Sklearn - SVM Report

Nicolò A. Girardini - 203265

10 February 2019

1 Introduction

This report is focused on tuning the parameters of an SVM classifier for predictions on the **SPAMBASE dataset**. There are many parameters for each entry, pointing features useful to understand whether or not a mail is to be considered **SPAM** (label **1**) or **NOT SPAM** (label **0**).

The percentage of the dataset being SPAM, as described in the documentation, is 39.4% on a total of 4601 entries.

2 The Algorithm

The algorithm chosen is an **SVM classifier**, in the sklearn implementation (namely SVC). I looked to tune the parameters starting from the kernel of the algorithm itself. The chosen one is **rbf** (Radial Basis Function), since the inputs have many features that could be hard to process for a linear kernel . The other parameters one would want to tune are **C** and **gamma**. The first is a regularization parameter, while the second is a parameter for the gaussian kernel.

The first operation was to split the training set into a training and a validation set, with the validation one being 20%. The whole training set will be used once the optimal parameters are discovered.

To tune all the parameters at once the process of **Grid Search** has been chosen. It performs the cross-validation task for a classifier (with one or more parameters) in an automated procedure. I used as cross-validation procedure a k-fold algorithm with three folds.

The GridSearch is a classifier itself, and after training it there is the possibility to see the tuned parameters for its best estimator.

The possible values for the parameters on which I performed the GridSearch are:
The possible values chosen for **C** are: 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}
The possible values chosen for **gamma** are: *1, 0.2, 0.1, 0.05, 0.02, 0.01, 0.001, 0.0005, 0.0001*

2.1 Optimal Parameters Found

The **rbf** kernel has been trained with the different parameters and the optimal ones, found by the GridSearch, are: $\mathbf{C = 10^3}$ and $\mathbf{gamma = 10^{-4}}$.

The accuracy obtained with the tuned parameters is around **0.9226** on the validation set and the following table shows the precision, recall and F1 scores for both spam and non-spam labels and the total average for each measure.

Class	Precision	Recall	f1-score
NON-SPAM	0.90	0.97	0.94
SPAM	0.95	0.86	0.90
TOTAL	0.92	0.92	0.92

Here we can also see the learning curve, in which it is evident the improvement of performances with more training examples (even if the total number of examples is still low). This happens while the training score decreases: it is index of a better generalization performance, implying also reduction of overfitting.

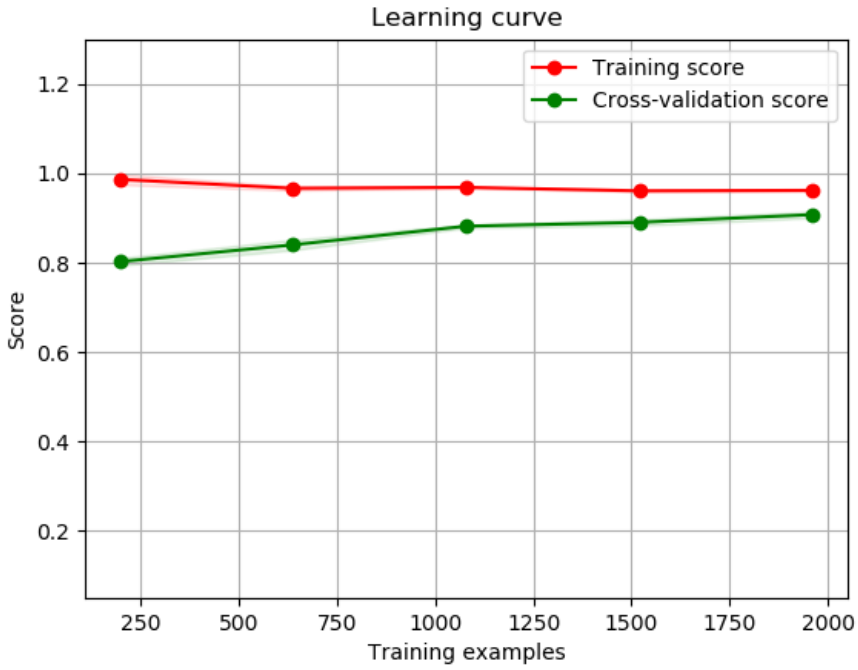

Figure 1: Learning curve for the **rbf** kernel

3 Final Results

The final results of the accuracies are obtained after training the optimal model for the kernel over the whole training set and comparing them to the test set labels.

The resuling accuracy for the resulting RBF kernel is **0.9304**.