Assignment NO 2
Course: Machine Learning

Movie reviews

Mohsen M. Kivi

513530

May – 2024

This report addresses the queries outlined in the homework PDF file. The analysis was conducted using Pycharm and Github version control, where coding was performed to achieve the objectives of the assignment

Please contact me for any potential problems with accessing the scripts and reports via Email.

Statement of Academic Integrity:

I affirm that this report is the result of my own work and that I did not share any part of it with anyone else except the teacher.

a. Train model

The first model has been developed so far. the process done so far is building a vocabulary with the size of 1000 from the dataset. then a process has been taken to extract the features of the vocabulary. Then, the NB classifier is trained using the train_nb function. The function computes the probability distributions for the positive and negative classes and returns the weights w and bias b of the NB classifier. The trained model is evaluated on the training set using the inference_nb function, which takes the feature matrix X, weights w, and bias b as input. The function computes the predicted sentiment labels and returns the accuracy of the model on the small training set.

The code reports the small training accuracy of the NB classifier, which is approximately 82%. The model is able to correctly classify 82% of the movie reviews in the small training set.

In the next step, we will try to find training accuracy for the real train set, validation set, and test set.

Result:

Training accuracy: 81.904

Testing accuracy: 50.0

Validating accuracy: 50.0

It seems that the model is not working well for the data.

b. Variants

1. Let’s make the model to ignore the common words in stopwords.txt

To do that after loading the common words as a list, we define *not_common_word(word)* function to see if a word is common or uncommon. Then we open and read our main list of words and by applying the above condition we exclude the common words from that then we build a vocabulary with a size of 1000, and then save it as a new vocabulary called, *vocabulary_not_common.txt*

2. S

c. Assignments

1. Vocabulary size

we increased the size of the vocabulary from 1000 to 10000, to see how the training accuracy would be affected.

Since the datasets are very big and time-consuming, we will just test it on the small training set.

The Training accuracy would be: 87.35199

2. SVM & LR

Let's see how SVMs work for this problem. We load the SVM training model from the pvml directory of Git Hub. And then we try to train the model for our small training set. That gives us:

Training accuracy: 73.712

For linear regression, after doing the same procedure we get:

Training accuracy: 74.99