Cornell University

ECE 4200 Fundamentals of Machine Learning

Model report: Kaggle Competition

December 2020

Nicolás Oriol Guerra (no227)

1. Introduction

The competition goal is to achieve the best possible performance in a model to predict different font types. As features, we are provided with inherent characteristics of observations of each of the different fonts. Overall performance will be measured by accuracy on provided testing data.

2. Methodology

The problem will be approached in two different ways: fitting of base models enforced by competition guidelines and fitting of improved models. Base models are naïve bayes and logistic regression. Code and performance will be recorded in the `BaseModels` notebook. Improved models are different types and combinations and parameter settings of neural networks. Their performance is recorded in `ImprovedModels` notebook.

Before starting with any model building, the data is normalized and split into training (70%) and validation (30%) sets. The validation set is used to predict model performance in the test set. Its predictions have been found to be very accurate and overall performance can be estimated with just the validation set.

3. Models, performance and prediction

The general approach to the problem consisted of three main steps: computing base models, training neural networks and performing PCA on the dataset before training the models.[1]

Regarding the performance of base models, it can be said that it is suboptimal. Naïve bayes had a predicted performance of 0.309 and the true score was 0.323. For the logistic model, these numbers were 0.470 and 0.466 respectively.In the case of neural networks, several approaches were taken. First, parameters were tried at random which initially brought significant performance improvements (up to 0.669)

A bottleneck was experienced when trying to train several different neural networks using a for loop and varying parameters. The reason for this was computing power. For this reason, only certain parameter settings were iterated over, thus creating 40 networks to compare and creating the need to use RandomizedSearchCV as well. Afterwards, deep neural networks were tried by hand by using Google Collab notebooks as additional computing power. Performance for each of these tests is found in `NNetworkTests1-2` notebooks (one for local and one for google collab models).

Finally, PCA was also attempted to reduce the number of significant variables. A study was also conducted to determine the optimal number of principal components with regards to model performance. This number was found to be 203 components. Different neural networks were tried with different parameters to try to improve performance. A gradient boosting classifier was also tried but its performance was found to be much worse than neural networks (~50%).

4. Conclusion and final model

In conclusion, the final model chosen was a neural network trained with alpha = 1.1, relu activation functions and adaptative learning rate and the following layer structure: (200, 100, 100, 50). The training data was previously processed with a PCA and the first 203 components were chosen. Final competition accuracy was 0.769.

---

[1] All code and processes can be found in github.com/noriolg/font-recognition-model