CO 544- Machine Learning and Data Mining
Project Specification

01.03.2020

1 Overview

Machine learning and Data mining can aid in solving problems which may include a large amount of heterogeneous data. In this project we will explore the applicaton of machine learning and data mining to solve a classification problem.

2 Dataset

In order to maintain the confidentiality, the attribute names have been altered. The attributes can be continuous, nominal with small numbers of values, and nominal with larger numbers of values. The attributes details are as follows.

Attributes Details

A1: b, a
A2: continuous
A3: u, y, l
A4: g, p, gg
A5: continuous
A6: c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff
A7: continuous
A8: TRUE, FALSE
A9: v, h, bb, j, n, z, dd, ff, o
A10: continuous
A11: TRUE, FALSE
A12: continuous
A13: TRUE, FALSE
A14: continuous
A15: g, p, s
A16: Success,Failure(class attribute)

3 Submission

Your task is to submit the predictions (the class label) for a given set of data. A randomly selected evaluation dataset (which is not included in the dataset provided) will be released in the last week of the project.

4 Assessment

The submission artifacts and the marks allocation is as in Table 1.

1. Report and Source Code

You have to submit a report outlining the work done in the project in the form of XX.pdf where xx is your group name. The report should include

- A brief introduction to the problem

Milestone	Due By	Marks(/35)
Report and source code	6^{th} of April 2020	20
Group Presentation	Week 10	10
Accuracy and Overall Rank	6^{th} of April 2020	5

Table 1: Marks Allocation

- Any explorations conducted on features.
- Description of your final approach(s) to predictions, the motivation and reasoning behind it.
- Alternatives considered, and why you choose the final approach.

The source codes of the analysis should also be submitted.

2. Group Presentation

Each group (member) has to present their work including final model to the class using one slide (common to the group) within 5 minutes allocated to their group.

3. Accuracy and Overall Rank

A randomly set out sample from the dataset will be provided to you in Week 9. You have to submit your predictions to the mentioned dataset. The accuracies of each group submission will be calculated and the groups will be ranked based on accuracy. Based on that, you will be given a mark (out of 5) as below: Assuming there are no ties, there are N submissions and your group gets the rank R, with accuracy A the mark is calculated as

$$4 \times \frac{max\{min\{A, 0.9\} - 0.4, 0\}}{0.5} + \frac{N - R}{N - 1} \tag{1}$$

This expression can result in marks from 0 to 5. The accuracy is weighed higher than the class rank. The component of 4 will lead to an accuracy of 0.9 or higher getting 4/4 while an accuracy of 0.4 or lower getting 0/4; and linearly scales over the interval of accuracies [0.4,0.9]

Plagiarism Policy

This is a reminder that all your submitted project work in this subject should of your own. Accordingly, copying from other teams is not permitted.