ISE 5103 Intelligent Data Analytics
Homework #3

Instructor: Charles Nicholson

See course website for due date

Learning objective: Dimension Reduction

Submission notes:

1. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader *may* view your R code, but should never *have* to in order to find your solutions.
2. In the PDF, clearly identify each problem (e.g. Problem 1a, Problem 2b, etc.) Also, note that only *relevant* and informative computer output should be provided.
3. Make sure to *provide comments* on what your R code is doing. Keep it clean and clear!
4. You will submit your complete R script. Note: include `library` commands to load *all* packages that are used in the completion of the assignment. Place these statements at the top of your script.
5. Do not zip your files for submission. Submit exactly two files. Name the files "LastName-HW1" with the appropriate file extension (that is, .pdf for the write-up and .R for the script)

1 Glass data (60 points)

The study of classification of types of glass is motivated by criminological investigations. At the scene of a crime, the glass left can be used as evidence... if it is correctly identified.

The data set we consider consists of 213 unique glass samples labeled as one of six class categories[1]:

type	description
1	building windows float processed
2	building windows non-float processed
3	vehicle windows float processed
5	containers
6	tableware
7	headlamps

There are nine predictors, including the refractive index and percentages of the following eight elements found in the glass: Na (Sodium), Mg (Magnesium), Al (Aluminum), Si (Silicon), K (Potassium), Ca (Calcium), Ba (Barium), and Fe (Iron).

The data is available here: `http://archive.ics.uci.edu/ml/datasets/Glass+Identification` and is also available in the `mlbench` package as the dataset `Glass`.

[1] I do not know why they skipped class "4" in the data.

Note: There is one duplicate row in the mlbench data. Please find the duplicate row and remove it. See the R function `duplicated` to help you find it.

(a) (20 points) Mathematics of PCA

i. Create the correlation matrix of all the numerical attributes in the `Glass` data and store the results in a new object `corMat`.

ii. Compute the eigenvalues and eigenvectors of `corMat`.

iii. Use `prcomp` to compute the principal components of the `Glass` attributes (make sure to use the `scale` option).

iv. Compare the results from (ii) and (iii) – Are they the same? Different? Why?

v. Using R demonstrate that principal components 1 and 2 from (iii) are orthogonal. (Hint: the inner product between two vectors is useful in determining the angle between the two vectors)

(b) (20 points) Application of PCA

i. Provide visualizations of the principal component analysis results from the `Glass` data. Consider incorporating the glass type to group and color your biplot.

ii. Provide an interpretation of the first two prinicpal components the `Glass` data.

iii. Based on the the PCA results, do you believe that you can effectively reduce the dimension of the data? If so, to what degree? If not, why?

(c) (20 points) Application of LDA

i. Since the `Glass` data is grouped into various labeled glass types we can consider linear discriminant analysis (LDA) as another form of dimension reduction. Use the `lda` method from the `MASS` package to reduce the `Glass` data dimensionality.

ii. How would you interpret the first discriminant function, LD1?

iii. Use the `ldahist` function from the `MASS` package to visualize the results for LD1 and LD2. Comment on the results.

2 **Facebook metrics** (40 points)

Use the data in the file "FB-metrics.csv" for the following problem.

The data is related to Facebook posts published during 2014 on the Facebook page of a renowned cosmetics brand. This dataset contains 495 rows and 19 features: the first 7 are known prior to post publication (e.g., current total page likes, post month, etc.) and last 11 features used for evaluating post impact (e.g., total number of impressions, etc.).

Additionally, please see the paper "Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach" by Moro et al. (2016) for more information on the original study.

In this task, we want to explore the 11 evaluation features using PCA and t-SNE for dimension reduction. A description of the 11 features are found in Table 2 of the Moro et al. (2016) (Note: I removed "total interactions" as it is simply the sum of other features.)

(a) (20 points) Use PCA to analyze the 11 evaluation features. Provide visualizations, interpretations, and comments as appropriate.

(b) (20 points) Use t-SNE from the `Rtsne` package in R to explore 2 or 3-dimensional representations of the data. Can you find a visualization you find interesting?

For both parts above, consider using different colors to highlight factors associated with the 7 input features, e.g., paid vs. not-paid, category, type, month, etc. I am interested in what you can find. Do not be afraid to play with the data, slice and dice as you see fit.

3 Extra credit: Uniform Manifold Approximation and Projection (UMAP) (10 points)

Uniform Manifold Approximation and Projection (UMAP) is another algorithm for dimension reduction. It is a relatively new technique (it came out in 2018) and may outperform t-SNE. For extra-credit, try it out and let me know what you think. Choose either the Glass data or the Facebook metrics data and compare the results to what you already found. The R implementation is in the `umap` package.

You can read more about UMAP here: `https://pair-code.github.io/understanding-umap/`.