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. Notation

To ease the comprehension of this work, this section summarizes the notation used in the presend paper and introduces some definititions.

Scalars, vectors, matrices are represented by lower-case $(a, b, \dots)$, boldface lower-case $(\mathbf{a}, \mathbf{b}, \dots)$ and boldface capital $(\mathbf{A}, \mathbf{B}, \dots)$, respectively. The matrix transpose operator is represented by $(\cdot)^T$ and the symbol $\hat{(\cdot)}$ represents an estimated value.

B. Data Overview

C. Data Preprocessing

Z-Score Normalization
Principal Component Analysis

D. Linear Regression

Linear regression is a simple and useful tool for supervised learning. It consists in an approach for predicting quantitative outcome $\mathbf{y}$ based on a single predictor $\mathbf{x}$. It is assumed a linear relationship between predictor and outcome [2].

$$\mathbf{y} \approx \beta_0 + \beta_1 \mathbf{x}\,. \tag{1}$$

The model in Eq. 1 presents two constants, β_0 and β_1, which stands for intercept and slope, respectively. Hence, the main goal is to estimate both coefficients to estimate an correlation between $\mathbf{x}$ and $\mathbf{y}$.

We use a set for training to fit our estimated parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, to predict any sort of model based on linear correlation with a mean-zero random error term ϵ, a catch-all variable to acumulate what the model misses, since in general its true relationship is not linear [2].

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x} + \epsilon\,, \tag{2}$$

Taking advantage on the estimated parameters, we compute Eq. 2 to predict a continuos outcome.

However, the coefficients are unkwnon in practice and the objective of the ordinary least squares (OLS) linear regression is to find a plane that minimizes the residual sum of squares (RSS) between the observed data and the predicted response, i.e, the variance (σ^2) of the error.

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2\,, \tag{3}$$

where the i-th term of e represents $y_i - \hat{y}_i$.

To assess the linear regression quality we may use the following indices: Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and R^2 Statistic.

$$\text{MSE} = \frac{1}{n}\text{RSS} \tag{4}$$

The error variance usually is unknown, so we may estimate it from the data taking advantage on the model in Eq. 4. The idea is extended by computing the square root of this model to obtain the RMSE, i.e, $\text{RMSE} = \sqrt{\text{MSE}}$. Nevertheless, the value os these indices range accordingly with the $\mathbf{y}$ unit [3]. In order to overcome this limitation, the R^2 statistic provides another meausure of quality between 0 and 1, independent of $\mathbf{y}$ scale. In general terms, it provides an index to measure the amount of variability that is left unexplained in the model fit. It implies that as the R^2 values approximates to 1, a large proportion of the data variability is explained by the regression.

E. Partial Least Squares

O modelo de regressão de mínimos parciais pode ser visto como uma junção das funcionalidades dos modelos de regressão linear, que buscam maximar a correlação dos

preditores com saída, e os em componentes principais, que capturam as maiores variâncias nos preditores. Assim, é um método supervisionado que gera novas componentes que tenham máxima covariância com a saída, assim permitindo um número menor de componentes necessárias em relação ao PCR. Entretanto o problema da interpretabilidade dos novos preditores ainda persiste. Existem alguns algorítimos para o calculo do PLS, como o NIPALS e o SIMPLS.

F. Penalized Models

The OLS regression approach provides an unbiased and low variance model. Although, this simple moodel present quite accurate predictions for proper data, its MSE perfomance can be improved by the addition of the sum of the squared regression parameters weighted by a penalization/regularization term (λ) [3].

$$\text{RSS}_{L_2} = \text{RSS} + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{5}$$

The goal with the model presented in Eq. 5 is to allow a small increase in bias, which results in a substancial drop in the error variance. It imposes a new constraint to observe, the experimental search for an optimal λ value, to obtain an overall MSE lower than unbiased model [2], [3].

G. Principal Component Regression

Como o conjunto de dados utilizado possui um grande número de preditores, seria interessante reduzir esse número para tornar o modelo mais simples e menos custoso computacionalmente. Para isso, uma das estratégias é achar as chamadas componentes principais, que são definidas pelos autovetores da matriz de covariância dos preditores. Assim projeta-se os dados em um número reduzido de preditores, aquelas atreladas aos maiores autovalores, ou seja as que apresentam uma variabilidade maior. Existem dois problemas com esse método, o primeiro é que se torna difícil a interpretação das componentes, o segundo é que o método não define as componentes pela sua relação com a saída, o que pode fazer que as componentes dominantes não apresentem correlação com a saída, o que prejudica o desenvolvimento de um modelo eficiente, pois não se pode ter controle sobre a relação dos novos preditores com a saída.

H. Cross Validation

A validação cruzada consiste numa técnica para avaliar o modelo dentro conjunto de teste, geralmente levando em conta a flexibilidade do modelo e o erro quadrático médio. Em suma divide-se o conjunto em k grupos distintos de tamanhos semelhantes, esses grupos um é removido e passa ser o conjunto de validação, então se produz os modelos a partir das amostras restantes, e para verificar eu funcionamento tenta-se prever o conjunto de validação. Então o grupo removido retorna ao conjunto de treino e o grupo seguinte é removido e se torna o conjunto de validação. Esse processo é repetido até que todos os grupos sejam utilizados como validadores.

Essa estratégia serve muitas vezes para indicar quais modelos terão uma previsão melhor no conjunto de teste, visto que permite comparar os níveis de erro e variância gerado pelos modelos. Importante ressaltar que usar k muito pequeno (como apenas em dois grupos k = 2) ou muito grande (k = número de amostras) gerará problemas. No primeiro caso poderá um alto enviesamento dos modelos, visto que muitas amostras serão deixadas de fora do grupo de treino, e no segundo ocorrerá muita variância nos modelos pois os grupos utilizados no método são muito semelhantes. Portanto para compensar ambos os efeitos costuma-se usar k = 5 ou 10, visto que experimentalmente apresentam níveis aceitáveis de variância e enviesamento.

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.

[2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. [Online]. Available: https://faculty.marshall.usc.edu/gareth-james/ISL/

[3] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer, 2013.