STINK3014 (NEURAL NETWORKS)
A251 – Assignment (#3) (10 %)

Instructor: Azizi Ab Aziz
Submission date: 23/ Dec / 2025 (before 11.59 pm) via UUM Learning Portal

"*You are not meant for crawling, so don't. You have wings. Learn to use them and fly.*" [Rumi]

PART I: THEORETICAL CONCEPTS

Questions: Self-Organising Map (SOM) / Adaptive Resonance Theory / Autoencoders

1. What is the primary objective of a Self-Organising Map (SOM), and how does it differ from traditional clustering algorithms like K-Means?
    a. How does SOM handle high-dimensional data?
    b. How can SOM be used for missing data imputation
2. What is the "dead neuron" problem in SOM, and how can it be mitigated?
3. Describe the basic two-layer architecture of an ART network: the F1 layer (input/comparison) and the F2 layer (category/recognition).
    a. How does increasing or decreasing the Vigilance Parameter (ρ)affect the categories formed by the network?
    b. What does it mean for an input vector to "resonate" with a learned category in ART?
4. How does ART handle novel inputs that do not sufficiently match any existing category (i.e., when the match criterion is not met)?
5. How do autoencoders learn useful representations without labels?
6. Describe the effects of the bottleneck layer size towards the autoencoder performance?

PART II: CASE STUDY

Sintok Bank faces a significant business challenge: its current customer segmentation, based solely on simple demographics (age, income), is failing to drive effective marketing and is contributing to high churn, particularly among younger clients (UUM students). The core problem is that the bank possesses massive amounts of high-dimensional, unlabelled behavioural data, encompassing over 80 different metrics that cover transactions, credit usage, and digital activity. This data is currently too complex and redundant for direct use in segmentation, resulting in the "Curse of Dimensionality." The bank's goal is to move beyond these outdated categories by implementing a two-phase data-driven strategy: first, a technique must be employed to reduce the 80+ behavioural metrics into a core set of 6 to 10 meaningful, independent features that capture the essence of customer value and behaviour; second, an effective method must be applied to automatically discover the optimal, natural number of distinct customer clusters within this reduced feature space, resulting in actionable segments like "Digital-First Users" or "Credit-Reliant Spenders." The final success of this project hinges on transforming these discovered segments into targeted business strategies, enabling the bank to personalize product offerings, reduce marketing waste, and significantly improve customer retention and lifetime value.

Questions (based on this case study, as in Part II)

1. What theoretical advantages does KSOM provide when used to pre-cluster customer data before training a Multi-Layer Perceptron (MLP) to predict segment labels?
2. How does using Autoencoder-generated latent features as inputs to an MLP improve classification accuracy compared to a traditional MLP trained on 80+ raw banking metrics?
3. Explain why ART/ART2's stable category formation can enhance the consistency of labels used to train an MLP classifier.
4. What preprocessing steps are required to ensure KSOM / Autoencoder outputs are compatible before feeding them into an MLP classifier?
5. How can an automated retraining system be designed so that KSOM/ART recalibration and MLP retraining occur consistently without breaking continuity of customer segment definitions?

PART III: PROTOTYPE DEVELOPMENT

HidupSihat Malaysia is a mid-size fitness and wellness company operating gyms, online coaching, and nutrition programs. As they expand their digital services, the management wants to better understand customer behaviour to design:

- Personalised membership packages
- Targeted marketing campaigns
- Adaptive recommendation systems

For this purpose, the management has collected data from 500 members with eight features. These features are: (refer to file STINK3014_Assignment03_Customer_Lifestyle.csv)

Feature	Description
Age	20–60
BMI	Body Mass Index
Workout_Freq	Days per week exercising
Avg_Steps	Average daily steps
Monthly_Spend	RM spent on fitness services
Stress_Level	1–10 scale
Sleep_Hours	Hours per night
Diet_Quality	1–10 nutrition score

Previously, the traditional segmentation analysis (age, gender, location) has produced mixed results, so the management has requested you to apply unsupervised learning:

- K-Means Clustering for basic grouping
- KSOM (Kohonen Self-Organising Map) for visual pattern discovery and nonlinear clustering

Tasks (*please provide your Python code for this task):

1. **Data Familiarisation**
   a) Describe each variable in the dataset and explain why it is relevant for lifestyle segmentation.
   b) Create a correlation heatmap (you can use a Seaborn library to generate this diagram) and based on the visualizations, suggest **THREE (3)** possible behavioural patterns exist among HidupSihat Malaysia members.

2. K-Means Clustering Tasks

a) Use the Elbow Method to identify the optimal number of clusters.
b) Justify your final choice of K
c) Run K-Means with the chosen number of clusters.
d) Provide the cluster centroids and interpret each centroid
e) Assign a descriptive label to each cluster for the best K-Means model and provide meaningful interpretation based on those optimal clusters (you are required to use any THREE (3) AI Large Language models (e.g., ChatGPT, DeepSeek, Gemini, Claude, etc)

3. Kohonen Self-Organising Map Clustering Tasks

a) Train to cluster neuron weight vectors with *K* from 2 to 8.
    i. Also, you can use the best K that you have obtained from the K-Means clustering tasks.
b) Choose appropriate initial values for learning rate and epochs.
    i. Explain how each initial value affects SOM training and results
c) For each experiment,
    i. Show the final quantisation error (QE).
    ii. Show the topographic error (TE).
    iii. Interpret what QE and TE say about model quality.
d) Assign a descriptive label to each cluster for the best KSOM model and provide meaningful interpretation based on those optimal clusters (you are required to use any THREE (3) AI Large Language models (e.g., ChatGPT, DeepSeek, Gemini, Claude, etc)

4. Analysis and Comparison

a) Compare the SOM Cluster with K-Means directly on the selected dataset. Are the same high-level segments found?
b) Which method (K-Means vs. SOM) discovered more meaningful customer profiles? Why?

Policy:

All grading of deliverables will be based on the standards indicated for each deliverable. Deliverables may not be turned in late, and no cheating! For this class, cheating will include: plagiarism (using the writings of another without proper citation), copying of another (either current or past student's work), working with another on individually assigned work, or in any other way presenting as one's work that which is not entirely one's work. The occurrence of plagiarism will result in removal from the course with a failing grade.