CSCI218 Foundations of Artificial Intelligence

Assessed Lab 1 (5%)

Due date:

Text Classification with Naïve Bayes Classifiers

Overview

Text classification is an important task in natural language processing. It aims to categorise a given document, paragraph, or a sentence into a set of predefined classes. Text classification has applications in business, social media, electronic health record, education, to name just a few. In this assignment, you will conduct text classification by using Naïve Bayes Classifier on a benchmark dataset.

You are provided with the 20 newsgroups dataset and the code explained in Week 6's tutorial. They can be obtained from Moodle by downloading the file "*20-newsgroups_text-classification- master.zip*." [1] under "Week 6 Lab."

What you need to complete this lab:

- The **20 newsgroups dataset** and the **code** *("Multinomial Naive Bayes- BOW with TF.ipynb")*.
- The lecture & tutorial content and recordings in Week 6.
- Python 3 programming environment with required libraries, packages, and modules.

Objectives

- Understand Bayes' formula and Naïve Bayes Classifier.
- Understand text classification and the pre-processing procedure in natural language processing.
- Learn to conduct text classification on a benchmark data set.
- Learn to use libraries, packages and modules related to text classification.

Questions (5 marks)

Run the code "*Multinomial Naive Bayes- BOW with TF.ipynb*" (further modifications in the code may be required) and answer the following questions:

1. Describe the key steps for data preparation and feature extraction **(1 mark).**

2. Report the overall classification results, including precision, recall, and f1-score. Explain the meaning of these criteria **(1 mark).**
3. Plot the confusion matrix for your classification result. Find the pair of classes that confuses the classifier most. Is this result consistent with your expectation? **(1 mark).**
4. Based on the confusion matrix, report the individual accuracy scores for each class **(1 mark).**
5. Train a Complement Naive Bayes classifier and compare its classification results with those of Multinomial Naive Bayes **(1 mark).**

Submission

- Submit **a single PDF file** which contains your answers to the questions of all tasks. All questions are to be answered. A clear and complete explanation needs to be provided with each answer.
- You must show your name and student number on the first page of the PDF report.
- Your PDF report should begin with a short introduction to the lab and the dataset.
- Submit the PDF file via the submission link on Moodle.
- The PDF file must contain typed text of your answers (**do not submit a scan of a handwritten document**. Any handwritten document will be ignored). The document can include computer generated graphics and illustrations (hand-drawn graphics and illustrations will be ignored).
- The PDF document of your answers should be no more than 4 pages including all graphs and illustrations. Appendix is allowed and will not be counted for the 4-page limit.
- The size limit for this PDF report is 20MB.
- Late submission will not be accepted without academic consideration being granted.

References

[1] https://github.com/gokriznastic/20-newsgroups_text-classification.

**** END ****