CLUSTERING EXERCISE

Web Mining & Retrieval 2013/2014

Main Objectives

- Learn to write unsupervised ML algorithms
  - Kmeans
  - QT-Kmeans
- Learn to manage weka datasets with Java

Material

- In clustering_ex.zip you can find
  - iris.arff: the IRIS dataset
    - We have already seen it in the classroom!
  - lib/weka.jar: the weka library
  - Kmeans.java: a first and simple implementation of the Kmeans algorithm
- To compile the code
  - `javac –cp lib/weka.jar KMeans.java`
- To execute the code
  - `java -cp .:./lib/weka.jar Kmeans`
- Or, import it in an Eclipse project

Exercise n° 1

Analyze the code in Kmeans.java. You can notice that it is a very simple implementation of the Kmeans algorithm that supports only bi-dimensional vector.

- Modify the algorithm to support N-dimensional (N>=3) vectors
- Apply the algorithm you write on the IRIS dataset
  - Be careful on the choice of the number K of clusters

Exercise n° 2


Starting from the implementation of the Exercise n° 1 your task now is to evaluate the algorithm with respect to the gold standard classes of the IRIS dataset.

Implement an «Evaluator» that, given the original dataset and the clustered one, produces the confusion matrix and computes the Error rate of your algorithm.
Notice that the correct labels are readable easily via Java.

Exercise n° 3



We have seen during classroom that Kmeans has some disadvantages, e.g. the need of a user-specified K.

Implement the QT-Kmeans version of the algorithm, and try different values of the threshold σ and τ.
Verify with the Evaluator from Exercise 2 if the algorithm is performing better.

Contacts

If you have problems or do you need help, do not hesitate to contact

- Giuseppe Castellucci: castellucci@ing.uniroma2.it
- Danilo Croce: croce@info.uniroma2.it
- Simone Filice: filice@ing.uniroma2.it

And (obviously)

- Prof. Roberto Basili: basili@info.uniroma2.it