****

Due Tuesday 04/16/2013 at start of class --- WILL TAKE TIME…A LOT OF IT

To be done individually or in groups of 2

****

Implement the K-means algorithm which operates over numeric data containing class labels (later on, you'll be doing both: *supervised validation* as well as *unsupervised validation*). Your program should run smoothly ON ***ANY DATASET*** – ***containing ONLY numeric attributes and a class label – formatted in ARFF***. Your program should take an ARFF file as input and try out different distance/similarity measures and values for K in order to figure out how the data clusters best. Use standard stopping conditions for K-means – unless you run into problems.

Your program should use all combinations of the following:

- **Distance/Similarity**: *Euclidian distance* and *cosine similarity* --- EXCLUDE THE CLASS LABEL (WILL ALWAYS BE THE VERY LAST ATTRIBUTE) FROM YOUR DISTANCE/SIMILARITY COMPUTATIONS. THE CLASS LABEL WILL ONLY BE USED FOR VALIDATION PURPOSES.
- **Value for K**: as many as there are class labels, twice the number of class labels, and three times the number of class labels

For every combination of the above (6 in total), your program should:

1. Find and display the K clusters in a meaningful way --- can use a text file per cluster or whatever you like
2. Report the cohesion and separation using WSS and BSS measures (*unsupervised validation*)
3. Report the Entropy per cluster as well as the total ***weighted*** Entropy (*supervised validation*) – VIEW EACH CLUSTER AS A NODE AND THE SET OF ALL CLUSTERS AS A SPLIT IN DTI (assume parent's entropy to be a 1)

In a separate report, create tables showing the effect of the value of K and the distance/similarity measure chosen on the quality of the generated clusters as defined by WSS alone, BSS/WSS, and total ***weighted*** Entropy (i.e., three tables in total). Plot your tables in meaningful graphs as you see fit.

Repeat the above for all three datasets INCLUDED IN THIS HW FOLDER: *iris.arff*, *AllGenes.arff* (a cleaned version from HW3 containing all genes; please use it instead of yours), & *SigGenes.arff* (a cleaned version from HW3 containing only significant genes; please use it instead of yours). **As aforementioned, each dataset contains numeric attributes followed by a classes label; the latter should not be used in distance/similarity computations but only for validation purposes.**

For each dataset, use your tables and graphs to draw meaningful conclusions such as (1) best measure/K combination per dataset according to WSS alone, BSS/WSS, and total ***weighted*** entropy, (2) when do you notice all three validity measures (i.e., WSS, BSS/WSS, and Entropy) agreeing, (3) overall, which dataset from HW3 (i.e., *AllGenes.arff* vs. *SigGenes.arff*) clusters better.

****

Please note that your program should work for ANY DATASET formatted in ARFF which contains numeric attributes along with a class label. I WILL RUN YOUR PROGRAM ON THE GIVEN DATA FILES AS WELL AS OTHERS THAT YOU HAVEN'T SEEN.

Document your code properly and include instructions on how to run it. Your email should contain a SINGLE zip file with the following:

(1) Self-contained source code for a program which performs K-means clustering along with instructions on how to run it (***PS: self-contained means that all needed external libraries must be included***)
(2) (prefer) an executable version of your code especially for Java folks, &
(3) your report.

Please make sure that your code runs on my environment:

- Windows 7 machine
- Supports: python 2.7, Java 1.7 & R
- No fancy editors

****