Universitat Politécnica de Catalunya
Barcelona Tech
Facultat d'Informética de Barcelona

Feature Selection

Daniel Natanael García Zapata

Master of Science in Artificial Intelligence
January 2018

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The main motivation for this report is to explore feature selection on a basic dataset. The Iris dataset is chosen because of it is easy to visualize and see the results of the methods analyze in the report. Most of the time in data mining is expend in the visualization and pre-processing of the data. These are key factors that are going to be focus on in this report. Also the motivation is to show how projecting the data and selecting good features can help improve the classification task and also to learn that in many cases there are only few features that are important to the classification problem.

1.2 Objectives

The objectives of these experiments are to see how the data projection changes from using PCA and LDA. For comparison two learning algorithms will be used, however the accuracy of these are not relevant since they are only for comparison. The importance is to show how the data is projected. For Filters and Wrappers, the objective is to learn how to choose the features using some methods and to learn that it is critical to learn how to choose features that contribute more to the prediction task.

Chapter 2

Datasets

2.1 Iris

The Iris dataset was first introduced by Fisher in 1936 [1]. The dataset contains 3 classes, each class contains 50 instances. The classses represent a type of iris plant; the types being setosa, versicolor and virginica. The dataset is well known for pattern recognition.

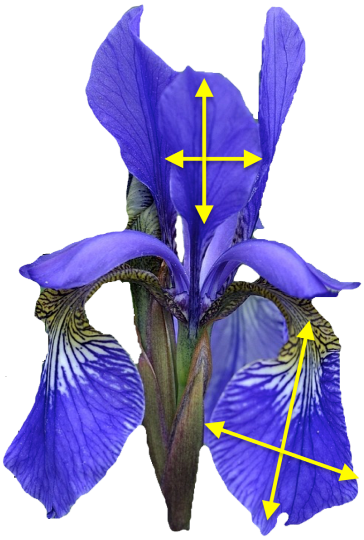

Figure 2.1: Iris flower

Chapter 3

Visualization

3.1 Data Exploration

One of the first things to do with any dataset is to visualize it, and be able to learn the distribution of the data. A pairplot is quite useful for this task, this graph shows the bivariate relation between each pair of features.

The pairplot figure shows the plot that is colour encoded representing the iris species. From the previous plot, it is important to note that setosa is linearly separable from the rest in all pair of features. However, versicolor and virginica are not linearly separable between each other. The first conclusion is that setosa can be classify with ease from the other two types of iris.

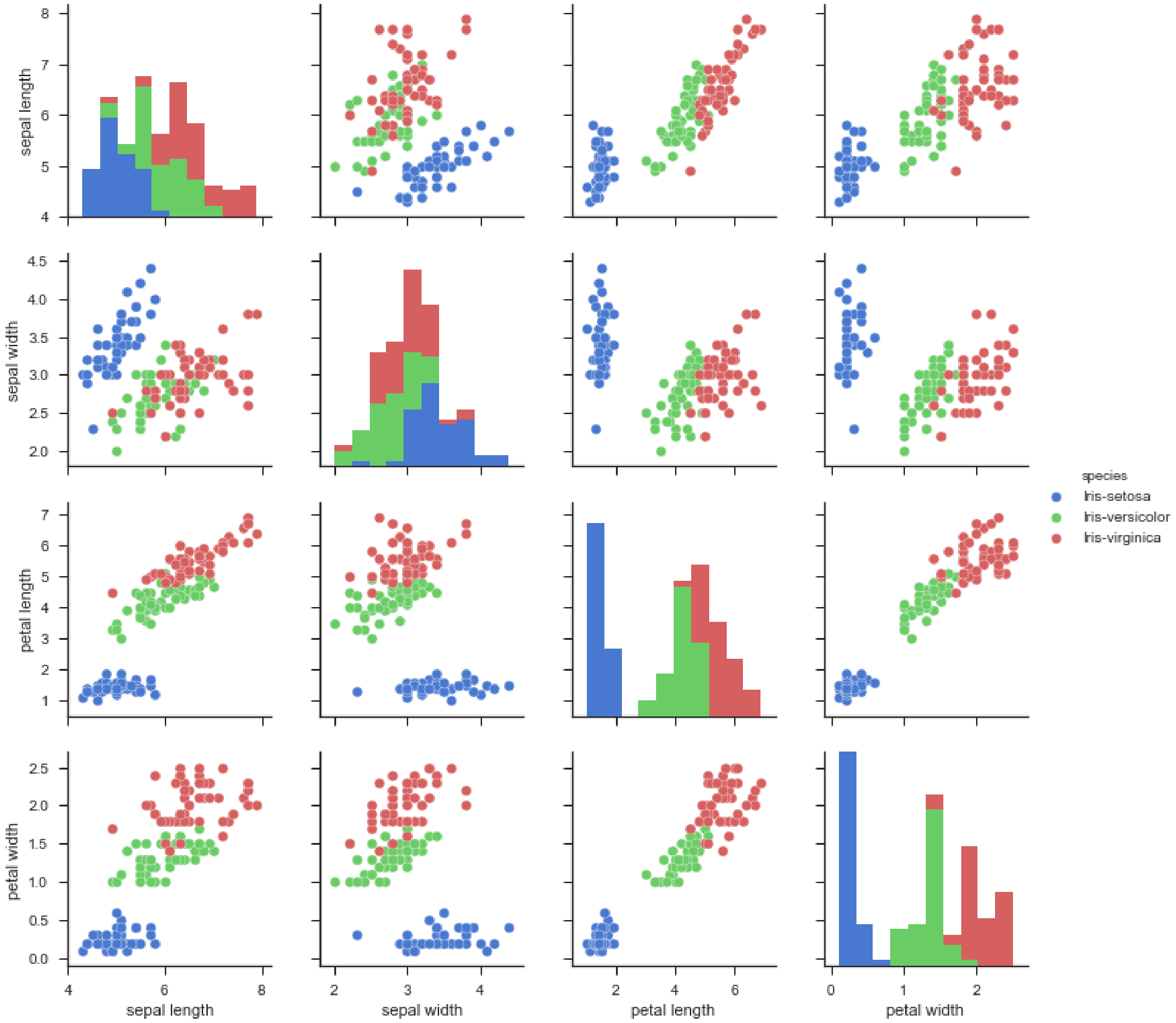

Figure 3.1: Plotting of every pair of features

Using Andrews plot is clear that setosa distribution is different from versicolor and virginica like demonstrated in the pairplot. Differentiating between latter two features is less easy. Andrew's curves have some really nice properties. These curves preserve the mean and variance of the function corresponding to the observation and also the distance between two functions [2]. It is interesting to note that although versicolor and virginica have similar curves, they do have a different variance. Hence, it is interesting to apply some techniques such as PCA and LDA.

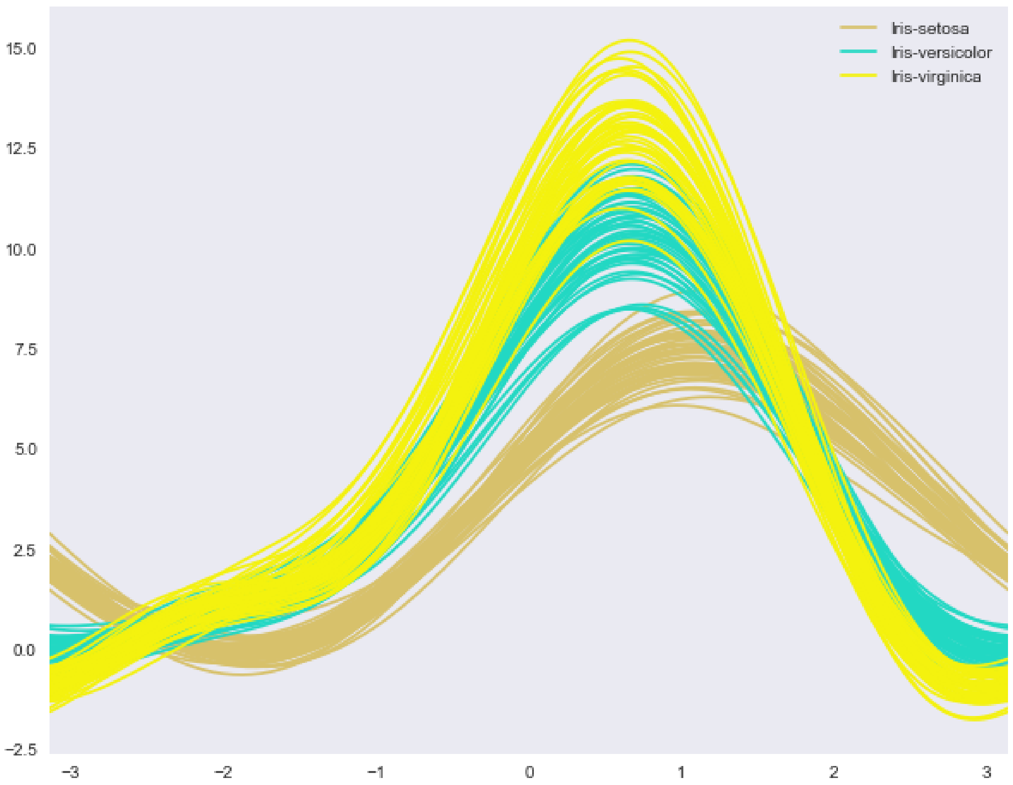

Figure 3.2: Andrews curves are based on Fourier series where the coefficients are the observation's values.

Chapter 4

Setup

Before performing any feature selection algorithm on the Iris dataset, it is important to setup the dataset so that these algorithms can perform better on the dataset.

4.1 Standardize

Any normalization process is important in PCA, this is because PCA maximizes the variance by projecting the data into another direction that seeks to maximize it. The key with PCA is to maximize variation, so it is important to let variables that have more variation to contribute more. It is then important to standardize the data so that the data has the same scale. In this case, standardization will be used to obtain the results needed.

$$z = \frac{x - \mu}{\sigma}$$

The outcome of standardization is that the features are rescaled so that they have the properties of a standard normal distribution, mean 0 and standard deviation of 1.

4.2 Explained Variance

PCA is defined as "an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on" [3].

PCA can use the correlation matrix or the covariance matrix. The covariance matrix is generally used when the scales are similar and the correlation matrix when it is on different scales. In this case, the covariance matrix is going to be used.

1	-0.10936925	0.87175416	0.81795363
-0.10936925	1	-0.4205161	-0.35654409
0.87175416	-0.4205161	1	0.9627571
0.81795363	-0.35654409	0.9627571	1

Table 4.1: Iris Covariance Matrix

Before performing PCA, the **Explained Variance** can be obtained with the covariance matrix. The explained variance ratio gives the percentage of variance explained by each of the components.

0.72770452	0.23030523	0.03683832	0.00515193

Table 4.2: Explained Variance Ratio

After obtaining the explained variance ratio, it can be seen that the first two components explain the 95.8% of the variance. For the next experiments, PCA and LDA, only the first two principal components are going to be used.

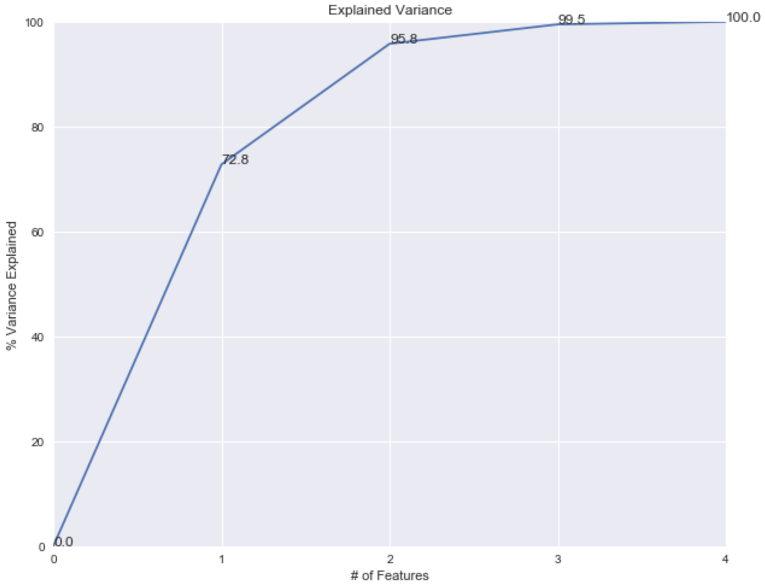

Figure 4.1: Variance Explained from the eigenvectors

Chapter 5

Experimental Results

This chapter presents the results from the different experiments held on the Iris dataset. Many times features do not contribute equally to make a prediction. Often, some features are irrelevant at the moment of making a prediction. There are different methods that can help choose features or make transformation to the data to improve the classification accuracy. This section explore some of these methods; the experiments consists of using PCA, LDA, Filters and Wrappers.

5.1 PCA

PCA tries to project the data into a new space where the maximum of the variance is preserved and maximized. For comparing PCA and LDA, both projections will be trained using logistic regression and SVM. The following table shows the test accuracy of the PCA projection using only the first two eigenvectors.

Algorithm	Score
Logistic Regression	0.84
SVM	0.93

Table 5.1: PCA Score

The following plot is using the two first principal components and shows the projected datapoints. Although PCA maximizes the variance, there are still datapoints from versicolor and virginica class that are close together, some even mixed with one another.

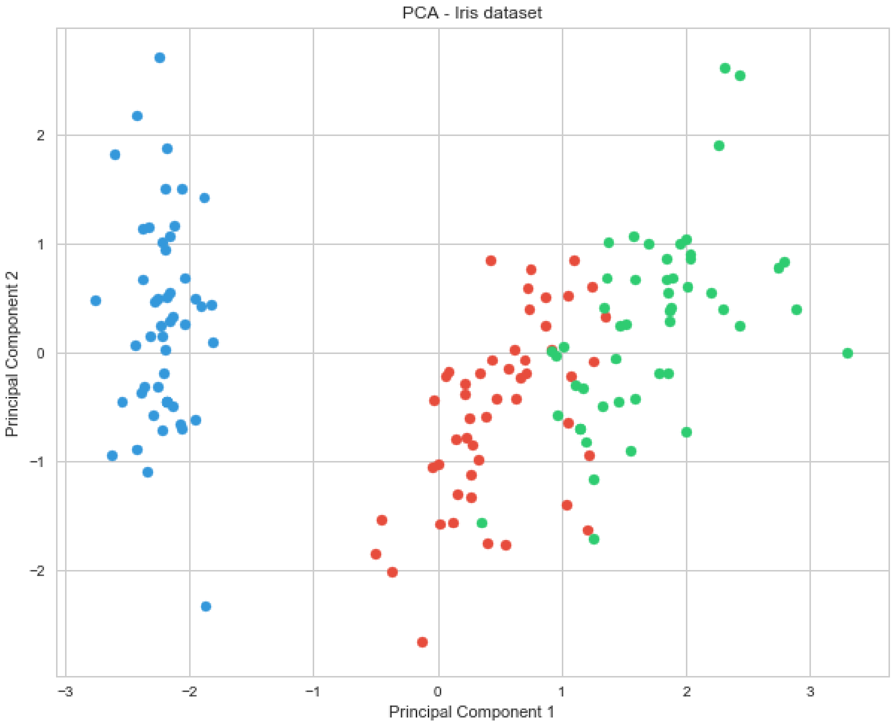

Figure 5.1: PCA first two principal components plot.

5.2 LDA

LDA projects the samples from the same class very close and the projected means of the different classes as far apart as possible.

Algorithm	Score
Logistic Regression	0.93
SVM	1.0

Table 5.2: LDA Score

LDA got higher accuracy in both logistic regression and SVM than PCA. This is because LDA also seeks to maximize the distance between classes as it can be seen in the plot portraying the first two principal components. In this graph versicolor and virginica are almost linearly seprable, except for one datapoint that is almost at the same point as an opposite class datapoint.

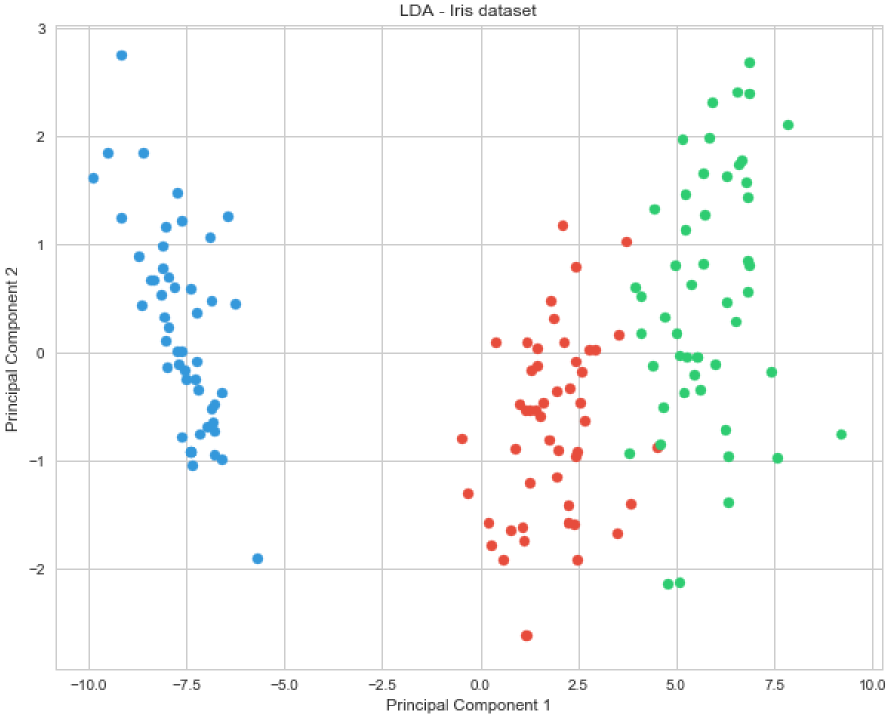

Figure 5.2: LDA first two principal components plot.

5.3 Filters

Filters are evaluated using non-learning algorithm, that is no machine learning algorithm is used to select the features. Instead these family of algorithms uses information obtained from the features. In this case, the selected algorithm to use is RELIEFF which uses local neighbourhoods information and is also used for multi-class classification problems. This method searches for the k near misses from each different class and averages their contribution to the prediction of the class.

Feature	Score
sepal length	794.81
sepal width	340.22
petal length	1938.89
petal width	2038.38

Table 5.3: Filters Score

Looking at the previous table, petal width and petal length are the features that contribute more the the correct prediction of the class. By recalling the pairplot, and looking at the diagonal plots that represent the histogram of each class. It can be shown that indeed these two features are more linearly separable, at least setosa is completely separable. Versicolor and virginica are together but not completely mixed as in the case of the first two features. Petal width contribute 2.5% more than sepal lenght and 5.9% more than sepal width.

5.4 Wrappers

Wrappers, on the contrary to Filters, evaluate the feature selection process using learning algorithms. In this report, the chosen method is the Random Forest algorithm and the Extra-Trees method. Decision trees are used to perform feature selection by using the split points. The more frequent a feature is used in the splitting points process then the more important the feature is.

Feature	Score
sepal length	0.096
sepal width	0.033
petal length	0.343
petal width	0.526

Table 5.4: Wrappers Score

The results obtain from using Embedded Random Forest for obtaining the feature importance is show in the previous table. The results obtained are quite similar with the results obtained from the Filters. Being petal width the feature that contributes more to the prediction and petal length being the second. Both sepal length and width contribute almost nothing to the classification. These results are constant with what the previous experiments have shown. As a conclusion, a model can be built using just the last two features and the prediction will be as good as using all of them.

Chapter 6

Conclusion

6.1 Summary of Achievements

The Iris dataset is good enough for pattern recognition. The fact that two of the classes are not linearly separable helps to learn and utilize methods such as PCA, LDA and feature selection. Also, this dataset has only four variables so it is easier to visualize what these methods are doing to the dataset. As for PCA and LDA, both are helpful for maximizing the variance. As versicolor and virginica are distributed similarly, these methods help separate both classes. LDA does have a better performance than PCA in this case since LDA takes into accounts also the classes. It tries to tries to maximize the distance between classes.

Both filters and wrappers have advantages and disadvantages. Luckily, in the Iris dataset both came to the same conclusion. Filters are good methods that does not take into account a learning algorithm so this makes the methods useful for quite a huge range of learning algorithms. The problem is that it is quite exhaustive to make a search for a good selection of features. In this case, since Iris only has four features the search is small. On the contrary, wrappers use a learning algorithm to choose the best features. The advantage is that the features obtained are particularly good in the chosen algorithm but may not be on any other learning algorithm. Because of this, wrappers generally obtain better recognition rates but at the cost of adapting to only the chosen machine learning algorithm.

Bibliography

[1] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, 7(2):179–188, 1936.

[2] C. Garcıa-Osorio and C. Fyfe. Visualization of high-dimensional data via orthogonal curves. *Journal of Universal Computer Science*, 11(11):1806–1819, 2005.

[3] I. T. Jolliffe. Principal components in regression analysis. *Principal component analysis*, pages 167–198, 2002.