ECE 20875 Final Project
Name: Yi En Gan (Andrew)
Project Path 2
Date: 12/6/2019

Project Data

The project analyzes how well can the behavior of students when watching videos be used to predict the average performance. The file 'behavior-performance.txt' contains observed features and the quiz outcome tied to a student ID and video ID. The features considered are as follows.

fracSpent	Fraction of time watching video
fracComp	Fraction of video watched
fracPaused	Fraction of time pausing video
numPauses	Number of times paused
avgPBR	Average playback rate
numRWs	Number of rewinds
numFFs	Number of fast forwards

Number of Students Considered for Each Analysis

Total Number of Students	3976	
Clustering	1535	>=3 quizzes
Predicting Average Performance	94	>= half of quizzes
Predicting for Specific Video	3976	All student-vid pairs

Analyses Chosen for Component 1 : Multivariate k-means

The first component attempts to cluster students based on their behavior when watching the videos. The analysis chosen was multivariate k-means. This method was chosen because it is intuitive, easier to implement and can easily account for many dimensions (or features) It also allows us to assign a point to a cluster, and each point can only belong to one cluster. The distance of each point to its cluster can be used in a graph to show the relationship between number of clusters and how well the students are clustered and show the number of clusters that best represents the data.

Analyses Chosen for Component 2 : Ridge Regression for Whole Dataset

The second component trains the model using the data provided to predict the average performance of a student based on the student's behavior. The ridge model is used. The data is split 9:1 for training and testing the model. This method of analysis was chosen because it accounts for multiple variables when predicting the outcome. It also returns a continuous value instead of a discrete value, which is useful in determining the mean squared error and accuracy of the model. By comparing the predicted scores and actual scores, we know how good the model is at predicting a student's performance.

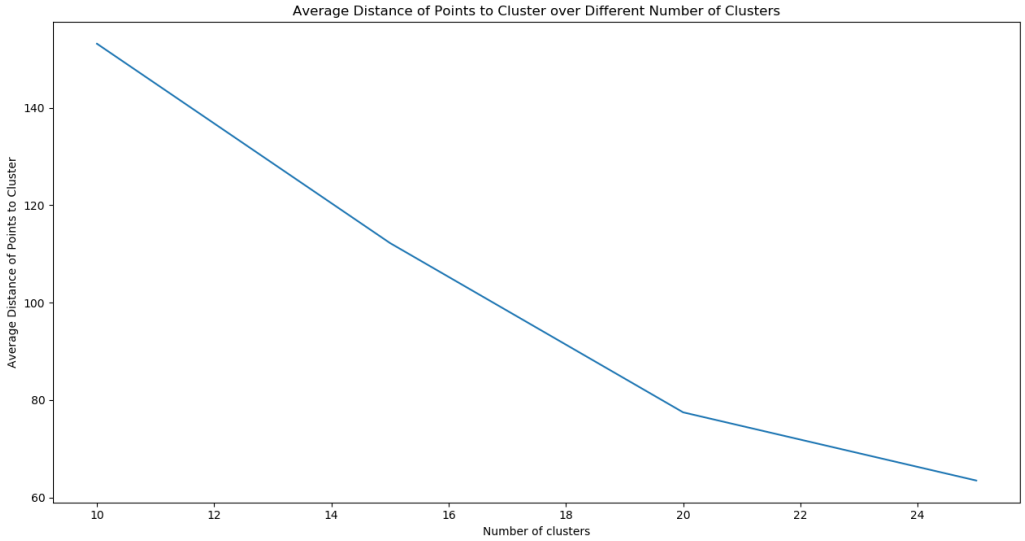
Analyses Chosen for Component 3 : Ridge Regression for Each Video

The last component uses ridge regression to make predictions. The method used for this component will be very similar for the second component, with the difference being that the data will be grouped and averaged by video ID instead of student ID. Instead of training one model for the entire dataset, there will be one model trained per video ID. This method of analysis will take the second component one step further and tell us how well behavior can be used to predict performance for different video types.

Collected Results

Output for Component 1:

How well can the students be clustered by their behavior?



The graph above shows the average distance of all points to its cluster. It is reasonable that as the number of clusters increase, the average distance of all points in one cluster to its center decreases.

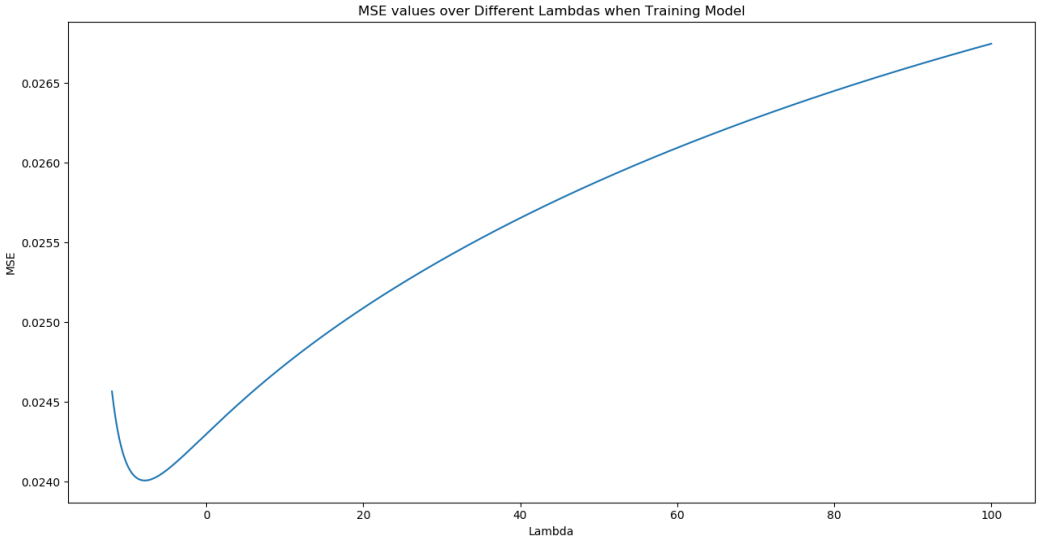
There is a change in gradient of average distance at numCluster = 20. This might be due to the points being over-clustered, i.e. a cluster being broken up into smaller clusters, instead of having points with similar features being in one cluster.

Therefore, the number of clusters that best represents the data is 20, with the average distance from points to its respective clusters being around 80.

The students are well clustered by their behavior. This allows instructors to predict student performance in clusters instead of on an individual basis.

Output for Component 2:

How well can you predict a student's average performance (i.e., average score s across all quizzes) in this course by the behaviors they exhibited while watching the course videos?



Predicting students by given behavioral features
Training Data: 96
Testing Data:10
Threshold Value: 0.2
Model Accuracy: 70.00 %
Model Coefficients:
[-0.0125 -0.0192 0.0143 0.0333 0.0218 0.1002 -0.1151]
MSE : 0.02
R^2 : 0.35
Best Lambda Value: -7.85

The 70% accuracy may be due to some factors not observed, or that the correlation between observed behavior and performance is not very strong. From this we know that the model was rather accurate in predicting the performance of student based on the behavior.

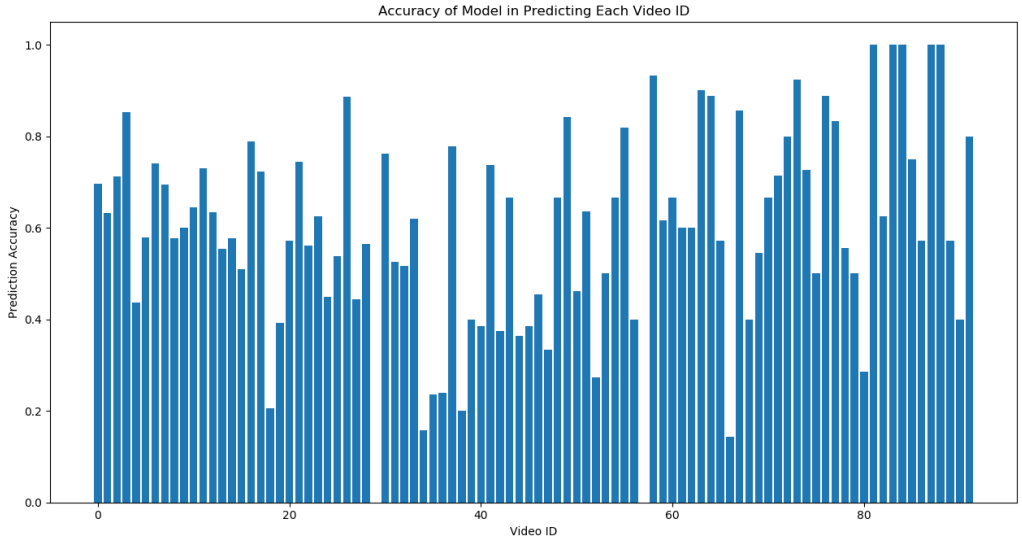
The best lambda value was a negative value. This encourages bias when training the model. An R-squared value of 35% indicates that the datapoints are slightly further away from the model regression line. This is because the data being analyzed is related to human behavior, which is less predictable.

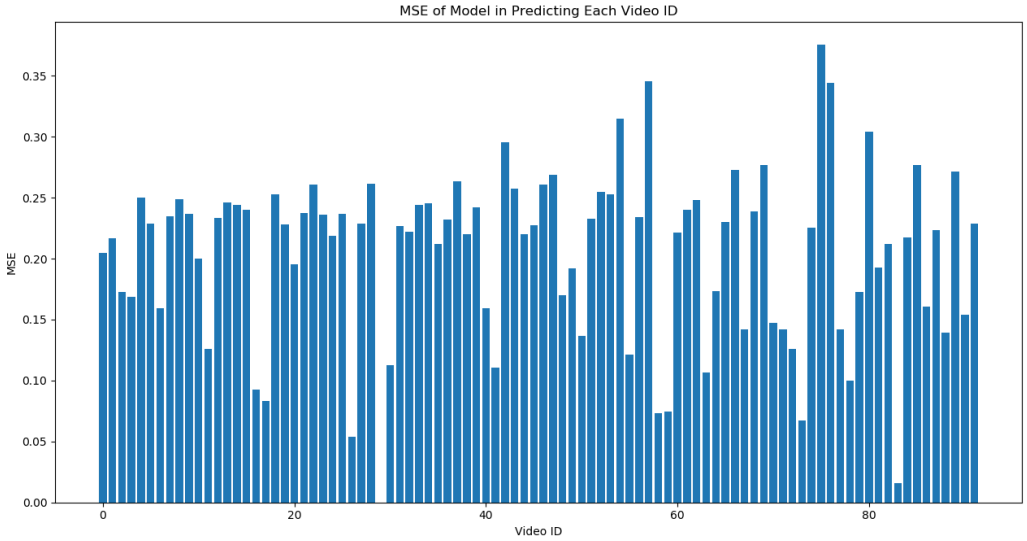
numRWs and numFFs seem to carry more weight than other features.

https://people.duke.edu/~rnau/rsquared.htm

Output for Component 3:

How well can you predict a student's performance on an in-video quiz question (i.e., whether they will be correct or incorrect) in this course by the behaviors they exhibited while watching the corresponding video?





The graph above shows how accurate can the watching behavior be used to predict the performance of students for each video. Videos with higher accuracy tend to have smaller mean squared error.

We can conclude that videos [3, 26, 49, 55, 58, 63, 64, 67, 72, 73, 76, 77, 81, 83, 84, 87, 88] have an accuracy rate above 80%, which means it is easier to predict student performance for these videos.

For videos [0, 1, 2, 6, 7, 10, 11, 12, 16, 17, 21, 23, 30, 33, 37, 41, 43, 48, 51, 54, 59, 60, 61, 62, 70, 71, 74, 82, 85, 91, 92], their accuracy rates lie within the 60~80% rate, which means they are fairly predictable.

Some videos are very difficult to predict through behavior alone. This may be because certain videos require background knowledge on the topic discussed to perform better in the quiz (video does not contain all information needed to do well on the quiz) or that certain videos have hidden information that require students to pay extra attention to capture the information necessary to do well in the quizzes.

Project Conclusion

The students can be clustered well based on the exhibited behavior, with optimal results when number of clusters is at 20, which provides an average distance of 80 between points and their respective clusters.

It is also proven that student behavior can be used to predict the average performance, with an accuracy of 70% with the given data and trained model.

However, when it comes to predicting student performance for each video, certain videos are more predictable than others. Therefore, should the instructor wish to assign quiz scores based on behavior alone, only videos with a model accuracy rate of >80% should be considered.

Additional Output

Terminal output is recorded in ‘output.txt’.