YS19 - Artificial Intelligence II
Project 1

Efstathios Chatziloizos - 1115201800212

November 2022

A sentiment classifier using logistic regression for imdb movie reviews has been developed. The classifier deals with 2 classes, negative and positive, as per instructions.

The given dataset is partitioned into training and validation datasets using sklearn's **train_test_split**. Before the split the data has been randomly shuffled.

Cross validation is used to find the best hyperparameters, as well as for the plotting of the learning curve. Regularization is utilized to avoid overfitting the model. More specifically, the C parameter is used (the lower the value of C, the higher the regularization is).

The classifier of choice is sklearn's **LogisticRegression**. For the vectorizing, the process used to convert a collection of text documents to a matrix of token counts, two vectorizers where used: sklearn's **CountVectorizer** and **HashingVectorizer**. The classifier is evaluated using precision, recall and F-measure.

The hyperparameters are chosen with the help of sklearn's **GridSearchCV** using cross-validation. However, it has to be noted that a plethora of examples and combinations of parameters have been tried, only a small sample of which has been added to this report. In this implementation the f1_macro is optimized, although there were no significant differences while optimizing different macros. A much wider range of parameters than those seen in the final source code were tested using **GridSearchCV**, but only some of the most important of said parameters are present in the source code to reduce the running time.

To plot the learning curves, sklearn's **learning_curve** using cross-validation has been used. The curves are plotted according to their f1_macro. The training score is in red, whereas the cross-validation score is in green. The curves are used to show whether the model underfits or overfits.

Extensive comments have been used throughout the whole source code, explaining every step of this implementation.

The code was written using VSCode, although the code was also tested on https://colab.research.google.com/

Model Chosen - Count Vectorizer with best parameters

The final model presented uses a **CountVectorizer**. The best hyperparameters using **GridSearchCV** are chosen and the model is trained using **LogisticRegression**. The model does not seem to overfit.
Parameters:
{C: 0.1, class_weight: None, max_iter: 500, multi_class: auto, penalty: l2, solver: liblinear}

Precision: 0.893929775998715
Recall: 0.8939736678415924
F-measure: 0.8939106716529297

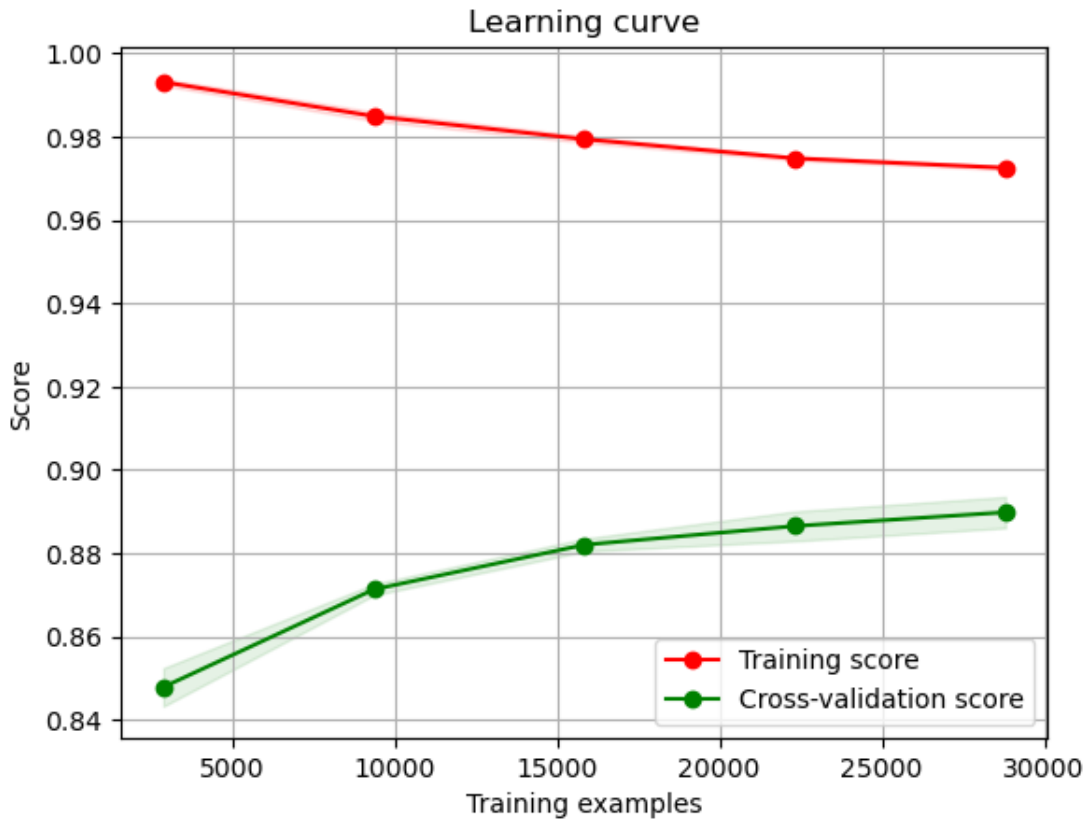

Figure 1: Model Chosen - Count Vectorizer with best parameters

Hashing Vectorizer with best parameters

The model does not seem to overfit.
Parameters:
{C: 2, class_weight: balanced, max_iter: 500, multi_class: auto, penalty: l2, solver: lbfgs}

Precision: 0.8965530444700367
Recall: 0.8964928467758657
F-measure: 0.8963555658819093

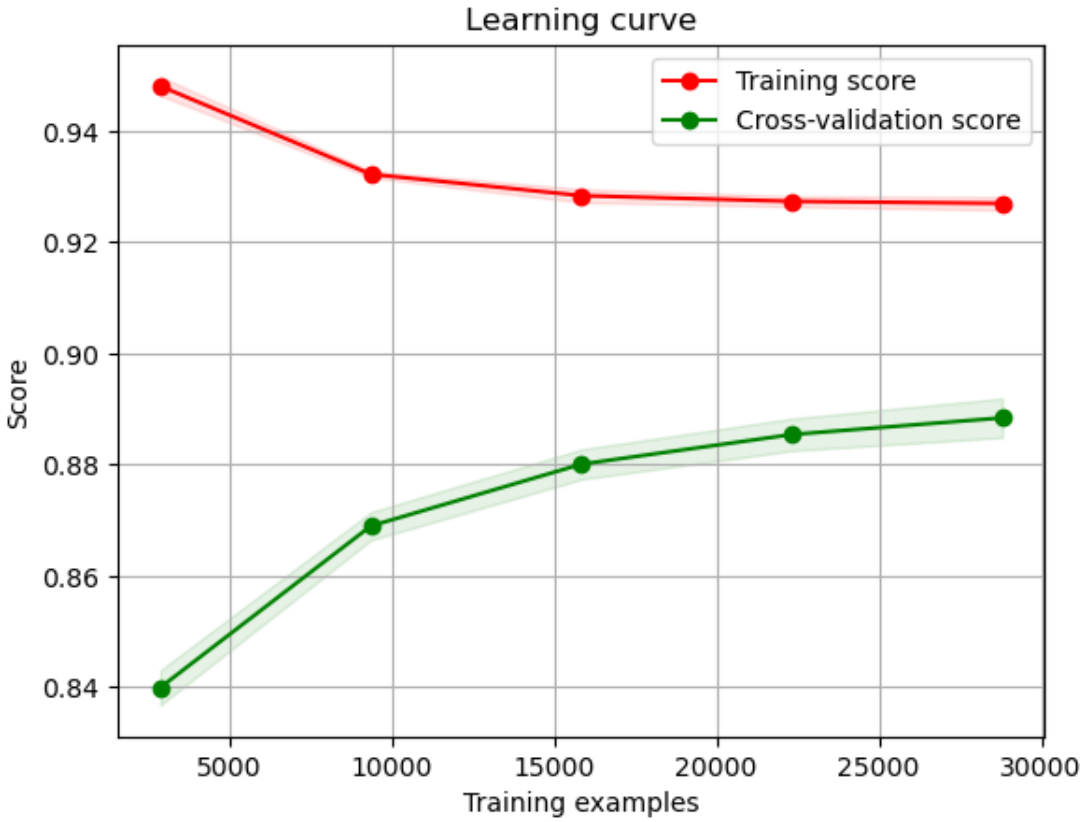

Figure 2: Hashing Vectorizer with best parameters

Notes

The models that were tested during the experimentation process of this assignment were highly prone to overfitting. A useful tool to avoid overfitting the training data was the C parameter (regularization).

Even though GridSearch is a great tool to find the best parameters it does not guarantee that the model avoids overfitting.

The two types of vectorizers were equally as good when the correct parameters were chosen.

The Count Vectorizer when paired with LogisticRegression in this specific dataset seemed to be much more prone to overfitting and therefore smaller values of C (more regularization) had to be chosen. This is highlighted in the below graph.

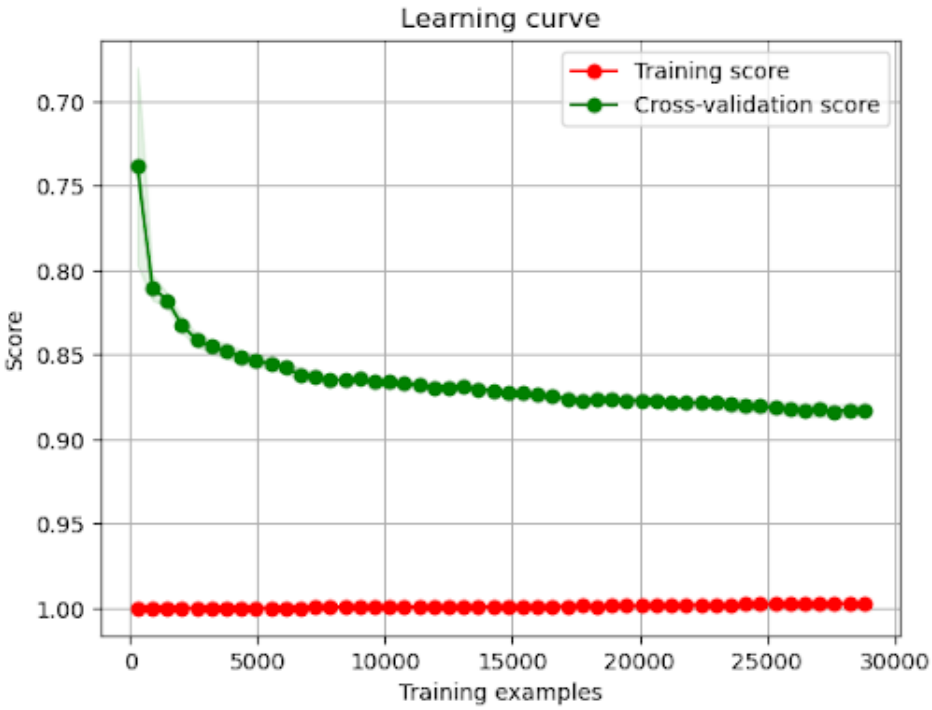

Figure 3: As it can be easily seen the model is greatly overfitting the training data, almost scoring 100%.

2021 Data Mining Assignment

The 2021 Data Mining 1st Assignment was very similar to this one. Snippets of code might have been used from my last year's project. The project was carried out with the help of Loukovitis Georgios - 1115201800100, who passed this class last year and therefore won't be handing in any projects this year.