Department of Computer Science
University of Pretoria

Data Mining
COS 781

Assignment 2

September 21, 2019

1 Objectives

This assignment aims to achieve the following general learning objectives:

- To gain experience with a real-world data set, in the context of data mining and exploratory data analysis;
- To gain practical experience in the application of various data cleaning techniques;
- To gain experience in formal, scientific report writing.

2 Plagiarism Policy

The Department of Computer Science considers plagiarism to be a serious offence. Disciplinary action will be taken against students who commit plagiarism. Plagiarism includes copying someone else's work without consent, copying a friend's work (even with consent) and copying material from the Internet. Copying will not be tolerated in this module.

For a formal definition of plagiarism, the student is referred to `http://www.ais.up.ac.za/plagiarism/index.htm` (from the main page of the University of Pretoria site, follow the `Library` quick link, and then click the `Plagiarism` link under the `Services` menu). If you have any questions regarding this please consult the lecturer to avoid any misunderstanding.

Note that all assignments submitted for this module implicitly agree to this plagiarism policy, and declare that the submitted work is the student's own work. Assignments may be checked using the Turnitin system. After plagiarism checking, assignments will not be permanently stored on the Turnitin database.

3 Submission Instructions

You will have to write a report for this assignment. The report should be in standard PDF format, preferably compiled using LaTeX. You will have to submit only the report. No additional files of any sort should be submitted. Do not submit files in any other format other than PDF. Failure to follow any of these instructions will result in a zero mark for the assignment.

Upload just your PDF file (named `s99999999.pdf`, where `99999999` is your student number), to the appropriate assignment upload on the course website. Multiple uploads are allowed, but only the last one will be marked. The deadline is **Friday, 4 October 2019, at 23:00**.

4 Data Set

The data set that you will work with for this assignment is a real-world data set produced by Statistics South Africa. The data set describes the General Household Survey for 2017, and contains data about South African households and individuals. You will use the data related to individuals.

The course website provides an archive containing an ASCII CSV (comma separated value) file, as well as a PDF file containing important data set meta information. Additional data file formats are also available at `http://nesstar.statssa.gov.za:8282/webview/` located in the sub-folder Household Surveys → General Household Survey → General Household Survey (Revised 2017) → 2017 → General Household Survey 2017 (Person file). From this folder, select the Download link on the top right of the page.

5 Data Preparation

Assignment 1 and Assignment 2 are closely linked. Assignment 2 requires you to perform data preparation on the data set described in Section 4. In Assignment 3 you will then apply a data mining approach to the prepared data set in order to extract useful knowledge. Your data preparation should focus on the following points:

- The data set is large and high dimensional, and there are many potential areas to focus on. Therefore, begin by deciding on the focus of the knowledge that you will extract from the data set. Try to choose a focus area that will produce interesting, non-trivial knowledge. Do not perform too broad an analysis by focusing on several aspects, and end up not delivering interesting insight on any of them. Take an objective view on the data, and disregard any preconceived ideas you have on the topic.
- Once you have decided on the focus of your analysis, you should decide on which data preparation steps are appropriate for the data set. You may have to perform analysis (possibly using EDA techniques) to determine what type of cleaning is necessary. You must consider every aspect of data preparation. Therefore, refer to the slides on Theme 2, and consider each of the separate aspects of data preparation. Your report will have to detail all your decisions. Even for data preparation steps that you do not perform, you must justify why you made the decision to not perform these steps.
- Additionally, you also need to decide on the data mining approach (or approaches) you intend to use in Assignment 3. You may choose between self-organising maps, data clustering techniques, rule induction methods, or decision trees. You may also choose to combine two of these approaches if you wish to. The data mining approach you select will dictate some of the data preparation steps that you will have to perform, because not all data preparation methods are appropriate for every data mining technique.

6 Report

You must write a report describing the data preparation you performed in Section 5. The report should be of an academic nature. This means that:

- The report's structure should include all the aspects typically required of an academic paper (these include a title, abstract, introduction, methodology discussion, and conclusions).
- An adequate background discussion should be provided (this means that you must broadly discuss every data preparation technique you use, and provide a reference to published sources that describe the techniques. It is **not** acceptable to cite the course slides or Wikipedia.
- Include a discussion on the data set you are analysing, with the target audience being a person who is not familiar with the data set. You need only focus on the overall data set characteristics, and then specific details on the data examples and data set attributes you choose to include in your preparation and analysis. Also provide a conclusion in which you summarise what you have achieved through all the data preparation steps you have performed.
- Adequate references must be provided (this means that you must cite appropriate sources for all techniques that you discuss, and all the reference details must be correct).
- The tone of language should be formal and scientific, and must use correct spelling and grammar.
- Any figures or graphs included should be clear and of a professional standard. Label all appropriate parts of graphs and other data visualisations, so that they can be easily interpreted (it is usually not sufficient to simply paste a screen shot of a visualisation from a data analysis program or package).

It is recommended that you consult several existing conference and journal papers, as a guide to the type of style you should adopt. There are many such sources freely available online.

Your report should be concise and to the point. Make sure that you leave no gaps in your descriptions of techniques or the procedures you followed. This is especially the case in relation to data cleaning and

preprocessing, where you should describe both the steps that you performed, and the ones you did not perform. For the steps that were not performed, you should explain why you performed no processing. This is the case even if it seems obvious to you.

It is very strongly recommended that you use the LaTeX template for IEEE conference papers (available as `LatexSample.zip` in the folder for this assignment, on the course website), to typeset your paper. If you use this template, you should aim for a report (including references) of no more than six pages.

7 Marking

The following general breakdown will be used during the assessment of this assignment:

Category	**Mark Allocation**
Writing style and report structure	10 marks
Background information	10 marks
Data cleaning discussion	20 marks
References	10 marks
TOTAL	**50 marks**