SAPIENZA
UNIVERSITÀ DI ROMA

Homework 2

Machine Learning

Students:
Flavio Maiorana
2051396

Academic Year 2023/2024

1 Data visualization and preprocessing

First of all, it could be useful to gain some insight one how the dataaset is made. The available data is already split into training and testing set, thus we will consider the test split only in the evaluation phase in order to prevent data leakage.

```
[Dataset for Training]
    N Examples: 6369
    N Classes: 5
    Classes: [0 1 2 3 4]
    - Class 0: 1000 (15.701051970482022)
    - Class 1: 1500 (23.551577955723033)
    - Class 2: 1500 (23.551577955723033)
    - Class 3: 2000 (31.402103940964043)
    - Class 4: 369 (5.7936881771078665)
```

Some comments: the dataset is highly imbalanced

1.1 Preprocessing for evaluation

2 Model

Different approaches can be used to solve this problem. We will treat mainly 4 models: KNN, SVM (linear and nonlinear), Gaussian Naive Bayes and Softmax Regression. They will be compared based on the same train-test split.

2.1 KNN

`[Dataset 1]`

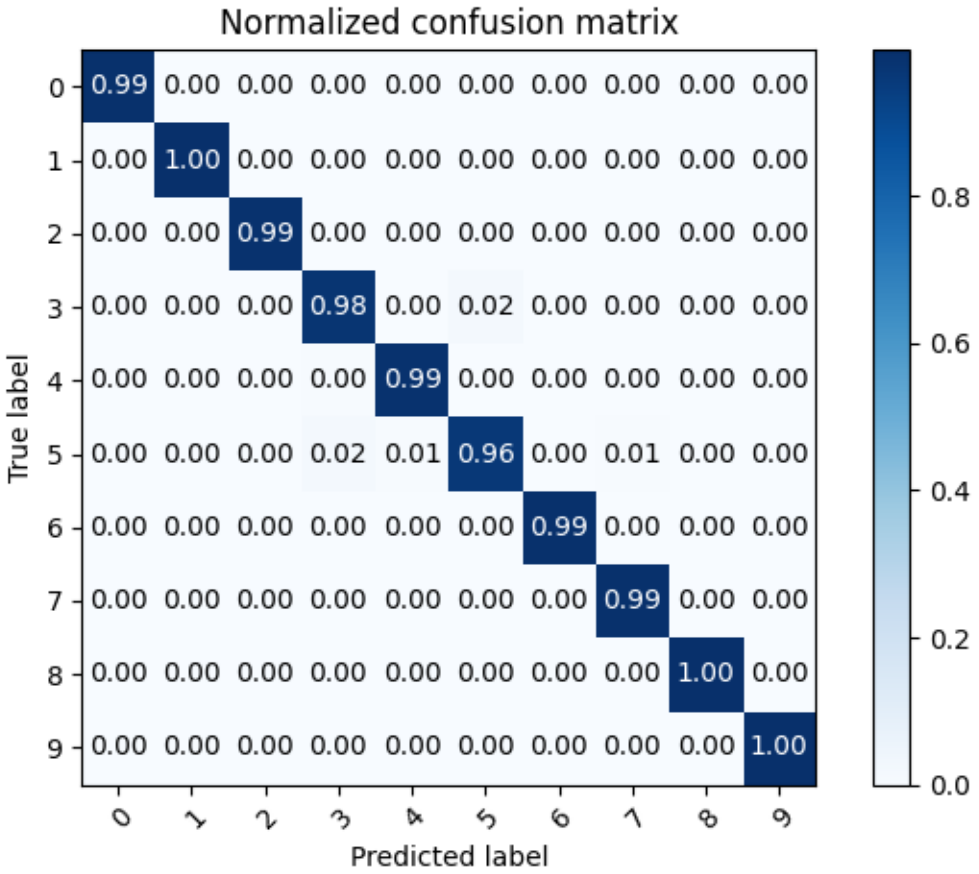

Figure 1: Knn Dataset 1

`[Dataset 2]`

2.2 Softmax Regression

`[Logistic Regression on Standardized Dataset 2]`

2.3 SVM

The last method that was tried is Support Vector Machines. This method is still parametric and linear and practically tries to maximise a margin between samples.

`[Linear SVM on Dataset 1]`

Linear SVM does slightly worse than KNN with 5 neighbors, and it takes also considerably more time to train the SVM classifier. Instead, by using a polynomial kernel of degree 3 we obtain a slightly better performance.

`[Poly SVM Dataset 1]`

SVM with 3-degree polynomial kernel has a better performance than linear SVM and also KNN. Although, it takes slightly more time to train it. Also, when increasing the degree of the polynomial, performance decreases, which means the classifier would overfit on training data.

`[Poly SVM Dataset 2]`

2.4 Considerations

In the end KNN and Poly SVM were the best methods on both datasets. KNN has the advantage of taking less computational during training time, while it costs slighly more during inference. Conversely, SVM has slighly higher accuracy but it takes a little bit more to train it. In both cases the main problem is the precision and recall of classes 3 and 5, which often get exchanged.