Mary Clay

Foundations of Machine Learning

27 October 2022

Midterm Solution

This document details the way in which my solution to the midterm project was constructed.

I designed a logistic regression model to complete this task. The first step I took, after reading the training set, was to tokenize the data. This involved using the regular expressions library to convert each sentence into a list of words. After this, in order to convert the tokenized sentences into something the logistic regression model can interpret, I used the CountVectorizer library to convert each word into a vector. This part of the program outputs a matrix of token counts. After doing this, I trained these vector arrays on the logistic regression model, and performed the same prior steps on the development set, and predicted the values of the development set. For my final product, I performed the same steps: tokenizing, vectorizing, and training on the development set, and predicting on the test set. I then created a function to output my predicted labels for the test set to a file, ‘MaryClay\_test\_result.txt’.

The libraries I used were Pandas, Numpy, Sklearn (Logistic Regression, Accuracy Score, and CountVectorizer), and Regular Expressions.

This project made me learn a lot. I had prior experience with Numpy, Pandas, and logistic regression, but had never had to convert categorical features or tokenize them. I learned a lot about how CountVectorizer works in particular. My biggest roadblock in completing this assignment was the inability to use deep learning techniques. I also had some trouble with overfitting (I got around 85% accuracy at first, but realized it was performing poorly on the development set).