Feature and variable selection in complex data classification

Manuel Oviedo de la Fuente, University of A Coruña, CITIC (manuel.oviedo@udc.es)
Manuel Febrero Bande University of Santiago de Compostela

Introduction

This study addresses the classification of complex data such as spectrometric curves, hyperspectral images and 3D point cloud. The study also focuses on the procedures for feature and variable selection through the recursive use of distance correlation (DC). For this, the functional data analysis (FDA) framework will be used through the R package fda.usc.

Feature and variable selection in logistic regression

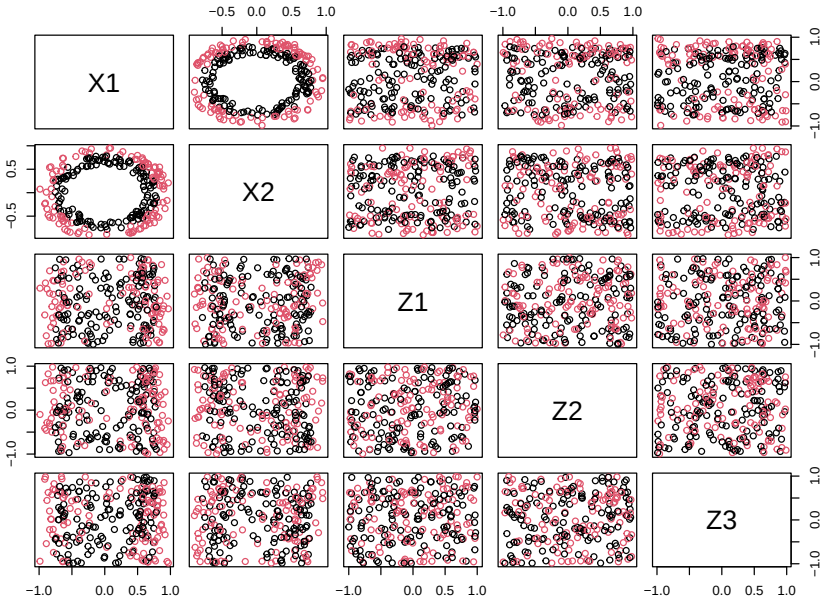
Febrero-Bande et al. (2019) consider the problem of variable selection in regression models in the case of functional variables that may be mixed with other type of variables (scalar, multivariate, directional, etc.). Their proposal begins with a simple null model and sequentially selects a new variable to be incorporated into the model based on the use of DC proposed by (Székely et al., 2007), $Y_i = \alpha + \sum_{j=1}^{J} f_j(X_i^{(j)}) + \varepsilon_i, \quad i = 1, \ldots, N$

We are interested in an automatic regression procedure capable of dealing with a large number of covariates of different nature. We adapt the variable selection procedure propossed by Febrero-Bande et al. (2019) from regression to classification.

Simulation Example

Multi class classification is implemented by training multiple logistic additive regression classifiers (one vs all scheme) using incoming function `classif.gsam.vs()`.

```
Nt=250; Np=100; nB=100;Nvar = 19 ; Xdat = simul2d(Nt,Np,Nvar) # data generation
Xtrain <- Xdat$Xtrain; Xtest <- Xdat$Xtest; pairs(Xtrain[,2:6],col=Xtrain$grupo)
```




```
gsam <- classif.gsam.vs(ldata(Xtrain[,1:(Nvar-1)]),"grupo")
gsam$i.predictor # Varable selected 1,  otherwise 0
```

```
##  X1  X2  Z1  Z2  Z3  Z4  Z5  Z6  Z7  Z8  Z9 Z10 Z11 Z12 Z13 Z14 Z15
##   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
```

```
pred <- predict.classif(gsam,ldata(Xtest[,1:(Nvar-1)]))
table(Xtest$grupo,pred) # Confusion matrix
```

```
##    pred
##       1  2
##   1 41  1
##   2  0 58
```

```
mean(Xtest$grupo==pred) # Accuracy
```

```
## [1] 0.99
```

Complex data classification

- **Hyperspectral image classification.** In this second case study, we also review different models classification algorithms for the prediction of the future class of pixel in a hyperspectral image that have in common that make use of FDA.
- **3D point cloud classification.** Finally, we will reproduce the examples of Oviedo-de la Fuente et al. (2021) to select optimum scales in multiscale classification problems with machine learning. A maximum of three scales for each feature was sufficient to obtain the best results in the classification, measured in terms of precision, recall and F1-index.

Funding and Financial Support:

CITIC funding Consellería de Economía, Empleo e Industria and FEDER Galicia 2014-2020) and MODES group by the Xunta de Galicia (ED431C-2020-14 and ED431G 2019/01).

References

Febrero-Bande, M., González-Manteiga, W., and Oviedo de la Fuente, M. (2019). Variable selection in functional additive regression models. *Computational Statistics*, 34(2):469–487.

Oviedo-de la Fuente, M., Cabo, C., Ordóñez, C., and Roca-Pardiñas, J. (2021). A distance correlation approach for optimum multiscale selection in 3d point cloud classification. *Mathematics*, 9(12):1328.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794.