Análise de técnicas de regressão linear na predição da solubilidade de compostos químicos

Matheus Ferreira Lessa
Universidade Federal do Ceará
Fortaleza, Brasil
mathewslessa@gmail.com

Talles Felix Gomes
Universidade Federal do Ceará
Fortaleza, Brasil
tallesfelixg95@gmail.com

Paulo Victor Felício Lage
Universidade Federal do Ceará
Fortaleza, Brasil
victorlage7@gmail.com

***Resumo*—Quando o assunto é a predição de informações a partir de um conjunto de dados, inúmeros são as técnicas existentes para a criação de modelos relacionando preditores e saídas. Uma das mais famosas famílias de técnicas com esse finalidade é a família das regressões lineares. O objetivo desse trabalho consiste em comparar a performance de diferentes tipos de regressões lineares na predição da solubilidade de compostos químicos. Ao todo três modelos são analisados - OLS, Ridge e PLS - e seus resultados são apresentados e comparados ao final deste trabalho.**

***Index Terms*—Predição, solubilidade, pré-processamento de dados, regressão linear, regressão penalizada L2, regressão Ridge, Regressão parcial dos mínimos quadrados (PLS)**

I. Introdução

O método dos mínimos quadrados foi publicado pela primeira vez por Carl Friedrich Gauss em 1809, após Adrien-Marie Legendre e outros matemáticos terem criado seus fundamentos teóricos. Este método é considerado um precursor da análise de regressão que atualmente conta com diversas variantes e tem aplicações nas mais diversas áreas como por exemplo ciências naturais, estatística, finanças e marketing digital.

A análise de regressão é um método estatístico usado para modelar as relações entre diferentes variáveis (dependentes e independentes). Ela permite fazer previsões usando modelos criados com base nos dados existentes. A ideia por trás da forma mais simples de regressão consiste na definição de uma variável a ser predita e uma variável (regressão simples) ou várias variáveis (regressão múltipla) chamadas variáveis preditoras. A regressão consiste em construir uma variável regressiva combinando as variáveis preditoras o mais próximo possível da variável que queremos prever.

Como citado acima, as regressões podem ser classificadas como simples ou múltipla quando respectivamente têm-se apenas um ou múltiplos preditores. Além disso, pode-se também classificá-las como sendo lineares, quando busca-se uma relação linear entre os preditores e o resultado, ou não linear, para os casos em que uma relação linear não for capaz de representar bem os padrões dos dados. Em todo caso, as motivações para o uso de uma analise de regressão são as mesmas: quantificar as relações entre variáveis e fornecer predições. Neste trabalho, o foco principalmente é nas regressões lineares múltiplas ordinárias e suas derivações.

Dentre essas derivações podemos citar o modelo dos mínimos quadrados parciais (PLS) e os modelos penalizados, como a regressão Ridge, Lasso e Elastic Net. Enquanto a regressão linear ordinária, em um extremo, encontra estimativas de parâmetros com viés mínimo, a regressão Ridge, Lasso e Elastic Net encontram estimativas com menor variação, sendo assim interessantes em várias aplicações. Já o PLS estabelece um compromisso entre os objetivos de reduzir a dimensão do espaço dos preditores ao mesmo tempo que busca uma relação preditiva com a resposta, por isso ele pode ser visto como um procedimento de redução da dimensão supervisionado, diferente por exemplo, do PCR (Regressão do componentes principais) que é um procedimento não supervisionado.

Inúmeras são as aplicações práticas dos métodos citados anteriormente nas mais diversas áreas do conhecimento. Temos, por exemplo, as regressões lineares múltiplas sendo usadas na identificação de fatores que afetam os resultados do time de futebol do Chelsea [1] e também na melhoraria da precisão de previsões para redução de dados de uma rede de sensores sem fio [2]. Esse último realizados por pesquisadores do GREat[1]/UFC em parceria com o laboratório francês LRSM/IBIC. No que diz respeito a regressão Ridge, ela é usada principalmente para resolver problemas de multicolinearidade em regressão múltipla ordinárias. Suas aplicações são bem variadas e permitem abordar problemas relacionados, por exemplo, a gestão de recursos hídricos [3] e a análise de dados relativos à sífilis [4]. Por fim, a regressão parcial dos mínimos quadrados (PLS) também mostra seu potencial no estudo sobre a ocorrência de cabelos grisalhos prematuramente [5] e na previsão climática sazonal [6].

O objetivo deste trabalho consiste em fazer uma análise comparativa de três diferentes tipos de regressões lineares múltiplas - a ordinária, a Ridge e a PLS - para a previsão da solubilidade de compostos químicos. Essa análise foi dividida em duas partes principais. Na primeira parte (seção II) os dados são analisado e pré-processados. Além disso, uma visão global das técnicas de análise e dos três modelos de regressão usados neste trabalho são apresentados. Já na segunda parte (seção III), os três modelos são aplicados na predição da solubilidade de compostos químicos e os resultados são apresentados juntamente com as conclusões que puderam ser

[1]Grupo de redes de computadores, engenharia de software e sistemas

tiradas da análise.

II. Métodos

O conjunto de dados usado neste trabalho foi inicialmente estudado por Tetko et al. [7] e Huuskonen [8] que usaram modelos de regressão linear e de redes neurais para estimar a relação entre a solubilidade de compostos químicos e um conjuntos complexos de descritores.

Para a nossa análise, utilizaremos 1267 compostos que são descritos por 228 preditores pertencentes a três diferentes grupos: 208 deles são "impressões digitais"binárias que indicam a presença ou ausência de uma determinada subestrutura química, 16 são descritores de contagem (como o número de átomos ou o número de átomos de carbono) e 4 são descritores contínuos (como o fator hidrofílico ou a área superficial). Tal conjunto foi dividido em dois outros subconjuntos, sendo um deles o conjunto de treino e o outro o conjunto de teste compostos, respectivamente, por 951 e 316 observações.

O estudo da solubilidade de um composto é de fundamental importância em um grande número de disciplinas científicas e aplicações práticas, que vão desde o processamento do minério, previsões ambientais, uso para medicamentos, desenvolvimento agroquímico, bem como o transporte de poluentes. Frequentemente, a solubilidade é usada para descrever, distinguir e servir como um guia das possíveis aplicações de uma determinada substancia. Tudo isso é possível graças ao fato da solubilidade ser considerada uma propriedade característica de uma substância.[2]

Além do mais, a capacidade de prever com precisão a solubilidade de um composto representa, por exemplo, uma potencial economia financeira em muitos processos de desenvolvimento de produtos químicos, tais como produtos farmacêuticos.

Uma vez apresentados os objetivos desse trabalho e o conjunto de dados utilizados, na seção II-A é feita uma análise mais detalhadas dos dados e um pré-processamento dos dados é realizado. Em seguida na seção II-B, são apresentadas as métricas e técnicas utilizadas para a avaliação dos modelos ao longo do trabalho, essas métricas são de grande importância na comparação dos modelos. Nas seções seguintes (II-C, II-D e II-E), as três regressões que estão sendo comparadas são apresentados e as ideias por trás cada uma delas são exploradas.

A. Análise e pré-processamento dos dados

Antes da utilização dos dados no aprendizado dos parâmetros de algum modelo é importante fazer uma análise exploratória deles e aplicar algumas transformações necessárias tendo em vista os modelos que se deseja utilizar.

Em uma primeira análise vemos que não é possível fazer muitas transformações sobre os preditores binários, portanto o foco será sobre os vinte preditores contínuos. Analisando-os percebe-se que eles não estão nem escalonados nem centrados e apresentam uma certa assimetria. Então, com o objetivo de melhorar a estabilidade numérica, reduzir assimetria e trabalhar com um conjunto de dados mais homogêneo, foi utilizada a transformação de Yeo-Johnson e aplicadas as técnicas de centralização e escalonamento aos dados.

[2]Informações extraídas de https://pt.wikipedia.org/wiki/Solubilidade

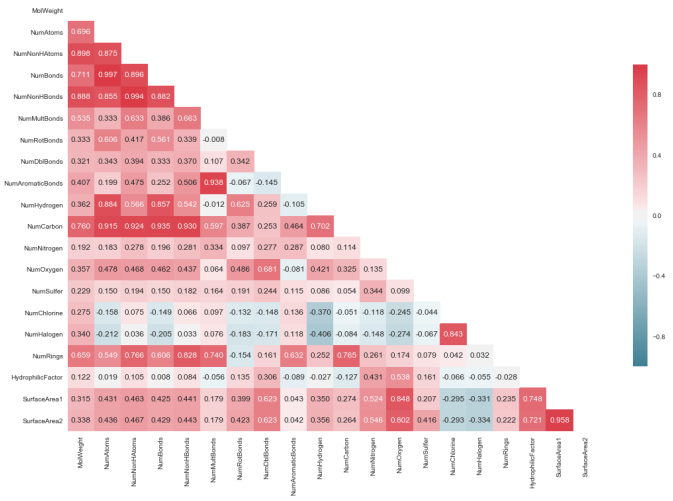

Figura 1: Matriz de correlação dos pares de preditores contínuos transformados

Tendo em mente que o objetivo deste trabalho gira em torno do uso de regressões lineares, é interessante analisar se os preditores são linearmente relacionados com a saída. Analisando a Fig. 2, percebe-se que alguns dos preditores parecem ser linearmente relacionados com a saída, enquanto outros não. De forma geral, a correlação dos preditores e o resultado é inferior a 0.6 para os vinte preditores contínuos. Um forma de superar esse problema seria expandindo o espaço dos preditores com termos não lineares para alguns dos preditores.

Para usar algumas técnicas de regressão linear é imprescindível ter uma atenção especial na correlação entre os preditores. Preditores muito correlacionados poderão gerar problemas no momento do cálculo dos coeficientes do modelo. A Fig. 1 mostra a matriz de correlação entre todos os preditores. Nela percebe-se que algum dos preditores são altamente correlacionados. Exitem 36 preditores com uma correlação superior a 0.9 com algum outro preditor do conjunto. Em alguns testes ao longo deste trabalho esses 36 preditores foram removidos.

B. Métricas para medir o desempenho dos modelos

Afim de verificar a performance dos modelos obtidos duas métricas foram utilizadas: a raiz do erro quadrático médio (RMSE) e o coeficiente de determinação (R^2). Além disso, a técnica da validação cruzada foi aplicada com o objetivo de obter valores mais significativos e realistas para as métricas utilizadas. As duas métricas e a técnica de validação cruzadas são descritas com mais detalhes a seguir.

1) Raiz do erro quadrático médio (RMSE): Essa métrica nada mais é do que a raiz quadrada da média dos valores dos resíduos ao quadrado, ou seja,

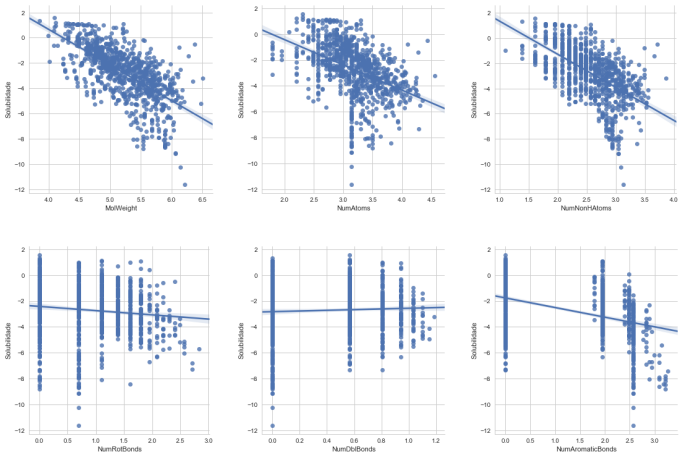

Figura 2: Gráficos de dispersão de alguns preditores contínuos transformados relacionados com a saída

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

onde y_i é a saída observada, $\hat{y}_i$ é a saída predita pelo modelo e i representa uma das n amostras.

Quando a saída é um número, o erro quadrático médio quadrático é o método mais comum para caracterizar as capacidades preditivas de um modelo. Seu valor é geralmente interpretado como a distância (em média) entre os resíduos e o zero ou como a distância média entre os valores observados e as previsões do modelo.[3]

2) Coeficiente de determinação (R^2): É uma medida que indica o quão ajustado um modelo de predição está em relação aos valores observados. Ela varia entre 0 e 1, indicando, em percentagem, o quanto o modelo consegue explicar os valores observados. Quanto maior o seu valor, mais explicativo é o modelo, ou seja, melhor ele se ajusta à amostra.

Existem múltiplas fórmulas para o cálculo dessa métrica, mas em sua versão mais simples ela consiste simplesmente no quadrado da correlação entre os valores observados e previstos.

3) Validação cruzada 10-fold: É uma técnica que auxilia na avaliação da eficácia do desempenho de um modelo de predição e é bastante usada para a estimação de hiperparâmetros.

Resumidamente, a validação cruzada 10-*fold* consiste em dividir aleatoriamente todo o conjunto de dados em dez partes iguais (*fold*). Então uma das dez partes é selecionada como o conjunto de validação e as outras nove partes constituirão o conjunto de aprendizagem. As métricas calculadas com o conjunto de validação são salvas e essa operação é repetida dez vezes para que finalmente cada parte tenha sido usada exatamente uma vez como um conjunto de validação. Por fim, a média das dez métricas registrados servirá como uma métrica global de desempenho mais confiável para o modelo.

[3]Embora não citado nesse trabalho, vale destacar que essa métrica é a base para a definição do famoso *Bias–variance tradeoff*

C. Modelo de regressão linear ordinária (OLR)

O modelo de regressão linear ordinária é um dos modelos mais simples que busca encontrar uma relação linear entre vários preditores (regressão múltipla) e uma saída numérica. Tal relação, pode ser expressa da seguinte forma

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_P x_{iP} + e_i \tag{2}$$

onde y_i representa a resposta numérica para i-ésima amostra; β_0 representa a interceptação estimada; β_j representa o valor do preditor j-ésimo para a amostra de i; x_{ij} representa o valor do j-ésimo preditor da i-ésima amostra; e_i representa o erro aleatório que não pode ser explicado pelo modelo.

Dada essa representação, o próximo passo consiste no aprendizado dos parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ visando obter um modelo que se ajusta o melhor possível aos dados. De forma mais precisa, os parâmetros β_i são aqueles que minimizam a soma dos quadrados da diferença entre os valores da resposta observa y_i e os valores da resposta predita usando o modelo $\hat{y}_i$,

$$\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Minimizando a função SSE com relação a β, obtemos uma expressão (4) para o cálculo dos parâmetros β.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

onde, com N sendo a quantidade de amostras,

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1P} \\ 1 & x_{21} & x_{22} & \cdots & x_{2P} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{NP} \end{pmatrix} \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

Além da estimação dos β usando o método dos mínimos quadrados é possível também estimá-los usando a técnica da máxima verossimilhança, onde assumimos que o erro $e_i \sim \mathcal{N}(0, \sigma^2)$ e i.i.d, consequentemente y_i também é normalmente distribuída com uma média que depende dos β e portanto, eles podem ser estimados através da máxima verossimilhança. Em ambos os casos o resultado é aquele apresentado em (4).

Vale destacar que pela forma em que a regressão linear ordinária é definida podemos concluir que os modelos obtido a partir dela terão um baixo viés e uma alta variância. É exatamente por causa dessa característica que as regressão penalizadas, apresentadas na próxima seção, se tornam interessantes.

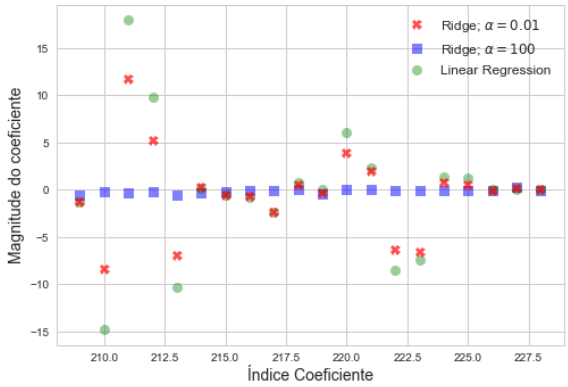

Figura 3: Comparação da influencia do parâmetro α na regressão Ridge sobre a magnitude dos coeficientes dos vinte últimos preditores para α = 0 (OLR), 0.01 e 100

D. Modelo de regressão linear penalizado L2 (Ridge)

Os modelos de regressão linear penalizados, em especial a regressão de Ridge que é usada nesse trabalho, são técnicas que têm como objetivo reduzir a complexidade do modelo e a multicolinearidade, além de prevenir o overfitting (dificuldade de previsão em novos conjuntos de dados) por meio do aumento do viés junto com uma possível minimização da variância.

A técnica funciona adicionando uma penalidade a soma dos quadrados dos resíduos, da forma:

$$\text{SSE}_R = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \qquad (5)$$

L2 significa que a penalidade que está sendo utilizada para a estimativa dos parâmetros é de segunda ordem (ao quadrado). Ao realizar a minimização da equação (5), percebemos que o termo que foi adicionado limita o crescimento dos parâmetros β_i. Em consequência, o modelo perderá um pouco sua capacidade de acompanhar grandes variações nos dados o que provoca uma redução da variância seguido normalmente por aumento do viés. Isso permite ter melhores resultados ao realizar predições com dados nunca vistos.

A característica de limitação do valor dos parâmetros pode ser vista na Fig. 3, onde é feita uma comparação com diferentes valores de α. Nela vemos que conforme o valor de $alpha$ aumenta, menores são os parâmetros do modelo de regressão linear. É importante destacar que embora a magnitude dos coeficientes se torne cada vez menor com o aumento do α ela será normalmente diferente de zero. Se a eliminação da influência de alguns preditores for desejada, é necessário usar outros modelos penalizados como o Lasso, em que o coeficiente de alguns preditores podem ser zerados.

A melhor forma para estimar o hiper parâmetro α consiste em escolher um conjunto de valores para α e com cada um deles realizar uma validação cruzada utilizando como métrica o RMSE, por exemplo. Ao final, para cada valor de α do conjunto teremos um valor do RMSE correspondente. O melhor α é simplesmente aquele que gerou o menor RMSE. Mais detalhes sobre esse processo são apresentados na seção III-C.

E. Regressão parcial dos mínimos quadrados (PLS)

A regressão parcial dos mínimos quadrados é uma técnica aplicada a conjuntos de dados com preditores altamente correlacionados, visto que com a grande variância dos dados, a regressão linear ordinária teria uma alta variabilidade e se tornaria instável. Além disso, ela pode também ser aplicada quando o número de preditores for maior do que o número de observações, caso onde a regressão linear ordinária não será capaz de encontrar uma resposta única.

Existem duas técnicas principais que podem ser usadas na presença de preditores altamente correlacionados, são elas a PCR (Regressão dos componentes principais) e a PLS (Regressão parcial dos mínimos quadrados). A primeira nada mais é do que a aplicação de uma regressão linear múltipla sobre os componentes obtidos através após aplicação da PCA[4].

No entanto, a redução da dimensão por meio da PCA não produz necessariamente novos preditores relacionados com a resposta, pois ele busca apenas resumir a variabilidade presente no espaço dos preditores. Portanto, se a variabilidade no espaço do preditor não estiver relacionada à variabilidade da resposta, então a PCR pode ter dificuldade em identificar uma relação preditiva mesmo quando ela realmente existir.

Exatamente devido essa característica da PCA, em que a variância do preditor pode não estar relacionada com a variância da resposta, a PLS se torna uma solução mais indicada para regressão linear quando existem preditores correlacionados e por isso ela foi escolhida para esta análise.

A PLS, assim como a PCR, encontra combinações lineares entre os preditores, com a diferença de que considera as respostas quando seleciona os componentes, sendo então um método supervisionado, diferente da PCR que apenas considera a variância presente no espaço dos preditores, ou seja, um método não supervisionado.

Em algumas aplicações a PLS e a PCR podem apresentar resultados semelhantes, mas mesmo nesses casos o PLS acaba, na maioria das vezes, levando vantagem por conseguir resultados semelhantes aos da PCR com um menor número de componentes. Isso acontece pois os componentes da PLS já possuem uma relação explícita com a resposta.

Vale destacar que a escolha do número de componentes é feita por meio da validação cruzada como no Ridge. Para mais informações sobre o algoritmo utilizado na PLS podem ser consultadas em [9].

III. Resultados e Conclusões

Tendo agora uma visão global do assunto com o qual estamos trabalhando, resta saber como as três regressões funcionam na prática. Para isso, cada uma delas foi usada para criação de um modelo de predição da solubilidade de compostos químicos. Os resultados obtidos são apresentados e

[4]O Homework 1 pode ser consultado para mais informações sobre a PCA

comparados nas seções a seguir. Nos três casos, o desempenho dos modelos são analisados por meio das métricas apresentadas na seção II-B.

A. *Modelo de regressão linear ordinária (OLR)*

Em um primeiro momento o modelo de regressão linear ordinária é treinado usando o conjunto de treino transformado com todos os preditores. Os seguintes resultados foram obtidos ao avaliar o conjunto de treino e o conjunto de teste

Conjunto treino

RMSE	R^2
0.4813	0.9446

Conjunto teste

RMSE	R^2
0.7456	0.8709

Como já era esperado, com o conjunto de treino os resultados foram melhores (menor RMSE e R^2 mais próximo de 1) do que com o conjunto de teste. No entanto, essa diferença pode indicar também um possível *overfitting* dos dados ou uma consequência da colinearidade entre os preditores.

É importante também verificar que os valores de RMSE e R^2 encontrados com os dados de teste são valores confiáveis e representam bem o modelo independente do conjunto de teste usado. Para isso, usamos a validação cruzada que nos fornece os seguintes valores

RMSE	R^2
0.7194	0.8675

Como os valores não são muito diferentes daqueles obtidos anteriormente, podemos afirmar que os valores encontrados com o conjunto de teste estão representando bem o comportamento do modelo.

De forma geral o modelo está cumprindo bem sua função de predição, como vemos na Fig. 4 esquerda, onde a nuvem de pontos está relativamente próxima da reta identidade. Na imagem a direita, observa-se que os resíduos parecem estar dispersos aleatoriamente em torno do 0 indicando que o modelo conseguiu identificar os padrões gerais dos dados.

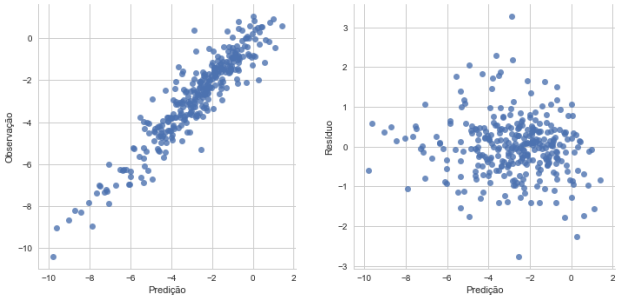

Figura 4: Resultados obtidos com a regressão linear ordinária. Esquerda : Valores observados versus valores previstos para o conjunto de teste de solubilidade. Direita : resíduo versus os valores previstos.

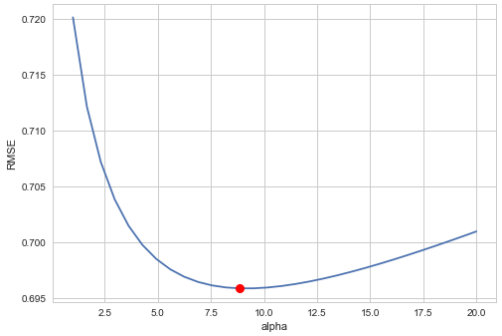

Figura 5: Resultado da métrica RMSE obtida por meio da validação cruzada aplicada a diferentes valores do hiper parâmetro α. O valor mínimo da RMSE foi com α = 9,16.

A fim de analisar melhor a suspeita de *overfitting* e suas possíveis causas, a seguir vamos remover os preditores altamente correlacionados, mais precisamente removeremos um dos preditores que possuir uma correlação com algum outro preditor superior a 0.9. Ao todo, 36 preditores foram removidos. Recalculando os coeficientes da regressão linear ordinária sobre esse novo conjunto de dados, temos

Conjunto treino

RMSE	R^2
0.4813	0.9446

Conjunto teste

RMSE	R^2
0.7456	0.8709

É possível perceber uma pequena redução na diferença entre os resultados das previsões aplicadas ao conjunto de treino e ao conjunto de teste. No entanto, ela ainda continua significativa. Aparentemente o modelo possui uma alta variância e um baixo viés, ou seja um sobre ajustamento ao dados de teste. Uma solução para isso é a utilização de um modelo de regressão penalizado que é analisado na seção III-B.

B. *Modelo de regressão linear penalizado L2 (Ridge)*

Utilizando a regressão penalizada, os coeficientes tendem a diminuir para zero à medida que a penalidade α aumenta, com isso aumentamos o viés e minimizamos a variância. Usando a validação cruzada podemos escolher o melhor valor para o hiper parâmetro α. Na Fig. 5 vemos os valores de RMSE para diferentes valores de α e temos que o melhor α é igual a 9.16.

RMSE	α
0.7025	9.16

A medida que α passa de 12.5, é introduzido um bias muito grande e a variância e o RMSE voltam a aumentar, fazendo com que o modelo entre em estado de *under-fit*.

Como os valores foram encontrados utilizando validação cruzada, também podemos utilizar essa métrica para comparar o modelo utilizando o conjunto de teste, onde encontramos os seguintes valores de RMSE e R^2 dado os sendo os valores de RMSE não muito distantes daqueles encontrados na validação

cruzada, mostrando que o conjunto de testes representa bem o modelo utilizando Ridge, sendo também obtidos valores próximos de 1 para R^2.

Tabela I: Resultados da validação cruzada para diferentes valores do hiper parâmetro α

α	RMSE	R^2
5.5862	0.7100	0.8829
6.2413	0.7096	0.8830
6.8965	0.7093	0.8831
7.5517	0.709112	0.8832
8.2068	0.709067	0.8832
8.8620	0.709118	0.8832
9.5172	0.709245	0.8831
10.1724	0.709433	0.8831
10.8275	0.709673	0.8830
11.4827	0.709953	0.8829
12.1379	0.710268	0.8828
12.7931	0.710610	0.8827

C. Regressão parcial dos mínimos quadrados (PLS)

Ao observar a Fig. 1 percebemos que temos um total de 36 preditores com correlação maior que 0.9 e devido a esta alta correlação as informações gerais estão contidas em um número menor de dimensões.

Durante a fase de treino ao aplicar a Regressão PLS em um intervalo de 1 a 30 variáveis, obteve-se o menor valor para RMSE com total de 19 componentes PLS. Ressalta-se que foi utilizada a validação cruzada com um K = 10. O número de componentes é obtido no valor onde o RMSE é mínimo.

Com 19 componentes PLS é obtido um RMSE = 0.7318 e um R^2 = 0.8756. A Fig. 6 confirma os bons resultados encontrados.

RMSE	R^2
0.7318	0.8756

Embora os resultados em termos de RMSE e R^2 não sejam muito diferentes dos demais métodos, o PLS apresenta uma grande vantagem, pois ele atingiu a mesma precisão dos demais com apenas 19 componentes. O que representa um ganho significativo em tempo computacional.

D. Conclusões

Dados os resultados obtidos concluímos que cada método abordado possui suas vantagens e desvantagens. Como esperado, a regressão linear ordinária por ser a mais simples apresenta resultados bons, mas um pouco piores que os outros dois métodos. No entanto, ela é base para a teoria dos demais e nos permite perceber as vantagens de usar um método penalizado ou então um método focado na redução da dimensão do espaço dos preditores.

Por fim, podemos afirmar que temos três ótimos modelos para a predição da solubilidade de compostos químicos, com dois deles se destacando mais, o Ridge e o PLS. O primeiro por ser um modelo com uma menor variância e maior viés e o outros por proporcionar uma grande redução na dimensão do espaço dos preditores.

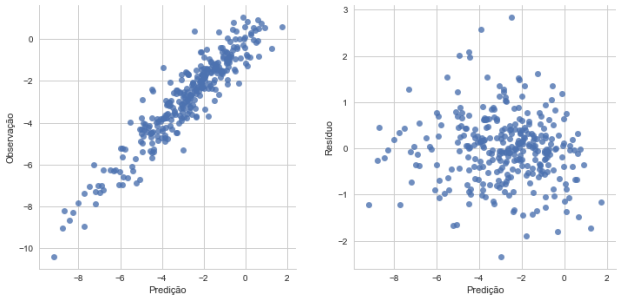

Figura 6: Resultados obtidos com a regressão parcial dos mínimos quadrados. Esquerda : Valores observados versus valores previstos para o conjunto de teste de solubilidade. Direita : resíduo versus os valores previstos.

Referências

[1] Castillo, Margarita Viloria, Amelec Parody Muñoz, Alexander Posso, Heidi. (2017). Application of Multiple Linear Regression Models in the Identification of Factors Affecting the Results of the Chelsea Football Team. International Journal of Control Theory and Applications. 10. 7-13.

[2] C. G. N. de Carvalho, D. G. Gomes, J. N. de Souza and N. Agoulmine, "Multiple linear regression to improve prediction accuracy in WSN data reduction,"2011 7th Latin American Network Operations and Management Symposium, Quito, 2011, pp. 1-8.

[3] S.F. Shih, W.F.P. Shih, Application of ridge regression analysis to water resources studies, Journal of Hydrology, Volume 40, Issues 1–2, 1979, Pages 165-174, ISSN 0022-1694,

[4] JOUR Hadgu, Alula. An application of ridge regression analysis in the study of syphilis data. Statistics in Medicine. doi:10.1002/sim.4780030311.

[5] Bastien, Philippe and Michel Tenenhaus. "PLS generalized linear regression. Application to the analysis of life time data." (2001).

[6] Abudu, Shalamu King, James Pagano, Thomas. (2010). Application of Partial Least-Squares Regression in Seasonal Streamflow Forecasting. Journal of Hydrologic Engineering - J HYDROL ENG. 15. 10.1061/(ASCE)HE.1943-5584.0000216.

[7] Tetko, Igor V. et al. "Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices." Journal of chemical information and computer sciences 41 6 (2001): 1488-93 .

[8] Huuskonen, Jarmo. (2000). Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. Journal of chemical information and computer sciences. 40. 773-7. 10.1021/ci9901338.

[9] Kuhn, Max and Kjell. Johnson. 2013. Applied Predictive Modeling. New York: Springer.