



Trường Đại học Khoa học tự nhiên, VNU-HCM

LINEAR REGRESSION

ĐỒ ÁN 3 - TOÁN ỨNG DỤNG & THỐNG KÊ

Thực hiện

19127216 - Đặng Hoàn Mỹ

August 2021

MÔ HÌNH 11 ĐẶC TRƯNG

Sử dụng các hàm thông dụng và xây dựng nên mô hình thông thường từ 11 đặc trưng cho sẵn để xác định chất lượng của rượu vang.

Những bước cơ bản của việc xây dựng mô hình:

- Nạp thư viện và tập dữ liệu
- Tách tập dữ liệu ra hai phần huấn luyện (training set) và kiểm thử (test set).
- Huấn luyện mô hình trên tập huấn luyện.
- Dự đoán kết quả tập kiểm thử.
- Lấy các hệ số cho phương trình hồi quy tuyến tính.

Đánh giá mô hình:

CVScore: 25.023743283113376

Mean Absolute Error: 0.5005810380439448

Root Mean Squared Error: 0.6206780310653043

RSquared: 0.3783638631118801

Adj RSquared: 0.3726079729555086

RSquared - Adj RSquared: 0.005755890156371457

The linear regression equation: $y =$

$32.184 + 0.0638 * \text{fixed acidity} + -1.1851 * \text{volatile acidity} + -0.4266 * \text{citric acid} + 0.0326 * \text{residual sugar} + -1.4253 * \text{chlorides} + 0.0025 * \text{free sulfur dioxide} + -0.0032 * \text{total sulfur dioxide} + -29.523 * \text{density} + -0.0397 * \text{pH} + 0.7318 * \text{sulphates} + 0.2799 * \text{alcohol}$

MÔ HÌNH 1 ĐẶC TRƯNG

Sử dụng phương pháp Cross Validation - kiểm chứng chéo để kiểm tra lần lượt các folds trong tập kiểm thử. Các folds này được chia một tập dữ liệu ngẫu nhiên nhằm mục đích đưa ra những thành phần huấn luyện đa dạng hơn.

Phương pháp sử dụng để tìm ra cột tốt nhất là xây dựng từng model trên từng cột để kiểm tra các giá trị như R-Squared cao nhất, RMSE thấp nhất hoặc độ chính xác - cross_val_value() cao nhất.

$$1.8412 + 0.3689 * alcohol$$

Những bước cơ bản khi xây dựng mô hình:

- Tách dữ liệu theo từng cột
- Xây dựng mô hình và tiến hành chia tập dữ liệu thành $k = 10$ phần.
Với Cross Validation thì các tập được chia ngẫu nhiên và $k - 1$ tập sẽ thành tập huấn luyện và tập còn lại để kiểm tra.
- Sau khi xây dựng mô hình, chạy bộ test của mô hình đó.
- Lấy các giá trị đánh giá mô hình như R-Squared, RMSE và CVScore cho từng cột và lấy cột có giá trị phù hợp.
 - VỚI R-Squared, đây là giá trị bình phương hiệu chỉnh, giá trị càng lớn càng cho thấy mức độ phụ thuộc của chất lượng rượu (quality) và đặt tính đó như thế nào (ví dụ: alcohol...)
 - VỚI CVScore (cross_val_score), đây là giá trị thể hiện độ chính xác của tổng thể mô hình mà chúng ta xây dựng. Có một số giá trị càng âm, điều này chứng tỏ là mô hình của chúng ta đưa ra thực tế sẽ càng sai.
 - VỚI MSE và RMSE, cả hai đều thể hiện mức độ gần của nhau
- VỚI những giá trị nêu trên, ta có thể kết luận cột alcohol đáp ứng được các yêu cầu đó, sau khi xây dựng mô hình qua từng cột.

Col: alcohol
 CVScore: 12.670040464670224
 Mean Absolute Error: 0.5031808152120568
 Root Mean Squared Error: 0.617049708770383
 RSquared: 0.2552381203994226
 Adj RSquared: 0.25461644938139205

The column we need for this question:

alcohol

The linear regression equation:

$$1.8412 + 0.3689 * alcohol$$

MÔ HÌNH TỰ XÂY DỰNG

Tiếp tục sử dụng phương pháp Cross Validation để tìm ra mô hình phù hợp. Tiến hành thực hiện xây dựng 2 mô hình để thử nghiệm.

Mô hình đầu tiên xây dựng dựa trên hệ số tương quan của các đặc trưng của rượu. Lấy các đặc trưng có mối tương quan lớn hơn 0.05 làm đầu vào cho x và chất lượng rượu quality làm biến mục tiêu y. Từ đây, chúng ta lấy 5 đặc trưng đầu tiên để tiến hành tạo ra mô hình.

Alcohol - Volatile Acidity - Citric Acid - Total Sulfur Dioxide - Sulphates

Mô hình thứ hai được xây dựng từ câu b, lấy ra 5 đặc trưng có CVScore cao nhất.

Alcohol - Volatile Acidity - Total Sulfur Dioxide - Citric Acid - Fixed Acidity

Cả hai mô hình đều cho ra kết quả khá tốt nhưng mô hình thứ hai cho ra những giá trị đánh giá mô hình thấp hơn. (CVScore thấp hơn, RSquared thấp hơn và RMSE cao hơn)

Vậy nên từ mô hình thứ nhất và sau khi tìm hiểu những thông tin về rượu vang thì nhận thấy giá trị Citric Acid không được nhắc đến nên đã tiến hành tạo ra mô hình không có Citric Acid.

Alcohol - Volatile Acidity - Total Sulfur Dioxide - Sulphates

Ở mô hình mới này, cho ra kết quả CVScore cao hơn nhưng MAE cao hơn 0.0003 và RSquared cho ra thấp hơn. Nhưng giá trị RSquared và Adjusted R Squared gần nhau hơn; CVScore cao hơn và độ lỗi sai là 0.0003, cùng với ít đặc trưng hơn nên quyết định giữ lại mô hình này là kết quả cuối cùng.

Ở những tài liệu khác, có những ý kiến cho rằng pH và tỉ trọng density cũng góp phần ảnh hưởng vào chất lượng rượu nhưng giá trị density cho ra giá trị khá tệ khi đưa vào mô hình chỉ một cột đó (câu b) và giá trị pH đã nằm trong giá trị cần của bất kì loại rượu nào và cũng như sự phụ thuộc của nó không quá đáng kể khi lấy hệ số tương quan 0.05. (pH: 3 ~ 4)

CVScore: 24.387365326606474

Mean Absolute Error: 0.44497654218663407

Root Mean Squared Error: 0.5796540951524313

RSquared: 0.3615747763884387

Adj RSquared: 0.35943778819224936

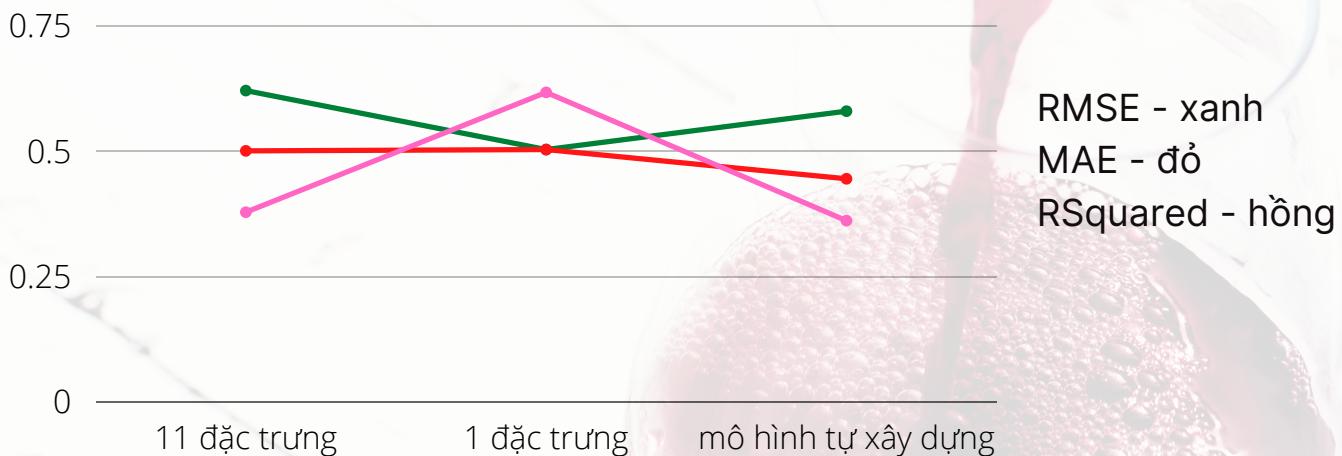
RSquared - Adj RSquared: 0.002136988196189349

The linear regression equation: $y =$

$5.468 + 1.9712 * \text{alcohol} + -1.3871 * \text{volatile acidity} + -0.8947 * \text{total sulfur dioxide} + 0.9393 * \text{sulphates}$

ĐÁNH GIÁ CÁC MÔ HÌNH

Ở bài toán hồi quy tuyến tính, có nhiều phương pháp để tiếp cận. Như những thuật toán Regression Tree hoặc Random Forest cũng có thể mang lại những hiệu quả cao hơn như trên các trang web đề cập.



Mô hình tốt nhất trong ba câu là mô hình 1 đặc trưng alcohol bởi vì có RSquared và RMSE thấp hơn (và RMSE bằng với MAE tương đương độ lỗi của các mẫu kiểm tra đều gần xấp xỉ nhau).

Mô hình 11 đặc trưng và mô hình tự xây dựng có sự tương đương, chỉ có chỉ số MAE thấp hơn nên sẽ phụ thuộc vào giá trị test ở thực tế.

THAM KHẢO

- A. (2021a, March 22). KNN Prediction For Red Wine Quality. Kaggle. <https://www.kaggle.com/aditianiknn-prediction-for-red-wine-quality/notebook>
- Bassey, P. (2019, September 19). Logistic Regression Vs Support Vector Machines (SVM). Medium. <https://medium.com/axum-labs/logistic-regression-vs-support-vector-machines-svm-c335610a3d16>
- Dave, A. (2019, February 23). Regression from scratch — Wine quality prediction - DataDrivenInvestor. Medium. <https://medium.datadriveninvestor.com/regression-from-scratch-wine-quality-prediction-d61195cb91c8>
- Diwan, A. (2020, September 5). The K-Fold Cross Validation in Machine Learning. Knowledge Hut. <https://www.knowledgehut.com/blog/data-science/k-fold-cross-validation-in-ml>
- J. (2019, August 20). Cross-Validation with Linear Regression. Kaggle. <https://www.kaggle.com/jnikhilsai/cross-validation-with-linear-regression>
- K. (2018, December 21). K-Fold Cross Validation - DataDrivenInvestor. Medium. <https://medium.datadriveninvestor.com/k-fold-cross-validation-6b8518070833>
- Li, E. C. (n.d.). Regression Tree & Model Tree for Analyzing Red Wine Quality. R Studio Pubs Static. Retrieved August 15, 2021, from https://rstudio-pubs-static.s3.amazonaws.com/274165_627a87883a534f15b42c4b879d369ac7.html
- Nguyen, D. (2020, November 25). Red Wine Quality Prediction Using Regression Modeling and Machine Learning. Medium. <https://towardsdatascience.com/red-wine-quality-prediction-using-regression-modeling-and-machine-learning-7a3e2c3e1f46>
- RPubs - Red wine quality analysis. (2018, January 26). RPubs. <https://rpubs.com/ashwinkashok/redWineQualityFin>
- S. (2020, October 9). Wine Quality Kfold cross validation and Prediction. Kaggle. <https://www.kaggle.com/suveesh/wine-quality-kfold-cross-validation-and-prediction#Train-and-Predict>
- S. (2021b, January 10). A Quick Overview of Regression Algorithms in Machine Learning. Analytics Vidhya. <https://www.analyticsvidhya.com/blog/2021/01/a-quick-overview-of-regression-algorithms-in-machine-learning/>
- Toàn P. V. (2021, August 15). Một vài hiểu nhầm khi mới học Machine Learning. Viblo. <https://viblo.asia/p/mot-vai-hieu-nham-khi-moi-hoc-machine-learning-4dbZNoDnIYM>
- Wine and its analysis. (2014). <https://eprints.ucm.es/id/eprint/29446/8/PIMCD%20N%C2%BA%20243.%20ANEXO%202.%20E-BOOK-%20WINE%20AND%20ITS%20ANALYSIS.pdf>