Predição de Solubilidade Através de Modelos de Regressão Linear

Daniel Lemos Simões
398985
daniellems@alu.ufc.br

Gustavo Filipe do Nascimento
402889
gusnas@alu.ufc.br

Pedro Cercelino Matos
399325
*p*edrocercelino@alu.ufc.br

Clailton Almeida Lopes
400091
*c*lailtonx2@gmail.com

Lucas Martins de Oliveira
398900
*e*ng.lucas@alu.ufc.br

***Resumo*—Neste artigo, foi estudado os diferentes modelos de regressão linear. Primeiramente foi realizado o pré processamento dos dados, removendo a skewness dos conjuntos de teste e de treino utilizando a Transformação de Yeo Johnson. Depois, fez-se uso de diferentes modelos de regressão linear para predição das solubilidades dos compostos, e logo em seguida foi testado cada resultado obtido por cada um dos modelos. Assim sendo, analisou-se as relações entre a estrutura dos compostos químicos com suas devidas solubilidades. Dentre os resultados, o modelo que se saiu melhor foi o linear.**

***Index Terms*—linear regression, solubility, regression models**

I. Introdução

Pode-se definir solubilidade como a propriedade que uma substância tem de se dissolver em algum líquido. A solubilidade é um ótimo indicador, e pode ser útil em diversas aplicações, por exemplo na separação de misturas e na síntese de compostos químicos.

Com as informações supracitadas, é possível enxergar a quantidade de dados que podemos extrair apenas com os níveis de solubilidade de um composto, assim como será visto futuramente.

Primeiramente, foi feito o pré-processamento dos dados a fim de alcançarmos o melhor resultado possível e removermos possíveis dados indesejados de nosso conjunto, posteriormente, foi utilizado modelos de regressão linear para a predição das solubilidades de cada composto. Ao final, os modelos foram analisados a fim de obtermos qual modelo se saiu melhor nas predições.

II. Metodologia

O conjunto de dados que será trabalhado contém 1267 amostras de um certo composto, e em cada amostra existem 228 preditores, dentre esses, 208 são indicadores de presença ou ausência de uma subestrutura, 16 mostram a quantidade de ligações e de átomos de bromo e os 4 restantes indicam peso molecular e área de superfície.

0) Pré-processamento
   Analisando a matriz de correlação (Figura 6) percebe-se uma alta correlação entre algumas variáveis, devido a isso, foi preciso removê-las para evitar um aumento no viés e consequentemente no erro do modelo.
   No pré-processamento, foi possível notar uma assimetria no conjunto de dados que fora fornecido, portanto, a fim de que os modelos a serem construídos sejam os mais precisos possíveis, devido aos valores negativos presentes no preditor de fator hidrofílico (**HydrophillicFactor**), foi utilizada a Transformação de Yeo-Johnson.

1) Ordinary linear regression.

   Um dos primeiros modelos de regressão a serem estudados rigorosamente pela estatística, a Regressão Linear é um método que propõe que a relação de uma resposta (saída) às suas variáveis pode ser descrita como uma equação linear, cujo os coeficientes β podem ser estimados através diversas formas, sendo a que usaremos chamada de Método dos Quadrados Mínimos (MQM):

$$RSS(\beta) = \sum_{i=1}^{N}(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

   Onde se o objetivo, é achar βs que façam a função se aproximar de 0, indicando uma boa adaptação do modelo. Foram utilizadas amostras de um conjuntos de teste e reamostragem de Validação-Crusada K-fold para avaliar a capacidade preditiva do modelo desenvolvido computacionalmente, obtendo no primeiro método um erro (RMSE) considerávelmente maior que no segundo. Abaixo podemos observar a comparação dos valores preditos após a validação crusada k-fold (onde devido à quantidade de amostras no conjunto de treino foi possível dividir o conjunto em 10 partes, ou seja k = 10) com os valores reais da saída.

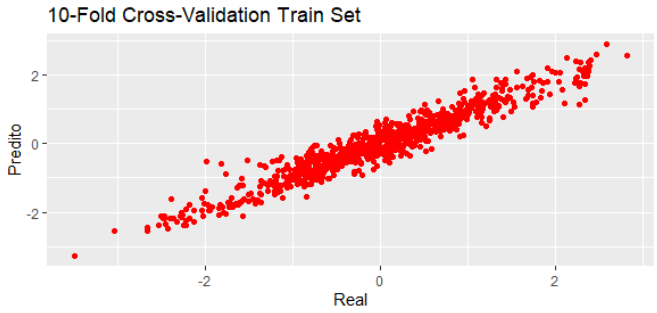

Figura 1. Aplicação da Validação Cruzada 10-Fold ao Conjunto de Treino

Analisando o gráfico , é notável o nível de linearidade entre os valores preditos e os valores reais, indicando boa adequação do modelo desenvolvido.

2) L2-Penalized linear regression.

Nos modelos desta categoria, tem-se a inclusão de um termo λ responsável por penalizar o modelo, diminuindo o erro geral, mas compensando no aumento da variância. L2 atua adicionando um limite que puxa e regulariza os parâmetros da regressão linear, suavizando o modelo como um todo. λ tem de ser escolhido atentando para que o seu valor não produza um bias muito grande nas predições do modelo. Para este trabalho, λ foi escolhido entre 10 valores posíveis. Sendo o final igual a 0.01778279.

O modelo escolhido foi o Ridge. Este é um método de encolhimento (shirinkage) que busca suavizar a colinearidade entre as partes do dataset. Seu algoritmo tenta minimizar a participação dos atributos que menos influenciam nas predições, dando mais espaço aos que possuem mais influência para guiarem a uma predição mais exata.

Para validação foi usado o método k-fold cross validation que consiste na divisão do dataset em uma parte para treino e outra para teste. Neste trabalho utilizou-se 10-fold para validar o modelo escolhido.

Como métricas de qualidade foram gerados o RMSE e Rsquared, respectivamente, iguais a 0.4015807 e 0.8408091.

As imagens a seguir foram plotadas pela aplicação do método Ridge ao conjunto de treino e em seguida ao conjunto de teste.

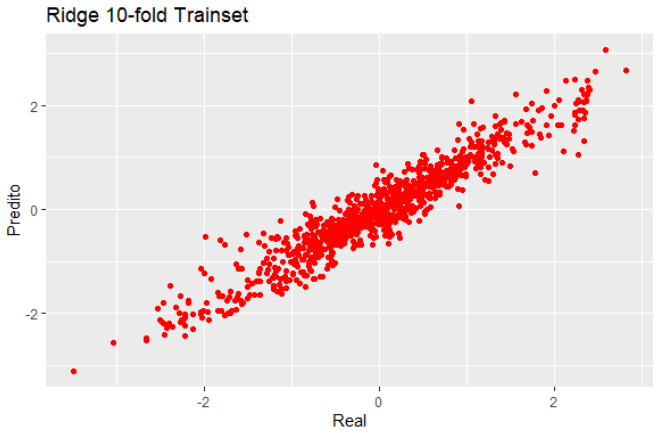

Figura 2. Aplicação do Ridge ao Trainset.

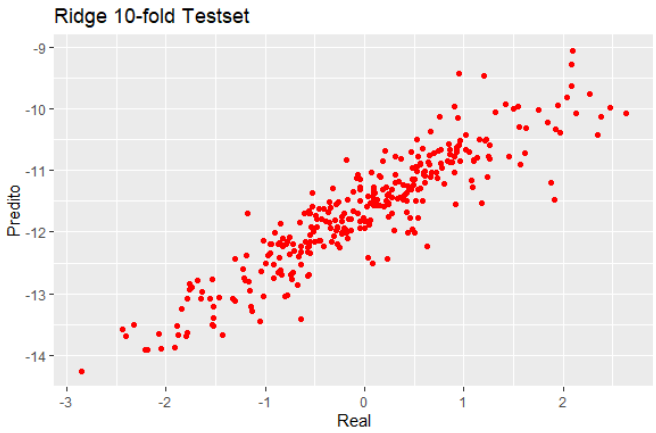

Figura 3. Aplicação do Ridge ao Testset.

3) Principal Component Regression and Partial Least Squares

Para modelar os dados e realizar predições sobre eles, também podemos utilizar dois métodos bastante úteis, a Regressão por componentes principais (PCR) e Regressão de mínimos quadrados parciais (PLS), ambas as regressões possuem conceitos semelhantes de reduzir a dimensionalidade das variáveis preditoras e posteriormente efetuar uma regressão linear para fazer previsões sobre os dados.

A diferença entre a PCR e a PLS é que a regressão de mínimos quadrados parciais é um método de aprendizgem supervisionada, ou seja, leva em consideração não só os preditores e como eles se organizam, mas também a resposta, preservando a direção dos componentes que melhor se relacionam com a resposta e com as variáveis preditoras. Neste trabalho, optamos por realizar a regressão por componentes principais e por isso, é nela que nos aprofundaremos para explicar os métodos de regressão linear por redução de dimensão.

$$Z1 = \sum_{j=1}^{p} \phi j1.Xj \quad (1)$$

A equação acima, denota o cálculo da direção Zp que é utilizada para construir o modelo de regressão. Na PLS esse cálculo é feito com ϕ sendo o coeficiente de regressão de Y em Xj, já na PCR o cálculo da direção Z é feita apenas com base nos preditores.

Como já dito no trabalho anterior, reduzir a dimensionalidade da matriz dos dados é importante quando estamos lidando com um número muito elevado de preditores ou quando esse valor embora pequeno, supere o número de amostras disponíveis. Para mitigar esse problema, na etapa de pós-processamento, utilizamos a técnica de PCA (Principal Component Analisys).

Na técnica de PCR, realizamos o PCA para diminuir a dimensão da matriz e eliminar preditores multicolinearizados, retendo apenas os preditores que não são correlacionados e em seguida realizamos uma regressão linear de mínimos quadrados. Uma vantagem de se utilizar a técnica de PCR é a atenuação do sobreajuste do modelo, pois como estamos retendo a quantidade mínina de preditores que representam os dados, novas amostras tendem a se adaptar melhor e ter maior acurácia do que uma regressão linear ordinária que leva em consideração todos os preditores.

Embora a PCR tenha vantagens consideráveis, temos que se ater aos pontos negativos também, como o fato de que a técnica do PCA é um método de aprendizagem não-supervisionada e por isso, nos baseamos apenas na forma como os dados estão organizados para eliminar preditores, sem o feedback da variável de interesse para nos guiar, fato que pode diminuir consideravelmente a precisão do modelo preditivo pois afeta a direção dos componentes que possuem uma maior variabilidade.

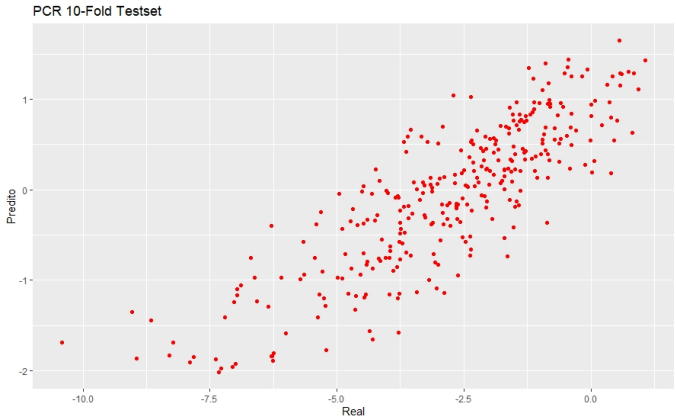

Figura 4. Aplicação da PCR ao conjunto de teste.

O gráfico de dispersão da figura 4, exemplifica a regressão linear por componentes principais aplicada aos dados, a predição foi prejudicada por conta da PCR ser uma técnica não-supervisionada como dito anteriormente.

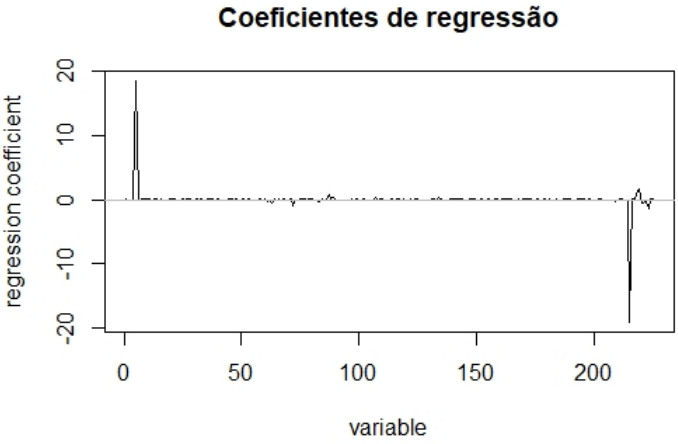

Figura 5. Coeficentes da regressão de componentes principais.

O gráfico da figura 5 mostra a estimativa dos coeficientes de regressão do modelo, percebemos que há valores fora do padrão, que acabam reduzindo a precisão da predição. Embora ainda possamos considerar que para um certo grupo de valores, a predição possa ser considerada assertiva.

III. Resultados

Com base na análise dos gráficos dos modelos de regressão linear, temos que o modelo mais eficaz em realizar a predição dos dados foi a regressão linear ordinária, visto que o parâmetro de desempenho RMSE (Root Mean Square Error) é pequeno, indicando que o coeficiente de correlação é próximo de 1, o que significa que a previsão possui um baixo viés (BIAS). Enquanto que o parâmetro R^2 ou coeficiente de determinação, foi de 0.8396, que mostra que 83.96 % da variável dependente consegue ser explicada pelo modelo. Logo, a regressão linear ordinária é a que melhor prevê a solubilidade dos compostos químicos neste conjunto de dados.

Referências

[1] **Kuhn, M., Johnson, K. (2013).** ***Applied Predictive Modeling*** **New York: Springer.**

[2] **Michela Mulas. (2019).** ***Data pre-processing*****. Slides de Aula.**

[3] **Cross-Validation Essentials in R. STHDA** ***(11/03/2018) Disponível em: Cross-Validation Essentials in R Acesso em: (10/10/2019)***

[4] **R Documentation.** ***Disponível em: R Documentation Acesso em: (06/10/2019)***

IV. Anexo

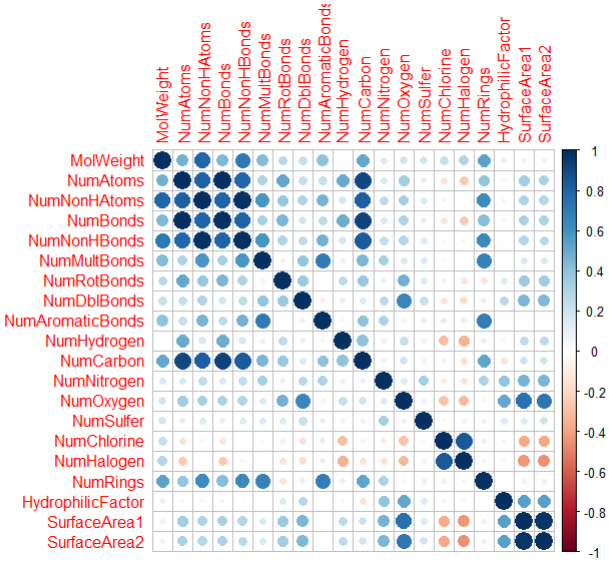

Figura 6. Matriz de Correlação