Statistical comparisons of multiple classifiers across diverse data sets

Davide Vettore 868855

Objective:

The objective of the activity is to perform a statistical comparison between different classifier across a collection of four different data sets.

Introduction:

We can start by observing the datasets upon which the experiment will be conducted. Each dataset consists of two quantitative variables, accompanied by a reference label which will allow us to assess the performances of our classifiers. Upon analyzing scatterplots of the data, we anticipate encountering diverse patterns and structures across the datasets, those different problem domains will require diverse decision boundaries to achieve optimal accuracy.

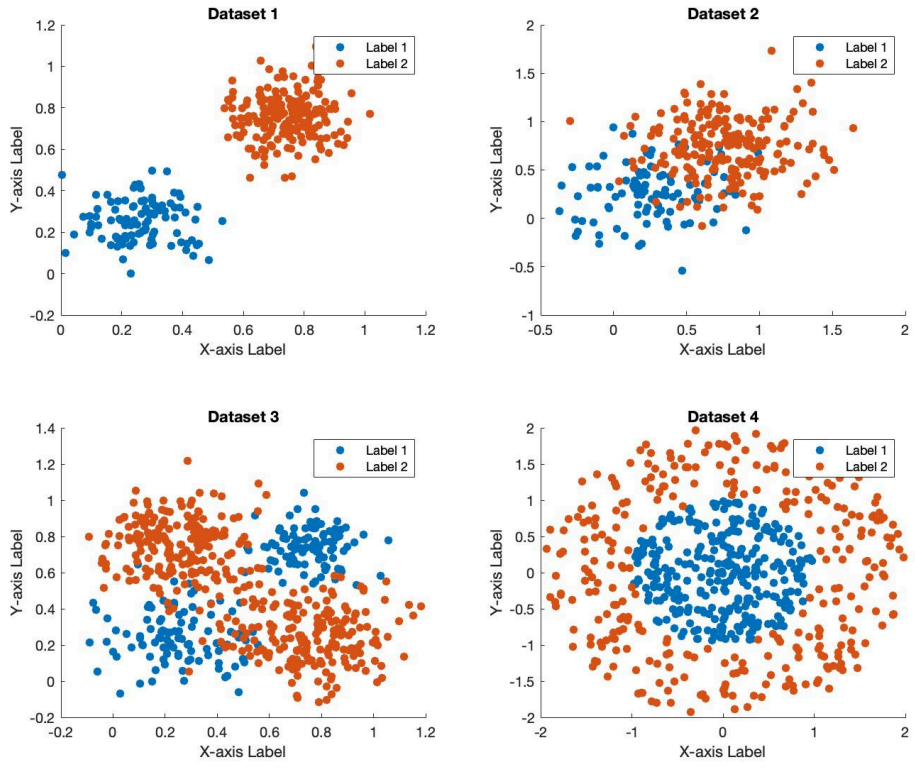

Figure 1: Scatterplots of the four data sets

The statistical comparison will be conducted across the following classifiers:

1. **Decision Tree** with *Gini's diveristy index* as splittin criterion and 10 max number of splits.
2. **Discriminant Analysis** with *Quadratic* discriminant type.

3. **Linear Support Vector Machine** with kernell scale set to 1.
4. **Medium Gaussian Support Vector Machine** with kernell scale set to $\sqrt{p}$.
5. **K-Nearest Neighbor** with K set to aproximately $\sqrt{n}$ and *Euclidean* distance metric.

Procedure:

In order to perform a statistical comparison between the performances of the different classifiers, a **Friedman test** is performed. The first step consists in ranking the performances of the classifiers across each data sets, the averange rank for each classifier is then computed. The performance of two classifiers is significantly different if the previusly computed average ranks differ by at least the Critical Difference:

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

With N number of considered data sets, k number of considered classifiers and q_α critical values based on the Studentized range statistic divided by $\sqrt{2}$.

Results:

The performance of each algorithm was assessed on the diverse datasets by following a 2-folds cross validation approach and the results were computed as an average over five different iterations.

Cross Validated Accuracy	Tree	QDA	Linear SVM	Gaussian SVM	KNN
Dataset 1	0.9947	1.0000	1.0000	1.0000	1.0000
Dataset 2	0.8380	0.8787	0.8853	0.8820	0.8767
Dataset 3	0.8763	0.9147	0.6667	0.7577	0.9263
Dataset 4	0.9600	0.9757	0.5680	0.9727	0.9457

It is now possible to rank the performances of each classifier over the datasets. The average rank is computed and will be utilized for the final step.

Ranks	Tree	QDA	Linear SVM	Gaussian SVM	KNN
Dataset 1	5	1	2	3	4
Dataset 2	5	3	1	2	4
Dataset 3	3	2	5	4	1
Dataset 4	3	1	5	2	4
Mean Rank	4.0000	1.7500	3.2500	2.7500	3.2500

In our settings, considering a critical value $q_{0.10} = 2.459$, a Critical Difference equal to 2.7492 is obtained. We can now visualize the critical difference diagram which shows us if we have any evidence that the performances of a classifier are significantly different from the ones of another classifier.

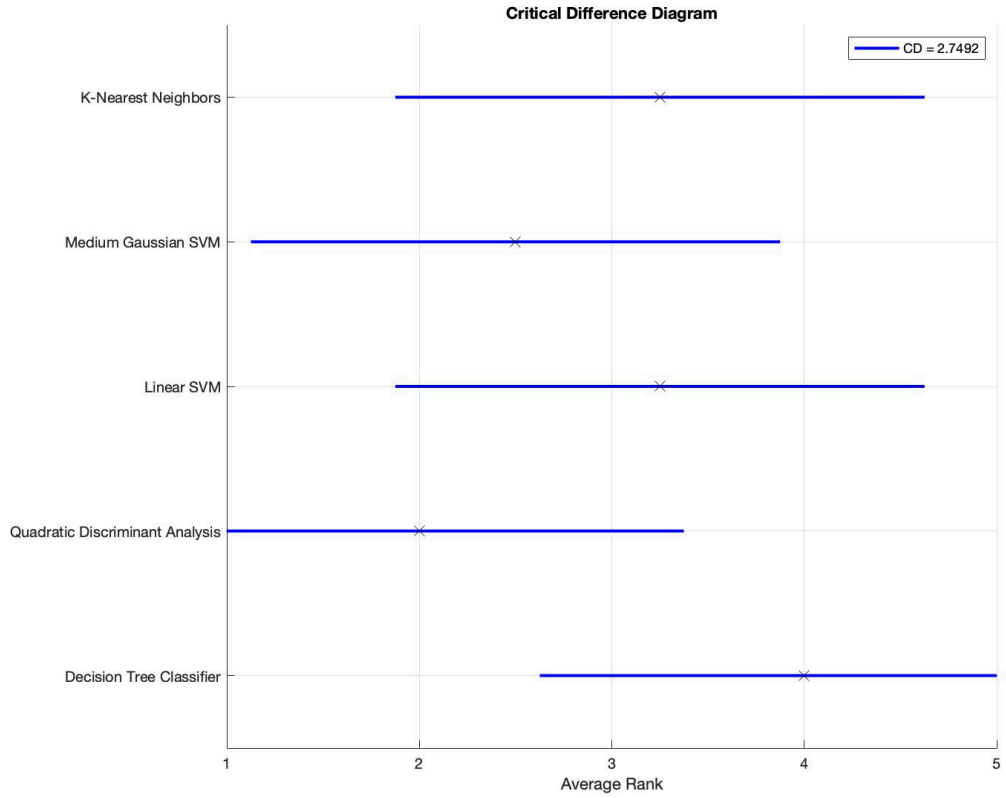

Figure 2: Critical Difference Diagram with $\alpha = 0.10$

Conclusion:

According to the diagram, with $\alpha = 0.10$, there is no significant evidence that the classifier performances differ from each other. This outcome aligns with our expectations, given the limited number of datasets utilized in the experiment. Notably, upon examining the Critical Difference formula, if our analysis had been conducted over 10 datasets instead of 4, the resulting CD would have been substantially smaller (1.7388), potentially leading to divergent conclusions.