Coursework - 2: Sentiment Analysis

Vinay P. Namboodiri

March 23, 2022

Coursework number	2
Percentage of unit	20%
Submission location	Moodle
Submission deadline	**29th April 2022 at 8 PM**
Submission components	PDF report and Jupyter notebooks in a zip file
Anonymous marking	No, in groups

1 Overview

In this coursework the aim is to understand about the specific machine learning techniques covered in the second half of the course. The topics relate to solving an NLP task of sentiment analysis using support vector machines and boosting techniques.

2 The Task

The task in this assignment is to solve an NLP task related to sentiment analysis using support vector machines and boosting techniques. The data is a subset of the IMDB movie reviews dataset. There are two sentiments - positive or negative associated with each movie review. The coursework is to be done *in groups of up to 4 persons*. A Jupyter notebook is being provided as a template to start the assignment and a train and a sample test file are provided for the assignment.

3 Groups

The task needs to be solved in groups of up to 4 persons. The same group for coursework 1 could be followed. Alternatively, you need to specify a group

you would like to work with and note that in the following Google form provided for the same `https://forms.gle/kaVaX9P9U763EhsBA`.

Similar to coursework 1, a common group mark will count for 80% of your allotted marks and the remaining 20% will be obtained from a group peer-review that will have a later deadline of 2^{nd} May 2022 8 PM.

4 Details about the tasks

4.1 Classification task

The assignment consists of two separate tasks for the dataset provided of predicting the sentiment given a movie review. The training dataset consists of a number of movie review sentences and the corresponding sentiment 'positive' or 'negative'. The test dataset has a similar format and a sample test file is provided. The final test file is an expanded test file with 1000 additional test sentences in addition to the 1500 sentences provided in the sample test file

4.2 Marking procedure

. The coursework will be evaluated in two phases. One is based on accuracy on a test dataset that is the expanded test file. The second is based on the analysis. Thus 20% of the marks are evaluated based on group peer-review. 40% of the marks are evaluated on the basis of accuracy (20% for each task) and 40% of the marks are evaluated based on the analysis (20% for each task analysis).

Distribution of marks for accuracy

- Half of the marks for accuracy are awarded if the classifier scores above 60% on the test set.
- Remaining marks are scored linearly between 60% to 90% based on the accuracy obtained for classification on the test set. Any score above 90% is given full marks. This evaluation is done through an automated test script.
- One Jupyter notebook must be submitted that has the final code you want to use for evaluation using the auto-marker for accuracy. You should follow the guidelines for the automarking notebook as much as possible. Especially, it should be possible to evaluate your code using the evaluation function of the notebook.

Distribution of marks for analysis:
The analysis is evaluated based on the report and the accompanying code. Note that the analysis includes graphs and needs to be shown through a separate analysis Jupyter notebook that is separate from the one used for automarking the assignment. You can freely modify the contents of the provided Jupyter notebook to obtain the analysis Jupyter notebook. Further details about analysis are provided in each classification task

4.3 Task1: SVM Classification task analysis

In this task, you need to use SVMs for classifying the sentiment. Especially, you need to experiment with different kernels. You need to write at least one custom kernel and evaluate your system with at least 3 kernels (including built-in kernels). You are allowed to use scikit-learn or numpy libraries for this assignment. Note that the custom kernel you write should not be a built-in scikit-learn function. The kernel you use for the final auto-marking can be either a built-in kernel or a custom kernel.

The marks for analysis of the classification task using SVMs (total 20 marks) consists of the following sections:

- Evaluation and comparison of 3 kernels on the train and test dataset. Further, correct use of cross-validation to tune the hyper-parameters for each kernel if applicable for the kernels. (13 marks)
- Analysis of mis-classified test examples to understand failures (for instance, based on length or specific word characteristics) (5 marks)
- Clarity of code (2 marks)

These will be evaluated through the report you provide (at maximum 2 pages for the SVM analysis) and the accompanying Jupyter notebook used for analysis

4.4 Task2: Boosting task analysis

In this task, you need to implement a Boosting algorithm to solve the binary classification task of sentiment analysis. You can use boosting built using decision trees and you are allowed to use the decision trees from scikit-learn library.

The marks for analysis of the classification task using boosting consists of the following sections:

- Evaluation and analysis of various parameters of the boosting algorithm, the pre-processing routine for the data and the decision tree routine using cross-validation (12 marks)
- Analysis of mis-classified test examples to understand failures. This could be for different kernels. (for instance, based on length or specific word characteristics) (5 marks)
- Clarity of code (3 marks)

These will be evaluated through the report you provide (at maximum 2 pages for the boosting analysis) and the accompanying Jupyter notebook used for analysis

4.5 Marking Rubric

The lab report will be evaluated on the basis of breadth and depth of the analysis carried out and will be graded using the following rubric in the range 0 to 5:

- 0 - No analysis provided
- 1 - Preliminary analysis provided for the task that is insufficient
- 2 - A basic analysis has been provided for the task. It is lacking in breadth and depth (for instance, 3 kernels not evaluated)
- 3 - Sufficiently clear analysis has been provided for the task. Some limitations in terms of either breadth or depth (for instance, parameters of the kernel not evaluated)
- 4 - Very good analysis has been provided for the task. Covers both breadth and depth. (for instance, all 3 kernels evaluated and parameters related to the kernel also evaluated)
- 5 - Excellent analysis with good insights being observed and sufficiently broad (for instance, a particular kernel has a sensitivity in a particular parameter range, but is otherwise broadly stable and supported with a graph)

5 Late submissions

The university policy will be followed on late submissions. If a piece of work is submitted after the submission date, the maximum possible mark will be 40% of the full mark. If work is submitted more than five days after the submission date, student will receive zero marks. If you need an extension, please contact your Director of Studies.

6 Plagiarism

Do not plagiarise. Plagiarism is a serious academic offence. For details on what it is and how to avoid it, please visit `http://www.bath.ac.uk/library/help/infoguides/plagiarism.html`