Artificial Intelligence II

IMDB Sentiment Analysis - Logistic Regression

Georgios - Alexandros Vasilakopoulos
A.M. 1115202000018

Data Preprocessing

- Numerical values were removed from the training data, since, in most cases, numbers do not imply sentiment.
- The tfidf vectorizer was applied to transform the training data and the 'english' stopword set was utilized in order to remove common english words that do not indicate sentiment.

  Tfidf was preferred over CountVectorizer because it takes into account the document frequency and, also, it reduces the impact of frequent, less informative tokens.

  By default, the tfidf vectorizer replaces upper-case letters with lower-case and removes punctuation in the final tokens. This, of course, is desirable, since punctuation usually does not indicate sentiment and the existence of feature-words with capital letters would confuse the model.

  Furthermore, the parameter max_df was set to 0.95, so that words with frequency > 0.95 are ignored. Such words are found both in positive and negative reviews and, therefore, it is safe to assume that they can be neglected.

  Also, the parameter min_df was set to 0.001. This value was produced by the gridsearch algorithm as optimal. As the values of min_df decrease down to 0, the number of features increases, which means that the complexity of the model increases.

Model Options

- Throughout all the tests that were conducted, the number of iterations never exceeded the common value of 1000, therefore, in all the tests, the value of max_iterations will be fixed to 1000.

- One important decision to be made is whether to use l_2 regularization, l_1 regularization, or a combination of the two (elasticnet).
- As mentioned previously, the tfidf's mid_df variable was set as a model parameter, as it effects the complexity of the model.

GridSearch

- The gridsearch algorithm was used in order to find the optimal values for the parameters, exhaustively.
- Due to the rather large set of possible parameters of the model, each possible model was tested on just 50% of the training data set, in order to shorten the training time.
- The models were evaluated through the f1 score, because it takes into account both precision and recall and, generally, it is considered a reliable all-in-one metric.
- The results indicate that the best performing regularizer is the l2-norm with $C = 2$, while the best value for min_df is 0.001 (which was the smallest value that was tested). It seems that the more features that we include, the better the performance, but there is a small tradeoff in the fit_times and score_times.
- I avoided to include elasticnet in the gridsearch algorithm because the total testing time would increase by a lot, due to the (relatively) slow convergence of the saga solver on the training set. Nevertheless, since the best performing model uses the l2 norm, if we included elasticnet in the gridsearch, the produced model would either use the l2 norm again, or elasticnet, but with an l1_ratio close to 0.

Evaluation of the Produced Model:

- The model that resulted from gridsearch produced the following metric scores, after cross validation on the entire data set:

  F1-score: 0.895

  Precision: 0.887

  Recall: 0.902

- The first and third graphs of the learning curves show that the model indeed does not overfit or underfit, as the performance of the model increases up to a certain amount, as the training examples increase.
- It seems that the relationship between the number of training examples and the time required to train is -somewhat- linear, as expected.

Comments on the predictions

- After examining some reviews and comparing the predictions to the actual sentiment classes, one can notice that the model is good at predicting reviews that are either heavily positive or heavily negative. That is probably because such reviews contain keywords that imply sentiment of a single class.
- On the other hand, the model sometimes misspredicts reviews that are somewhat mixed, due to the existence of keywords that belong to both sentiment classes
- In addition, there are some cases where the review also contains details of the plot. This may also confuse the model if the sentiment of the movie differs from the sentiment of the review.
- If our model was able to combine somehow the meanings of words in sequence, then it might have predicted such reviews more accurately. For example, many lengthy reviews contain expressions like ”all in all”, ”overall” etc, which describe clearly the overall sentiment of the review.