K-MEANS CLUSTERING AND DIMENSIONALITY REDUCTION USING PRINCIPAL COMPONENT ANALYSIS

ABSTRACT

PURPOSE: The purpose of this experiment is to explore principal component analysis by finding pertinent relationships between data variables to describe observations and to cluster our data.

METHODS: Methods included three separate tests. One to find the two variables that provide the best clustering of our data, another to show the clustering after dimensionality reduction, and a final test to show the clustering after dimensionality reduction and standardization of the data. Clustering is found using a *k*-means algorithm and assessed by the Davies-Bouldin Index.

RESULTS: Results show a table of the Davies-Bouldin Indices from each test along with the variables that best describe the data from test one. Three figures are displayed with the clustering from each test.

CONCLUSIONS: The first test shows that the observations can be described by two components. Tests two and three show that the standardization and dimensionality reduction of the data produces much more accurate results.

INTRODUCTION

The objective in this assignment is to find a correlation between variables and assign clusters to groups of data of similar properties.

Principal components analysis (PCA) is a method used to establish the difference from the mean of variables in a data set. To implement PCA, we need to find the zero-mean matrix of the data. Often this requires us to first transpose the data so that the columns represent the values of a single observation. The zero-mean matrix M from a data matrix A can be represented as $M = A - \vec{1}A$. From the zero-mean matrix, the principal components of the matrix are found by computing the singular-value decomposition (SVD) of the zero-mean matrix. In this data set, we are trying to reduce the dimensionality to two components, or two dimensions. To do this, each observation is scored by multiplying the i^{th} observation of the zero-mean matrix by the corresponding right singular vector from the SVD. The resulting matrix is of the form $Z = MV$ and contains the scores of the principal components to the desired dimension. In our case there will be two scores. Once the scores of the data are found, the clusters are visualized by a scatter plot with the horizontal and vertical axes representing the scores of the first and second principal components, respectively. Dimensionality reduction can show how many components are needed to cluster data and allows a *k*-means algorithm to more accurately describe this clustering. The *k*-means algorithm is an unsupervised machine learning method that aims to find centroids that partition the data into subsets called clusters. The clustering algorithm is assessed by using a Davies-Bouldin index. Standardization allows different variables to be measured on the same scale by reducing them to be on the same scale. We can calculate the PCA and cluster the data from standardized data in the same way it is done with unstandardized data.

A testable hypothesis for this experiment is to cluster data in three different ways using the *k*-means algorithm. One method is by using the two variables that provide the best clustering, another by PCA using SVD, and finally by PCA using SVD from standardized data.

METHODS

To implement principal component analysis and dimensionality reduction, we first need to load the data set and remove the first column that contains the variable names of the chemicals that were measured in the experiment. We then transpose the data to have the variables represented as columns instead of rows. From there, we can set yvec to be the first column of the data and remove that column.

The next task is to find the pair of data that provides the best clustering. For this, we need to find the Davies-Bouldin Index for every two-column sub matrix possible in the data. First, we need to initialize two containers, one that represents the DB scores, and the other representing the variables that correspond to each score. Next, we can create a nested for loop where the first index iterates through the columns of the matrix, and the second index iterates through the columns that follow the first index. By this method we can generate all the possible 2 column sub matrices without repetition. With each iteration, we can create an Xmat that represents the two columns. Then the DB scores container is updated by passing Xmat and yvec into the dbindex function, which will loop over the indexes and find the db score of each cluster. The score is then calculated as the mean of the scores of the clusters. The DB indexes container is then updated by adding the indexes of the variables in which the scores were just calculated for as a row to the container. After this nested loop iteration the lowest DBI score can be found with the corresponding variable indices. Now that the two best cluster variables are found we can plot the yvec along with these variables using the gscatter function to visualize the clusters.

Our next task is to reduce the data from 13 dimensions down to two dimensions. We achieve this by using PCA and score reduced clustering. To calculate the SVD of the data matrix, we need to first find the zero-mean matrix. This is found by taking the data matrix and subtracting the column-wise invocation of the mean on the data times the ones vector of the same length. From there, we can find the singular value decomposition of the matrix by using the svd function in MATLAB. We then take the zero mean matrix and multiply by the two PCA loading vectors, which are the first two right singular vectors of the zero matrix. From there, we can then find the DBI using this two-column matrix with yvec and plot the clustering to visualize as it was done before.

Our third task is to standardize the data and then produce the clustering that the standardization produces. This was a relatively straightforward calculation, we simply needed to standardize the data using the zscore function in MATLAB. The remaining calculations are the same as before, where we reduced the data using SVD and PCA, starting by calculating the zero-mean matrix. These results are then presented visually as clusters once again using the gscatter function.

RESULTS

Table 1. Shows the test used with the corresponding results from the implementation of the test. The data columns test finds the two best clustering vectors, with the Davies-Bouldin Index along with the index of the variable columns. Index 1 belongs to the Ethanol column and index 7 belongs to the Flavonoids column. The raw PCA Scores row shows the DB index of the clustered reduced data to 2D. The standardized PCA column shows the DB Index of the clustered standardized data using PCA.

TEST	DB INDEX	VARIABLES
Data Columns	0.7875	[1 7]
Raw PCA Scores	1.5148	
Standardized PCA	0.6392	

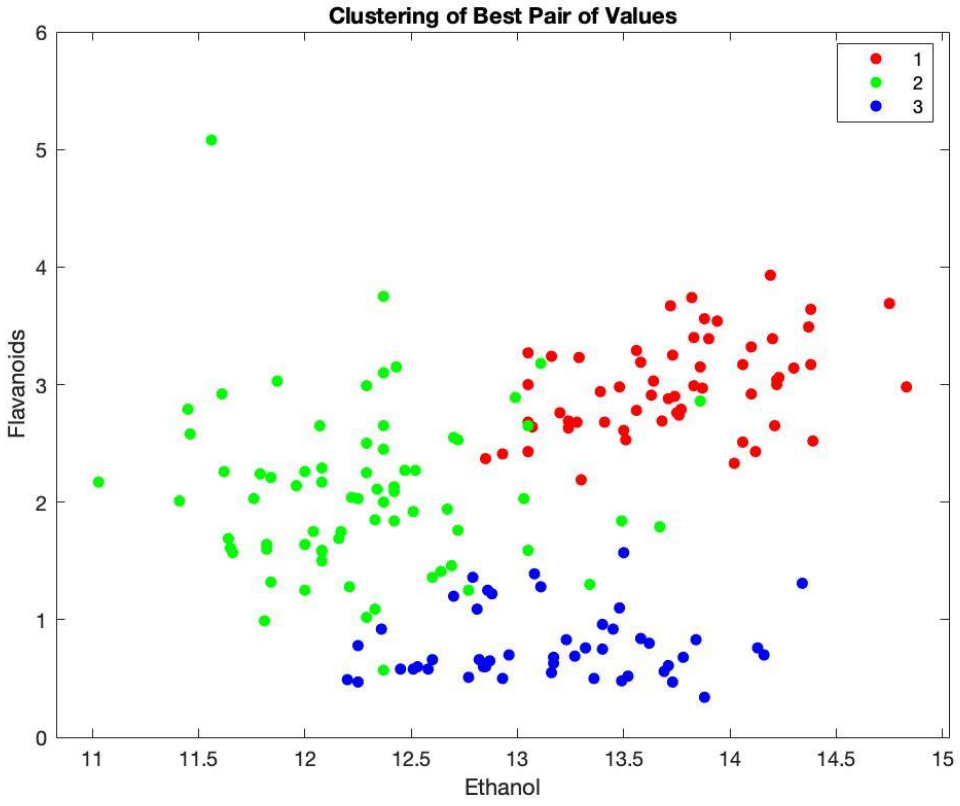

Figure 1. Shows a visualization of clustering of the best columns for clustering, which was determined by using the lowest DB Index from all of the possible column combinations. The x-axis shows ethanol levels and the y-axis shows falvanoid levels.

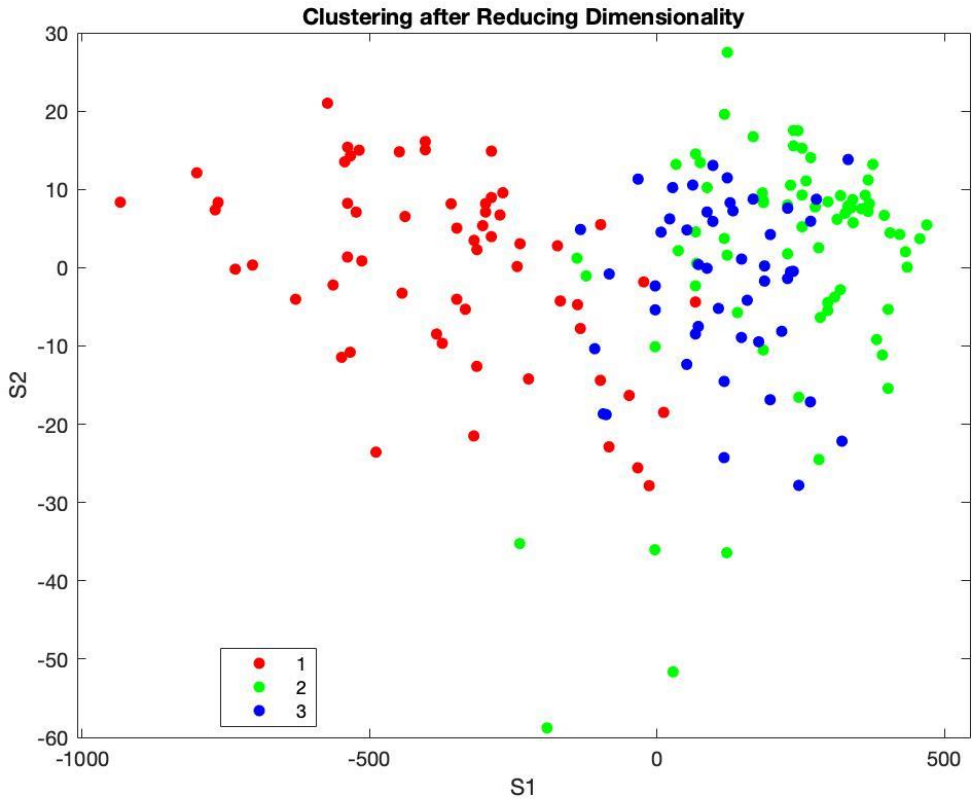

Figure 2. Shows a visualization of clustering after reducing the dimensionality of the data from 13D to 2D. S1 and S2 represent the first two right singular vectors from the SVD of the zero-mean matrix.

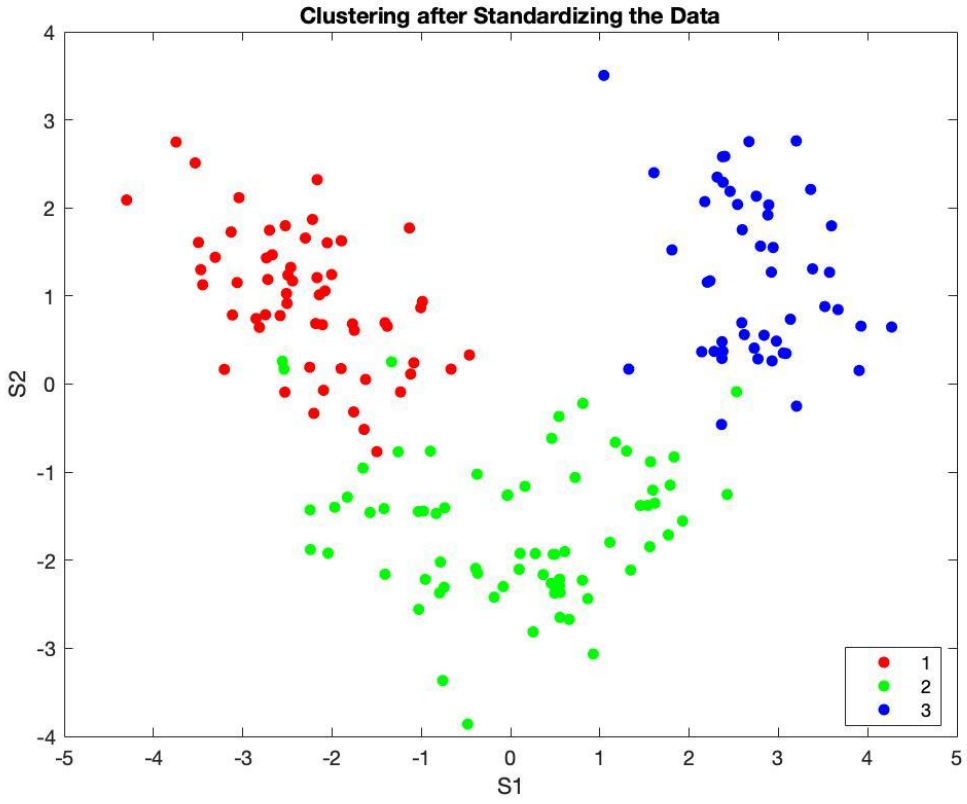

Figure 3. Shows a visualization of clustering after standardization of the data and reducing the dimensionality of the data from 13D to 2D. S1 and S2 represent the first two right singular vectors from the SVD of the zero-mean matrix.

DISCUSSION

In the implementation of the hypothesis, we can see that three different and varying results are produced by the three methods of clustering described earlier. Table 1 shows a summary of the methods used to cluster data with the corresponding DB index that resulted from *k*-means algorithm.

In our first test, we can see that our algorithm produced a DBI of 0.7875, with the variables of index 1 and 7 that correspond to this DBI. We can observe that the DBI is fairly low from this test, which is to be expected, since we are calculating the DBI for every possible combination of variables in the dataset. Since the indexes 1 and 7 correspond to Ethanol and Flavanoids, we can say that the observations can be most accurately described by these two specific variables in comparison to the others. By observations, we are referring to the cultivars–or grape types–in the wine. With this same test, looking at Figure 1 shows the clustering from the k-means algorithm. By observation, we can see that the k-means algorithm was able to predict the clustering of the data with moderate accuracy, with the exception of overlapping between some of the clusters.

In our second test, we aimed to first reduce the data from 13 dimensions down to two dimensions and then produce the clustering of the reduced data. In Table 1, it shows that by PCA as described in the methods section, the DBI from the k-means clustering algorithm is 1.5148. This is relatively inaccurate compared to the other algorithms that we produced. This can be seen visually in Figure 2, where we can see that there is much overlap between the clusters.

Finally, we performed a third test, which included a standardization of the data before the dimensionality reduction. This test produced a DBI of 0.6392, which is the lowest of the tests. By looking at Figure 2, it is quite clear that the k-means clustering algorithm was able to cluster the data to a fairly high accuracy. Further, in comparing tests one and three, we can see that dimensionality reduction can improve the accuracy of the k-means algorithm, meaning that only two principal components are needed to describe the data.

In comparing tests two and three, we can see that standardization of the data produced a much more accurate result than unstandardized data. This essentially reveals the importance of data standardization. If we have data variables with different variances, our principal component analysis will be inaccurate since PCA is looking to project to maximize the variance of the data variables. This means that variables with higher variance might mistakenly contribute more as a principal component and explain most of the variance in the data. If we standardize, then we will see that other variables can contribute as well.

PCA is a powerful tool that allows us to see many relationships in our data linearly in lower dimensions by dimensionality reduction, that otherwise would be impossible to see in higher dimensions. *k*-means clustering allows us to group data in an unsupervised fashion. There are many applications that PCA and *k*-means clustering can be useful for, such as image classification applied in Magnetic Resonance Imaging (MRI) to detect maladies like tumors.

In summary, our first test showed us that our data observations be accurately be described by only two variables out of the 13. Moreover, the second and third tests show the importance of data standardization to ensure the variance of a variable will not dominate the other components. We were able to observe relationships in the data and provide reasonably accurate clustering of the observations. To conclude, our testable hypothesis is reasonable to produce our objectives.

REFERENCES

Ellis, R. Class 17 Principal Components Analysis – PCA. Retrieved Mar 11, 2021, from https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class17.pdf

Ellis, R. Class 18 Classification – K-Means Clustering. Retrieved Mar 11, 2021, from https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class18.pdf

Ellis, R. Class 20 PCA Revisited – Scatter Matrix and Dimensionality Reduction. Retrieved Mar 11, 2021, https://onq.queensu.ca/content/enforced/505646-CISC271W21/notes/Class20.pdf

C. Saha and M. F. Hossain, "MRI brain tumor images classification using K-means clustering, NSCT and SVM," 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON), Mathura, 2017, pp. 329-333, doi: 10.1109/UPCON.2017.8251069.