Kennesaw State University
Department of Computer Science

CS4742:Natural Language Processing

Fall 2024

Assignment 1 (Due: Oct 14th,2024) Instructor: M. Alexiou

As we discussed during the lecture, logistic regression is a popular statistical model used for binary classification tasks, where the goal is to predict one of two possible outcomes. It estimates the probability of a given input belonging to a particular class by fitting data to a logistic function, which outputs a value between 0 and 1. In sentiment analysis, logistic regression is commonly used to classify text data into categories such as positive or negative sentiment.

You are asked to form teams of maximum 2 students and implement the logistic regression algorithm in Python to predict either negative or positive reviews (binary classification), training and testing it using the Amazon product review dataset available in the following link: https://github.com/MuhammedBuyukkinaci/TensorFlow-Sentiment-Analysis-on-Amazon-Reviews-Data/blob/master/dataset/

Implementation Instructions:

(a) You are recommended to base your implementation on the tutorial available on the lectures. Please keep in mind that the goal of this exercise is to experiment with the concept of logistic regression. Therefore, you are asked to experiment with different learning and optimization parameters (including activation functions) for comparison and provide the results of at least 2 implementations in terms of (1) accuracy and (2) efficiency. In this case, efficiency is defined as training and inference speed.

(b) The attached dataset is large and not necessarily all datapoints are needed to train and evaluate your model. This provides you with the opportunity for additional experimentation starting from 40k reviews, until 80k reviews to measure the impact of training size to the accuracy.

(c) Please keep in mind that you cannot feed directly raw text to the logistic regression model as discussed in the class. Therefore, you should convert it first to a vector with numerical features using techniques such as Bag of Words to enable your logistic regression models to learn patterns in the data that correlate with different sentiments.

(d) Prepare a report in the form of a PDF file were you describe the different parameters used during your experimentation and discuss their corresponding accuracy results. Specifically for the accuracy results, you are also asked to include confusion matrices with True Positive, True Negative, False Positive and False Negative numbers.

(e) Upload one zip archive per team (with a README file that contains team member names and email addresses, source code file(s) and PDF report, to HW1 on Assignments on D2L by October 14th, 2024. (Only one submission per team!). Your assignment will be graded based on the aforementioned implementation instructions. Late submissions will be penalized at the rate of 10% / day.