Assignment 1

Machine Learning II
Session: 2017-2018
MSc Computer Science and Msc Big Data Analytics
Ramakrishna Mission Vivekananda University

You have to work on a problem of text classification. The description about the data and the tasks are given below. Subsequently, the guidelines of submission are provided. **The full marks of this assignment is 50**.

Data

We will consider the Renters-21578 data set. It is new wire of 1987.The documents were originally assembled and ordered with categories by Carnegi Group Inc. and Returns Ltd. The corpus originally contains 135 categories and the categories are overlapped i.e., one document may exists in several categories. Hence we will consider the Mod Apte version of Renters.The ModApte version contains 12902 documents with 90 categories and the corpus is divided into training and test sets. In this assignment we will consider only the following 10 categories of the ModApte version of Returns-21578 corps:

alum, barley, coffee, dmk, fuel, livestock, palm-oil, retail, soybean, veg-oil

The class label of a document is the name of the directory to which it belongs to. A link to download the corpus will be given to you over email.

Tasks

You have to perform the following tasks using the given Reuters Corpus with 10 categories

1. Remove the stop words from the raw text after tokenization and then map each token to its stem using a stemming algorithm.

   [5]

2. Create the term-document matrix by using an efficient tf-idf weighting scheme.

   [5]

3. Perform Naive bayes, logistic regression, support vector machine and multilayer perceptron classifiers to categorize the documents in the test set of the given Reuters corpus. The aim is to achieve the best performance for each of these classifiers by properly tuning its parameters and by choosing an efficient version of the classifiers using the training set of the given corpus. Properly explain your selection using experimental results. You may consider different types of feature combinations e.g., unigrams , bigrams etc.

   [15]

4. Evaluate and compare the performance of the classifiers using the actual class labels of the test samples. Properly explain and analyze the results. Discuss about significant findings from these results. Conclude with future scope of research (if any) of this assignment.

   [25]

Submission Guidelines

Write a report to describe the given tasks. The report should be submitted in PDF. The result should contain the following information:
A) The first page should contain the following information:

Name

Registration no/Roll no

University name

Program Name - BDA/CSE

Problem Release date:- 12/09/2017

Date of Submission:- 12/10/2017

B) From the second page onwards you have to provide the following information.

- A suitable title
- **Introduction:** This section should address the problem and the plan of actions. Describe about the dataset here.
- **Methods:** Explain the background of the classifiers in few sentences and with proper references. Discuss the idea of parameter tunning with proper references. You may also describe about other experimental settings here e.g., the tools used.
- **Evaluation Criteria:** Describe about Precession, Recall, F-measure, ROC curve etc. You may also discuss the micro-averaged and macro-aeraged technique.
- **Analysis of Results:** You may create different tables to show the experimental results. Moreover, you should explain (in simple sentences) your observations from these tables. Therefore analyze the results based on these observations as disucussed in the class.
- **Discussions and Conclusion:** Significant findings and scopes of future works may be explained here in few sentences.

Other Relevant Information

- You may send the codes in separate files along with the report, but this is not compulsory.
- **Deadline of the submission is 12/10/2017, 23:59 IST**.
- Multiple submissions are allowed within the deadline, but only the last submission will be graded. You should send the submission to welcometanmay@gmail.com.
- **For everyday that your submission is late your score gets multiplied by 0.8.**