Dimensionality Reduction and Clustering Techniques

Dr. Matthieu cisel

September 2021

1 Goals

From the data mining point of view, learners delve on the concepts of dimensionality reduction and clustering algorithms, in either R or Python, with a focus on basic techniques – Principal Component Analysis, and k-means and hierarchical clustering. Students must submit both a report and a carefully written code notebook.

1.1 k-means and Hierarchical Clustering

The k-means algorithm is one common approach to clustering. Learn how the algorithm works under the hood, implement k-means clustering in R, visualize and interpret the results, and select the number of clusters when it's not known ahead of time. Hierarchical clustering is another popular method for clustering that we will explore on real-life datasets.

1.2 PCA and MCA

Principal Component Analysis (PCA) is a common approach to dimensionality reduction. Learn exactly what PCA does, visualize the results of PCA with biplots and scree plots, and deal with practical issues such as centering and scaling the data before performing PCA. We also focus on Multiple Correspondence Analysis (MCA) to learn how to deal with qualitative data.

2 Presentation of the dataset

You are provided with an artificial dataset containing data on user profiles, and data on conversations from an imaginary dating app. You will use dimensionality reduction and clustering techniques to explore this artificial dataset (artificial as in created by the instructor for the needs of the class)

1. userid : id of the user

2. date.crea : date of the creation of the account
3. score : score of the profile (how liked is the profile by other users)
4. n.matches : total number of matches the user has had since account creation (with conversation)
5. n.photos : number of photos on the profile
6. last.up.photo : last time the user updated profile pictures
7. last.pr.update : last time the user updated profile text
8. last.connex : last time the user updated profile text
9. gender. O is male, 1 is female. 2 is "other" (notably if the user did not want to specify gender)
10. sent.ana : sentiment score for the text of the profile
11. length.prof : Number of words in the profile text
12. voyage : Keyword voyage found in the profile text
13. laugh : Keyword laugh found in the profile text
14. photo.keke : one of the profile pics comports a photo without a T-shirt / with sunglasses / selfie in an elevator
15. photo.beach : one of the profile pics comports a photo taken on the beach

3 Instructions

3.1 Identifying correlations in the variables

Search for correlations among (cor.test). Use both parametric and non-parametric techniques (Pearson vs. Spearman for correlation between continuous variables). Design a couple of plots featuring correlated variables.

3.2 Dimensionality Reduction

Perform a PCA on relevant continuous variables of the dataset

Use one of the variables as a supplementary one (often in blue in the circle of correlations).

Create a variable circle of correlations. Describe it (but not thoroughly). How would you name the principal components ?

Create an individual map with a sample of individuals.

Create a biplot with a limited number of individuals.

Create a table describing the loadings of each Principal Component. Describe this table.

Perform a MCA on relevant discrete variables of the dataset. Present the relevant graphs and describe the data.

3.3 k-means and Hierarchical Clustering

Perform a k-means clustering on principal components of the analysis you just did. Explain how k-means work. You must justify the choice of the number of clusters in the most detailed manner, notably through a scree plot that you must explain.

Perform and HCPC on continuous variables. Explain how HC works. Explain what are the pros and cons of HC compared to k-means.