Assignment 2

ENN543, Data Analytics and Optimisation

This document sets out the three (3) questions you are to complete for Assessment 2. The assignment is worth 30% of the overall subject grade. Weights for individual questions are indicated throughout the document. Students are to work in small groups (2-3) and submit their answers in a separate, single document (either a PDF or word document), and upload this to TurnItIn.

The submitted answers should be written in the style of a research report, and outline the methods investigated (and the rationale behind selecting them), how the evaluation was structured (i.e. training and testing splits, etc), the results obtained and the conclusions. One submission should be made per group.

Further Instructions:

1. Data required for this assessment is available on blackboard in *ENN543_Assessment_2_Data.zip*, alongside this document.

2. Answers should be submitted via the TurnItIn submission system, linked to on Blackboard. In the event that TurnItIn is down, or you are unable to submit via TurnItIn, please email your responses to `enn543query@qut.edu.au`.

3. As part of the submission, students should provide a brief table of contributions to outline the contribution of each group member. This table should be signed by all group members to signal agreement. Note that all group members will be assigned the same mark unless one or more members of the group explicitly request that marks be moderated based on contributions.

4. Matlab code or scripts (or equivalent materials for other languages) may be submitted as supplementary material or appendices, however note that this will not be directly marked, and will only be used if there are ambiguities.

5. Figures and outputs/results that are critical to the answer should be included in the main response.

6. Students who require an extension should lodge their extension application with HiQ (see `http://external-apps.qut.edu.au/studentservices/concession/`). Please note that teaching staff (including the unit coordinator) cannot grant extensions.

Problem 1. Clustering (30%). Understanding power use in the home is increasingly important as society strives to improve energy efficiency. Understanding what is normal power use, what is abnormal, and what typical patterns of use are can help owners identify ways to improve their own power efficiency.

The Task

The Household Power Consumption dataset captures energy use in a single home over a period of several years, and can be used to analyse usage patterns and detect periods of abnormal power use. You have been provided data covering a single year (2007) in *household_power_consumption_2007.csv*. The columns in this data correspond to the following variables (in order):

- `date`: Date in dd/mm/yyyy format.
- `time`: Time in hh:mm:ss format.
- `global_active_power`: Household global minute-averaged active power (in kilowatts).
- `global_reactive_power`: Household global minute-averaged reactive power (in kilowatts).
- `voltage`: Minute-averaged voltage (in volts).
- `global_intensity`: Household global minute-averaged current intensity (in ampere).
- `sub_metering_1`: Energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave.
- `sub_metering_2`: Energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry, containing a washing-machine, a tumble-drier, a refrigerator and a light.
- `sub_metering_3`: Energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

Using this data, you are to investigate if usage patterns can be identified in the data, and if abnormal behaviours can be detected. In particular you are to:

1. Cluster the data considering the three sub-meter readings only (`sub_metering_1`, `sub_metering_2`, `sub_metering_3`), using the clustering method (and clustering hyperparameters) of your choice. Justify your selection for the clustering method and parameters based on the requirements of this problem, the nature of the data, and the capabilities of the clustering method.
2. With the clustered data investigate:
   (a) Are trends visible in the clustered data? For example, can changes in use be seen at different times of the year (i.e. summer vs winter), or from a weekday to a weekend, or between different times of the day?

(b) Can any abnormal usage be detected? If abnormalities can be found, show a visual comparison between the abnormal time period and a nearby normal time period (i.e. the previous or next day). The method to select abnormal samples should classify approximately 1% of the data as abnormal. For the purposes of this problem, a period of abnormal usage is a period of 2 hours (or more) where 50% or more of the samples within that two hour block are abnormal.

In completing this question you may also like to consider:

1. Is it reasonable (or practical) to learn the clusters on all the data?
2. Can the data be aggregated in any way to reduce the volume of data? Does such aggregation alter the findings?

Problem 2. Subject Invariant Activity Recognition (35%). A common problem when performing classification and/or recognition of human behaviours, for example speech or activity recognition, is that different people will perform the same behaviour in different ways. In the case of speech this may manifest as different accents, while in the case of activity recognition differences between subjects may be driven by differences in how they move their body to perform the target action. To cope with such variations, models are typically trained on many subjects in the hope that they can capture the different possible manners in which the target behaviours are performed, such that the model can generalise to unseen subjects.

The Task

You are to investigate how feasible it is to recognise the type of activity being performed by a person in a subject invariant manner, i.e. the test set is such that it contains subjects who are not in the training or validation sets. You have been provided with data for 10 subjects that shows 6 activities (walking, walking upstairs, walking downstairs, sitting, standing, laying). The data is split across three files as follows:

- `wearables_signal.csv`, which contains 3,237 samples of 561 dimensional wireless sensor data captured as subjects perform actions;
- `weatables_activity.csv`, which contains the ground truth activity for each sample in `wearables_signal.csv`;
- `wearables_subject.csv`, which contains the subject ID for each sample in `wearables_signal.csv`.

Using this data, you are to develop two models to recognise the activities of an unseen subject, i.e. having trained the model on a selection of subjects, evaluate how well the model recognises the activities of a subject who as been held out of the training set. Your chosen approaches must:

- Include one approach that uses a dimension reduction method of your choice;
- Use a five-fold evaluation scheme, such that each fold contains 8 subjects in the training and validation sets, and the remaining 2 subjects in the test set, and the accuracy of a given model is the average of the performance across the 5 folds.

In addressing this problem you are free to select your own classification approaches from those covered in lectures and tutorials (though one must employ dimension reduction). Your answer should explain the methods you have chosen to use, and provide justification your choices. You may optionally wish to include small scale experiments to support your decisions.

Your answer should provide an analysis and discussion of your models performance, should identify situations where the model fails (and if possible reasons for this failure), and should consider if performance is consistent across all subjects.

Problem 3. Recognising a Person's Age from Their Face (35%). Age estimation is a widely studied task relating to facial recognition, with applications in domains such as biometrics, and human computer interaction. Estimating age from facial images suffers from many of same challenges as face recognition, such as variations in appearance caused by pose, lighting, and facial accessories (i.e. glasses) or facial hair. Much like facial recognition, a critical pre-processing step for age estimation is to localise and align the face, such that all examples are as consistent as possible in terms of the location of major landmarks such as the eyes and nose.

The Task

You are to develop two methods to estimate the age of a subject from a facial image. The file `UTKFace.zip` in the `Q3` directory within the data archive contains the aligned and cropped face images from the UTKFace dataset[1]. This archive contains 20, 000+ colour face images, all of which have been cropped and aligned in preparation for further processing. A selection of example raw images (i.e. uncropped images) are shown in Figure 1.

Figure 1: Example raw images from UTKFace. Note that the supplied cropped and aligned images contain only the face regions.

Faces in the archive are named as follows: `[age]_[gender]_[race]_[datestamp].jpg`, where:

- `[age]`: an integer from 0 to 116, indicating the age;
- `[gender]`: either 0 (male) or 1 (female);
- `[race]`: an integer from 0 to 4, denoting White, Black, Asian, Indian, and Others (like Hispanic, Latino, Middle Eastern);
- `[datestamp]`: date and time stamp, in the format of yyyymmddHHMMSSFFF, corresponding to the date and time an image was collected to UTKFace.

[1] see `https://susanqq.github.io/UTKFace/` for more details

The `[age]` value is to be the primary response of your model. You may use or ignore the other variables as you choose. A simple MATLAB script has also been provided on blackboard with this assignment to load and resize the images.

In addressing this problem you are free to select your own approaches from those covered in lectures and tutorials to determine age. Of your two approaches, one must use a dimension reduction method of your choice. You may also consider the task in one of two ways:

- A *regression* problem, where the task is to regress from the image to the age;
- A *classification* problem, where the task is to classify the image as the correct age. For this approach, you may wish to consider whether it is appropriate to classify the age exactly, or to classify the age within a small range ($5 - 10$ years).

Given the large size of the database, you are welcome to down-sample the images to a lower resolution, though be aware that if you are too aggressive in your down-sampling you may lose the ability to estimate age (images smaller than 32×32 pixels are not recommended).

Your answer should explain the methods you have chosen to use, and provide justification for your choices. You may optionally wish to include small scale experiments to support your decisions.

Any approaches used to modify the data (down sampling, colour conversion, etc) should be documented, as should the training and testing splits. Your answer must also provide an analysis and discussion of your models performance, including identifying situations where the model fails and possible reasons for the failures.