Machine Learning Assignment 2

Dr Burcu Can

February, 2021

1 The Dataset

A paper reviews dataset is given in Spanish language. The dataset involves reviews for scientific papers along with the reviewers' evaluations for each paper.

Each paper has the following information in JSON format:

```
timespan, paper id, preliminary decision, review ID, text, remarks, language,
orientation, evaluation, confidence.
```

You will only use the evaluation and the review columns in the dataset.

2 Task 1

You will build a linear regression model preferably using Scikit-learn library. You will use unigram bag-of-words features as binary features. Make up few test reviews which do not exist in the training set and discuss the review evaluation prediction for each test review. Discuss your findings in your report.

3 Task 2

Give the most important 10 featues (i.e. words) in positive and negative reviews.

4 Report

Write a report on your findings. Enclose your source code.

5 Submission Information

- Length: Maximum 1,000 words in total
- The source code also needs to be submitted along with the report.
- Format of the files to be submitted: a zip folder including your report in .pdf format, and your source codes.
- Deadline: 10 June 2021, 12pm (GMT)
- Submitted via Canvas

https://archive.ics.uci.edu/ml/datasets/Paper+Reviews

6 Marking

- Task 1: 50%
- Task 2: 30%
- Overall report: 20%

https://archive.ics.uci.edu/ml/datasets/Paper+Reviews