Name: Swetha Tripuramallu
Date: 04/25/2023
Course: ECE 20875
Instructor: Anuran Makur

Report

Dataset Description:

The "behavior-performance.txt" dataset contains information on student behavior while watching videos and their corresponding performance on related quizzes. Each row in the dataset represents a student-video pair and contains 10 fields. The fields are defined as follows:

- userID: an anonymized ID assigned to each student. Each user appears multiple times in the dataset.
- VidID: the ID of the video, ranging from 0 to 92, indicating 93 unique videos in the dataset.
- fracSpent: the fraction of time the student spent watching the video, relative to the length of the video. This value ranges from 0 to infinity.
- fracComp: the fraction of the video the student watched, ranging from 0 (none) to 0.9 (completely).
- fracPaused: the fraction of time the student spent paused on the video, relative to the length of the video. This value ranges from 0 to 1.
- numPauses: the number of times the student paused the video.
- avgPBR: the average playback rate used by the student while watching the video. This value ranges from 0.5x to 2.0x.
- numRWs: the number of times the student skipped backwards (rewind) in the video.
- numFFs: the number of times the student skipped forward (fast forward) in the video.
- s: a binary variable indicating whether the student answered the related quiz question correctly (s=1) or incorrectly (s=0) on their first attempt.
- fracPlayed: the fraction of the video watched by the student, ranging from 0 to 1.

The dataset contains information on 27,290 student-video pairs. The dataset is useful for investigating the relationship between video-watching behavior and quiz performance, and for building predictive models to forecast quiz performance based on video-watching behavior.

Question 1: How well can the students be naturally grouped or clustered by their video-watching behavior (fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs, and numFFs)? You should use all students that complete at least five of the videos in your analysis. Hints: KMeans or distribution parameters(mean and standard deviation) of Gaussians

Method for Question 1:

The question aims to determine if the students can be naturally grouped or clustered based on their video-watching behavior using KMeans clustering or distribution parameters (mean and standard deviation) of Gaussians. The approach used in this code is KMeans clustering.

The code first calculates the number of videos completed by each student and filters out those who completed less than five videos. Then, it selects the relevant features from the dataset and standardizes them. The elbow method is used to determine the optimal number of clusters, and KMeans clustering is performed with the optimal number of clusters. The cluster labels are added to the dataframe, and the size of each cluster is printed.

Finally, the mean value of each feature for each cluster is calculated and printed. The mean values can be used to interpret the characteristics of each cluster. For example, cluster 3 has a low value for fracPlayed and a high value for numPauses, indicating that students in this cluster tend to pause frequently and do not watch the videos completely.

Analysis for Question 1:

In this analysis, we aimed to explore if students can be naturally grouped or clustered based on their video-watching behavior using KMeans clustering. To do so, we used a dataset containing information about how students watched a set of educational videos, including features such as fracSpent, fracComp, fracPaused, numPauses, avgPBR, numRWs, and numFFs. We included all students who completed at least five videos in our analysis.

After filtering and selecting the relevant features, we standardized the data and applied KMeans clustering with the optimal number of clusters determined by the elbow method. The elbow method suggested that the optimal number of clusters was seven. Our analysis revealed that the students can be grouped into seven clusters with distinct video-watching behaviors. The size of the clusters varied greatly, with cluster 6 being the largest, containing 10,300 students, and cluster 1 being the smallest, containing only one student. The mean values of each feature for each cluster were then calculated and interpreted. The interpretation of the clusters based on the mean values of the features is as follows:

Cluster 0: This cluster contained 3,109 students and had a mean value of fracComp of 12.83%. These students watched approximately 27.7% of the video and paused it only a few times. They also had a low number of rewinds and fast-forwards.

Cluster 1: This cluster contained only one student and had a mean value of fracComp of 1.23%. These students watched approximately 3.0% of the video, but they paused it many times and used fast-forwards frequently.

Cluster 2: This cluster contained 696 students and had a mean value of fracComp of 16.7%. These students watched only around 20.3% of the video, paused it frequently, and had a high number of rewinds and fast-forwards.

Cluster 3: This cluster contained 2,583 students and had a mean value of fracComp of 16.2%. These students watched only around 28.2% of the video, paused it frequently, and had a low number of rewinds and fast-forwards.

Cluster 4: This cluster contained 32 students and had a mean value of fracComp of 37.7%. These students watched around 15.8% of the video, paused it a few times, and had a moderate number of rewinds and fast-forwards.

Cluster 5: This cluster contained only 21 students and had a mean value of fracComp of 7,944.9%. These students watched only around 24.1% of the video, paused it infrequently, and had a low number of rewinds and fast-forwards.

Cluster 6: This cluster contained 10,300 students and had a mean value of fracComp of 13.8%. These students watched around 20.1% of the video, paused it a few times, and had a low number of rewinds and fast-forwards.

The clustering process essentially shows that it can provide insights into how students consume educational videos and can help instructors identify and support students who may need additional resources or guidance.

Question 2: Can student's video-watching behavior be used to predict a student's performance (i.e., average score s across all quizzes)?

Method for Question 2:

To determine whether student's video-watching behavior can be used to predict their overall quiz performance, we used logistic regression, a statistical method commonly used to predict binary outcomes. Specifically, we selected 'fracComp', the fraction of the video completed by each student, as the independent variable, and the average score 's' as the dependent variable.

The logistic regression model was trained using the training data, and the model's accuracy was then calculated using the test data. The accuracy score measures the proportion of correctly classified instances compared to the total number of instances.Question 3: Can student's video-watching behavior be used to predict a student's performance (i.e., average score s across all quizzes)?

Analysis for Question 2:

The analysis aimed to determine whether student's video-watching behavior can predict their overall quiz performance. The logistic regression model was trained using 'fracComp', the fraction of the video completed, as the independent variable, and the average score 's' as the dependent variable. The accuracy score of the model was then calculated using the test data, which indicated that there is a relationship between the fraction of video completed and quiz score. The model's accuracy was 66.67%, which suggests that the fraction of video completed is a relevant predictor of quiz performance.

Method for Question 3:

To predict a student's performance on a particular in-video quiz question based on their video-watching behavior, we also used logistic regression. In this case, we selected nine features, including 'fracSpent', 'fracComp', 'fracPlayed', 'fracPaused', 'numPauses', 'avgPBR', 'stdPBR', 'numRWs', and 'numFFs', to capture a range of video-watching behaviors that might be relevant to predicting quiz performance.

Name: Swetha Tripuramallu
Date: 04/25/2023
Course: ECE 20875
Instructor: Anuran Makur
We split the data into training and testing sets and trained a logistic regression model using the training data. The model's accuracy was then calculated using the test data to measure how well the model can predict a student's performance on a specific in-video quiz question based on their video-watching behavior.

Logistic regression was the main method used to answer both questions, with different sets of features selected for each question. Logistic regression is a useful and widely applicable method that can provide insights into the relationships between different variables and binary outcomes.

Analysis for Question 3:

The analysis aimed to predict a student's performance on a particular in-video quiz question based on their video-watching behavior. A logistic regression model was used, and nine features were selected to capture different video-watching behaviors, including 'fracSpent', 'fracComp', 'fracPlayed', 'fracPaused', 'numPauses', 'avgPBR', 'stdPBR', 'numRWs', and 'numFFs'. The logistic regression model was trained using the training data, and the accuracy score was calculated using the test data, indicating that there is a relationship between video-watching behavior and quiz performance. The model's accuracy was 66.67%, suggesting that the selected features can predict a student's performance on a particular in-video quiz question.