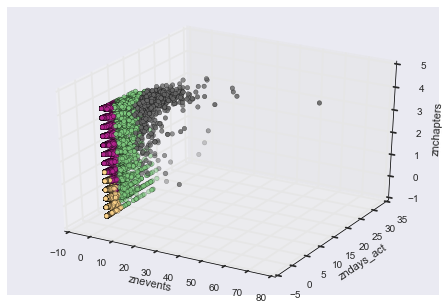
**CS50 K-means clustering with two clusters:**



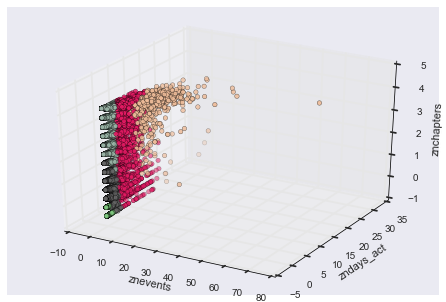
**CS50 K-means clustering with three clusters:**



**CS50 k-means clustering with four clusters:**

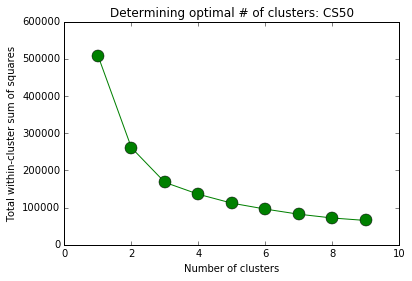


**CS50 k-means clustering with five clusters:**



While it would be hard to definitively choose the appropriate number of clusters using these visualizations, it is apparent that the clusters are fairly distinct (i.e. they do not overlap). Moreover, it is also significant that the move from four clusters to five results in very little change in the groupings; that leads me to conclude that the number of clusters should be less than five.

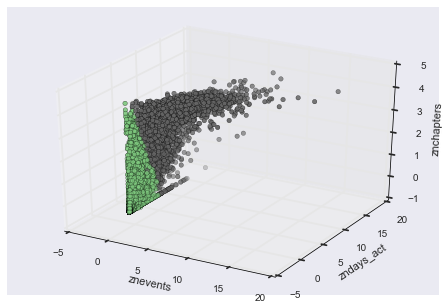
In order to obtain a more definitive answer, the elbow method was also used to identify the clusters. For this test, the “Total within-cluster sum of squares” was plotted against the number of clusters. By definition, the total within-cluster sum of squares will always decrease with the number of clusters; with this plot, one can hopefully identify the number of clusters at which the **marginal** decrease in the total within cluster sum of squares plateaus. This point is the optimal number of clusters.



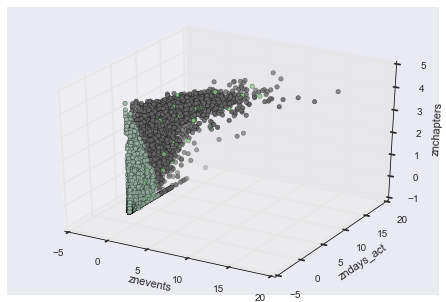
Though it is not clearcut (the line is fairly smooth), I conclude that the optimal number of clusters falls at 3 for the CS50 class.

The same visualizations and tests are displayed below for the Justice class.

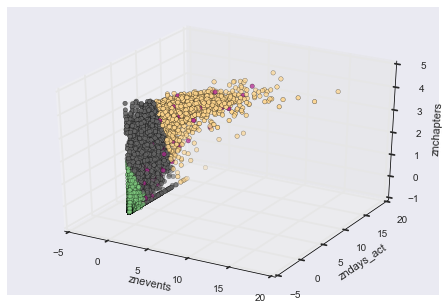
**Justice k-means clustering with two clusters:**



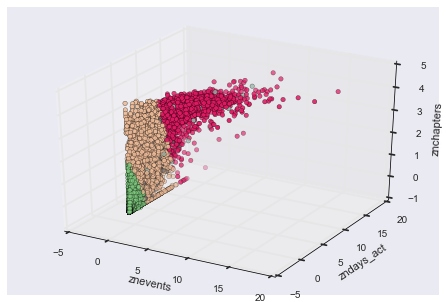
**Justice k-means clustering with three clusters:**



**Justice k-means clustering with four clusters:**



**Justice k-means clustering with five clusters:**



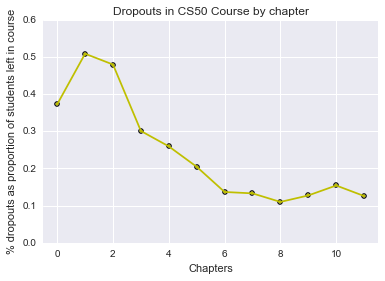
The conclusions that can be drawn from the Justice clustering visualizations are more clearcut: after two clusters, the groups start overlapping, and the additional groups, seen in the plots as separate dots, are interspersed among the other clusters. This leads me to conclude that the Justice class data should be clustered into two groups.

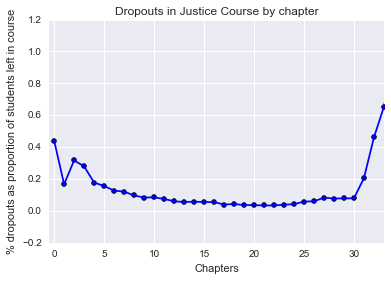
What are the distinct groups in the CS50 and Justice datasets? The below tables display the mean nevents, ndays\_act, nchapters and nforum\_posts variables:

|  |  |  |  |
| --- | --- | --- | --- |
| Mean | CS50 Course | | |
| Group 1 | Group 2 | Group 3 |
| nevents | 3 | 67 | 350 |
| ndays\_act | 1 | 7 | 27 |
| nchapters (total 12 chapters in course) | 1 | 8 | 10 |

|  |  |  |
| --- | --- | --- |
| Mean | Justice Course |  |
| Group 1 | Group 2 |
| nevents | 51 | 1239 |
| ndays\_act | 2 | 26 |
| nchapters (total 32 chapters in course) | 2 | 27 |
| nforum\_posts | 0.03 | 0.25 |

And where are students leaving the course? Below displays course dropout rates by chapter:





**Prediction Analyses:**

K-nearest neighbor and logistic regression were employed to answer the second research question: “Can supervised learning be used to predict class outcomes?”. For the predictive analyses, the pre-defined outcomes will be used as the response variables.

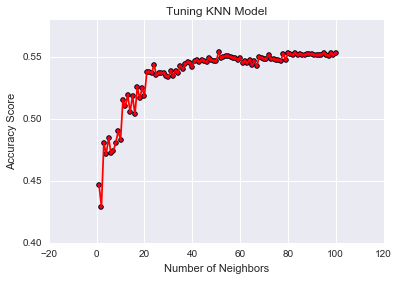
As discussed above, unfortunately, the dataset includes very courses measures of the students’ demographic backgrounds. Moreover, as shown below, very few students actually competed the course, so there is proportionally very little data to train the model.

|  |  |  |
| --- | --- | --- |
| CS50 Student Outcomes | | |
|
| registered only | 63,535 | 37% |
| viewed | 95,058 | 56% |
| explored | 9,741 | 6% |
| certified | 1,287 | 1% |
|  |  |  |
| Total | 169,621 |  |

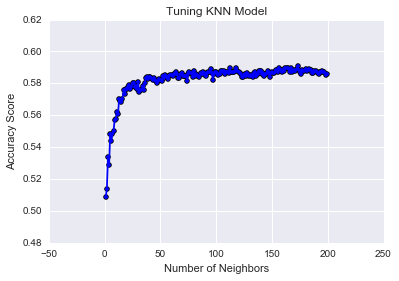
|  |  |  |
| --- | --- | --- |
| Justice Student Outcomes | | |
|
| registered only | 25,245 | 44% |
| viewed | 28,175 | 49% |
| explored | 1,640 | 3% |
| certified | 2,346 | 4% |
|  |  |  |
| Total | 57,406 |  |

Nonetheless, I ran the k-nearest neighbor model using (gender, age, country name, highest level of education completed) as predictors. As might be predicted, the results were disappointing, with an train-test split accuracy score of 0.50. Moreover, the model could not predict either the explored or certified outcomes and left those categories out of the predictions. Tuning the model resulted in a slight improvement in the result, as shown below:

**Tuning KNN Model: CS50**



**Tuning KNN Model: Justice**



These tests led me to conclude that 62 was the optimal number of neighbors for the CS50 model, and 88 was the optimal number of neighbors for the Justice model (holding the train-test random seed fixed; I have not yet played with the initial values to see if this makes a difference).

Resulting accuracy scores:

|  |  |  |
| --- | --- | --- |
|  | Accuracy Scores | # of neighbors |
| CS50 | 0.55 | 62 |
| Justice | 0.59 | 88 |

The logistic regression model fared better than the KNN model: resulting in an initial accuracy score of 0.56. Unfortunately, again, the model could not predict either the explored or certified outcomes and left those categories out of the predictions. Although there is not much one can do to vary the predictor variables, I tested three different combinations of the predictor variables to see if the train-test accuracy score would improve. (Note: I would have liked to have used cross-classification but could not find a module that could do it for the 4-category model.) The results did not vary at all for the CS50 model but varied slightly for the Justice model, as shown below:

|  |  |  |
| --- | --- | --- |
| Justice Course | Predictor Variables | Accuracy Score |
| Version 1 | gender, age, highest degree, all country dummies (excl. US) | 0.60 |
| Version 2 | gender, age, highest degree, US dummy, other country dummy | 0.59 |
| Version 3 | gender, age, highest degree, US dummy | 0.49 |