CSE 5243

Instructor: Jason Van Hulse

Homework 1

Due Date: 2/1/2017 5:30pm (hand in the written report in class).

In this lab, you will write a program which will take a dataset as input (*n* rows and *m* columns) and provide the following output dataset:

1) An *n x (2k+1)* data frame which outputs the *k* row ids for those examples which are closest or most similar to the given example (as measured by a chosen proximity function - see below) with their proximities. So if *k = 4*, then the first two rows of the output might look like:

Transaction ID	1st	1-prox	2nd	2-prox	3rd	3-prox	4th	4-prox
1	45	0.134	13	1.33	8	1.54	103	2.33
2	18	0.13	33	0.155	1	0.564	27	2.02

The interpretation of this output is that for the first record in the input dataset, the most similar record is #45, with a proximity measure of 0.134. Similarly, record # 13 is the second most similar record to record #1, with a proximity measure of 1.33. Note that this output should have *n* rows, which is the same as the input dataset.

If you prefer, you can also output this data as two separate data frames.

The two input datasets that your algorithm must run on are:
1) **Iris data**
2) **Income** - See below for a brief description of this dataset

Both datasets are posted in the Carmen site under the “Homework” section.

Program Features:

Your algorithm should have the following features and be able to handle these data issues:

1) Both categorical and continuous attributes which nominal, ordinal, ratio and interval.
2) Missing values and outliers
3) Attributes of different scales - please implement appropriate attribute transformation methods
4) Implement 2 different proximity measures - for example, you may implement a Euclidean distance measure and a cosine similarity. *Please pay careful attention to whether your metric is a similarity or dissimilarity when outputting the 'closest' or 'most similar' objects.*
5) Both datasets have a 'class' attribute - please do NOT use this attribute in the proximity function
6) The parameter *k* should be variable, with *k = 5* as the default value.

Report:
With this assignment, you will turn in a written report with the following information:

Section 1: Exploratory analysis of the Income dataset. Based on this analysis, what observations of the Income data do you have? Are there any interesting patterns or trends? Please elaborate and provide supporting results.

Section 2: A description of your program, including discussions on design choices made. For example, how did you choose to handle missing values or outliers for these datasets? Did you transform any of the attributes? For the income dataset, you should justify your choices based on the results of the exploratory analysis above.

Section 3: Analysis of the results. Some examples of analysis you might conduct (feel free to add other ideas):

A. How do you describe the distribution of proximities between each example and its first nearest neighbor? How does this distribution change as *k* increases?
B. You did not use the class attribute in the proximity function - but for each class, do you observe any differences for part A above?
C. Is there one example which is the closest to the largest number of other examples?

D. Do any of these results differ when you change the proximity measure?

Teams of up to 2 are highly encouraged, however, significantly more work is expected from teams. In particular, Section 2 and 3 above should be much more extensive.

What you need to turn in:

1) Code
2) Makefile (if applicable)
3) Readme - should describe how to run the code
4) Written Report
   A. The report should be a maximum of 9 pages (14 pages for teams of 2).
   B. The report should be well-written. Please proof-read and remove spelling and grammar errors and typos. *Writing and presentation will be part of your grade for this assignment.*

Please hand in the written report in class on the assignment due date.

You do *not* need to turn in the output datasets, rather these will be obtained by running your code.

Note: It is expected for you to code these from scratch, and not to use existing functions.The only built in *mathematical* or *statistical* functions you should use are mean, median, standard deviation, minimum and maximum.

How to hand in your work:

Please choose one of the programming languages from: C/C++, JAVA, Python, MATLAB, R. All the related files except for the data will be tarred in a *.zip file or *.tgz file, and submitted via Carmen. Please use this naming convention: ”Project1_Surnames_DotNumber.zip" or “Project1_Surnames_DotNumber.tgz.” The submitted file should be less than 5MB.

On Linux System (Mac OS, Ubuntu, RedHat, etc.)

[Source Code, BashScript (and Makefile), Readme.txt, Report] should be submitted. Do not submit raw dataset.

The program should be able to run on a standard Linux system. Readme.txt tells me where to put input data (raw dataset) and where to find output data and how to interpret the output.

When I type command "bash BashScript", the output would be generated.

On Windows system

[Source Code, Readme.txt, Report] should be submitted. Do not submit raw dataset.

The program should be able to run on a Win-7 system. Readme.txt tells me where to put input data (raw dataset) and where to find output data and how to interpret the output. Readme also tells me how to compile and run the program.

If you use C/C++, please make sure your program can get compiled and run on VS2010 or later version, or you can choose to use GNU C++ compiler.

If you use JAVA, please make sure your program can get compiled and run on Eclipse.

About the Income dataset: Extraction was done from the 1994 Census database. The first column in this dataset is the row-id.

Attribute Information:

- age: continuous.
- workclass
- fnlwgt: continuous. Meaning is ambiguous.
- education
- education-num: continuous.
- marital-status
- occupation
- relationship
- race
- gender
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country.
- class: >50K, <=50K.