Assignment 1 - Comparison of different classifiers on multiple datasets

March 12, 2024

Simone Vaccari
915222

Alessio De Luca
919790

1 Objective

When solving a task, it is often necessary to choose among different learning algorithms, and in order to select the best one (a process called "model selection") we need to estimate the performance of each learner. The objective of this experiment is to see how different classifiers perform on different datasets and compare the results.

2 Procedure

In order to evaluate the performance of each learner on a given dataset we applied the **5x2 cross-validation**: this method consists in applying a 2-fold cross-validation 5 times. Therefore, as a first step we randomly split our data into training set (50%) and testing set (50%). Secondly, we trained our model on the training data and used it to make predictions on the testing data, calculating the accuracy (defined as the ratio between the number of correct predictions and the total number of predictions). Then, we inverted the roles of the training and testing set, fitting the model on the second one and evaluating it on the first one, obtaining a second value of accuracy. The two values were then averaged to obtain a final value of accuracy of the model, relating to the first 2-fold cv.

This process was repeated 5 times and the 5 accuracy results were averaged to obtain a final number expressing the performance of a given learner on a given dataset. All these steps were carried out for each model and for each dataset.

For our experiment we had at our disposal 4 datasets, that can be visualized in Figure 1.

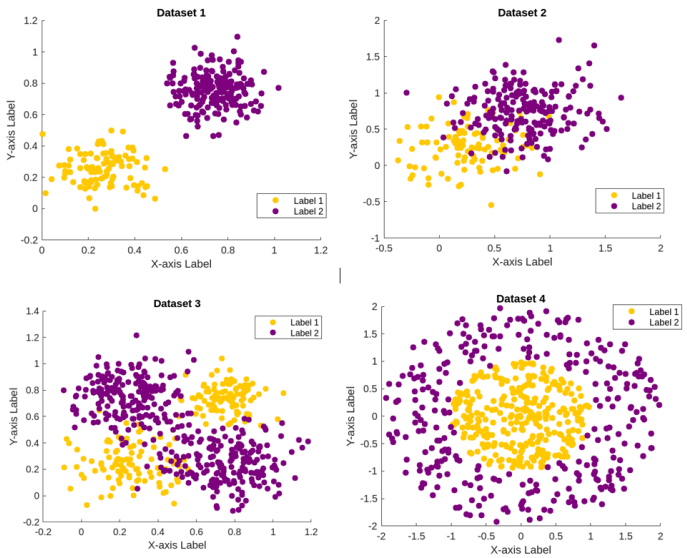

Figure 1: Datasets used for the experiment

We decided to use the following 4 different algorithms, with the specified hyperparameters:

- SVM, with linear kernel and scale factor 1;
- SVM, with gaussian kernel and scale factor 0.1;
- K-Nearest Neighbors, with K = 10 and using Euclidean distance;
- Decision Tree, with 15 as the maximum number of splits and Gini Index as splitting criterion;

To compare these algorithms we used the **Friedman Test**: for each dataset the methods were ranked from 1 to 4 according to their performance (the highest rank corresponds to the worst method). In case of ties, average ranks were assigned. Then, we averaged the ranks of each algorithm over all data sets, and compared the results with the *critical difference value*:

$$\mathrm{CD} = q_\alpha \cdot \sqrt{\frac{k(k+1)}{6N}} = 2.0914$$

where:

- $q_\alpha = 2.291$ is the critical value corresponding to a significance level $\alpha = 0.10$;
- $k = 4$ is the number of algorithms to compare;
- $N = 4$ is the number of datasets;

Two algorithms are considered significantly different if the difference between their average ranks is larger than the CD.

3 Results

When performing the experiment the following accuracies were obtained:

	Linear SVM	**Gaussian SVM**	**KNN**	**Tree**
Dataset1	1.0000	0.9993	1.0000	0.9947
Dataset2	0.8853	0.8287	0.8640	0.8253
Dataset3	0.6667	0.9210	0.9203	0.9010
Dataset4	0.5680	0.9727	0.9440	0.9600

Therefore, dataset by dataset the ranks were assigned as follows:

	Linear SVM	**Gaussian SVM**	**KNN**	**Tree**
Dataset1	1.5	3	1.5	4
Dataset2	1	3	2	4
Dataset3	4	1	2	3
Dataset4	4	1	3	2

Averaging the ranks of the four methods yielded these results:

	Linear SVM	**Gaussian SVM**	**KNN**	**Tree**
Average Rank	2.625	2.000	2.125	3.250

To visualize these results we plotted the so-called *Critical Difference Diagram* (Figure 2), where the dots represent the average ranks and the width of the lines corresponds to the CD value.

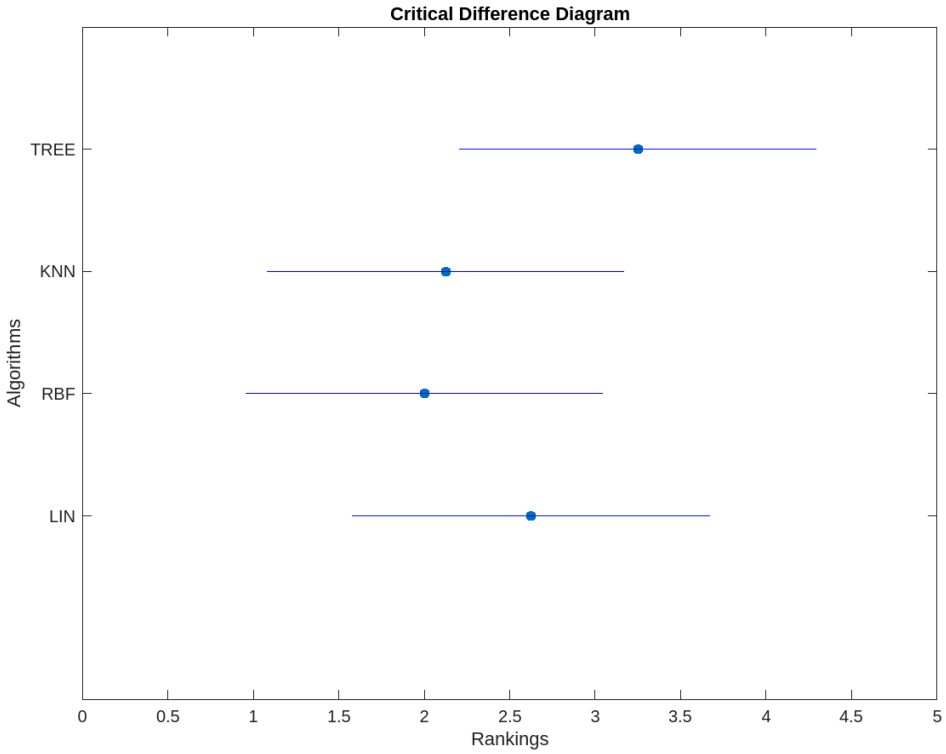

Figure 2: Critical Difference Diagram

4 Conclusions

As we can see from the diagram, there is not enough statistical evidence to draw conclusions on which method is significantly superior in this particular case.
By running the experiment multiple times the final values slightly change, and sometimes the KNN algorithm performs better than the SVM with Gaussian kernel. However, in all cases we cannot assert that this difference is significant. This is mostly due to the fact that the CD value is quite high: with a smaller CD, obtained by either considering a larger number of datasets or fewer algorithms to compare, we could expect to see some significant differences among the chosen algorithms.
It is important to say that here we didn't focus on finding the best values for the hyperparameters: in a hypothetical continuance of the experiment such a task could be carried out for each model in order to possibly improve its performances.