Topic

In this tutorial, we show that in certain circumstances, it is more convenient to use the factors computed from a principal component analysis (from the original attributes) as input features of the linear discriminant analysis algorithm.

The new representation space maintains the proximity between the examples. The new features known as "factors" or "latent variables", which are a linear combination of the original descriptors, have several advantageous properties: (**a**) their interpretation very often allows to detect patterns in the initial space; (**b**) a very reduced number of factors allows to restore information contained in the data, we can moreover remove the noise from the dataset by using only the most relevant factors (it is a sort of regularization by smoothing the information provided by the dataset); (**c**) the new features form an orthogonal basis, learning algorithms such as linear discriminant analysis have a better behavior.

This approach has a connection to the reduced-rank linear discriminant analysis. But, instead to this last one, the class information is not needed during the computations of the principal components. The computation can be very fast using an appropriate algorithm when we deal with very high-dimensional dataset (such as NIPALS). But, on the other hand, it seems that the standard reduced-rank LDA tends to be better in terms of classification accuracy.

Dataset

We use the famous WAVEFORM dataset (Breiman and al., 1984; http://archive.ics.uci.edu/ml/datasets/Waveform+Database+Generator+%28Version+1%29). It is very interesting because it is very popular. We know what we can obtain on this dataset. Especially, we know that the best generalization error rate is known in advance, we cannot make better than 0.14 (Breiman and al., 1984, page 55).

Steps of the experiment

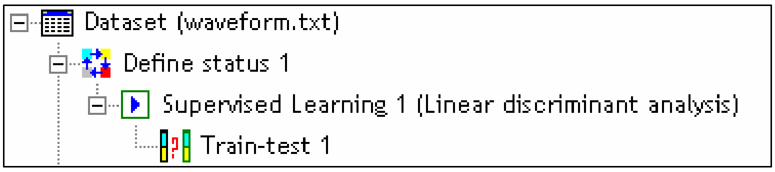
In a first time, we launch the linear discriminant analysis (LDA) on the original predictive attributes.

1. We open (**FILE / OPEN**) the data file "dr_waveform.bdm" (".bdm" is the binary file format; it is especially useful when we deal with very large dataset). The file contains 5000 instances and 22 attributes.
2. By using the DEFINE STATUS component (the shortcut into the toolbar), we set CLASS as target, the other attributes as input.
3. We add the "Linear Discriminant Analysis" component (SPV LEARNING tab). We click on the VIEW menu. Le learning phase is launched on the whole available instances i.e. the 5000 examples.
4. The resubstitution error rate (computed on the learning set) is 0.1350.

We know that the resubstitution error rate is often optimistic. It is more suitable to assess the performance of the classifier with a resampling approach. In this tutorial, we use the hold-out evaluation method.

5. To recreate the assessment conditions described in the Breiman's book (page 49), we set the training sample size to 300 observations; for the test sample, we set 4,700 observations. We want to repeat 10 times this process.

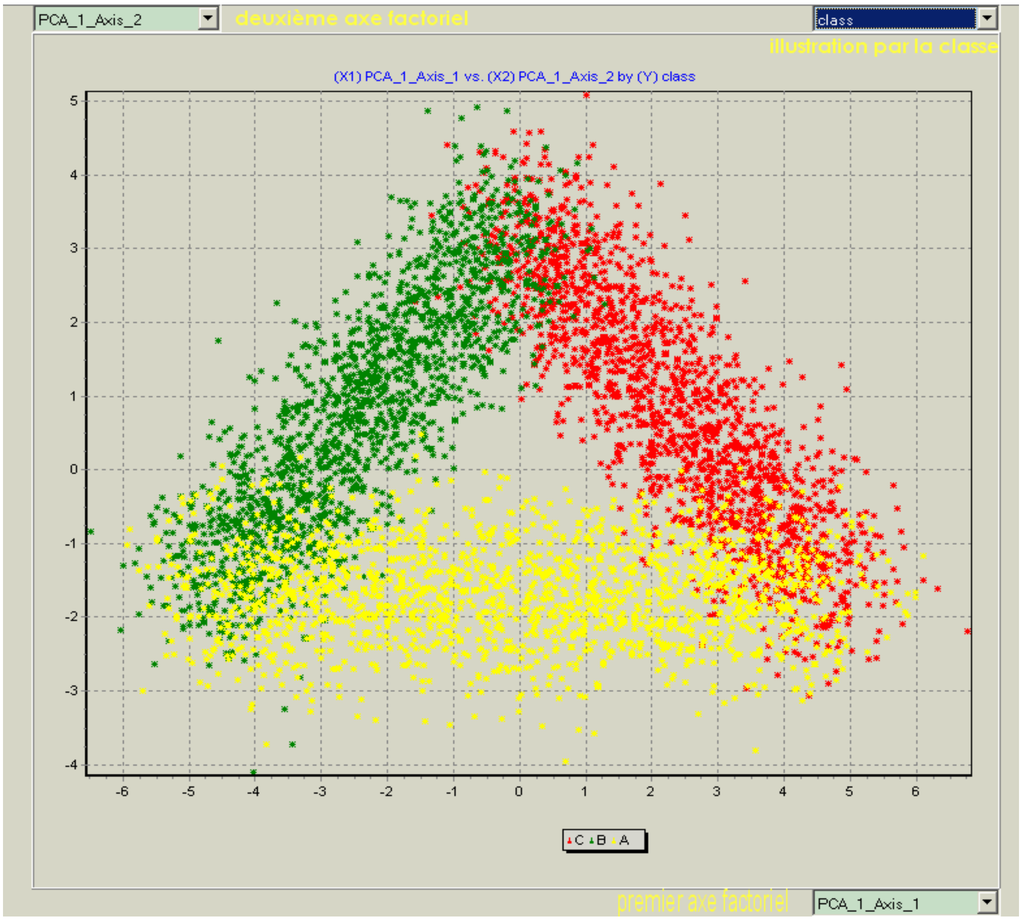
6. To do this, we insert the "Train-Test" component into the diagram, behind the LDA learning method. We set the following settings (PARAMETERS menu): Train set proportion = 0.06 (0.06 x 5000 = 300) and Repetition = 10.
7. **The measured test error rate is ≈ 0.20**. It is more in accordance to the results described into the literature about this dataset.
8. The diagram for this first analysis phase is the following



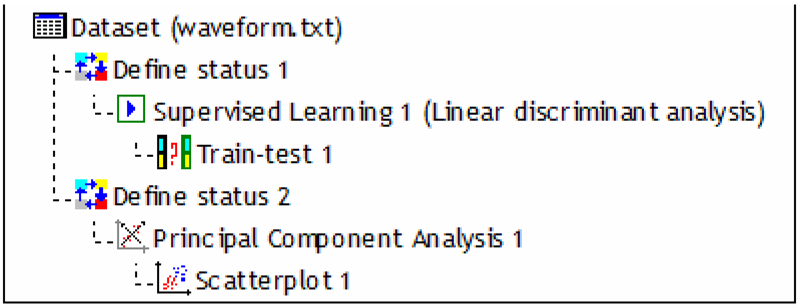
```
Dataset (waveform.txt)
  Define status 1
    Supervised Learning 1 (Linear discriminant analysis)
      Train-test 1
```

In a second time, we want to compute the factors of a principal component analysis (PCA).

9. We insert again the DEFINE STATUS component behind the data source. We set all the continuous descriptors as input.
10. Then we insert the "PRINCIPAL COMPONENT ANALYSIS" tool (**Factorial Analysis** tab).
11. We click on the VIEW menu to obtain the results. We note that the two first factors summarize the half of the available information (53.17%).
12. Tanagra computes automatically the first 5 factors (we can ask more by modifying the settings). We can project the available instances into the representation space defined by the 2 first factors. We use the "SCATTERPLOT" tool (**Data Visualization** tab). We note that the region associated to each value of the target attribute can be identified easily.



The used diagram to obtain this scatter plot is the following one.

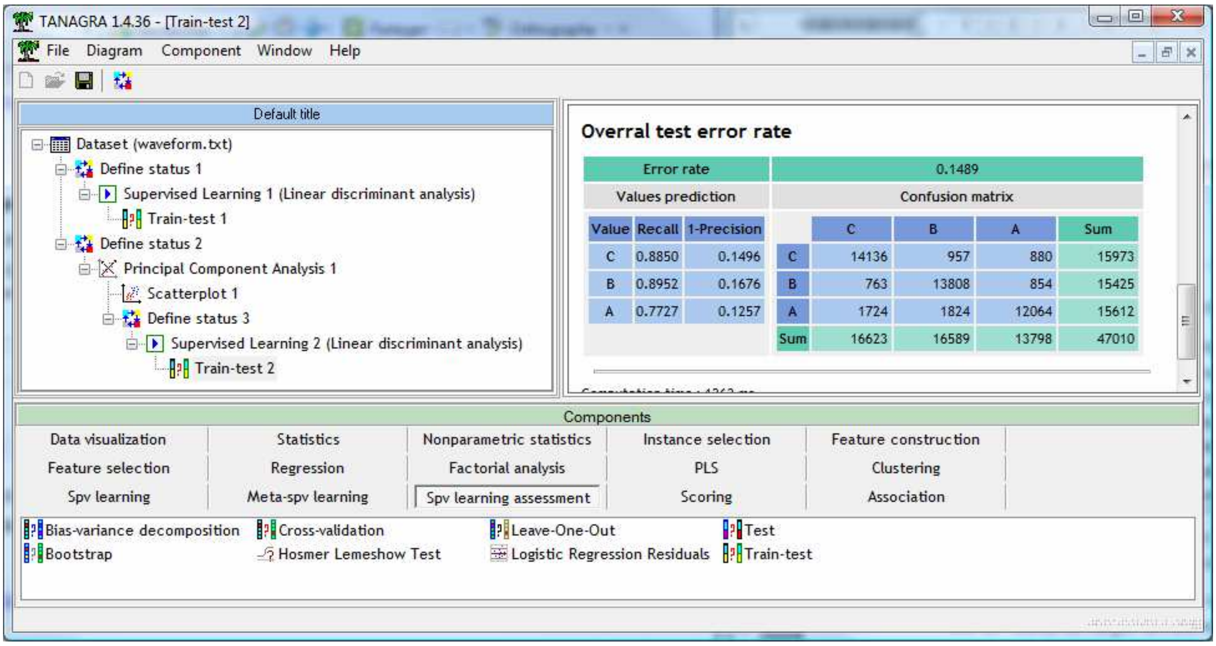


```
Dataset (waveform.txt)
 ├─ Define status 1
 │   └─ Supervised Learning 1 (Linear discriminant analysis)
 │       └─ Train-test 1
 └─ Define status 2
     └─ Principal Component Analysis 1
         └─ Scatterplot 1
```

In a third and last time, we want to use the two first computed factors of PCA as predictive variables for the LDA learning method.

13. We insert a new DEFINE STATUS component behind the "Principal Component Analysis 1". We set CLASS as target attribute. The two factors computed with the PCA (PCA_1_AXIS_1 and PCA_1_AXIS_2) are defined as input features.
14. Then we add the LINEAR DISCRIMINANT ANALYSIS tool (**Spv Learning** tab). We click on the VIEW menu. As before, the resubstitution error rate is 0.1350.
15. We add the TRAIN-TEST tool (**Spv Learning Assessment** tab) to obtain a more honest estimation of the error rate. We use the same settings as above (train = 0.06; repetition = 10). We click on the VIEW menu. We note that the whole path from the root of the diagram to the LDA is executed in each learning step. Especially, the characteristics of factors (eigen vectors) of the PCA are computed on the learning sample i.e. 300 instances.

16. **The test error rate is ≈ 0.15**. It is much better than this one obtained when we use the original descriptors as input features for the LDA.



```
Dataset (waveform.txt)
 ├─ Define status 1
 │   └─ Supervised Learning 1 (Linear discriminant analysis)
 │       └─ Train-test 1
 └─ Define status 2
     └─ Principal Component Analysis 1
         ├─ Scatterplot 1
         └─ Define status 3
             └─ Supervised Learning 2 (Linear discriminant analysis)
                 └─ Train-test 2
```

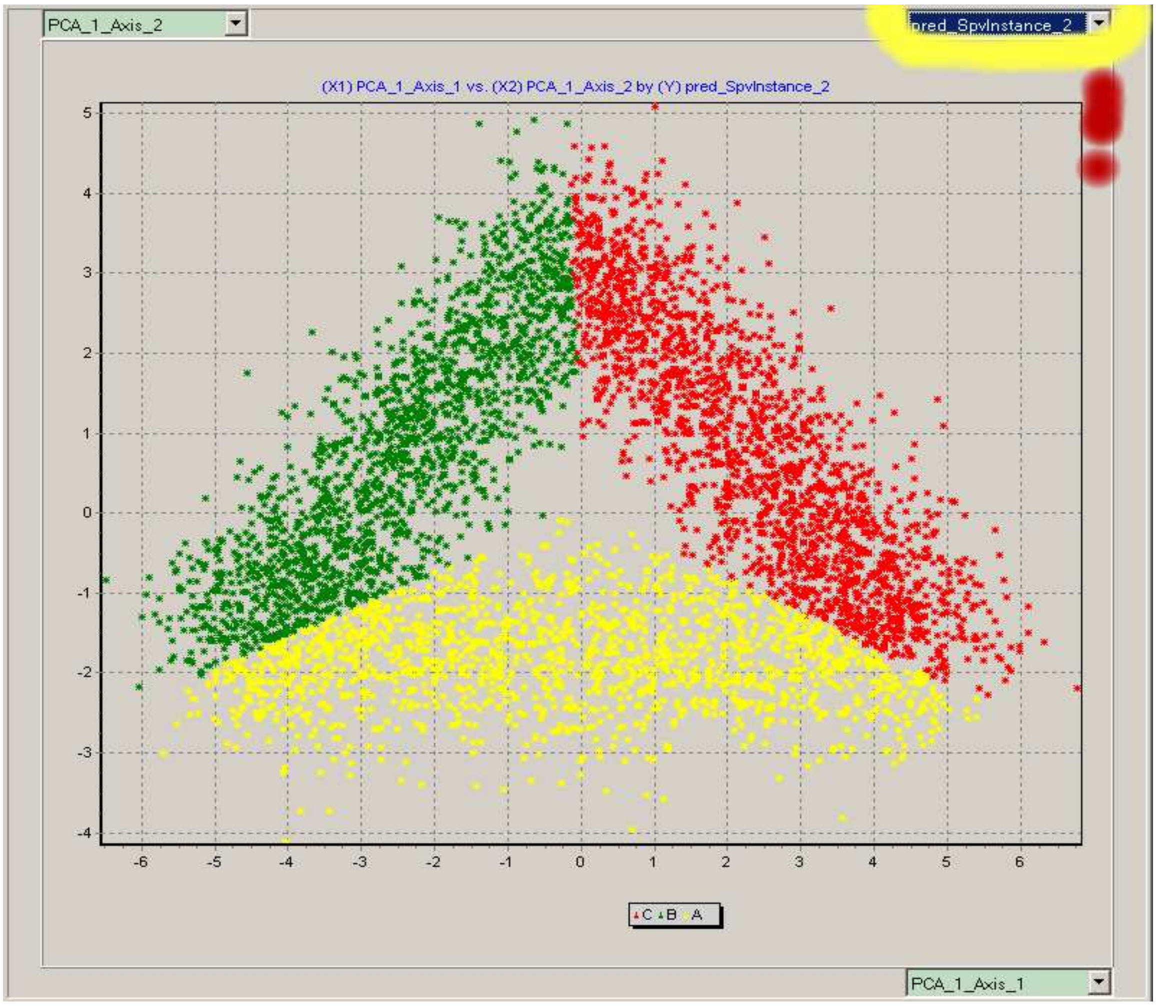
Overral test error rate

Error rate	0.1489
Values prediction	Confusion matrix

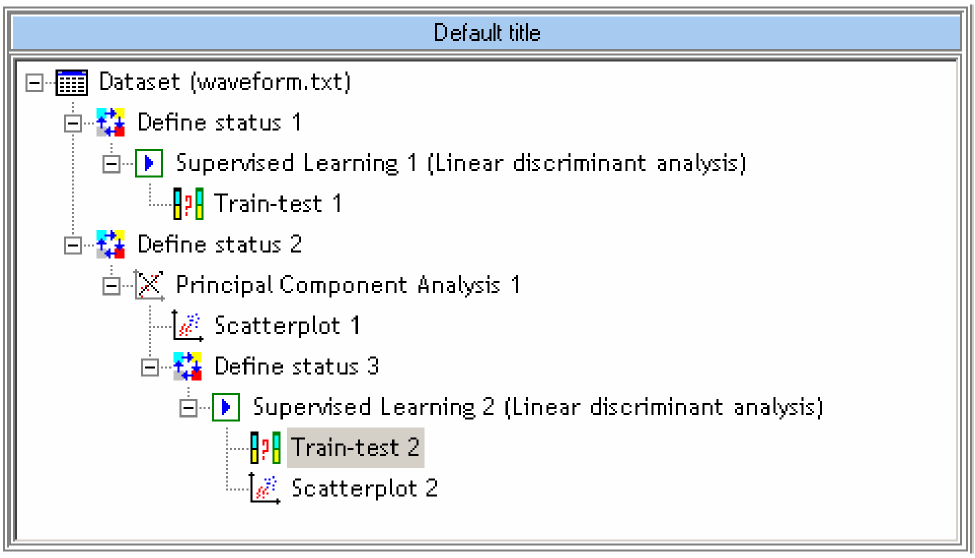
Value	Recall	1-Precision
C	0.8850	0.1496
B	0.8952	0.1676
A	0.7727	0.1257

	C	B	A	Sum
C	14136	957	880	15973
B	763	13808	854	15425
A	1724	1824	12064	15612
Sum	16623	16589	13798	47010

17. To understand the working of the LDA, which is a linear classifier, we project the regions delimited by the predictions of the classifier (PRED_SPV_INSTANCE_2) in the representation space defined by the factors of the PCA.



18. We understand than the errors of the model (misclassified instances) correspond to the configurations where the instances with different labels are overlapped.
19. We have defined the following diagram to obtain this result.



Default title

- Dataset (waveform.txt)
  - Define status 1
    - Supervised Learning 1 (Linear discriminant analysis)
      - Train-test 1
  - Define status 2
    - Principal Component Analysis 1
      - Scatterplot 1
      - Define status 3
        - Supervised Learning 2 (Linear discriminant analysis)
          - Train-test 2
          - Scatterplot 2