Fundamentals of Data Science

Pages 1–3
2021

Wine Quality Project

Matteo Basile, Simone Chieppa, Vito Collica, Giovanni Montobbio, Francesco Pinto

Sapienza University of Rome

26^{th} December

Keywords: Sensory preferences, Regression, Classification, Model Selection, Softmax Regression, Multivariate Linear Regression

ABSTRACT

Wine quality prediction is an interesting machine learning problem that fits both regression and classification, allowing the comparison of the two approaches. We used multivariate linear regression as our regression method and softmax regression as our classification method, on two datasets containing the chemical properties of a number of vinho verde samples, one for red wines and one for white wines, regression yelded better performance in our testing.

1 INTRODUCTION

For the final project we decided to work on wine quality estimation. This kind of work is important in the field of wine quality certification, where the assessment is usually done by human experts, and as such it has a very high cost, also the opinions of different experts may have a high level of variability, making the process more complicated. Being able to infer the quality of the samples based on easily available physiochemical tests, by making use of Machine Learning models such as those described in this paper, helps improving the speed and the accuracy of such estimates, making wine certification and quality assessment more controlled and decreasing costs. We chose this topic as we found it interesting for a machine learning application, and as the dataset works with both classification and regression we wanted run both and compare their performance.

2 RELATED WORK

Research on the topic was done by the a Department of Information Systems of the University of Minho, a research paper was published and made available to everyone [1]. The researchers used different methods such as SVM, multivariate regression and NN, our work was in good part inspired by what they've done

Work on the topic was also made by Dexter Nguyen[2], who worked on it as a Data Science Project during his studies at Duke University, and tackled the same problem with the LASSO method and random forest(RF), using K-Fold Cross Validation and reaching better performance with RF.

Another article was written by Nataliia Rastoropova and published in *Analytics Vidha*[3], a community of analysts and data scientists. Meant as a tutorial on how to use the various ML models, it compares 8 different models: support vector classifier, stochastic gradient descent classifier, random forest classifier, decision tree classifier, Gaussian Naive Bayes, K-Neighbors classifier, Ada Boost classifier and logistic regression, highlighting support vector classifier and random forest classifier as the best algorithms to apply on the model.

3 PROPOSED METHOD

We chose to use two models we learned during the course, namely Softmax Regression and Linear Regression. Linear regression estimates the output function by forming an hypothesis of its behaviour based on the features, while Softmax Regression estimates a decision boundary for each output class, separating it from every other class.

We decided to implement the code used to train the model ourselves, starting from the work done for Assignment 2, and we applied 5-fold cross validation on the training set, which splits it in 5 folds: 4 used for training and 1 used as a dev set, cycling 5 times so that each fold gets used as a dev set once. We made use of the dev set(s) to tune the hyperparameters.

To perform our comparisons between linear regression and softmax regression we computed several metrics, we decided to adopt the concept of tolerance as in the original paper[1], computing the various metrics for tolerances up to 1 (up to 2 for the REC curve), as we felt like even when off by one point the prediction would still be useful to potential users of the system. We decided to compare the models in various ways: First of all we rounded the output of Linear Regression (same as applying a tolerance of 0.5) to make it predict actual classes, we computed its accuracy and precision metrics and the mean square error(MSE) of the predictions, as well as those for softmax classification, we then turned softmax regression into a regression method by computing the weighted average of the probabilities of each class regressed by it to estimate the output class, which allowed us to plot REC curves for both models and compare them.

4 DATASET AND BENCHMARK

Our datasets were taken from the UCI's Machine Learning Repository[4] and are composed by a relatively large number of entries (4898 samples for white wines, 1599 samples for red wines), with 11 features regarding the chemical composure of each wine sample and the respective quality as evaluated by a minimum of three sensory assessors. They have already been pre-processed and are ready for use as-is.

5 EXPERIMENTAL RESULTS

We ran several experiments (more than 200) to find the hyperparameters that would give us the highest values of accuracy and precision in Regression and Classification, though with more time, trying out more hyperparameters, further improvements may be achieved.

At the end of the tests we chose the values for learning rate and number of iterations that yelded the highest values of accuracy and precision for the model, which are shown in the table below.

Results				
Wine Type/Model	Learning Rate	Iterations	Accuracy (T=1)	Precision (T=1)
White/Regression	7.5e-05	2000	0.809	0.551
White/Classification	7.5e-05	2000	0.934	0.613
Red/Regression	0.0002	4000	0.865	0.631
Red/Classification	0.0002	4000	0.944	0.814

The confusion matrices for Linear Regression (first row) and Softmax Regression (second row) follow.

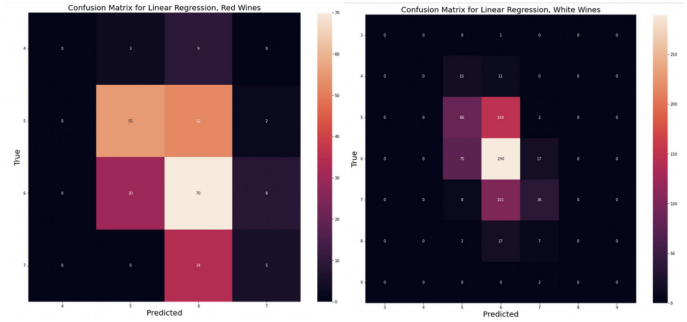

Figure 1. Confusion Matrices for Linear Regression on both datasets

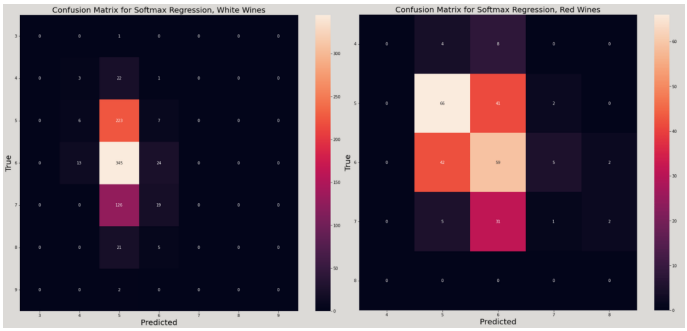

Figure 2. Confusion Matrices for Softmax Regression on both datasets

After interpolating the predictions of Softmax Regression to turn it into a regression model, we computed the mean square error for both models to see which one was more accurate, getting results of around 0.82 (white wines), 0.46 (red wines) for Linear Regression and 1.11 (white wines), 0.54 (red wines) for Softmax Regression (Interpolated), showing better performance for Linear Regression. We also plotted the REC curves for better visualization of the relative performance between the two models (Linear Regression and Interpolated Softmax Regression in particular) for different values of tolerance, which can be seen in figure 3.

On the dev set Softmax Regression outperformed Linear Regression, but when we ran the final tests things turned around, with Linear Regression achieving better results. We initially thought this would be the case, but analyzing the related research it would seem that Classification should perform better here, so it's possible that we ran into some form of overfitting during training.

Perhaps using a different train/test split or maybe trying out 10-fold cross-validation would have improved performance, as well as trying out even more sets of hyperparameters, but due to the limited time we had for the Project we didn't find time to explore further.

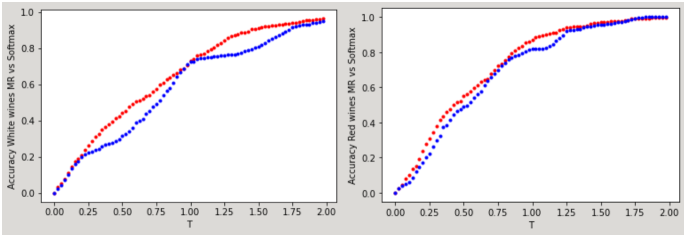

Figure 3. REC Curves for Linear Regression(in red) and Softmax Regression (Interpolated)(in blue), for red and white wines

6 TEAM ASSIGNMENTS

We divided the project into a series of subtasks, which we tackled by dynamically dividing our team in two subteams, with more team members (3 vs 2) being assigned to the harder tasks, this allowed us to achieve a very good efficiency, working on multiple tasks at the same time.

The coding part has been done by everyone at different times, Matteo and Giovanni worked mostly on coding and contributed to the entire code, giving particular attention to the Classification part; Vito, Francesco and Simone mainly worked on the Regression part.

Some light testing was performed through Google Colab, but most of the testing was handled by Vito, Francesco and Simone, who parallelized the tests between their computers.

The presentations and the report were penned by Francesco and Vito with the help and information from the entire team.

We all had different backgrounds, with different bachelor degrees: two of us came from computer science, one of us came from engineering in computer science and two of us came from statistics, this brought a large pool of knowledge that favored our productivity overall, and by the end of the work we all had learned something new from each other.

REFERENCES

[1] F. Almeida T. Matos P. Cortez, A. Cerdeira and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

[2] Dexter Nguyen. Red wine quality prediction using regression modeling and machine learning.

[3] Nataliia Rastoropova. Step-by-step guide for predicting wine preferences using scikit-learn. Analytics Vidha.

[4] Center for Machine Learning and Intelligent Systems, University of California. Machine learning repository.