Natural Language Processing – CS 4120

Assignment 4 – Section I

Due Date: 11:59pm, 02/19/2024

Total Mark: 100%

Machine Learning

Introduction

In previous lecture, you learned about supervised and unsupervised machine learning. This assignment is going to evaluate your gatherings on these two topics.

PART I: (50%)

Text Classification Using TF-IDF and Logistic Regression

Objective:

In this section, you will develop a text classification model using TF-IDF (Term Frequency-Inverse Document Frequency) for feature extraction and a Logistic Regression classifier for categorization. The goal is to understand how to process textual data, transform it into a suitable format for machine learning algorithms, and apply a classification model to predict the category of unseen documents.

Dataset:

You will use the "20 Newsgroups" dataset, a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. This dataset is widely used for experiments in text applications of machine learning techniques, such as text classification and text clustering.

Tasks:

Data Preparation:

Download the "20 Newsgroups" dataset. You can use sklearn's fetch_20newsgroups function for this purpose.

Perform basic text preprocessing steps including tokenization, removing stop words, and stemming or lemmatization. You can use built-in functionalities in NLTK library.

Feature Extraction:

Convert the preprocessed text data into numerical features using TF-IDF vectorization. Utilize TfidfVectorizer from sklearn.

Model Training:

Train a Logistic Regression classifier on the training set. You can use LogisticRegression from sklearn.linear_model.

Model Evaluation:

Evaluate the performance of your model on the testing set. Report metrics such as accuracy, precision, recall, and F1-score as well as confusion matrix.

[Optional] Perform error analysis by identifying the types of errors your model makes. Discuss possible reasons for these errors and suggest ways to improve the model.

NOTE: A jupyter notebook is provided for part I.

PART II: (50%)

As one of the unsupervised learning approaches, we talked about Clustering algorithm. The first step of cluster algorithm is initializing the clusters (in terms of their location).

Conduct research on what are the most common approaches for the initialization step of a clustering algorithm. You can refer to literature survey or come up with your own approaches. Provide a short summary for each approach (no less than two methods) and discuss pros and cons of each approach in terms of performance. If you are referring to a prior work, include a reference.

Deliverables

- A Python script or Jupyter notebook containing all the functions and their documentation.
- A report explaining your methodology, results and a discussion on the results.