G0B16: Robust Statistics: Project 2

Prof. Mia Hubert, Prof. Stefan Van Aelst

Academic year 2022 – 2023

The project consists of analyzing a more extensive version of the stars and galaxies data set. The training data are in the "StarGalaxy_train_RobSta" csv file while the file "StarGalaxy_test_RobSta" contains clean test data. Both datasets contain 22 variables. Each star/-galaxy has been observed in the F, J and M color bands which yields seven variables for each band. For example, for the F band, three variables measure light intensity (MAperF, MTotF, and MCoreF). The variable AreaF measures the size of the star based on its number of pixels, while the variables IR2F, csfF, and EllipF combine size and shape. Corresponding names are used for the variables in the J and M color bands. The first variable in both datasets is a classification of each observed light object as `star` (Class=0) or `galaxy` (Class=1) based on the expert opinion of astronomers.

Each student draws an individual random data set of 1000 observations from the training data. You use the following code, where you change 0012345 by your student number.

```
set.seed(0012345)
traindata <- read.csv("StarGalaxy_train_RobSta.csv", row.names=1)
nsamp <- 1000
mytraindata <- traindata[sample(nrow(traindata), nsamp), ]
```

You answer the questions by performing an appropriate analysis with `R`. The results including figures and tables, and the discussion of the results are reported in a written text (pdf) that consists of a maximum of 6 pages (12pt font size). Put the figures and tables when they are discussed, not at the last pages of the report. Only report results and interpretations, do not repeat theory from the course. Additionally a separate file with the full `R` code should be provided. You are allowed to use `R` Markdown to produce your report. Please, upload both files using the naming convention ''`familyname_firstname_Project2.extension`".

Your report and `R` script should be uploaded on Toledo before **June 11, 2023, 23:59h**. This project is graded on 5 points.

Good luck!

Assignment

1. The goal is to construct a robust classifier which can predict the class label (star or galaxy) for each light object based on its 21 measurements in the three color bands. The classification methods that can be used depend on the date for your examination.

   - Students with examination on **20 June 2023** use a logistic regression model to classify the objects. A logistic regression model assumes that the response (the class label) follows a Bernoulli distribution with success probability p_i which depends on the covariates $\mathbf{x}_i$ through the relation

     $$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \boldsymbol{\beta}^t\mathbf{x}_i \qquad i = 1, \ldots, n$$

     with inverse relation

     $$p_i = \text{expit}(\boldsymbol{\beta}^t\mathbf{x}_i) = \frac{1}{1+\exp(-\boldsymbol{\beta}^t\mathbf{x}_i)} \qquad i = 1, \ldots, n.$$

     The standard estimator for the coefficient vector $\boldsymbol{\beta}$ is the maximum likelihood estimator which is not robust. As alternatives the Mallow's type M-estimator of (Cantoni and Ronchetti, 2001) or the M-estimator studied by (Croux and Haesbroeck, 2003) can be used to fit the model. Both estimators can be fitted in `R` by using the function `glmrob` in package `robustbase`.

     Choose one of both estimators to construct a classification model using your training data. Can you identify outliers in the training data? Use the classification model to predict the probabilities for the objects in the testdata and classify the test objects based on these predicted probabilities. What is the best cutoff point for the predicted probabilities to separate stars from galaxies?

   - Students with examination on **27 June 2023** apply principal component analysis to estimate for each group a best fitting affine subspace for the objects in that group. For any object $\mathbf{x}_i$ we can then consider its projection on either subspace

     $$\hat{\mathbf{x}}_i^{(r)} = \hat{\mu}^{(r)} + \mathbf{P}_{p,k_r}^{(r)}\mathbf{t}_i^{(r)} \qquad r = 1, 2$$

     with $\mathbf{P}_{p,k_r}^{(r)}$ a loading matrix of size $p \times k_r$ and $\mathbf{t}_i^{(r)}$ the k_r dimensional score of $\mathbf{x}_i$ for this subspace.

     An object $\mathbf{x}_i$ can then be assigned to the group r for which the Euclidean distance between $\mathbf{x}_i$ and its projection $\hat{\mathbf{x}}_i^{(r)}$ is minimal.

     As PCA estimators you can choose between the robust sparse LTS estimator of (Wang and Aelst, 2020) or the cellwise robust MacroPCA estimator of (Hubert et al., 2019). The robust sparse LTS estimator is available in `R` by using the package `ltsspca` (functions `ltsspca` and `ltsspcaRw`). MacroPCA can be fitted in `R` by using the function `MacroPCA` from package `cellWise`.

Can you identify outliers in the training data? Construct the classification model on the training data and classify the objects in the test data.

References

Cantoni, E. and E. Ronchetti (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association 96*(455), 1022–1030.

Croux, C. and G. Haesbroeck (2003). Implementing the bianco and yohai estimator for logistic regression. *Computational Statistics & Data Analysis 44*(1), 273–295.

Hubert, M., P. J. Rousseeuw, and W. V. den Bossche (2019). Macropca: An all-in-one pca method allowing for missing values as well as cellwise and rowwise outliers. *Technometrics 61*(4), 459–473.

Wang, Y. and S. V. Aelst (2020). Sparse principal component analysis based on least trimmed squares. *Technometrics 62*(4), 473–485.