University of Colorado at Colorado Springs

Home Work Assignment 2

Due 04-27-2010

---

1 Clustering

Write a program in a language of your choice that clusters input points.

1.1 Dataset

You will work with a dataset provided by Dr. Fred Coolidge of the Department of Psychology at the University of Colorado at Colorado Springs. The dataset is available at `http://www.cs.uccs.edu/~kalita/work/cs586/2010/CoolidgePerpetratorVictimData.csv`. This dataset pertains to scores on personality disorder tests given to inmates in the State of Colorado. Dr. Coolidge's inventory of personality disorder tests are given to all inmates in the State of Colorado. Here we have a little sampling of data. The data contains information on 25 repeated victims and 75 repeated perpetrators of sexual abuse among inmates. The first column of the data says if someone is a victim or a perpetrator, but for the purposes of this assignment, please ignore this column.

The data for each individual row has values for 14 attributes, each of which is numerical. The attributes correspond to scores on 14 personality disorder tests. The tests are:

- antisocial (AN)
- avoidant (AV)
- borderline(BO)
- dependent(DE)
- depressive(DP)
- histrionic(HI)
- narcissistic(NA)
- obsessive-compulsive (OC)
- paranoid (PA)
- passive-aggressive (PG)
- schizotypal (ST)
- schizoid(SZ)
- sadistic(SA), and
- self-defeating (SD).

They are all measured by T scores. For T-scores, the mean of is always 50 and a standard deviation is always 10. Thus, an inmate with a T score on the antisocial scale of 65 is 1.5 standard deviations above the normative mean.

1.2 Things to Do

You will write programs for the following. You can do more if you want for extra credit, but make sure you do what's asked before you start doing more.

1. Think of a couple of distance measures to use. Use Pearson's correlation coefficient as one of the measures. You can choose any other measure you like. Please describe the distance measure you use. Here is a link for Pearson's correlation coefficient: `http://www.childrens-mercy.org/stats/definitions/correlation.htm`. A detailed writeup on Pearson's correlation coefficient is available in a manuscript by Coolidge at `http://www.cs.uccs.edu/~kalita/work/cs586/2010/CoolidgeChapter6.pdf`.

2. Implement a partitional algorithm. Increase the number of clusters from 2 to say 10 and see what happens. Interpret the results.

3. Implement an agglomerative clustering algorithm of your choice. Describe the algorithm briefly. Interpret your results.

4. Please read about how to measure quality of clusters and provide at least one metric to quantify the quality of clusters you produce.

1.3 What to Hand in

You will submit a 3-5 page paper with a title and your name. Use the IEEE Author style you have been using for the semester project papers you have been writing for the class. In this paper, you will have a short section with an appropriate heading for each of the most important steps in your program. Describe in this section how you carry out the step in question. You will also have a section called *Results*. Here, you will carry out the clustering experiments for the dataset, and report the results in the form of tables and graphs, as appropriate.

Please create an appendix where you provide visualization of which individuals belong to which clusters that you produce in all (some of) your experiments. Tables or simple diagrams showing sets and subsets will suffice.