Classification of samples with overlap in their feature-distribution : A case study using the Iris data set

H Hurchand, Continuing Education, Concordia University
Dec. 2019

Abstract

In this project, we examine the accuracy of three theoretically distinct classes of estimators to classify the Iris dataset, for situations where the distribution of variables overlap. The first estimator employed is the GaussianNB, which is probabilistic. The other two: the K-nearest neighbour and the support vector machine (SVM) are deterministic in nature. Three different kernels (radial, polynomial and linear) were used for the SVM. When applied to the original dataset, the three classes of estimators produced comparable accuracy, with a mean of 98%. When the dataset was subjected to dimension reduction,using PCA, it was the SVM methods and the GaussianNB which produced better accuracy, with a 100% success rate.

1. Introduction

There exists a plethora of estimators that can be used for classification purposes [1]. The type of estimators to be employed is very much a function of several considerations, among which the characteristic of the data set is a non-negligible one. For example, a dataset which depicts clearly isolated clusters could easily be treated with the K-NN classifier. The latter classifier, though computationally costly when the sample size increases, can yet be a preferred choice in many instances. There is often a balance to be struck between speed and accuracy for our training models.

In real-life, the kind of data set we have in hand varies largely. When one classifier can work well on a given set of data, the same classifier may provide less satisfaction of a different set of data, due to some different characteristics that this dataset present. One possible characteristic that datasets could exhibit is an overlap between different classes they contain. Classification near the overlapping region can be tricky and sensitive to the choice of classification methods applied.

Although, an arbitrary method can still yield a reasonable level of accuracy, for certain situations, like medical applications, an estimator which produces 100% is highly desirable. There is not much work seen which studies the choice of classification methods for these kind of datasets. In this work, we intend to provide some answers to this question using the Iris dataset.

Preliminary exploratory observations show that the Iris dataset display some overlaps for the versicolor and virginica species. The data also suggest the potential for dimension reduction, which could produce some interesting effect in reducing overlaps.

In this work, we use the K-NN classifier as a benchmark classifier. We intend to compare the accuracy of this classifier on overlapping datasets with the probabilistic GaussianNB model and the deterministic SVM model, with three distinct kernels. This comparison will take place in two phases, one in which there is no dimension reduction and the second one with dimension reduction, which is achieved through the application of PCA.

The objective of the research can be encapsulated in these two research questions.

1.1 Research Questions

1. Which one of the K-NN, GaussianNB or SVM (with radial base function, linear and polynomial kernels) produces better accuracy for classifying Iris species, found in the Iris-data set?

2. Does dimension reduction contributes to enhancing the accuracy achieved?

2. Methodology

The following approach was used with the expectation to produce valid and reliable conclusions to the above research questions.

2.1 Comparison of estimators

The three classes of estimators which we have considered were the k-NN, the GaussianNB and the SVM. All the estimators were applied to the same dataset. We ensured that this was the case by identically parametrizing the random-state of each run.

2.2 The k-NN : Parameter tuning

An accuracy test was run to find the best k-value for the k-NN classifier. The k-value was found to change for the training set being used. We employed the k-fold splitting method, with 30 splits. In Figure 1, we show the accuracy profile for the k-NN on the Iris dataset. The highest mean accuracy occurred for k = 7, which is the value we used for the k-NN estimator in the comparison part.

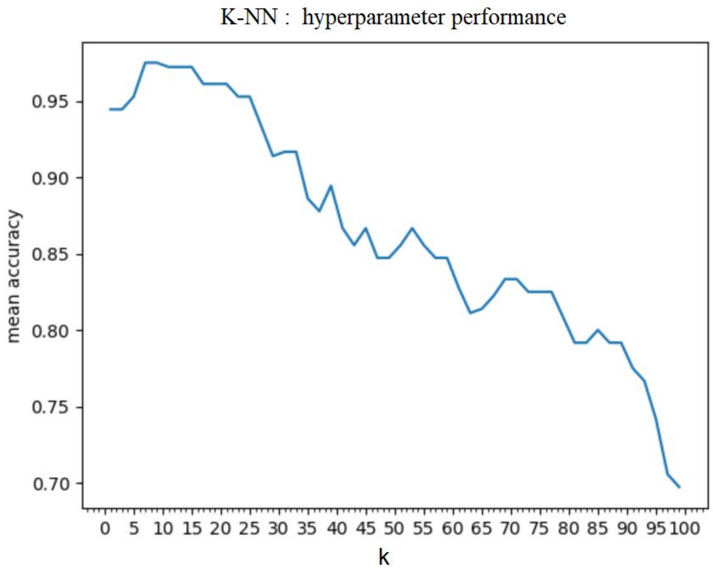


Figure 1 : The accuracy profile of the k-NN for the Iris dataset.

2.3 SVM-classifier

The SVM classifier can be used with a number of different kernels. There is no indication on the kind of kernel which can best be suited for the Iris dataset. Relying on visual analysis of the pairplots shown in Figure 2, these three kernels were found to be plausible choices: the radial base function (rbf), the linear kernel and the polynomial of degree 5. The polynomial of degree 5 was chosen, because it was believed to spline better the profile of the shape of the boundaries observed.

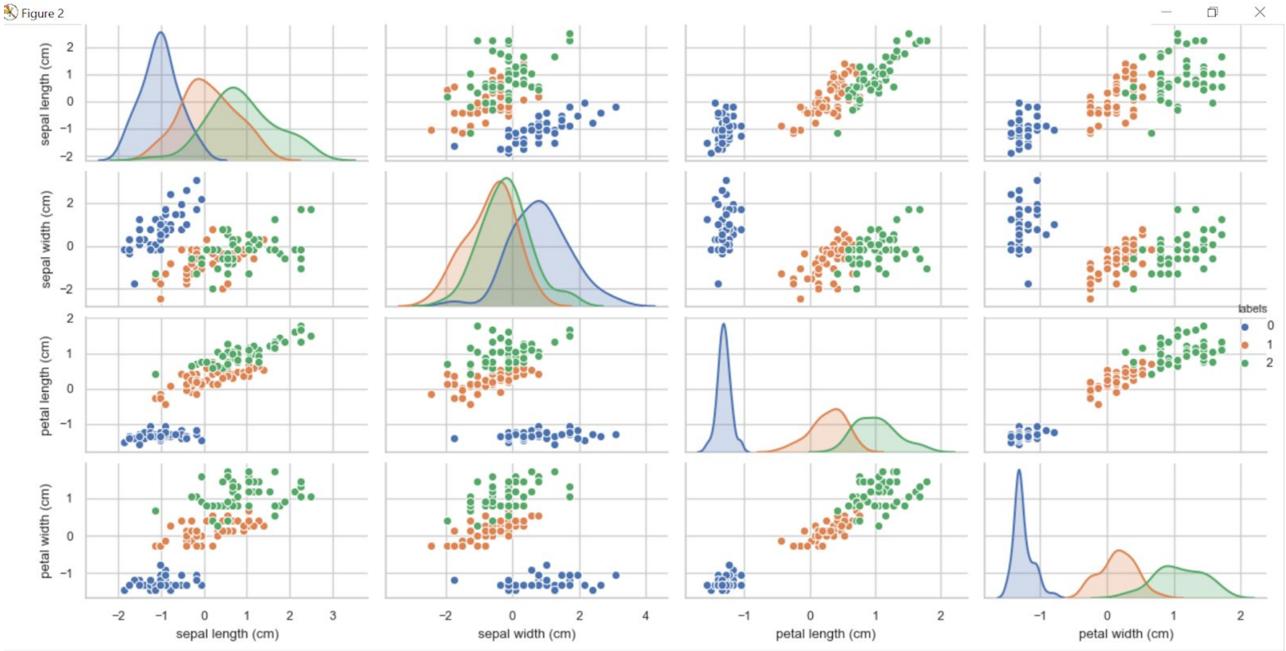


Figure 2 : Pairplot of the Iris dataset.

2.4 PCA

We applied the PCA to the Iris dataset. We found that 98% of the variations were explained by two components solely. Therefore, we used the 2-component model.

The whole process of estimator comparison was run both with the original Iris dataset and the PCA reduced dataset.

2.5 The Iris dataset

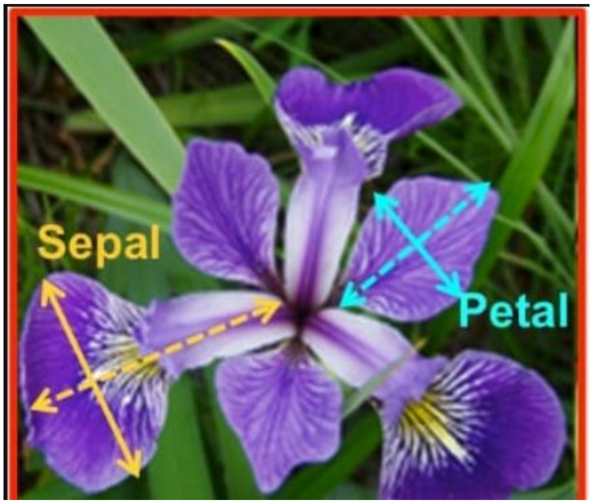
The Iris dataset is a collection of data on four features for three kinds of Iris flowers.

2.5.1 The data features

1. Petal length
2. Petal width
3. Sepal length
4. Sepal width

The length of these flower parts are measured in centimetres.

In the diagram below [2], the features are illustrated on a random Iris-type.



2.5.2 The classes

The dataset consists of these three flower types

1. Iris-setosa
2. Iris-Versicolour
3. Iris-Virginica

2.5.3 Samples

- The collection consists of 150 samples (i.e, Iris flowers).
- The dataset is balanced, with each of the three kinds being equally represented.

Limitations

- The samples were collected in 1963. It is assumed that predictions based on models constructed on these data will be valid for Iris flowers coming from a period in which variations are insignificant compared to the collected specimen.

Assumptions

- There is no complete information about the geographical origin of the samples. It is known that two of the three types originate from the Gaspé Peninsula. It assumed here that the variation in Iris features in location independent and the prediction will be valid for Iris flowers coming from any regions.

2.6 Data cleaning and manipulation

The Iris dataset contained no null nor missing values. There were no blatant outliers. We did not need to clean the data in these respect.

However, data was grouped in three blocks of species. The data was shuffled to ensure that there is a proper mix of the data. Additionally, we normalised the data because we used k-NN as one of the estimators. The same normalised data was used for the other estimators.

3. Results

When ran on the original data set, all the estimators produced comparable results with a mean accuracy of 94%. The SVM with polynomial kernel was less accurate than the other classifiers. This may be explained by an unoptimized selection for the degree of the polynomial.

When the estimators were applied to the PCA data, there was a significant improvement in the accuracy produced by all estimators. The GaussianNB and the SVM estimators (with rbf and linear kernels) all produced 100% accuracy. Nevertheless, the SVM with fifth degree polynomial kernel still produced less accurate classifications when compared to other methods.

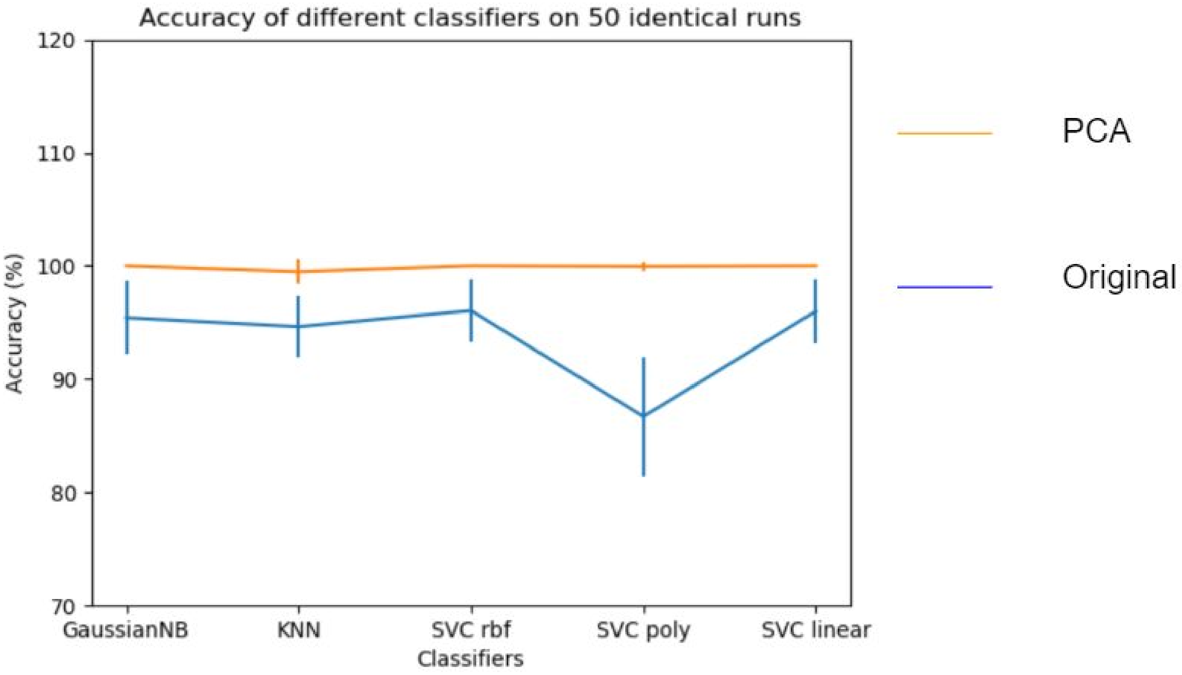


Figure 3: Accuracy produced by the three classes of classifiers using the original Iris dataset and the PCA dataset

4. Conclusion and Discussions

The SVM and GaussianNB methods were not found to produce better classification compared to the k-NN estimator, when we used the original Iris dataset. However, when we applied PCA to the data the former methods produced 100% accuracy and were better than the k-NN accuracy. It is believed that the transformation of data that occurs during PCA projection reduces the overlap between the component distributions. Such reduction in overlap brings in gain, in both the GaussianNB and the SVM methods. The GaussianNB relies on the probabilistic decision of properly matching a sample to its distribution. This is made easier with the reduced overlap created by PCA transformation. In the same manner, it becomes easier to construct hyperplanes which clearly demarcates the different clusters.

It should be highlighted that the overlaps noted in the Iris dataset was only marginal. It is unclear whether the same conclusion will be reached concerning estimator accuracy for a dataset presenting much pronounced overlap. In the future, such datasets will be investigated with this approach to provide get some insights in this area.

5. References

[1] https://www.greycampus.com/opencampus/machine-learning/different-types-of-classifiers

[2] https://en.wikipedia.org/wiki/Iris_flower_data_set