Classifying style and predicting quality of *vinho verde*

Jonathan Calles (jecalles@stanford.edu)

Vinho verde comes from north Portugal in two varieties, red and white, both delicious. PC: Erremoura - Own work, CC BY-SA 4.0,

Wine is exquisite,

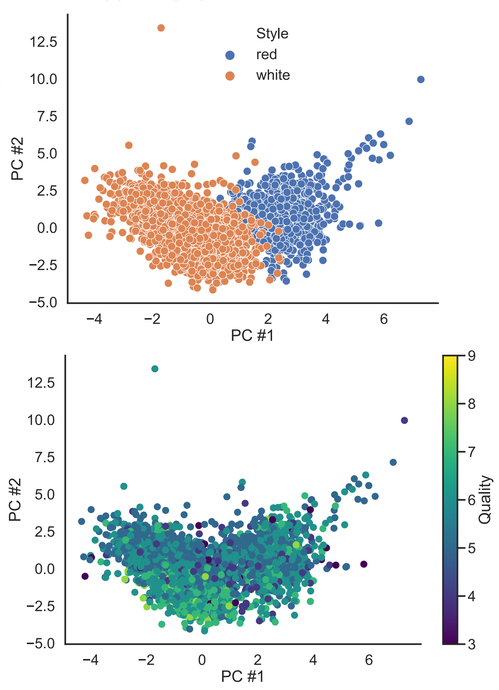
but often expensive. Moreover, graduate students are often poor. Sommeliers, on the other hand, are paid reasonably TO drink wine [1]. I've clearly misapplied my skills and training in ML. Or have I?

Cortez and colleagues [2] collected a dataset collating 11 physical measurements of each of 6497 examples of *vinho verde* from north Portugal. They also labelled each wine's style, red or white, and quality, as assessed by a panel of sommeliers.

Can I train models to predict the style and quality of a given wine, knowing only its physical properties, and become a robo-sommelier?

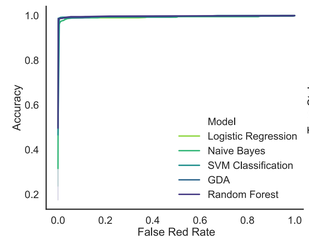
Using Principle Component Analysis (PCA), we can see that wine style can be described accurately with two principle components and a linear decision boundary.

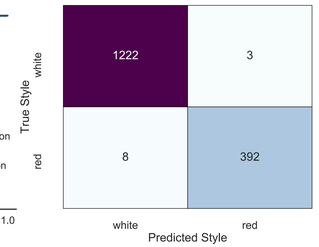
Quality, on the other hand, seems trickier. It appears as if different styles of wine have different ideal qualities, represented by areas with higher scores (yellow). This makes sense, but there is no discernable trend describing the difference between mediocre (blue) and bad (purple) wines.



*Wine style (**top**) and quality (**bottom**) plotted against the first two principle components of input features from the training set*

Classifying wine by its style is simple…





	Accuracy	Balanced Acc	F score	Cohen's Kappa	ROC AUC	# PC Used
Logistic Regression	99.38%	99.00%	0.9874	0.9668	0.9942	11
Naive Bayes	98.58%	97.55%	0.9708	0.9237	0.9928	2
SVM Classification	99.57%	99.13%	0.9912	0.9768	0.9971	11
GDA	99.63%	99.25%	0.9924	0.9801	0.9956	11
Random Forest	99.32%	98.88%	0.9862	0.9635	0.9977	11

***Top Left:** Receiver Operating Characteristic (ROC) curve [3] describing the performance of each classifier on the test set. **Top Right:** confusion matrix summarizing the performance of the Random Forest [4]. **Bottom:** table describing the performance of each classifier. Balanced accuracy weighs each class by its prevalence, F score [5] summarizes the precision/recall of a model, Cohen's Kappa [6] accounts for the predictor agreeing by chance, and the area under the curve (AUC) summarizes the ROC curve for each model.*

Each classifier was trained on data that was standardized to mean zero and unit variance, then transformed by PCA. The number of principle components used, as well as the hyperparameters for each model, were determined by 5-fold cross validation on the training set data.

Before testing, all models were retrained on both the test and dev sets.

All tested classifiers were able to robustly predict the style of a given wine from its physicochemical properties. Most models benefited from no data reduction (*c.f.* Naive Bayes).

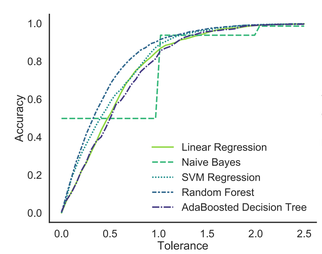
…but predicting its quality is harder

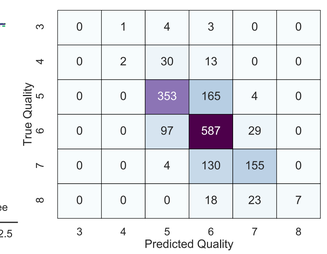
Each regressor was similarly trained and optimized to the classifiers above.

Predicting subjective quality proved much harder. The best published model by Cortez *et al.* 2009, an SVM regressor with a radial kernel, gave a mean absolute error (MAE) of 0.45 for white wines and 0.46 for reds, but required information on the style of wine.

The Random Forest model implemented here outperforms the authors' model, giving an MAE of 0.44 and generalizing to both wine styles.

Still, I probably couldn't pass for a sommelier with these stats…





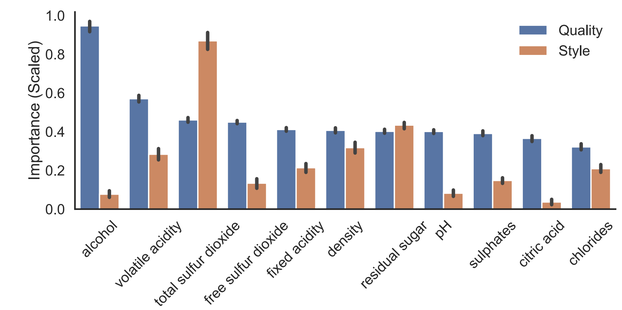
	Accuracy	Balanced Acc	MAE	R^2	REC AUC	# PC Used
Linear Regression	52.43%	24.38%	0.5688	0.2878	1.933	11
Naive Bayes	49.85%	28.16%	0.5797	-0.0311	1.931	9
SVM Regression	58.22%	29.08%	0.5077	0.4004	1.994	11
Random Forest	67.94%	37.10%	0.4374	0.5202	2.063	11
AdaBoosted Decision Tree	49.91%	21.99%	0.5876	0.2572	1.913	10

***Top Left:** Regression Error Characteristic (REC) curve [7] describing the performance of each regressor on the test set. **Top Right:** confusion matrix summarizing the performance of the Random Forest. **Bottom:** table describing the performance of each classifier. Mean absolute error describes the average deviation of each prediction from the true value, R^2 describes how much variance quality is described by the model, and AUC summarizes the REC curve for each model.*

Feature analysis suggests…

… that experts care quite a bit about the alcohol content of their wine ;) Harsh chemicals (acidity, sulfur dioxide, etc.) also weigh heavily in quality assessment.

The wine style can be most easily determined by the total SO_2 content of the wine, or by its sugar content.



Feature importances for Random Forest models in Quality (blue) and Style (orange) prediction problems, respectively. Importances were determined by feature permutation [4]

What didn't work?

- Using different scalings for input data (transform to normal/uniform, min/max scale)
- Clustering wines into sub-styles (e.g., grape varieties) using t-SNE, spectral clustering, or locally linear embedding
- Clustering excellent wines (quality >8) to see if they represent distinct "ideal" subtypes
- Training separate models on subclusters
- Using AdaBoosting, GradBoosting, or other boosting methods for quality prediction
- Using dimensionality reduction to mitigate variance in data
- Using DBSCAN to detect and eliminate outliers in training data
- Train separate models on inliers and outliers (this did worse than random guessing)

What's next?

More data. 6500 points is insufficient for more expressive models (e.g., neural nets)

More feature rich data. E.g., spectroscopy (color, absorbance, turbidity), more chemistry (e.g., tannin content), more physics (e.g., viscosity)

More wine!

References: [1] J. Smith, "What it takes to become a master sommelier," Business Insider, 2015. [2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," Decis. Support Syst., vol. 47, no. 4, pp. 547–553, Nov. 2009. [3] T. Fawcett, "An introduction to ROC analysis," Pattern Recognit. Lett., vol. 27, no. 8, pp. 861–874, Jun. 2006. [4] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001. [5] C. J. Van Rijsbergen, "Foundation of Evaluation," J. Doc., vol. 30, no. 4, pp. 365–373, Jan. 1974. [6] J. Cohen, "A Coefficient of Agreement for Nominal Scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960. [7] J. Bi and K. P. Bennett, "Regression Error Characteristic Curves," ICML 2013, 2003.