Group Project
COMP9417 Machine Learning and Data Mining
T3, 2019

Introduction

This is a group project that will be done by a team of 4-5 students and the aim is to apply machine learning techniques to predict some specific outputs in a dataset.

The first step is to go to the course Moodle page and in the Moodle/Homework & Assignment/Assignment (Group Project)/Group_Project_Member_Selection create your groups.

Group project contributes to 30% of the total mark (**30 marks**). The deadline to submit your report is **Tuesday 26 November, 5:00 pm**.

Submission will be via the Moodle page.

Recall the guidance regarding plagiarism in the course introduction: this applies to this report as well and if evidence of plagiarism is detected it may result in penalties ranging from loss of marks to suspension.

Dataset

The *StudentLife* dataset will be used in this project. This dataset is a collection of sensing data from the phones of 48 Dartmouth students over 10-week term to assess their mental health, academic performance and behavioural trends.

The objective of this project is to predict two psychology-related phenomena using the "*sensing data*" from mobile a mobile app. The first variable to predict is the flourishing scale, which is a measure of self-perceived success, and the second is PANAS scores, which is a measure of positive and negative affect. These two measures are collected through self-reported questionnaires. These scores can be treated as continuous variables, however, in this project we aim to do classification as well, therefore you can use a threshold to divide the scores into two groups of "*High*" if the value is higher than the threshold, and "*Low*" if the value is less than the threshold, and then perform classification. The expected predictions can be in the form of regression and/or classification.

The features that will be used in this project are sensing data which has been collected using automatic sensors. These include physical activity, audio activity, conversation start/end time, GPS location, Bluetooth data, WiFi, WiFi location, light start/end time, phone lock start/end time, phone charge start/end time.

One paper which describes the data collection with some preliminary analyses has been included in the data folder and the dataset is publicly available through: https://studentlife.cs.dartmouth.edu/

A detailed description of different sensor data, their representation in the dataset and their values are provided in: https://studentlife.cs.dartmouth.edu/dataset.html which you can use to understand the data structure and meaning.

Project implementation:

Each group has to implement a minimum of three methods. Each method can be a classification or regression. You are free to select the features, pre-process the features or create new features from the available ones. You are also free to choose your method for classification or regression even if the method has not been covered in the course. You can use any open-source library you need for your implementation.

The data for this project does not include all the collected data, therefore you can download the project data from course Moodle page where the *Inputs* folder includes all variables that you can potentially use as your features/attributes and the *Outputs* folder contains the answers to survey questionnaire for each participant (please note that, participant ID is coded as "u_xx").

To compute the scores for your output variables, you can consult the provided .pdf files in the *output* folder as your data to calculate the score for each measure. Flourishing score gives one measure and PANAS includes two measures: one for positive affect and one for negative affect. Therefore, in total, there are three measures to predict. For binary classification, you need to divide the scores into two groups ("*high*" vs "*low*") using a threshold. You can choose this threshold to be the median value in the entire dataset for each measure separately. Using the median value as your threshold divides your data into two balanced classes of almost same size, but if you choose to divide your data into two or more than two classes in another meaningful way, that is still fine.

You are free to use all the provided features or a subset of features or your engineered features, however you are expected to give a justification for your choice. You may run some exploratory analysis or some feature selection techniques to select your features. There is no restriction on how you choose your features as long as you can justify it.

Each implemented method has to be applied on both Flourishing scale and PANAS scales and results have to be compared. You have to use cross validation method to tune the hyperparameters of your models and evaluate it on unseen data. You are free to choose the number of folds if you use k-fold cross validation. You are also expected to discuss briefly the importance of features in each of your models.

Report:

Each group has to submit one report which contains introduction, dataset, methods and evaluation, results, discussion and conclusion. The report is expected to be 12-15 pages (with single column, 1.5 line spacing).

Here is guideline for the report:

- Title page: title of the project, name of the group and group members

- Introduction: a brief explanation of the problem, the aim of the project and methods
- Dataset: description of the dataset, binarization method (how you create your classes)
- Methods: A detailed explanation of all methods developed, features used/engineered, hyperparameter tuning method, cross validation, evaluation metrics, design choice, etc.
- Results: Presenting the results of each method, important features and the selected hyperparameters
- Discussion: Compare different methods, their features and their performance on different output variables.
- Conclusion: Give a summary of the project and the findings
- Reference: list of all literature that you have used in your project

Peer review:

Individual contribution to the project will be assessed through a peer-review process which will be announced later, after the reports are submitted. This will be used to scale marks based on contribution.

Anyone who does not complete the peer review by the Thursday of Week 12 (5 December) will be deemed to have not contributed to the assignment. Peer review is a confidential process and group members are not allowed to disclose their review to their peers.

Project help:

General questions regarding group project should be posted in the Group project forum in the course Moodle page.