Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Verteilte System

Prof. Dr. Eirini Ntoutsi
Tai Le Quy and Damianos Melidis

Data Mining I - SS17
Project 2- Unsupervised Learning

Announcement date: **28.06.2017**

1. The scope

The goal of these projects is to gain hands on experience on the whole pipeline of data mining, from dataset preparation up to pattern extraction and interpretation. In Project 2, we will focus on unsupervised learning.

2. Dataset

The dataset consists of different instances of glass segments, each object/instance containing information on the chemical composition of the segment. Our goals is to group those instances into groups/clusters of similar instances. There is also a class attribute (*Type of glass*) which must not be used for clustering but can be used for the evaluation of the clustering results.

More information and link to the data is available at:
https://archive.ics.uci.edu/ml/datasets/Glass+Identification

3. Tasks

3.1 Dataset understanding

The first important step is getting to know your data and to this end you can explore key statistics and plots. In particular, compute the following and describe and interpret your results:

- Uni-variate analysis
  - Plot the value distribution of each feature (boxplot, histograms). What do you observe?
- Checking for correlations: Bi-variate analysis
  - Plot the correlation matrix. Are there correlated attributes?
  - Hint: Can you find features are functions of others?
- Normalization
  - Check the attribute value ranges. How can you deal with different ranges?

3.2 Preprocessing/ transformation

Based on the results thus far, consider whether you want to filter out some features (*feature selection*) or whether you want to create derived features (*feature transformation*).

What is your final dataset to be used for mining? List the updated feature list, if you created new features and describe how they were derived.

3.3 Clustering

- **Select the clustering algorithm**
  - Experiment with at least 3 clustering algorithms, from those discussed in the lecture
- **Parameters setting**
  - Experiment with different distance functions (e.g Manhattan or Euclidean distance)
  - How can you set the parameters for each algorithm?
- **Calculate performance measures**
  - Internal measures: cohesion, separation and silhouette coefficient (when applicable)
  - External measures: clustering purity
  - Compare the measures for each algorithm
- **Clustering visualization**
  - Choose the most representative features and plot the clusters, what are your observations?
  - Interpret the visualization for each algorithm
- **Performance measures vs. visualization**
  - Can you compare the performance evaluation by the numerical measures and the interpretation by the visualization

4. Submission Logistics

- **Deadline**: **26.07.2017** 23:59 (Berlin time).
- **Submission email**: Use the following title for your submission email *[DM 1 - Project 2]*.
- **Deliverables**: i) project report in pdf format, ii) code (in zip or tar.gz) and iii) a readme file to guide us on how to run your code and reproduce your results.
- **Programming language**: You can use the language of your choice. We recommend Python libraries, SciPy and SciKit-Learn, covered in the tutorials.
- **Working in groups**: We recommend you work in groups (up to 4 people per group) but you can also submit alone. In case of a group, please enlist information for all group members.

Good luck!