INF436 Machine Learning / Homework 3-1

K-means

Jae Yun JUN KIM*

February 19, 2019

Due: Before the next lab session.
Evaluation: Interrogation during the next lab session about:

- code (in group of up to 3 people)
- (theoretical, practical) questions (individual)

Remark:

- Only groups of one/two/three people accepted. Forbidden groups of larger number of people.
- No late homework will be accepted.
- No plagiarism. If plagiarism happens, both the "lender" and the "borrower" will have a zero.
- Code yourself from scratch. No homework will be considered if you solve the problem using any ML library.
- Do thoroughly all the demanded tasks.
- Study the theory for the interrogation.

---

1 Tasks

A) **Clustering some synthetic data**

1. Download from the course site the 2D data stored in `data_kmeans.txt` file.
2. Cluster them using the K-means algorithm using the formulas seen in class.
3. Test your model with some new data.
4. Plot both training and test results in a 2D graph.

B) **Clustering some real data**

Download from the course site the 6D data stored in `grade_students.csv` file. The source of this dataset is the **The Student/Teacher Achievement Ratio (STAR) Project** organized by the Tennessee State Department of Education in the USA. The reference is the following:

`https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/10766`

This dataset contains 6 features of 5500 students from 79 schools in the state of Tennessee: students' free or reduced-price lunch status, number of absence days, the standarized Stanford Achievement Test Scores for reading, Math, listening and word study.

1. Using the given dataset, cluster the students in 3 clusters (weak, average and gifted clusters) using the K-means algorithm.
2. Interpret your results.

*ECE Paris Graduate School of Engineering, 37 quai de Grenelle 75015 Paris, France; jae-yun.jun-kim@ece.fr