CS 584 – Data Mining

HW 4 – Due 4/12/16

Total: 100 points

Part 1 (20 points)

Use the distance matrix in the table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged. Show your work.

	P1	P2	P3	P4	P5
P1	0	0.9	0.59	0.45	0.65
P2	0.9	0	0.36	0.53	0.02
P3	0.59	0.36	0	0.56	0.15
P4	0.45	0.53	0.56	0	0.24
P5	0.65	0.02	0.15	0.24	0

Part 2. Weka (80 points)

Use datasets from the UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html)

Choose one from each column of the following table:

One from this column	**One from this column**
Arrythmia	Abalone
Image Segmentation	Auto-MPG
Isolet	Housing
Nursery	Ionosphere

What you need to do for this problem is:

1. Get familiar with the clustering algorithms in Weka. I suggest you try clustering the Iris dataset (in UCI Machine Learning Archive, see below), or the Spam dataset from the previous homework. You may use the Explorer for this part. This will allow you to see what parameter options there are for the various algorithms, and you can visualize the clusters for the Iris dataset. However, for the actual homework described below, you are required to use the command line.

2. Learn how to run Weka from the command line. Read the Weka manual, and/or go to the following links for a quick reference:

   http://weka.wikispaces.com/How+do+I+use+WEKA+from+command+line%3F

   https://weka.wikispaces.com/Using+cluster+algorithms

Weka API, extracted from Javadoc, can be found here: http://weka.sourceforge.net/doc.dev/.

The clusterers package contains various clustering algorithms:
http://weka.sourceforge.net/doc.dev/weka/clusterers/Clusterer.html

Use the command-line option for the following.

3. Choose two datasets, one from each column above. You can load the data file using one of the following procedures:
    a. Change the .data file to .arff format.
    b. Load the data file as CSV files. Change the .data file to .csv, and choose the CSV file option. Check the "Invoke options dialog" checkbox. Pick the desired file and when the dialog box pops up, change the "NoHeaderRowPresent" to true.
4. Determine how you will measure the quality of the clusters produced.
5. Choose two algorithms to compare. For each algorithm, compare the different parameter options (e.g. different "k" for k-means, different linkage techniques for hierarchical clustering, etc).
6. Set up and run a comparison experiment using the command line, obtaining the quality measures you determined above. You will need to write a script automating the experimental process (e.g. pre-processing, trying different seed values to randomize the initial centroids for k-means, etc). *You may find that some algorithms cannot be meaningfully applied to some datasets. If so, you can explain why in lieu of the experiment. However, saying "the data has continuous values, the algorithm only applies to nominal values" isn't good enough - you should instead discretize the continuous attributes. "Not applicable" is only valid if there is no reasonable way of preprocessing the data to make the algorithm apply. Each data set and algorithm you choose must be used at least once.*
7. Explain which algorithm you would use for what types of data and why.

Answer the following questions:

1. (12 points) Description of how you measured the cluster quality (this will include a brief overview of the datasets).
2. (48 points) Discussion of each of the four experiments, consisting of:
    a. How you prepared the data
    b. Parameters for the algorithm. See #4 above. Outline the results obtained from different parameters.
    c. Experimental result summary.

For each, you should include a brief discussion of why you made the decisions you did.

3. (20 points) Conclusions: General discussion of the appropriate conditions for use of each algorithm. You may instead want to frame this as a discussion of appropriate type of algorithm for a general category of data (probably a more difficult task, but also more interesting.)

You should also include the output from your sample runs.

Scoring

Scoring will be based on:

- Appropriateness and correctness of cluster quality measurement
- Experiment and discussion
- Knowledge displayed in conclusions

Report and Submission

- Collect output from your experiments. **Submit Part 1, detailed Weka output/screenshot, script, and report electronically on Blackboard.**

- For the following items, also submit a hardcopy in class: Part 1 and the Weka report from Part 2.