Solubilidade de compostos Químicos

1st Rodrigo Hiury Silva Araujo
Dept. de Eng. de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
rodrigohiury@alu.ufc.br

2nd Ian Rabelo Lopes
Dept. de Eng. de Teleinformática
Universidade Federal do Ceará
Fortaleza, Brazil
ianrlopes@alu.ufc.br

3rd Cauê Vinícius Carvalho Melo
Dept. de Eng. Elétrica
Universidade Federal do Ceará
Fortaleza, Brazil
cauevinicius@alu.ufc.br

***Resumo*—Este trabalho explorou diversos modelos de regressão, incluindo Regressão Linear Ordinária (OLS), Regressão Ridge, Regressão Lasso, Mínimos Quadrados Parciais (PLS), Regressão de Componentes Principais (PCR) e Redes Neurais, com o objetivo de prever propriedades moleculares a partir de variáveis descritivas. O OLS foi eficiente em problemas lineares simples, enquanto a Regressão Ridge e a Lasso mostraram-se superiores em cenários com multicolinearidade, com a Lasso realizando seleção de variáveis. Técnicas como PLS e PCR foram aplicadas para reduzir a dimensionalidade dos dados e as Redes Neurais demonstraram capacidade de modelar relações complexas e não lineares.**

***Index Terms*—Solubilidade, Regressão linear, Redes neurais**

I. Introdução

A aplicação de técnicas de inteligência computacional tem se mostrado fundamental na modelagem de propriedades químicas e físicas de compostos, contribuindo significativamente para a otimização e o desenvolvimento de novas substâncias em diversas áreas, como a indústria farmacêutica e de materiais. Dentre as propriedades de interesse, a solubilidade é um parâmetro crucial, influenciando a biodisponibilidade e a eficácia de diversos compostos. Para abordar esse desafio, métodos de regressão oferecem ferramentas estatísticas poderosas para a previsão e análise das relações entre variáveis preditoras e a variável resposta.

A regressão, de forma geral, é uma técnica que busca modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. O método de Mínimos Quadrados Ordinários (OLS) é a abordagem clássica, onde se busca minimizar a soma dos quadrados dos resíduos entre os valores observados e os valores previstos pelo modelo. Entretanto, quando lidamos com conjuntos de dados que apresentam alta dimensionalidade ou multicolinearidade, técnicas de regressão penalizada, como Ridge e Lasso, se mostram vantajosas ao introduzirem um termo de penalização que controla a magnitude dos coeficientes.

Além disso, métodos como Regressão por Componentes Principais (PCR) e Partial Least Squares (PLS) são frequentemente empregados para reduzir a dimensionalidade do conjunto de variáveis, extraindo fatores latentes que capturam a maior parte da variabilidade dos dados. Esses métodos são especialmente úteis em problemas onde as variáveis preditoras possuem altos níveis de correlação, proporcionando modelos mais robustos e interpretáveis.

Paralelamente, o avanço dos métodos de redes neurais artificiais (RNA) tem permitido modelar relações não lineares complexas entre as variáveis, o que é particularmente relevante em aplicações de química computacional, onde interações e efeitos sinérgicos podem influenciar fortemente a solubilidade dos compostos. Estudos recentes demonstram que abordagens híbridas, combinando técnicas de regressão clássica e redes neurais, podem aprimorar ainda mais a capacidade preditiva dos modelos para dados químicos.

Neste trabalho, realizamos uma análise exploratória detalhada de um dataset contendo informações químicas e físicas de diversos compostos e suas respectivas solubilidades. Com base nessa análise, serão construídos modelos preditivos utilizando OLS, regressão linear penalizada, PLS/PCR e redes neurais, com o objetivo de desenvolver um modelo matemático robusto para a previsão da solubilidade. Essa abordagem integrativa não apenas amplia o entendimento sobre as variáveis que influenciam a solubilidade, mas também exemplifica a aplicabilidade dos métodos de inteligência computacional em problemas reais da área química.

II. Métodos

Neste trabalho, estamos analisando dados relacionados a propriedades moleculares, com foco na previsão de características específicas, como o peso molecular, a partir de um conjunto de variáveis descritivas. Os datasets utilizados, como ‘solTestX‘, ‘solTrainX‘, e suas versões transformadas, contêm informações detalhadas sobre as moléculas, incluindo descritores químicos e físicos que podem influenciar diretamente as propriedades de interesse. A análise exploratória inicial envolve a visualização de distribuições, correlações e relações bivariadas entre as variáveis, permitindo identificar padrões, tendências e possíveis outliers nos dados. Essa etapa é crucial para compreender a estrutura do dataset e orientar a escolha das técnicas de modelagem mais adequadas.

O objetivo principal do trabalho é desenvolver modelos de regressão capazes de prever com precisão a solubilidade de um dado composto químico. Para entendermos os métodos utilizados na resolução do problema, primeiro vamos definir, de forma genérica, o que é regressão. A regressão é uma técnica estatística que busca modelar a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (ou preditoras). O objetivo principal é identificar como mudanças nas variáveis independentes influenciam

a variável dependente, permitindo tanto a explicação quanto a previsão de comportamentos observados nos dados. Em suma, a regressão fornece um modelo matemático que quantifica essa relação, podendo ser aplicado a problemas simples com uma única variável preditora ou a cenários mais complexos envolvendo múltiplos fatores. Existem múltiplos métodos de regressão que podem ser usados na predição de dados. Nesse trabalho utilizamos métodos como regressão linear ordinária (OLS), regressão linear penalizada (Ridge e Lasso), e técnicas mais avançadas, como redes neurais (MLPRegressor). Além disso, foram aplicadas técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), para simplificar a complexidade dos dados e melhorar a eficiência dos modelos. Cada um dos modelos foi construido usando validação cruzada, e avaliado por meio de métricas de erro, como o método da raíz quadrática média do erro (RMSE) e o coeficiente de determinação (R^2). Essa análise não apenas busca prever as propriedades moleculares com precisão, mas também entender as relações subjacentes entre as variáveis, contribuindo para avanços em áreas como a química computacional e a área de desenvolvimento de fármacos. Vamos definir cada um dos métodos aplicados.

1) **Validação Cruzada k-fold**: A validação cruzada k-fold é uma técnica de avaliação de modelos que divide o conjunto de dados em k subconjuntos de tamanho igual. O modelo é treinado k vezes, utilizando $k-1$ subconjuntos para treinamento e o subconjunto restante para teste. Esse processo é repetido até que cada subconjunto tenha sido usado como subconjunto de teste. A validação cruzada k-fold é amplamente utilizada para garantir que o modelo generalize bem para dados não vistos, evitando problemas como *overfitting* (quando o modelo se ajusta demais aos dados de treinamento).
   - **Prós**: Reduz o risco de *overfitting* e fornece uma avaliação mais robusta do desempenho do modelo. - **Contras**: Pode ser computacionalmente caro, especialmente para grandes conjuntos de dados ou modelos mais complexos.
2) **Raiz Quadrática Média (RMSE)**: O RMSE (*Root Mean Squared Error*) é uma métrica utilizada para medir a precisão de um modelo de regressão. Ele calcula a raiz quadrada da média das diferenças ao quadrado entre os valores previstos pelo modelo e os valores reais. Quanto menor o valor do RMSE, mais próximo o modelo está de prever os valores reais. Essa métrica é especialmente útil para comparar o desempenho de diferentes modelos, pois penaliza erros maiores de forma mais significativa.
   - **Prós**: Fácil de interpretar e sensível a grandes erros, o que ajuda a identificar modelos com previsões ruins.
   - **Contras**: Pode ser muito sensível a *outliers*, distorcendo a interpretação do desempenho do modelo.
3) **Coeficiente de Determinação (R^2)**: O coeficiente de determinação, representado por R^2, é uma medida que indica a proporção da variância da variável dependente que é explicada pelo modelo de regressão. Ele varia de 0 a 1, onde 1 significa que o modelo explica toda a variabilidade dos dados, e 0 indica que o modelo não explica nada. Um valor de R^2 próximo de 1 sugere que o modelo tem um bom ajuste aos dados, enquanto valores mais baixos indicam que o modelo pode não estar capturando adequadamente as relações entre as variáveis.
   - **Prós**: Facilita a comparação entre modelos e fornece uma medida intuitiva de qualidade do ajuste.
   - **Contras**: Não indica se o modelo é adequado para previsões futuras e pode ser enganoso em modelos com muitas variáveis.
4) **Mínimos Quadrados Ordinários (OLS)**: O método dos Mínimos Quadrados Ordinários (OLS) é uma técnica clássica de regressão linear que busca encontrar os coeficientes do modelo que minimizam a soma dos quadrados dos resíduos (as diferenças entre os valores observados e os valores previstos). O OLS é amplamente utilizado devido à sua simplicidade e eficiência, sendo uma das abordagens mais comuns para ajustar modelos de regressão linear. Ele assume que a relação entre as variáveis é linear e que os erros são normalmente distribuídos e independentes.
   - **Prós**: Simples de implementar e interpretar, além de ser eficiente para problemas lineares.
   - **Contras**: Sensível a *outliers* e multicolinearidade, além de não funcionar bem para relações não lineares.
5) **Regressão Linear Ridge**: A Regressão Linear Ridge é uma variação da regressão linear ordinária (OLS) que adiciona um termo de regularização para lidar com problemas de multicolinearidade (quando as variáveis independentes estão altamente correlacionadas). Esse termo de regularização, controlado por um parâmetro λ, penaliza os coeficientes do modelo, reduzindo sua magnitude e evitando que eles assumam valores extremos. Isso ajuda a melhorar a generalização do modelo, especialmente em conjuntos de dados onde o número de variáveis é grande em relação ao número de observações. A Regressão Ridge é particularmente útil quando se deseja evitar *overfitting*, mantendo a capacidade preditiva do modelo.
   - **Prós**: Reduz o impacto da multicolinearidade e melhora a generalização do modelo.
   - **Contras**: Não realiza seleção de variáveis (todos os coeficientes permanecem no modelo, embora reduzidos).
6) **Regressão Linear Lasso**: A Regressão Linear Lasso (*Least Absolute Shrinkage and Selection Operator*) é uma técnica de regressão que, assim como a Regressão Ridge, adiciona um termo de regularização para evitar *overfitting*. No entanto, o Lasso utiliza uma penalidade $L1$, que não apenas reduz a magnitude dos coeficientes, mas também pode zerar alguns deles, efetivamente realizando uma seleção de variáveis. Isso torna o Lasso particularmente útil quando se deseja identificar as variáveis mais relevantes para o modelo, simplificando a interpretação e reduzindo a complexidade.

- **Prós**: Realiza seleção de variáveis, simplificando o modelo e melhorando a interpretação.
- **Contras**: Pode ser instável quando há alta correlação entre variáveis, e a seleção de variáveis pode não ser única.

7) **Mínimos Quadrados Parciais (PLS)**: O método de Mínimos Quadrados Parciais (*Partial Least Squares*, PLS) é uma técnica de regressão que combina aspectos da regressão linear e da análise de componentes principais. Ele é especialmente útil quando há multicolinearidade entre as variáveis independentes ou quando o número de variáveis é maior que o número de observações. O PLS cria componentes latentes (não observáveis) que maximizam a covariância entre as variáveis independentes e dependentes, permitindo uma previsão mais robusta e interpretável.
   - **Prós**: Lida bem com multicolinearidade e conjuntos de dados com muitas variáveis.
   - **Contras**: Menos interpretável do que a regressão linear tradicional, e os componentes latentes podem ser difíceis de explicar.
8) **Regressão de Componentes Principais (PCR)**: A Regressão de Componentes Principais (*Principal Component Regression*, PCR) é uma técnica que utiliza a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos dados antes de aplicar a regressão linear. A PCA por sua vez é uma técnica que transforma o conjunto de dados original em um conjunto de dados com um numero menor de componentes, que são combinações lineares das variáveis originais e capturam a maior parte da variância dos dados. A PCR é útil para lidar com multicolinearidade e reduzir o número de variáveis, melhorando a eficiência do modelo.
   - **Prós**: Reduz a dimensionalidade e lida bem com multicolinearidade.
   - **Contras**: Os componentes principais podem não ter significado físico, dificultando a interpretação do modelo.
9) **Modelo de Redes Neurais**: As Redes Neurais são modelos computacionais inspirados no funcionamento do cérebro humano, capazes de aprender padrões complexos e não lineares nos dados. Um modelo de rede neural consiste em camadas de neurônios interconectados, onde cada neurônio aplica uma transformação matemática aos dados de entrada. Esses modelos são altamente flexíveis e podem ser usados para tarefas de regressão, classificação e muito mais. No contexto de regressão, as redes neurais são particularmente úteis quando a relação entre as variáveis é complexa e não pode ser capturada por modelos lineares tradicionais.
   - **Prós**: Altamente flexíveis e capazes de modelar relações não lineares complexas.
   - **Contras**: Gera modelos difíceis de interpretar, que exigem grandes quantidades de dados e podem ser computacionalmente caros.

RESULTADOS

Os resultados obtidos neste trabalho refletem a aplicação de diversas técnicas de análise de dados e modelagem estatística para prever propriedades moleculares, como o peso molecular (*MolWeight*), a partir de um conjunto de variáveis descritivas. A análise exploratória inicial, realizada por meio de histogramas, box-plots e mapas de calor, revelou a distribuição das variáveis e suas correlações, permitindo identificar padrões e possíveis *outliers*. Por exemplo, a função `plotar_histograma_box` mostrou que algumas variáveis apresentam distribuições assimétricas, o que pode indicar a necessidade de transformações ou normalizações para melhorar o desempenho dos modelos.

Na etapa de modelagem, foram implementados e avaliados diversos algoritmos de regressão. A regressão linear ordinária (OLS), tanto implementada manualmente quanto utilizando a biblioteca `scikit-learn`, apresentou resultados promissores, com valores de RMSE (*Root Mean Squared Error*) e R^2 (coeficiente de determinação) que indicam um bom ajuste aos dados. A validação cruzada *k-fold*, realizada com a função `k_fold_validation`, confirmou a robustez do modelo, com médias consistentes de RMSE e R^2 ao longo das diferentes partições dos dados. Além disso, a regressão penalizada (Ridge e Lasso) foi aplicada para lidar com possíveis problemas de multicolinearidade, e o melhor valor de λ (parâmetro de regularização) foi identificado com base no menor RMSE.

A aplicação de técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), permitiu simplificar a estrutura dos dados, reduzindo o número de variáveis enquanto preservava a maior parte da variância. Isso foi especialmente útil para melhorar a eficiência dos modelos e reduzir o risco de *overfitting*. Por fim, a utilização de redes neurais (*MLPRegressor*) demonstrou ser uma abordagem viável para problemas complexos, com resultados competitivos em termos de R^2 e RMSE, tanto na validação cruzada quanto no conjunto de teste.

Em resumo, os resultados mostram que as técnicas aplicadas foram eficazes para prever as propriedades moleculares, com modelos que apresentaram bom desempenho e generalização. A análise exploratória e o pré-processamento dos dados foram fundamentais para garantir a qualidade das previsões, enquanto a comparação entre diferentes métodos de regressão permitiu identificar as abordagens mais adequadas para o problema em questão. Esses resultados contribuem para o avanço de pesquisas em áreas como química computacional e desenvolvimento de fármacos, onde a previsão de propriedades moleculares é essencial.

A. Mínimos Quadrados Ordinários (OLS)

O **modelo de Regressão Linear Ordinária (OLS)** foi implementado tanto manualmente quanto utilizando a biblioteca `scikit-learn`. Tanto o código da OLS criado pela equipe, quanto o código implementado utilizando bibliotecas prontas do `scikit-learn` apresentaram um **RMSE (Raiz Quadrática Média)** de aproximadamente **0.7456** e um **Coeficiente de Determinação (R^2)** de **0.8709**. Tal ocorrido

implica que o modelo implementado e criado pela equipe performa bem comparado a um modelo já pronto da biblioteca `scikit-learn` e, devido aos valores encontrados de RMSE e Coeficiente de Determinação (R^2), temos que o modelo faz previsões precisas e explica bem a variabilidade dos dados. Ademais, utilizando a técnica de **k-fold** obtemos que a **RMSE (Raiz Quadrática Média)** encontrada foi de aproximadamente **0.7073** e o **Coeficiente de Determinação** (R^2) foi de **0.8726**, mostrando um incremento na acuracia devido a utilização da técnica de k-fold. No entanto, o OLS é sensível a multicolinearidade, o que pode prejudicar sua generalização em conjuntos de dados com variáveis altamente correlacionadas. Além disso, o modelo não inclui regularização, o que pode levar a coeficientes com magnitudes excessivas em cenários complexos.

B. Regressão Linear Ridge

O **modelo de Regressão Linear Ridge** foi aplicado para lidar com problemas de multicolinearidade, com o parâmetro de regularização λ sendo ajustado para minimizar o RMSE. Foi encontrado um valor de **lambda** igual a **10** e um **RMSE** de aproximadamente **0.7094**, resultado de RMSE esse semelhante ao encontrado na regressão por OLS. No entanto, a grande vantagem do Ridge é sua capacidade de reduzir a magnitude dos coeficientes, evitando valores extremos e melhorando a generalização do modelo. A principal limitação do Ridge é que ele não realiza seleção de variáveis, mantendo todos os coeficientes no modelo, o que pode dificultar a interpretação em casos com muitas variáveis.

C. Regressão de componentes principais (PCR)

Para realizar a regressão de componentes principais(PCR), antes foi feita a análise de componentes principais (PCA). A análise de componentes principais (PCA) mostrou que o número ótimo de componentes para realização da PCR é de 24 componentes, conforme mostra o gráfico na figura **1**. Após isso, foi feita a regressão de componentes principais (PCR) e foi obtido um **RMSE** de **0.9456** e um **Coeficiente de Determinação** (R^2) de **0.7923**.

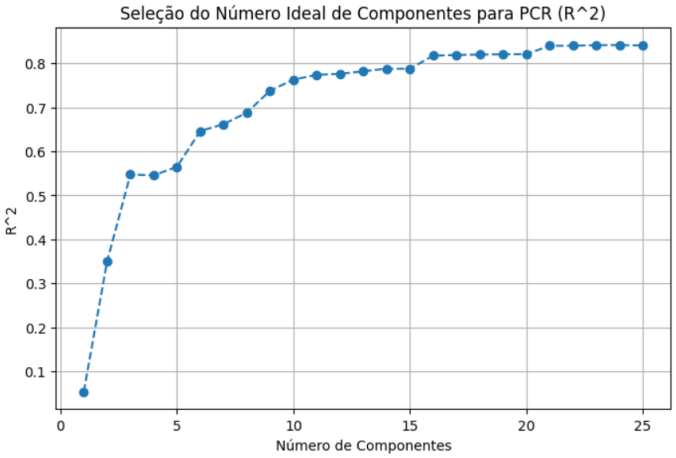

Figura 1: Número de componentes pelo erro R^2

D. Redes Neurais

Por fim, implementamos um modelo em redes neurais para a resolução do problema. Nossa rede possui uma camada de entrada contendo 228 neurônios, uma camada intermediária contendo 12 neurônios e uma camada de saída contendo 1 neurônio. Para as camadas intermediárias, foi escolhida somente uma de 12 neurônios pesando os fatores de custo computacional e performance do código. Treinando nossa rede, obtemos uma das melhores taxas de erro de todas as técnicas usadas, com **RMSE** de **0.6837** e R^2 de **0.8898**. Tal performance reflete uma possível não-linearidade nos dados, o que pode favorecer o uso das redes neurais.

III. Conclusão

Em conclusão, a escolha do modelo ideal depende da estrutura dos dados e do problema em questão. A combinação de técnicas de pré-processamento, modelagem e validação cruzada foi essencial para garantir a robustez e a generalização dos modelos. A aplicação desses métodos, aliada à análise de métricas como **RMSE** e R^2, permitiu obter previsões confiáveis e *insights* valiosos, contribuindo para a tomada de decisões informadas em problemas de regressão. Além disso, cabe salientar que nem sempre a escolha de modelos mais complexos em prol de modelos mais simples vai trazer resultados significativamente melhores, podendo inclusive ser algo negativo e que diminui a acurácia da predição. Tal comportamento pode ser observado nesse trabalho ao ser feita a comparação dos resultados (RMSE e R^2) obtidos na regressão pelo método dos mínimos quadrados ordinários (OLS) e pela regressão de componentes principais (PCR).

Referências

[1] UFC. ”Inteligência computacional aplicada- ”Data pre-processing”, por Michela Mulas.

[2] Max Kuhn and Kjell Johnson, ”Applied predictive modeling”, Springer, 2013.