ECE20875: Python for Data Science Spring 2022

1. Project team information

   Name: Yi-Hsiang Chang

   Github: SeanoChang

   Email: chang755@purdue.edu

   Name: Li Chung Yang

   Github:LiChungYang

   Email: yang2010@purdue.edu

2. Descriptive Statistics

   The dataset we are working with is for the behavior-performance. In this text file, it shows ten different fields of behaviors that may affect students' performance. In question one, we will deal with all the students that complete at least five of the videos. For question two, we will deal with the data with the average score s across all quizzes and the two selected behaviors fracComp, numRWs. For question three, we will further use the data we analyzed in question two but also consider the userID and videoID.

3. Approach

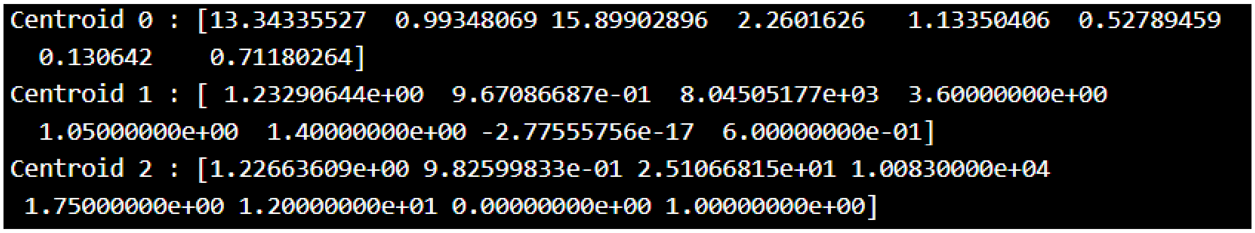
   For question one, we need to understand how well can the students be naturally grouped or clustered by their video-watching behavior. To find out the appropriate data that we want, we first got students who fulfilled the criteria that have video completion rate >= 0.9 with at least 5 videos watched. Then, we grouped and clustered them by their behavior using the method of k means and Gaussian Mixture Models. In order to search the center point of the different clusters for k mean, we used the silhouette score and find out the number of centroids from different clusters. The difference between the method of Gaussian Mixture Models and k means is that k means consider the mean to find the centroid of different clusters whereas Gaussian Mixture Models consider both mean and variance. Since there are many different datasets in this question, we expect we will find many centroids if we use the method of Gaussian Mixture Models. Also, since the raw dataset has been grouped by different behavior, we expect the number in each behavior will not be very different. So, we think we will not need to consider the variance. As a result, we think we expect to use the method of k means.

For question two, we expected the answer for this question is yes. We think the student's video-watching behavior can be used to predict a student's performance. After our consideration, we think the fields of fracComp and numRWs are the two behaviors that can best predict a student's performance.For fracComp, if the fraction of the video a student watches more, that means the student learns about the content more. For numRWs, the number of times a student skipped backward shows the student wants to understand the content of the video more. The other reason why we chose these two behaviors for analysis is that there are fewer outliers in these two fields. So, the data we got from these two fields will be more useful. As the exmaple of fracSpent, we can find the number of 69.5978688827 which means that the student spent around 69 times of the video time length on that particular video. This indicates that we can not predict the larger the number of fracSpent we get the better the performance of that student on that particular video will be. After selecting the behaviors, we used ridge regression to predict the relationships between the selected behaviors and s. The reason we use the method of ridge regression instead of polyfit is that we can not predict whether the data is linear or not. By using ridge regression, we can get a more accurate result. We also wrote a function which ask the user to enter the number of the students you want to predict. This function will randomly select the user input numbers of students from the dataset and perform prediction of how accuracy of our method.

For question three, we will use the two behaviors we chose in question two to predict a student's performance on a particular in-video quiz question. we need to find the nearest center average score s. To do this, we need to calculate the distance from the data to all the centroids by using distance to centroid of zeros divide by the distance to centroid of ones plus the distance to centroid of zeros and add all the distances and return the list. After that, we predicted the score for a student based on the kmeans model and ridge regression model. We use the fractional distance between each data and all the centroids to perform prediction. Then, use the distances from each centroid to the data to predict the score. We then find out all the predictions for the student and then find the average correctness or accuracy of the predictions for both kmean and ridge method. Finally, we will return the accuracy of the prediction. Just like question two, we will ask the user to enter the number of the students you want to predict and show the predicted students’ performance in different videos.

4. Analysis

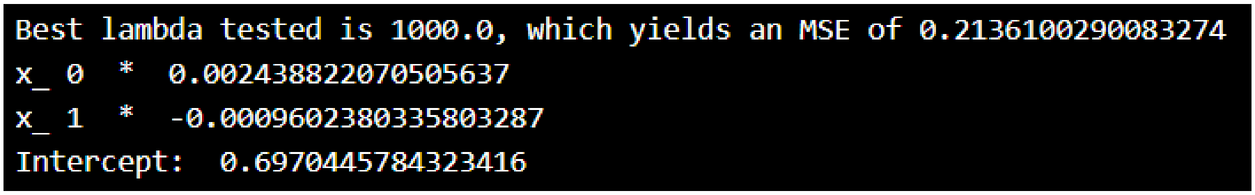
To answer question one, we grouped and clustered the data by their behavior using the method of k means. In order to find the center point of different clusters, we used the silhouette score and find out 3 center point from three clusters as figure 1 showed. However, when we use the methods of Gaussian Mixture Models, we found out the best number of clusters is 10. The way of Gaussian Mixture Models is a more accurate method. However, since the best number of clusters we found for Gaussian Mixture Models is a little too large, we decided to use k mean to clustered the video as we expected.



```
Centroid 0 : [13.34335527  0.99348069 15.89902896  2.2601626   1.13350406  0.52789459
  0.130642    0.71180264]
Centroid 1 : [ 1.23290644e+00  9.67086687e-01  8.04505177e+03  3.60000000e+00
  1.05000000e+00  1.40000000e+00 -2.77555756e-17  6.00000000e-01]
Centroid 2 : [1.22663609e+00 9.82599833e-01 2.51066815e+01 1.00830000e+04
 1.75000000e+00 1.20000000e+01 0.00000000e+00 1.00000000e+00]
```

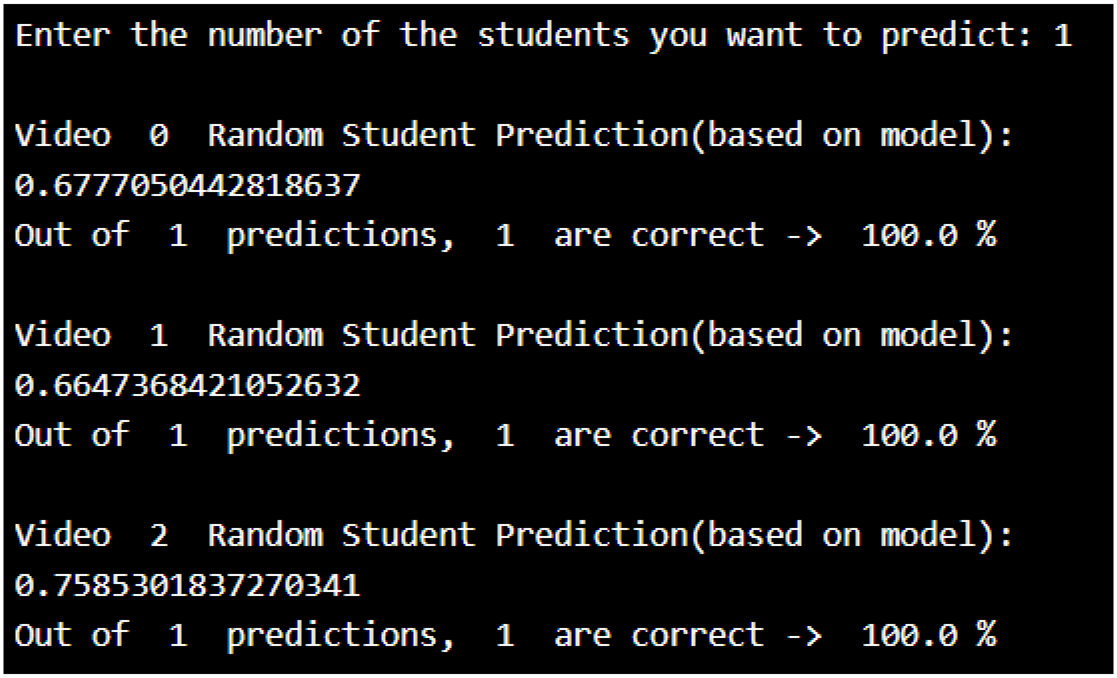
Figure 1: The centroids found using k mean

As discussed in the approach section, student's video-watching behavior can be used to predict a student's performance. As figure 2 shows, the values of X_ 0 is the parameter for fracComp, the values of X_ 1 is the parameter for numRWs, and the values of Intercept is the constant of the model function. As shown in figure 3, If we enter the number of the students you want to predict as 1, the prediction will be 100% for every video. This is because we only have one data to act as trained and tested data. If we use a number other than 1 the values of percentage will obviously be between 0 and 100.



```
Best lambda tested is 1000.0, which yields an MSE of 0.2136100290083274
x_ 0  *  0.002438822070505637
x_ 1  *  -0.0009602380335803287
Intercept:  0.6970445784323416
```

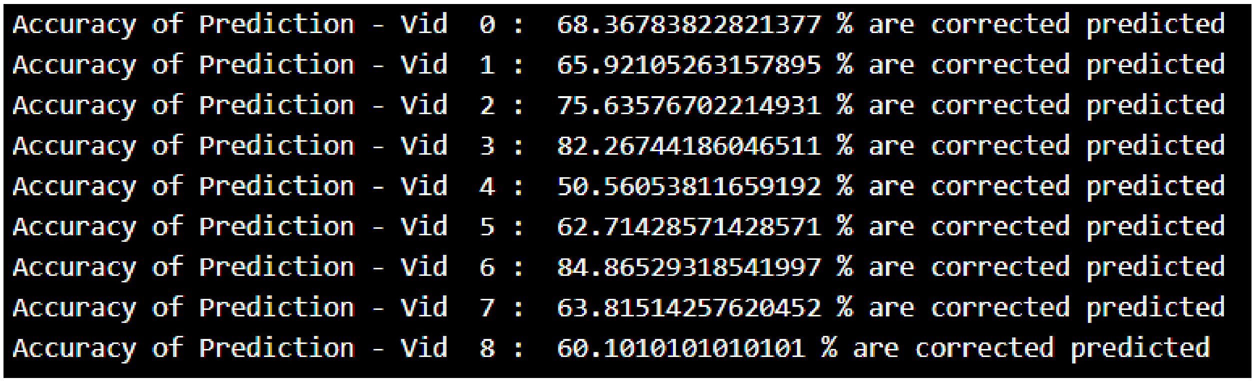
Figure 2: The parameter for the ridge regression model



```
Enter the number of the students you want to predict: 1

Video  0  Random Student Prediction(based on model):
0.6777050442818637
Out of  1  predictions,  1  are correct ->  100.0 %

Video  1  Random Student Prediction(based on model):
0.6647368421052632
Out of  1  predictions,  1  are correct ->  100.0 %

Video  2  Random Student Prediction(based on model):
0.7585301837270341
Out of  1  predictions,  1  are correct ->  100.0 %
```

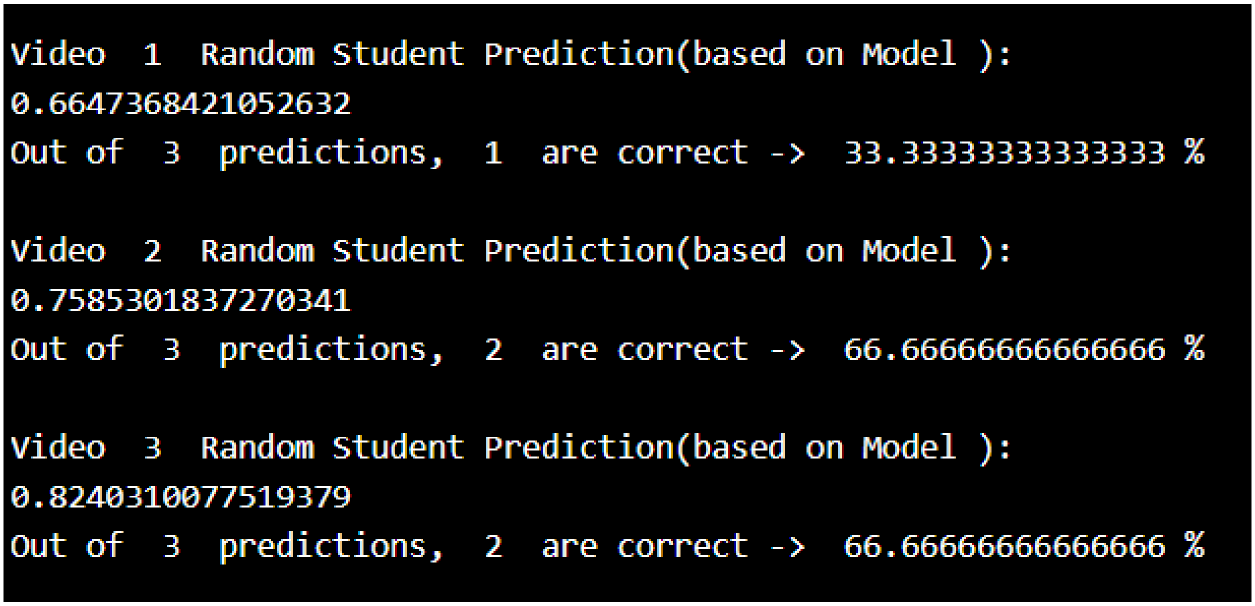
Figure 3: Random student prediction

To answer question 3, we can use the two behaviors (fracComp and numRWs) to predict a student's performance on a particular in-video quiz question. Figure 4 shows the accuracy pf the prediction in different video when consier all student-video pairs in our analysis. As shown in figure 5, if we enter the number of the students you want to predict as 3, we can predict students' performance on a particular in-video quiz question.



```
Accuracy of Prediction - Vid  0 :  68.36783822821377 % are corrected predicted
Accuracy of Prediction - Vid  1 :  65.92105263157895 % are corrected predicted
Accuracy of Prediction - Vid  2 :  75.63576702214931 % are corrected predicted
Accuracy of Prediction - Vid  3 :  82.26744186046511 % are corrected predicted
Accuracy of Prediction - Vid  4 :  50.56053811659192 % are corrected predicted
Accuracy of Prediction - Vid  5 :  62.71428571428571 % are corrected predicted
Accuracy of Prediction - Vid  6 :  84.86529318541997 % are corrected predicted
Accuracy of Prediction - Vid  7 :  63.81514257620452 % are corrected predicted
Accuracy of Prediction - Vid  8 :  60.1010101010101 % are corrected predicted
```

Figure 4: Accuracy of the prediction in different videos.



```
Video  1  Random Student Prediction(based on Model ):
0.6647368421052632
Out of  3  predictions,  1  are correct ->  33.33333333333333 %

Video  2  Random Student Prediction(based on Model ):
0.7585301837270341
Out of  3  predictions,  2  are correct ->  66.66666666666666 %

Video  3  Random Student Prediction(based on Model ):
0.8240310077519379
Out of  3  predictions,  2  are correct ->  66.66666666666666 %
```

Figure 5: shows the percentage of the accuracy if there are 3 students