SINGAPORE UNIVERSITY OF SOCIAL SCIENCES

ANL305e
Association and Clustering

Group-based Assignment

July 2019 Presentation

GROUP-BASED ASSIGNMENT

This assignment is worth 30% of the final mark for ANL305e Association and Clustering.

The cut-off date for this assignment is **14 October 2019, 2355hrs**.

This is a group-based assignment. You should form a group of **3 members** from your seminar group. Each group is required to upload a single report via your respective seminar group site in Canvas. Please elect a group leader. The responsibility of the group leader is to upload the report on behalf of the group. Those submitting individually will be given a 10 marks deduction.

It is important for each group member to contribute substantially to the final submitted work. All group members are equally responsible for the entire submitted assignment. If you feel that your group members did not contribute sufficiently to the work submitted, please highlight this to your instructor as soon as possible. Your instructor will then investigate and decide on any action that needs to be taken. It is not necessary for all group members to be awarded the same mark.

Up to 25 marks of penalties will be imposed for inappropriate or poor paraphrasing. For serious cases, they will be investigated by the examination department. More information on effective paraphrasing strategies can be found on https://academicguides.waldenu.edu/writingcenter/evidence/paraphrase/effective.

Note to Students:

You are to include the following particulars in your submission: Course Code, Title of the GBA, SUSS PI No., Your Name, and Submission Date.

Write a report containing your responses to the following questions:

Question 1

Twelfth grade or Grade12 is the final year of high school in most of North America. As one of the disciplines the Grade-12 students need to take an exam in SAT or ACT, math is essential for them to apply to the university. In the final exam of Grade 12s, High School ABC would like to analyse the possible factors that affect students' math scores. A student dataset has been collected from several high schools in the same district, and stored in a file named *Math12_sample.csv*. The dataset includes the information of 2,699 Grade-12 students. All the fields are listed out in the following table.

Field	Type	Description
id	Nominal	Each student has a unique student id.
math12	Nominal	This variable indicates whether the student passed the Grade-12 math test. Two possible values: fail, pass.
math8	Nominal	This variable indicates the student's performance in the Grade-8 math test. There are three levels: good, pass, and fail.
fight8	Nominal	This is the number of fights involved when the student was in Grade 8. There are three possible values: never, once or twice, more than twice.
read12	Nominal	This variable indicates the student's performance in the grade-8 reading test. There are three levels: good, pass, and fail.
catholic	Nominal	This variable indicates whether the student's high school is a catholic high school: Yes or No.
female	Nominal	Yes = female, No = male.
race	Nominal	Race code of the student: ai: American Indian/Alaska Native, api: Asian or Pacific Islander, black, hispanic, white.
parmar	Nominal	Parents' marital status. Six possible values: divorced, married, never married, not married but cohabit, separated, widowed.
familyin	Nominal	Family income of the student. There are twelve levels from "none" to around $75,000.
fatheredu	Nominal	Student father's highest education. Eight different levels: not finish hs: not finish high school, hs grad: high school graduate, junior coll: junior college graduate, coll<4: not finish 4-year college, coll grad: college graduate, masters, doctorate, dont know: the education information is missing.
motheredu	Nominal	Student mother's highest education. Possible values are the same as the field named fatheredu.

(a) Evaluate the suitability of applying Association Rule Mining on the dataset.

(4 marks)

(b) Construct an ARM solution using the Apriori node. The model details and interpretation of results should include the following:

(i) Report the role settings of the data attributes and the parameter settings of the Apriori algorithm: Minimum support and Minimum confidence. Explain the settings. Report the number of rules generated, and give a screenshot of the rules. (Note: to observe both "fail" and "pass" scores, set the "Measurement" of all fields as "Nominal",)

(ii) Report ***two (2)*** interesting rules which include the field "catholic" and interpret their implications.

(16 marks)

(c) Based on your understanding of the two rules you listed out in (b) (ii), can we conclude that "cathalic=no" is the cause of why a student failed math12? Discuss the difference between association and causation.

(10 marks)

(d) Compare and contrast the use of CARMA and Apriori to solve this problem (Note: there is no need to run CARMA in IBM SPSS Modeler). If we would like to use CARMA, is there any data preparation that we need to do beforehand? Take one field as an example to explain how to prepare the data.

(10 marks)

Question 2

Pulsars are a rare type of Neutron star that produces radio emission detectable on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. Since pulsars are found in the wreckage of a collapsed supernova, they help us understand what happens when astral collapses. The mystery of the birth and evolution of the universe can also be revealed through research on them.

Dataset "*Pulsa_Candidates.csv*" includes 17,899 pulsar candidates collected during the High Time Resolution Universe Survey. Among these candidates, 1,639 are confirmed to be real pulsars based on their moving features in a variety of ways. The dataset has nine (9) fields. Eight (8) of them are the candidate's moving features (Note: you don't need to spend too much time on understanding the technical jargons used in this question). There is one (1) field named "Class" that shows whether the candidate is a real or fake pulsar.

(a) High Time Resolution Universe Survey would like to figure out how the moving features determine groupings of pulsars. Your team is to help them to evaluate the suitability of applying clustering on this problem.

(5 marks)

(b) Design a K-Means model to study this problem. An expert suggested that a suitable number of clusters K is between 2 and 4. Use the Silhouette measure and Matrix node

(that cross-tabulates the generated clusters and the "Class" field) to verify the performance of clustering results generated based on different K-values.

(i) Give a screenshot of the stream of the constructed K-Means model.
(ii) For each K value, show the screenshot of the Matrix node's result, and calculate the mislabelling rates of "class = 0" and "class = 1".
(iii) In your opinion, which K-value is better? Justify your answer.

(Note: 1. To implement K-Means, the role of the field "Class" should be Target, not Input; 2. To generate the performance matrix, you can link a Matrix node to the clustering result node.)

(25 marks)

(c) The given dataset has imbalanced class distribution issue (1,639 real vs. 16,259 fake). The inability of detecting small clusters is a drawback of K-Means. Discuss this issue, and propose one possible solution to address it. Keep the length of your discussion to one page.
(Note: you can refer to a solution proposed in research papers or articles, but you need to give a clear description of the method used and also a clear citation of the reference)

(20 marks)

Another 10 marks are allocated for your writing.

Your writing should be succinct but not at the expense of excluding relevant details. Highlight only the points that are relevant to your discussion. Use plain and simple language. Some questions may not come with absolutely right or wrong answers. For such questions, you have the liberty to express your views about the problem. However, your points have to be supported by evidence and good reasoning. It's the quality and not the length that counts. Make sure you follow the report guidelines and style specified in this assignment.

Make sure you indicate your name and student number on the cover page of the report.

The topics in the main report should be presented in the order according to the sequence of the tasks/questions listed in the assignment; that is, in the order of (a), (b), ..., etc. You can have several sub-sections within a section if you deem appropriate.

The report must be self-contained. It is important to include all relevant tables and figures in the report as evidence to support the answers given.

The followings are some details of report format:

- Length: **should not exceed 12 pages** (including the relevant graphs, tables, references, screenshots and appendices (if any), but excluding the cover page)
- Font Style: Times New Roman
- Font size: 12
- Line spacing: 1.5
- Margins: 1" for the top, bottom, right and left
- Include the page number on each page

Some further suggestions:

- Ensure minimal grammatical and typographical errors
- Write clearly in plain English
- Write appropriately to the context
- Cite appropriate sources
- Provide a reference or bibliography at the end of the main report
- Include less relevant details in the Appendix
- Good overall presentation of the report

---- END OF ASSIGNMENT ----