Individual Project ARI5102
Data Analysis Techniques

Konstantinos Makantasis, Charlie Abela

March 3, 2025

This document contains the details for the individual ARI5102 project, which is marked out of 100%; however, it is equivalent to 80% of the total mark for this unit. Questions related to the project should ideally be discussed with the lecturer. While discussion between yourselves is considered healthy, the final deliverable must be produced by you and not plagiarised.

The deadline for this project is 23:59 Monday, 2nd June 2025. Deliverables and the plagiarism form must be uploaded on the VLE in the relevant areas that will be made available. Projects submitted late will be penalised and/or may not be accepted.

1. Introduction

This assignment aims to assess the student's capability in executing a tidy and orderly modelling task, employing a methodologically robust approach. It is important to note that the focus is not on attaining the highest possible values for performance metrics, nor is there an expectation for a particular solution. There is not a definitive right or wrong answer; various strategies may be applied to each subtask. The focus is on your ability to recognize when a decision or assumption is necessary and offer a rationale for the chosen method.

1.1 Dataset Description

You will use the affect modelling dataset in the file named “project_dataset.csv”, available on the VLE. The objective of any affect model is to predict an individual’s emotional state based on affect measurements (such as facial expression features, audio features, and physiology measurements). The dataset employed in this project was developed to analyse and model spontaneous human behaviours in computer-mediated communication settings. It captures interactions between participants in a naturalistic environment, enabling the exploration of human affect and emotional responses.

The dataset is organized as follows:

1. Participant Identification: The first column contains the participant ID.
2. Audio Features: The dataset includes 130 audio features similar to those used in the COMPARE challenge dataset Schuller et al. (2014)[1]. These features capture various vocal characteristics including:
    a. Voice intensity metrics,
    b. Pitch measurements,
    c. Mel-frequency cepstral coefficients (MFCCs).

[1] Schuller, Björn, et al. "The interspeech 2014 computational paralinguistics challenge: Cognitive & physical load, multitasking." *Proceedings INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore*. 2014.

3. Emotional State Annotations: The last two columns contain:
    a. Arousal scores (measuring emotional intensity),
    b. Valence scores (measuring emotional positivity/negativity).
4. Temporal Structure: The data follows a chronological sequence:
    a. Each row represents measurements from a 3-second time window.
    b. Rows for each participant are arranged sequentially.
    c. Example: For participant 1, the first row covers 0-3 seconds, the second row covers 3-6 seconds, and so on.

2. Specifications

The project consists of the following four (4) tasks. You should use the **Python** programming language.

2.1. Task 1 (10 marks)

If you were asked to build a model for predicting arousal and valence, using the provided audio features as explanatory variables:

1. Which performance metrics would you use to evaluate your model's predictions? **(2 marks)**
2. Does the selection of the performance metric depend on the type of the response variables? Explain your reasoning. **(3 marks)**
3. Which validation protocol (e.g., holdout set, k-fold cross-validation, etc.) would you use given that the objective is to build a predictive model able to generalise across participants (i.e., make accurate predictions for unseen participants)? Justify your choice. **(5 marks)**

Note: You don't have to implement anything in Python for this task.

2.2. Task 2 (50 marks)

Using the provided audio features, build predictive models for arousal and valence:

1. Develop a predictive model for each response variable (arousal and valence) using some or all of the provided audio features as explanatory variables. **(10 marks)**
2. Evaluate the implemented models using the metrics and validation protocol you proposed in Task 1. **(7 marks)**
3. Interpret the trained models (if the selected approach allows for interpretation) and the obtained results. **(3 marks)**

A common approach in affect modelling is arousal/valence binarization. Towards this direction, a threshold value is defined and emotional states with arousal/valence values larger than or equal to that threshold are denoted as "high" arousal/valence, while emotional states with arousal/valence value lower than the threshold are denoted as "low" arousal/valence.

1. Select and justify appropriate threshold values for binarizing both arousal and valence annotations (the threshold for binarizing arousal should not necessarily be equal to the threshold for binarizing valence). **(5 marks)**

2. Implement a predictive model for each binarized response variable. **(5 marks)**
3. Select appropriate metrics to evaluate the performance of the model in this scenario using the validation protocol you proposed in Task 1. **(5 marks)**

Emotions are inherently subjective, which can introduce bias during the annotation of affect datasets. One way to mitigate or reduce this subjectivity bias is to formulate affect modelling as a ranking problem. To achieve this, continuous arousal/valence labels are discretised into "high," "medium," and "low" categories.

1. Select and justify appropriate threshold values for discretising ("high", "neutral", "low") both arousal and valence annotations (the threshold for discretising arousal should not necessarily be equal to the threshold for discretising valence). **(5 marks)**
2. Implement a ranking predictive model for each response variable. **(5 marks)**
3. Select appropriate metrics to evaluate the performance of the models in this scenario using the validation protocol you proposed in Task 1. **(5 marks)**

2.3. Task 3 (30 marks)

In this task, you will identify similar observations captured from the first participant (participant ID = 1). Complete the following steps:

1. Create groups of similar observations from the first participant by proposing and implementing two suitable algorithms. **(10 marks)**
2. Evaluate the clusters quality using appropriate metrics. **(5 marks)**
3. Compare the algorithms you implemented and select the best one. **(5 marks)**
4. Create visualisations for the clustering results. **(5 marks)**
5. How would you assign a new observation from the first participant to an existing group? **(5 marks)**

2.4. Task 4 (10 marks)

Based on the results obtained from Task 3:

1. Explain whether the clustering information could be used to build more accurate models for Task 2 and describe what you would do to build such models. (**10 marks**)

Note: You don't have to implement anything in Python for this task. You must discuss what your approach would be.

3. Deliverables

1. **Python Jupyter Notebook**: submit your work via a Python Jupyter Notebook with the following filename *"yourFirstname_yourSurname.ipynb"*. You may import and use any publicly available Python packages you deem necessary.
2. **PDF:** The pdf file generated by the Python Jupyter Notebook.
3. Filled in and signed **plagiarism form.**

4. Final Remarks

Final suggestion: if you have difficulties, do not hesitate to contact the lecturer. Any issues, including technical difficulties, should be identified and highlighted as early as possible to ensure timely resolution.

Good luck!!!