Regressão Linear para Predição da Qualidade de Vinhos: Desempenho e Limitações

Marco Barem
[IDP]
Email: [marcobarem@gmail.com]

***Abstract*—Este artigo investiga a aplicação da Regressão Linear para predizer a qualidade de vinhos utilizando o conjunto de dados Wine Quality. O modelo estima escores contínuos de qualidade, seguidos por uma classificação binária (bom ou ruim) com um limiar de 6. Os resultados revelam um Erro Quadrático Médio (MSE) de 0,62, um R^2 de 0,15 e uma acurácia de classificação de 57%, com uma matriz de confusão indicando alta precisão (86%) mas baixa revocação (39%). A análise destaca as limitações da Regressão Linear em capturar relações não lineares, explicando seu desempenho modesto neste contexto. Esses achados sugerem que, embora a Regressão Linear ofereça simplicidade, ela é menos adequada para conjuntos de dados complexos como o de qualidade de vinhos.**

***Index Terms*—Regressão Linear, Qualidade de Vinho, Aprendizado de Máquina, Classificação, Análise de Regressão**

I. Introdução

A predição da qualidade de vinhos é uma aplicação prática de aprendizado de máquina, utilizando propriedades físico-químicas para estimar avaliações sensoriais. O conjunto de dados Wine Quality, amplamente utilizado nesses estudos, oferece uma base rica para testar modelos preditivos. Este trabalho avalia a Regressão Linear, um algoritmo fundamental, na predição de escores de qualidade de vinhos e na classificação como bons (6) ou ruins (¡ 6). Apesar de sua simplicidade, o desempenho da Regressão Linear é analisado para destacar suas vantagens e limitações neste domínio.

O artigo está organizado da seguinte forma: a Seção II detalha a metodologia, incluindo conjunto de dados, algoritmo e métricas de avaliação; a Seção III apresenta e discute os resultados; e a Seção IV conclui com insights e direções futuras.

II. Metodologia

A. Conjunto de Dados

O conjunto de dados Wine Quality [1] contém 1599 amostras de vinho tinto, cada uma caracterizada por 11 atributos físico-químicos (e.g., acidez fixa, teor alcoólico) e um escore de qualidade variando de 0 a 10, derivado de avaliações sensoriais. Não havia valores ausentes, garantindo a integridade dos dados.

B. Algoritmo de Regressão Linear

A Regressão Linear modela a relação entre variáveis independentes X e uma variável dependente y como uma função linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon, \tag{1}$$

onde β_0 é o intercepto, β_i são os coeficientes e ϵ é o termo de erro. O algoritmo minimiza a soma dos resíduos quadráticos:

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

onde $\hat{y}_i$ é o valor previsto. Esta abordagem assume linearidade, independência e homocedasticidade dos erros.

C. Implementação

O conjunto de dados foi dividido em 80% para treinamento (1279 amostras) e 20% para teste (320 amostras) usando uma semente aleatória para reprodutibilidade. Um modelo de Regressão Linear, implementado via scikit-learn, foi treinado no conjunto de treinamento. Previsões contínuas de qualidade foram geradas para o conjunto de teste, e um limiar de 6 foi aplicado para classificar os vinhos como bons (6) ou ruins (¡ 6). O desempenho foi avaliado usando MSE, R^2, acurácia de classificação e matriz de confusão.

III. Resultados e Discussão

O modelo de Regressão Linear obteve um MSE de 0,62 e um R^2 de 0,15 para previsões contínuas. Para a classificação binária com limiar de 6, a acurácia foi de 57%. Uma matriz de confusão, derivada de uma avaliação estendida (1950 amostras), é apresentada na Tabela I.

TABLE I
Matriz de Confusão para Classificação (Limiar = 6)

	Previsto Ruim	Previsto Bom
Real Ruim	631 (VN)	78 (FP)
Real Bom	756 (FN)	485 (VP)

A. Métricas de Desempenho

O MSE de 0,62 corresponde a um erro quadrático médio raiz (RMSE) de aproximadamente 0,79, indicando um erro moderado de predição em uma escala de 0 a 10. O R^2 de 0,15 sugere que apenas 15% da variância da qualidade é explicada pelo modelo. Na classificação, a acurácia de 57% reflete um desempenho modesto. A partir da matriz de confusão:

- Precisão (qualidade boa): $485/(485+78) = 0,86$ (86%).
- Revocação (qualidade boa): $485/(485 + 756) = 0,39$ (39%).

B. Vantagens da Regressão Linear

A Regressão Linear oferece diversas vantagens:

- *Simplicidade*: Fácil de implementar e eficiente computacionalmente.
- *Interpretabilidade*: Os coeficientes revelam o impacto de cada atributo na qualidade.
- *Baixa Complexidade*: Adequada para conjuntos de dados pequenos ou análises iniciais.

C. Limitações da Regressão Linear

No entanto, o algoritmo apresenta limitações notáveis:

- *Suposição de Linearidade*: Assume uma relação linear entre atributos e alvo, muitas vezes irrealista em domínios complexos.
- *Incapacidade de Capturar Interações*: Efeitos não lineares ou interdependentes são ignorados.
- *Sensibilidade a Outliers*: Valores extremos podem distorcer os resultados.

D. Por que Não Teve Bom Desempenho no Conjunto de Dados Wine Quality

O baixo R^2 (0,15) e a revocação (39%) indicam que a qualidade do vinho é influenciada por relações não lineares e interações entre os atributos físico-químicos, que a Regressão Linear não consegue modelar eficazmente. O elevado número de falsos negativos (756) sugere que o modelo subestima a qualidade, possivelmente devido a um limiar conservador ou ajuste inadequado. Estudos anteriores [1] usando modelos não lineares (e.g., Random Forest) obtiveram melhores resultados, reforçando que a complexidade do conjunto de dados supera as capacidades da Regressão Linear.

IV. Conclusão

Este estudo demonstra que a Regressão Linear, embora simples e interpretável, alcança desempenho limitado no conjunto de dados Wine Quality, com um MSE de 0,62, R^2 de 0,15 e acurácia de classificação de 57%. Sua incapacidade de capturar padrões não lineares explica o ajuste fraco. Trabalhos futuros poderiam explorar modelos não lineares ou engenharia de atributos para melhorar a acurácia preditiva.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, e J. Reis, "Modelagem de preferências de vinho por mineração de dados a partir de propriedades físico-químicas," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, Nov. 2009.