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

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.

D. Linear Regression

E. Notation

The symbol $\hat{(\cdot)}$ represents an estimated variable. The transpose operator is represented by $(\cdot)^\intercal$.

F. Linear Regression

Linear regression is a simple and useful tool for supervised learning. It consists in an approach for predicting quantitative outcome Y based on a single predictor X. It is assumed a linear relationship between predictor and outcome [2].

$$Y \approx \beta_0 + \beta_1 X\,. \tag{1}$$

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 X and 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.

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

Taking advantage on the estimated parameters, we compute eq. 2 to predict an continuos outcome.

Os modelos de regressão linear são um conjunto de técnicas de análise estatística de dados, as quais tem por objetivo prever uma tendência de comportamento de um dado, a partir de outros dados já obtidos, tendo como pré-suposto que cada um dos dados já conhecidos, os preditores, tem uma relação linear com o dado de saída.

Portanto esses modelos são definidos por seguirem a seguinte equação:

$$y_i = \beta_0 + \sum_{i=1}^{n} (\beta_i x_i + e_i) \tag{3}$$

Sendo yi o valor de cada saída, xi cada preditor, β_i os coeficientes que determinam a relação de cada um dos preditores

com a saída, e ei os erros que não podem ser previstos, ou seja, o típico ruído. Os parâmetros β_i não podem ser definidos com abso- luta certeza, em virtude dos elementos de erro ei. Entre- tanto, sendo o erro suficientemente pequeno, é possível obter uma boa aproximação desses valores. O objetivo se torna definir o plano que minimiza a soma dos er- ros quadráticos entre os valores reais e os estimados, a ponto que a diferença seria apenas gerada pelo fator de erro ei. Sendo SSE a soma dos erros quadráticos e yi o valor de cada predição do modelo. Assim, podemos obter esti- mativas coerentes a partir da seguinte equação vetorial: β = (XT X) - 1 XT y (3) Onde X = (X1, X2, . . . , Xn), y o vetor com cada res- posta, e β o vetor com a estimação dos parâmetros β_i. Métodos desse tipo são aplicáveis em diversos tipos de previsões, desde observações químicas a efeitos econômi- cos. Isso se deve ao fato de ser facilmente compreensível e de fácil aplicação. Porém, caso haja uma relação não linear relevante entre os preditores e a saída, esses mode- los não serão capazes de prever corretamente a variável desejada.

G. Penalized Models

Os modelos de regressão simples costumam não ter enviesamento, porém podem podem apresentar um certo nível de variância em relação aos resultados finais com suas previsões. Então, para tal visa-se aumentar levemente o enviesamento dos dados a fim de obter um decrescimento mais expressivo na variância do modelo. Para isso altera-se a equação de minimização da somas dos erros quadráticos adicionando um novo elemento: Sendo λ um parâmetro a ser determinado experimentalmente. A equação 4 serve para controlar as grandezas dos parâmetros β_i, permitindo que seus valores sejam altos apenas se contribuírem para a minimização da equação. Então a medida que λ aumenta os fatores β_i tendem a diminuir, podendo muitas vezes tornar o preditor associado desprezível. Por isso esse método faz parte do conjunto de métodos denominados métodos de diminuição. Portanto sendo feito os ajustes do parâmetro λ pode-se chegar a um modelo com variância menor e com baixo enviesamento, o que pode tornar o modelo mais atrativo que a regressão simples.

H. 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.

I. 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.

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/