Assignment 1: Cluster Analysis

1. Introduction

This assignment aims to:

- provide students the opportunity to consolidate your understanding of the basics of cluster analysis and to use SAS Enterprise Miner to conduct cluster analysis, and
- assess students' learning outcome of cluster analysis.

This assignment is an individual assignment. Each student is required to conduct cluster analysis with the given data set and write a report to present your understanding of cluster analysis and how the clustering is done, and discuss your findings

This assignment is worth 25% of the total marks of the course.

2. Deadline and submission

- The assignment (report) is due on Sunday 10 September 2023, 11:59PM Adelaide time.
- The report must be a single MS Word or PDF file and named as: <yourEmailID>_UMA_A1.
- The report must be submitted online via the link on Learnonline course site.
- Extensions to the assignment will only be granted under unexpected circumstances (e.g., illness). In this case, a student must provide supporting documents and submit their extension request (via Learnonline) at least a day before the assignment is due. A request of extension made on or after the due date of the assignment will not be considered. Any late submission of the report without a pre-arranged extension, a penalty of 20% of the assessed mark per day (including weekend) will be incurred. For example, if a student's submission is 2 days late and the originally assessed mark is 80 out of 100, then after the late penalty is applied, the actual mark the student is awarded will be 48 out of 100.

3. Detailed information about the assignment

3.1 The given data set

The given data set (filename: assign1-data.csv) is sourced from the UCI Machine Learning Repository (*with modifications*). The data set contains the measurements from digitised histopathological images of fine-needle aspiration (FNA) biopsies of 699 participants (breast cancer patients and healthy participants) in the studies of breast cancer. More information about the attributes of the data set and the relevant background information can be found, e.g. at the UCI site https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29 and sites like https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis.

3.2 Cluster analysis using SAS Enterprise Miner

Before clustering the data set, first understand the background about the data set, especially the meaning and data type of the attributes from sources such as the above listed sites.

Then in SAS Enterprise Miner, explore the data set, and inspect the role and level of each variable that have been specified by SAS when the data set is imported (right click on the File

Import node and choose Edit Variables to see/update the roles and levels). If necessary, change variable roles and levels to keep consistency with the meaning and types of the variables. Next, preprocess the data set if necessary, either using the preprocessing functions available within the Cluster tool or similar functions provided by other tools of SAS Enterprise Miner.

When running the Cluster tool & the Segment Profile tool:

- First set the number of clusters manually, e.g. to 6 and then reduce the number of clusters to 5,…, till 2 and run the Cluster node and the Segment Profile node with the different number of clusters respectively.
- Next use Automatic as the Specification Method to let SAS Enterprise Miner determine the number of clusters automatically, and explore the impact of different Selection Criterion of the Cluster tool and different data preprocessing options (e.g., data normalisation methods, missing value imputation methods etc.) on the number of clusters determined by SAS Enterprise Miner.
- Analyse and compare the results obtained in different cases.

Note that for this assignment, students must use SAS Enterprise Miner for data exploration, preprocessing and cluster analysis. You may include extra analysis/result done using other software tools, but it is not required.

3.3 Report

You must include the following five parts in the report, plus a reference section listing all the references cited in your report. Structure your report in the way that readers can clearly see the five required parts, e.g., by having five sections corresponding to the five required parts. You may have additional contents in your report, if you consider they are necessary.

1. Background [15 marks]

Provide a description *in your own words* to explain:

- [5 marks] what cluster analysis is,
- [10 marks] how the basic *k*-means algorithm works.

You must assume that your readers do not know about clustering, hence you need use easy to-understand language and simple examples in your description. Use figures and/or tables to help with the description and cite any references used.

2. Cluster analysis with SAS Enterprise Miner [40 marks]

Provide a description of the analysis that you have conducted in SAS Enterprise Miner following the instruction given in Section 3.2, including:

- [10 marks] Data exploration and preprocessing.
    - Describe what data exploration, variable role/level changes, and preprocessing that you have done and how they are done. Use diagrams/screenshots to help with the presentation of what you do and how you do them.
    - Justify on the changes made and data preprocessing that you have done. If no role/level changes or preprocessing are done, you must also provide justification why you don't think the changes or preprocesing are needed.
- [30 marks] The analyses conducted and the results.
  For **each** of the experiments/analyses you have done,

  - Describe the experiment settings (e.g., the number of clusters specified by you and other options)
  - Present the results obtained with the different settings (i.e., the results of running the Cluster node and the Segment Profile node with the settings). You must use diagrams/screenshots in the presentation of the results and provide text description of the results.

3. **Result discussion** [20 marks]
   Provide a discussion of the analysis results, covering at least the following two aspects and cite any references used:
   - [10 marks] Interpretation of the results, including the results obtained by running the Cluster node and Segment Profile node.
   - [10 marks] Comparison of the results obtained using different user specified numbers of clusters and the number of clusters determined automatically by SAS Enterprise Miner. Base the discussion on the results of running the Cluster and Segment Profile nodes, and explain why one clustering result may be better than or similar to another.
4. **In-depth discussion of the results with respect to the real world problem** [15 marks]
   Based on the clustering and segment profiling results, provide insights into the data set and the practical problem. For example, what is the common characteristics of the samples in the same cluster and how the different clusters are distinct from each other, and what implications do these commonalities and difference provide regarding the practical problem? To provide an in-depth discussion of the results with respect to the real world problem (breast cancer diagnosis), you need to explore and understand the meaning of the attributes in the data set and do some investigation related to the real world problem to support your discussion. Cite all references used.
5. **Reflection** [10 marks]
   Discuss the pros and cons of using k-means clustering and SAS Enterprise Miner to analyse the given data set, and the lessons learned. The discussion here needs to be put in the context of this assignment and your own experience, instead of a general description of the pros and cons of k-means or SAS Enterprise Miner. Cite all references used.

4. Marking of the report

Your assignment will be assessed according to the marking rubrics given on the next page.

Additionally, you need restrict the length of your report between 1,500 and 2,500 words, and part 1 (i.e., the description on cluster analysis and how basic k-means works) must not be over 400 words, excluding references, figures/diagrams, tables, figure/table captions, coversheet and table of contents. A report does not meet the word limit requirements may attract deduction of marks.

There is no strict requirement on the format of the report, but it needs to have a clear structure and the contents need be presented clearly. Poor quality of presentation may attract deduction of marks.

	85-100	75-84	65-74	50-64	0-49
Part 1 (15%)	An accurate and concise introduction to cluster analysis and *k*-means, suggesting a thorough understanding of the topic. The writing is easy for general audience to understand and well supported by high quality references and example(s).	An informative and well-written introduction to cluster analysis and *k*-means, suggesting clear understanding of the topic. The writing is understandable to general audience and supported by suitable references and example(s).	An informative introduction to cluster analysis and *k*-means for general audience, indicating good understanding of the topic. References are cited and some examples are provided.	A more or less informative introduction to cluster analysis and *k*- means, suggesting reasonable understanding of the topic. References are cited and some examples are provided.	The writing indicates incomplete and/or incorrect understanding of the topic.
Part 2 (40%)	A thorough exploration of the data set is done and presented clearly. All necessary preprocessing steps/aspects are considered, decisions on why a preprocessing step is or is not needed is well justified, and all necessary preprocessing steps are done correctly. All the required cluster analysis and segment profiling tasks are done correctly, the description of the steps and settings are complete and well presented, and all necessary results,diagrams/screenshots are included and described clearly.	Key data exploration and preprocessing steps are considered and done, and data preprocessing decisions overall are justified well, although some justifications may not be completely reasonable. All the required cluster analysis and segment profiling tasks are done and presented clearly and all necessary results,diagrams/screenshots are included and described clearly although a small portion of the details may be missing or inaccurate.	Key data exploration and preprocessing steps are considered and done, and data preprocessing decisions are justified, but some of the preprocessing done or the justification may have issues. Most of the required cluster analysis and segment profiling tasks are done and presented and most necessary results, diagrams/screenshots are included and described clearly.	Some data exploration and data preprocessing steps are done, and some justifications are provided, but may not be convincing. A large portion of the required cluster analysis tasks are done and reasonable amount of results are shown, which can be used to support useful discussion.	Limited data exploration and preprocessing are done, and lack justifications. Some cluster analysis experiments are done and presented, but the description does not clearly indicate good understanding of what has been done.
Part 3 (20%)	A clear and accurate description of the meaning of the results, indicating a thorough understanding of the results of the Cluster node and Segment Profile node of SAS EM. A comprehensive discussion and comparison of the clustering results using different settings, fully supported by the evidence provided by the Cluster node and Segment Profile node, indicating a thorough understanding on how to judge the clustering results by SAS EM.	A clear description of the meaning of the results, indicating a good understanding of the results of the Cluster node and Segment Profile node of SAS EM. A meaningful discussion and comparison of the clustering results using different settings, supported by the evidence provided by the Cluster node and Segment Profile node, indicating a good understanding on how to judge the clustering results by SAS EM.	Overall a meaningful interpretation of the results, indicating reasonable understanding of the results of the Cluster node and Segment Profile node of SAS EM. The discussion and comparison of the clustering results with different settings overall are sensible and supported by the evidence provided by the Cluster node and Segment Profile node, indicating a reasonable understanding on how to judge the clustering results by SAS EM.	A large portion of the interpretation of the results is meaningful, indicating essential understanding of the results of the Cluster node and Segment Profile node of SAS EM. The discussion and comparison of the clustering results with different settings show some understanding, and some discussions are supported by evidence.	The interpretation of the results is not reasonable, indicating poor understanding of the results of the Cluster node and Segment Profile node of SAS EM. No discussion or comparison are done, or what's presented is incorrect.
Part 4 (15%)	Insightful and inspiring views/conclusions are derived by making sense of the clusters/results obtained using SAS EM and exploring literature in the context of the real world problem, such that readers can gain useful understanding of the characteristics of different clusters of the participants of the studies to guide breast cancer diagnosis.	A good amount of insights are provided in the context of the real world problem, which could be useful for understanding the real world problem and/or guide breast cancer diagnosis. It can be clearly seen that the student has done good exploration of the literature and linked the cluster analysis result to the real world problem.	Some insights are provided in the context of the real world problem, which could be useful for understanding the real world problem and/or guide breast cancer diagnosis. The discussions are supported by evidence (literature and SAS EM results).	Clear indication of the effort in linking the cluster analysis results with the real world problem and some insightful discission in the context of the real world problem is provided.	No or minimal indication of the effort in linking the cluster analysis results with the real world problem and no or minimal insightful discission in the context of the real world problem is provided.
Part 5 (10%)	A comprehensive and meaningful reflection, which can serve as an excellent guide for using k-means and SAS EM for cluster analysis. It can be clearly seen that the reflection is based on personal experience of doing the assignment.	A high quality reflection based on personal experience of doing the assignment. The reflection can provide a good guidance on using k-means and SAS EM for cluster analysis.	A genuine reflection based on personal experience of doing the assignment, but the contents may not be adequate. The reflection can provide useful guidance on using k-means and SAS EM for cluster analysis.	Some meaningful reflection based on personal experience of doing the assignment, but some important aspects are not covered. The reflection can provide some guidance on using k-means and SAS EM for cluster analysis.	The reflection has no or minimal connection to the work the student has done for the assignment, and does not help with the use of k-means or SAS EM.