MC886/MO444 Machine Learning and Pattern Recognition

Assignment #3 — 2017s2 — Prof. Sandra Avila

Objective

Discover organizational structures in a dataset with unsupervised learning techniques.

Activities

1. Discover the number of groups present in the data or a reliable range of possible values. Do some experiments in this regard.
2. Analyze the medoids of some groups and their closest neighbors in the groups. Do they make sense? Are they talking about the same type of documents?
3. Think of possible ways of checking the validity/quality of your clusters.
4. Re-do the best experiment above considering the PCA dimensionality reduction. Consider different energies (variance) to cut and reduce dimensionality. What are the conclusions when using PCA in this problem?
5. Prepare a 4-page (max.) report with all your findings. It is UP TO YOU to convince the reader that you are proficient on Unsupervised Learning Techniques, and the choices it entails.

Dataset

There are 19,924 documents ('docs'). The bag-of-words feature vectors (with 2,209 dimensions) representing each document are also available ('data.csv' and 'ids').
The data is available at: `https://www.dropbox.com/s/fjtbwf7f5p9f3lx/documents.zip`

Deadline

Friday, **November 10**, in the beginning of the class, 7pm.
Penalty policy for late submission: You are not encouraged to submit your assignment after due date. However, in case you did, your grade will be penalized as follows:

- November 11 7pm : grade * 0.75
- November 12 7pm : grade * 0.5
- November 13 7pm : grade * 0.25

Submission

On the deadline day, bring your 4-page printed report. The template for report is available at `https://www.dropbox.com/s/nc6d89otr8ekvjd/report-model.zip`. Please, print on both sides of the page.
The report should be written in Portuguese or English.
Submit a zip file, with the code and the report (PDF file), via Moodle.
This activity is **NOT** individual, it must be done in pairs (two-person group).