Regression Analysis Applied to Breast Cancer Data: From Linear to Penalized Models

Brewton Morais
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
brewtonlmorais@gmail.com

Lucas Abdalah
DETI
Universidade Federal do Ceará
Fortaleza, Brazil
lucasabdalah@alu.ufc.br

Abstract—Breast Cancer is one of the most aggressive cancer types and has a great impact in cancer mortality, mainly in women. This work analyzes tumor cells characteristics in order to provide an effective method to preprocess and reduce the dimension of the data, since the presence of several predictors may provide redundant information, what increases the cost of a Machine Learning-based techniques. In this paper, we present a framework based in Principal Component Analysis (PCA) and data normalization to clean the data and extract only the more relevant parameters, that can preserve original data characteristics and feed a predicting model to provide a final diagnosis.

Index Terms—Breast Cancer, Dimensionality Reduction, Linear regression, Machine learning, Principal Component Analysis.

I. INTRODUCTION

[1].

II. METHODS

A. Data Overview

B. Data Preprocessing

Z-Score Normalization

C. Cross Validation

D. Linear Regression

E. Penalized Models

F. Principal Component Regression

G. Partial Least Squares

III. RESULTS

Linear Regression Model
Penalized Ridge Model
Principal Component Regression
Partial Least Squares

IV. DISCUSSION

V. CONCLUSION

VI. FURTHER WORK

REFERENCES

[1] Dummy, *Dummy Title*, 2nd ed. publisher, 2008.