Regression Analysis Applied to Solubility Data: From Ordinary 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*—Chemical and physical properties may be exploited to predict molecular interactions of solubility for drugs investigation. On this work we take advantage on mathematical tools, such as linear regression model analysis to perform prediction for data with information related to the chemical structure of compounds to predict their solubility. Based on data preprocessing techniques a data reduction approach is presented for the model. Linear relationship observed for some predictors is summarized with PCA and correlation matrix computing, to assess predictors relationship. In addition, a brief analysis of linear regression models and comparison to the penalized models, such as Lasso and Ridge. Finally, an analysis of the prediction accuracy of the developed model was performed with cross-validation techniques assessed by error parameters such as root mean squared error (RMSE) and coefficient of determination R^2.**

***Index Terms*—Chemical solubility, Linear regression, Machine learning, Penalized models, Principal component analysis.**

I. Introduction

The solubility is defined by the International Union of Pure and Applied Chemistry (IUPAC) as "the analytical composition of a saturated solution, expressed in terms of the proportion of a designated solute in a designated solvent, is the solubility of that solute." [1], i.e, the physical property of substances to dissolve, or not, in a given liquid. The solubility may be expressed as a concentration, molality, mole fraction, mole ratio, etc.

This characteristic property of a specific solute-solvent combination plays a fundamental role in scientific research and practical applications, observed in medicines and drugs field, as an index to characterize chemical substances and compounds, and to assist in the identification of a compound polarity.

The popular Kosher or kitchen salt (NaCl) is formed by ionic bonds between Sodium (Na^+) and Chlorine (Cl^-) ions, resulting in a polar salt that is soluble in aaqueous solution. This is due to its structure, composition, charge density, intermolecular attraction forces, carbon chain size (in the case of organic compounds).

For an analysis where the exact influence of each chemical and physical characteristic of the experiment compounds on the solubility of the product is not known, it is convenient to study the predictors and its impact to the output at each observation. However, different approaches may impose great computational cost, in such way that a study exploiting all the data description available may keep redundant information, and lead to drawbacks in data representation and visualization. The main goal is build framework capable to provide preprocessed data to use in a linear regression-based prediction model able to predict the data solubility. These steps precede more complex analysis such as feature extraction, dimensionality reduction, clustering and class separation [2]–[4].

The importance of these proposals is related to the applications, which can be in several areas, for example: with the financial market, economy in the large-scale use in the chemical industry, in the production in pharmaceutical laboratories, and even in the study of the influence of pollution on coral destruction.

Multiples approaches for the scenario are exploited in the literature with the same dataset or very similar, which linear regression models and neural networks are applied. Taking advantage on these mathematical tools, some paper explore the relationship of molecular topology predictors to infer the solubility of organic compounds in water [5], or in a similar approach, in addition to molecular topology, electronic interactions are explored in the so-called "E-state" [6]. For another dataset, yet working with the same chemical context and characteristics that propose to relate mineral solubility as a function of ionic strength and temperature [7].

In order to assess this method and overcome these limitations, we propose the application of statistics and linear algebra techniques to observe the relationship between predictors (chemical structure of compounds) and solubility. A model proposition that aims to obtain some pattern of behavior of this relationship, in order to support the prediction for unknown compounds and reduce complexity.

II. Methods

A. Data Overview

For the development of the predictive models we use the Solubility data set [1] provided by the Springer book 'Applied Predictive Modeling' [3] available as a built-in package in the R Project for Statistical Computing [8].

[1] https://cran.r-project.org/web/packages/AppliedPredictiveModeling/

The data consist of 1267 observations (samples) of various chemical compounds, with the solubility of the compound as the output. While the predictors are divided into three groups, the first with 208 binary predictors referring to the presence of a specific chemical substructure, the second with 16 predictors counting chemical bonds and certain types of atoms, and the third with four continuous predictors with characteristics of the compounds, such as molecular mass.

The binary predictors are labeled with *FP001* until the predictor *FP208*. The remaining 16 continuous predictors are identtified as: *MolWeight*, *NumAtoms*, *NumNonHAtoms*, *NumBonds*, *NumNonHBonds*, *NumMultBonds*, *NumRotBonds*, *NumDblBonds*, *NumAromaticBonds*, *NumHydrogen*, *NumCarbon*, *NumNitrogen*, *NumOxygen*, *NumSulfer*, *NumSulfer*, *NumChlorine*, *NumHalogen*, *NumRings*, *HydrophilicFactor*, *SurfaceArea1*, *SurfaceArea2*, *solubility*.

B. 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.

C. Data Preprocessing

Continuos and binary predictors contribute in different ways to the construction of the model, to define the best approach for each one.

For continuous data, we assess data mean, variance and skewness, to verify the existence of a significant presence of a left or right bias. In order to mitigate scale distortion bewteen continuos predictors and others associated issues, we normalize the data before generate and inferences from the data. We choose the z-score normalization, to perform a centering and scaling of the data to have a fair comparaison of the histograms, correlation plots and further operators.

For normal distributions, only the knowledge of two statistics is enough to describe the set: mean (μ) and standard deviation (σ) of the samples, summarized as $\mathcal{N}(\mu, \sigma^2)$. The z-score, or standard score, is a technique that shift and rescale the data, to provide a simplified version of each predictor representation, centered on zero, i.e, zero-mean distribution, and unitary standard deviation. So for a normal population distribution, it is enough to describe as $\mathcal{N}(0, 1)$.

We compute the sample mean ($\bar{x}$) for each predictor:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \,, \tag{1}$$

From the Eq. 1, we may obtain the standard deviation:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})}{N}} \,, \tag{2}$$

Finally, from results of both equations 1 and 2, we obtain the transformed data z_i:

$$z_i = \frac{x_i - \bar{x}}{\sigma} \,. \tag{3}$$

A model constructed from a non-normalized data set may present a biased outcomes, since the model is sensitive to variance. Moreover, in a non-normalized data, for large variance difference between the predictors, those with the greatest values misrepresent its relevance and distort the model, since it can be interpreted as delivering more relevant information, e.g, signal power/energy in digital signal processing field.

From the normalization we may use the covariance matrix and principal component analysis (PCA) to identify the predictors and its linear combinations that actually contribute to the model, producing a correlation analysis. It aims to identify those that carry redundant information, and in some cases eliminate it from the assessment. It also support the identification of the best set of regression tools applied, because in case that too many predictors have non-linear or more complex relationship, more powerful tools may be applied due its capacity of dealing with such complicated problems.

It was checked if there were any missing values, but it was found that all data is fully completed, which eliminates the need of any data compensation technique.

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 [4].

$$\mathbf{y} \approx \beta_0 + \beta_1 \mathbf{x} \,. \tag{4}$$

The model in Eq. 4 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 [4].

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x} + \epsilon \,, \tag{5}$$

Taking advantage on the estimated parameters, we compute Eq. 5 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.

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2 \,, \tag{6}$$

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.

$$\text{MSE} = \frac{1}{n}\text{RSS} \tag{7}$$

The error variance usually is unknown, so we may estimate it from the data taking advantage on the model in Eq. 7. 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].

$$\text{RSS}_{L_2} = \text{RSS} + \lambda \sum_{j=1}^{P} \beta_j^2 \tag{8}$$

The goal with the model presented in Eq. 8 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 [3], [4].

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

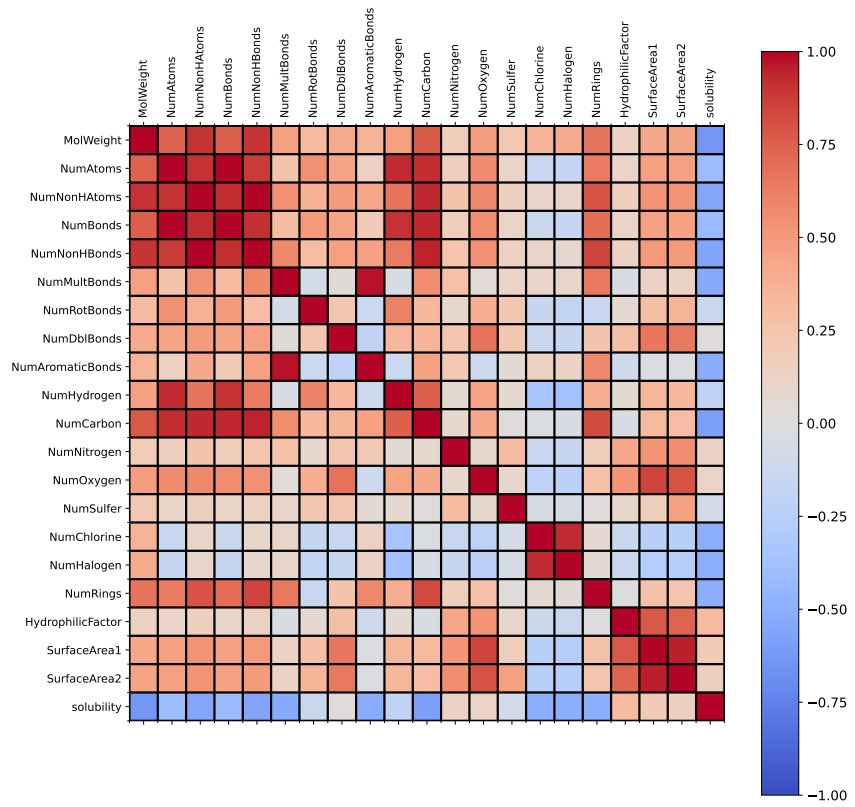

Fig. 1: Correlation Matrix presented as a Heatmap.

IV. Discussion

V. Conclusion

Dentre os resultados obtidos o não foram constatadas grandiosas diferenças dentre os resultados dos modelos, isso favorece a utilização de modelos mais simples e de menor custo computacional, em detrimento da pouca melhora da predição com o aumento do custo. Portanto ao menos que de fato haja uma necessidade em obter o melhor resultado possível, a solução com melhor custo-benefício foi a regressão linear simples, visto que é pouco custosa e obteve resultados não muito inferiores aos seus correntes, mas caso o custo computacional não precise ser considerado, o melhor modelo para o conjunto de dados fornecido é o de regressão "ridge", pois esse foi o que obteve os melhores resultados tanto na validação quanto no conjunto de teste.

VI. Further Work

Appendix

References

[1] A. D. McNaught and A. Wilkinson, *Compendium of Chemical Terminology: "The Gold Book"*. Oxford, 1997.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, 2nd ed. Springer, 2009.

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

[4] 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/

[5] I. V. Tetko, V. Y. Tanchuk, T. N. Kasheva, and A. E. P. Villa, "Estimation of aqueous solubility of chemical compounds using e-state indices," *Journal of Chemical Information and Computer Sciences*, vol. 41, pp. 1488–1493, 11 2001.

[6] J. Huuskonen, "Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology," *Journal of Chemical Information and Computer Sciences*, vol. 40, pp. 773–777, 05 2000.

[7] S. M. Javed, "A regression model for mineral solubility as a function of ionic strength and temperature," Master's thesis, The University of Arizona, United States, 1991.

[8] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/