CSE 5243

Instructor: Jason Van Hulse

Homework 1

Due Date: 9/13/2017 5:30pm (submitted through Carmen).

In this lab, you will work with the **Income** dataset (posted in Carmen). You are required to write a program which will take the **Income** dataset as input (520 observations), along with an adjustable parameter *k,* and provide the following output *for each* input observation:

1) 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)
2) For each of the *k* row ids that are most similar from part 1, output the proximity.

For example, 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 Income dataset, the most similar record is #45, with a proximity measure of 0.134. Record #13 is the second most similar record to record #1, with a proximity measure of 1.33, and so forth. Note that this output should have *520* rows, which is the same as the input dataset.

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

Program Features:

Your code only needs to run using the Income dataset; it does not need to be generalized to handle any type of data. Keep the following concepts in mind as you explore the Income dataset, as you will need to make reasonable design choices and justify those choices in the written report:

1) Do you need to handle different data types? If so, what methods should you use?
2) Do you need to deal with either missing values or outliers?

3) Do you need to deal with attributes of different scales? If so, how do you deal with this issue?
4) Is any other data preprocessing (for example, variable transformations) needed?

Proximity Measures:
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.*

Notes:
- The parameter *k* should be easily changeable, with *k = 5* as the default value.
- The Income dataset has a 'class' attribute - please do NOT use this attribute in the proximity function.
- It is expected for you to build code from scratch, and not to use existing functions. You should only use basic/simple mathematical or statistical functions - like mean, standard deviation, etc.
- There are a number of design choices you can make in this assignment. I would like you to make reasonable choices, and justify those decisions in the report. Points will be lost for making unreasonable choices (like trying to compute the mean of a categorical variable) or for failing to justify a choice.

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) Readme - should describe how to run the code
3) Written Report
   A. The report should be a maximum of 8 pages (13 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.

How to hand in your work:

Please choose one of the programming languages from: JAVA, Python, 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, 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 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.