COS 424 Final Project: Machine Learnt Sommelier

Joon Seo
Princeton University
Precept P04
jsseo@princeton.edu

Abstract

What makes a great wine? We use topic modeling, classification, and regression methods in order to examine this question through looking at 130,000 reviews of wines from WineEnthusiast. We use Latent Dirichlet Allocation and Non-Negative Matrix Factorization to break down wine reviews to different topics. We apply Bernoulli Naive Bayes, Logistic Regression, and Linear Support Vector Machine classification to classify wines to "above average" and "below average" after observing various factors in the production of wines. Lastly, we explore relationships between factors surrounding the creation of a wine and its review score, applying Linear Regression, ElasticNet Regression, and Ridge Regression. The dominant topics of wine reviews dealt with the again process, the notes, the brightness, the scent, and the fruits within a wine. This task of topic modeling was best completed with Non-Negative Matrix Factorization. Classification of wines by quality showed that Linear SVM and Logistic Regression were the most promising for this task. Meanwhile, regressions showed that higher rated wines originated from countries with well-regarded terroir such as France and Italy, along with American wines. Lower rated wines were associated with less known origins of production and niche varieties. Through the combination of latent topic analysis, classification of wine quality, and study of factors related to ratings, we are able to obtain a better understanding of great wines and to evaluate wines in the future knowing certain background information surrounding its production.

1 Introduction

Much of the intricacies of wine tasting remain shrouded in mystery, decreasing the accessibility of this ancient beverage to many who wish to enter the world of fine wines. We wish to uncover what factors are relevant to great and mediocre wines through applying machine learning methods to 130,000 different reviews from Wine Enthusiast that were scraped on November, 2017. We look at the factors surrounding the creation of the wine and what characteristics sommeliers look for in rating a great wine. We create an insightful set of machine learning models that can be applied to factors surrounding a wine to arrive at a rating similar to how a sommelier would rate a given wine. This would shed some light upon what specific factors these sommeliers look for in great wines and what we as consumers should look for when shopping for our next bottle.

In order to accomplish this task, we look to apply unsupervised learning methods such as Latent Dirichlet Allocation and Non-Negative Matrix Factorization to reviews of wines by sommeliers in order to find top topics in wines that are rated highly or rated poorly. We also apply supervised learning methods such as Logistic Regression, Bernoulli Naive Bayes, and Linear Support Vector Machines to find how to best classify wines by their quality. Finally, we explore relationships between factors such as region of production, variety of wine, and winery of origin to the rating of the wine by the sommelier. We use linear regression, ElasticNet Regression, and Ridge Regression

to examine how each of these factors impact the rating of the wine, looking specifically towards which components affect the rating of a wine the most.

2 Related Work

We look towards some of the literature on the application of machine learning methods to the evaluation of alcoholic beverages. Cortez et al. (2009) applied support vector machines, similar to the linear support vector machines we apply in our analysis, and other regression techniques in order to attempt to predict the quality of red and white wines [4]. Aguilera et al. (2012) look towards applying neural networks and an electronic nose based upon independent component analysis to classify varieties of wine [3]. It is interesting to see classification of wines through machine learning, as that is similar to our goal of scoring wines through some of the same characteristics. On the other hand, Viejo et al. (2018) analyze the foam and the color of beer to arrive at a rating through principal component analysis and other methods [5]. We note the applications of classification and other supervised learning methods to evaluating alcohol. However, we also wish to look towards unsupervised learning methods, such as Latent Dirichlet and Non-Negative Matrix Factorization, to contribute to the literature in terms of understanding what goes into great wine.

2.1 Data

The data are comprised of 130,000 reviews from WineEnthusiast that were scraped on November, 2017 by user zackthoutt on kaggle. The data contain a description of the wine, title of the review, country of origin, vineyard of origin, score, price, province, region, taster name, variety, and winery of origin. We drop entries that are missing a rating or a dollar value. We also split this data into an 80:20 split randomly to training and test sets for evaluations of the performances of different methods.

We predict that the price of the wine will be one of the more important factors in determining a rating, although we are aware that there are great, cheap wines on the market. We also predict that the province of origin and winery of origin are important factors to the rating of a wine, as such factors are mentioned as important to the terroir of the grapes.

2.2 Preprocessing methods

We preprocess the contents of reviews prior to topic modeling using the Count Vectorizer method in SciKitLearn. We use a preprocessor similar to the one provided by the COS 424 in order to stem, extract stop words, and convert the reviews into bag-of-words representations.

We use one-hot encoding on categorical variables. We look specifically at the variables of country of production, province of production, the taster's name, and variety of the wine. We also break apart wine scores for the Naive Bayes and SVM classifiers into "above average" and "below average", breaking apart the approximately normally distributed data into a binary classification of reviews. This allows us to apply more intuitive sentiment analysis methods to the data by having a binary classification of each wine.

2.3 Classification methods

We use topic modeling, classification, and regression methods from the SciKitLearn library.

1. *Latent Dirichlet Allocation*: we generate 5 topics for n components, use batch learning methods, learning decay of 0.7, maximum iterations of 60 to convergence, and learning offset of 10.
2. *Non-Negative Matrix Factorization*: we generate 5 topics for n components, coordinate descent solver, random initialization, maximum iterations of 60, l_1 ratio of 0, frobenius beta loss, and tolerance of 0.0001.
3. *Logistic Regression*: with l_2 penalty and liblinear solver.
4. *Bernoulli Naive Bayes*: with Grid Search cross validation of alpha of 0.1, binarize set to 0.0, and fit to priors.

5. *Linear Support Vector Classification*: with Grid Search cross validation of alpha of 0.0001, no class weights, epsilon of 0.1, eta of 0.0, l_1 ratio fo 0.15, hinge loss, maximum iterations of 1,000, and tolerance of 0.0001.

6. *Linear Regression*: with Grid Search cross validation of K best selection of 250 using f-score.

7. *ElasticNet*: using Grid Search cross validation with respect to mean squared error to find alpha of 0.0001, l_1 ratio of 0.5, maximum iterations of 1,000,000, and tolerance of 0.0001.

8. *Ridge Regression*: using Grid Search cross validation with respect to mean squared error to find alpha of 0.001, maximum iterations of 1,000,000, and tolerance of 0.001.

2.4 Evaluation

We measure performance when relevant for these different classifiers. We analyze topics for coherence and look closely at the implications behind the topics regarding quality of wine. We note that this is a rather subjective method of evaluation, but one must bear in mind the goal of the project to create an insightful model of evaluating different wines. We also look towards the distribution of weights within the different topics for our latent topic models. For unsupervised learning methods, we look towards precision-recall scores, ROC curves, and F1-scores. For the regressions, we look at Mean Squared Error, Residual Sum of Squares, and R-squared. We also analyze the different coefficients of the regressions in order to look at coherence and to better understand what factors add to and detract from wine ratings.

3 Results

3.1 Topic Modeling

3.1.1 Topic Generation

Table 1: LDA Topic Modeling of 5 Topics

Latent Dirichlet Allocation	
Topic 1: Mouth Feel	wine fru y it drink acid ripe s the rich
Topic 2: Wine Finish	y flavor finish palat appl c acid the fru e
Topic 3 : Wine Flavor	wine flavor it fru cherri oak y the thi rich
Topic 4 : Wine Notes	cherri palat the black aroma tannin spice red offer note
Topic 5 : Fruit Usage	flavor finish aroma fru thi palat plum berri cabernet blend

Looking at the topics generated by Latent Dirichlet Allocation, we are able to observe the typical topics discussed within wine reviews. It is worth noting that the results seem to be fairly disappointing in this case, as the topics are fairly muddled, and the categories we derived are not mutually exclusive nor exhaustive. The letters and occasional articles are due to Porter stemming leaving these fragments. One extension worth considering in the case of LDA would be to allow a greater number of iterations, as 60 left us with not as clear topics, but the runtime was already quite long. With more computing power, there is the possibility of having clearer topics through LDA than NMF.

The last topic is particularly interesting in that the usage of fruit and other berries aside from grapes is seen typically as a way of reducing the cost of the wine [1]. When thinking in the context of viticulture (how grapes are grown), topics 4 and 5 would be indicative of most likely wines on the lower end of the spectrum, as they involve flavors and notes that would not be as pronounced within a purely grape wine. Topics 1, 2, and 3 seem to deal with more of the typical red or white wine.

Table 2: NMF Topic Modeling of 5 Topics

Non-Negative Matrix Factorization	
Topic 1: Aging Process	wine fru it drink s ripe rich thi age the
Topic 2: Notes of the Wine	the palat e aroma note wh finish nose fresh appl
Topic 3: Brightness of Wine	y acid crisp fru fresh bright miner it balanc drink
Topic 4 : Scent of Wine	flavor finish thi aroma it oak feel a sweet berri
Topic 5 : Fruit Usage/ Taste	cherri black tannin palat aroma spice red the berri plum

In the case of Non-Negative Matrix Factorization, we see that the five topics are able to break down what would be the typical topics discussed in a review. The topics are a bit more comprehensible than those of LDA which is most likely due to NMF having enough iterations (60) to converge more. Perhaps it is not surprising that the topics present within the reviews of wine deal with the tastes and the scents of the wines, but in the context of gaining familiarity with what to look for within wines, the components within each topic presented offer some insight. These components are the top ten components in terms of weights within a given topic.

It is interesting to see the topic regarding the usage of fruits can be observed again in topics 4 and 5. It is surprising that tannin is associated with topic 5, considering it would probably be associated more so with grapes than other fruits. It is interesting to observe that much of the jargon deals with the scent and aftertaste of the wine, as can be seen with topic 2 and topic 4. The general impression of these reviews is that great care must be taken to distinguish between subtle hints and notes between different wines, with small differences leading to variation in scores.

3.1.2 Topic Weightings

We look towards weightings of the topics to see how the two models fit the reviews to the topics they generated. We are able to see that the LDA has a few reviews that have high weights with the five categories, but also have many reviews that have almost no weight in some of the categories. In the case of NMF, we see that many of the topics have lower weights, with none of the topics being very high matches with the reviews. This is probably due to the more mutually exclusive nature of the topics that have been generated.

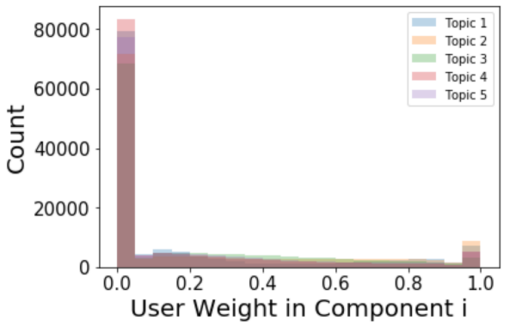

Figure 1: Topic Weighting of LDA

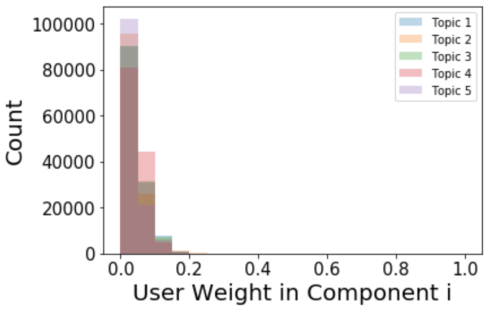

Figure 2: Topic Weighting of NMF

3.2 Classification

3.2.1 Comparison of Scores

We looked towards Bernoulli Naive Bayes, Linear SVM, and Logistic Regression to see how to best classify a wine as "above average" based upon the factors of country of origin, province of origin, wine taster name, variety of wine, and price of the wine. Our goal is to determine which of these three is the best classifier for this task by looking at precision, recall, and f1-scores.

Table 3: Comparison of Classifiers

	Precision	Recall	F_1-score	Support
Bernoulli Naive Bayes				
False	0.82	0.82	0.82	13,718
True	0.80	0.79	0.80	12,276
Linear Support Vector Machine				
False	0.83	0.87	0.85	13,718
True	0.84	0.80	0.82	12,276
Logistic Regression				
False	0.83	0.86	0.85	13,718
True	0.84	0.81	0.82	12,276

We are able to observe that the Linear Support Vector Machine has the best Precision and Recall scores for both true and false positive rates. It is interesting to note that the Bernoulli Naive Bayes classifier performs the worst across the board, which is surprising considering how well Logistic Regression performed in this task. Logistic Regression performs well at this task, with scores close to that of the Linear Support Vector Machine. However, this may be not as surprising due to the fact that the ratings of the wines was almost normally distributed, meaning converting the scores to below and above average would make the logit model appropriate for classification.

3.2.2 Comparison of Curves

In order to gain further insight into the relative performances of the three classifiers, we look towards precision-recall curves and ROC curves, after splitting the data into training and test sets. We can see that Linear SVM and Logistic Regression performed similarly and have similar areas under the precision-recall curve. This also holds true with Receiver Operating Characteristic curves, with Bernoulli NB having the worst performance. These curves support the previous conclusions over a broader range of values.

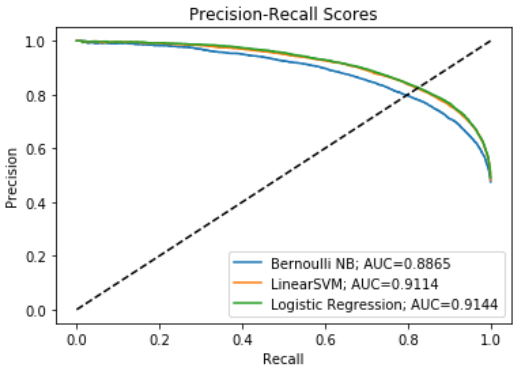

Figure 3: Precision-Recall Scores

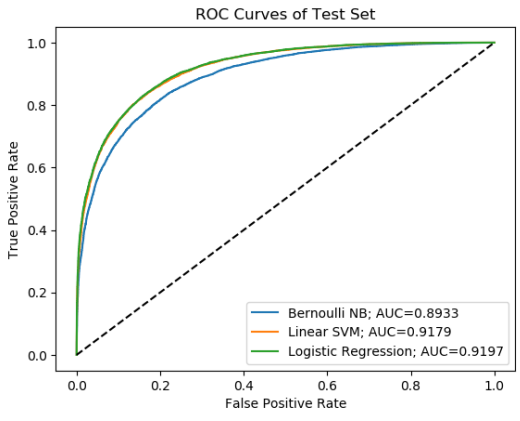

Figure 4: Receiver Operating Characteristic Curves

3.3 Regression

3.3.1 Model Comparison

We look towards regression models to attempt to predict the score of a given wine and in order to understand how different factors affect the score of a wine. We look towards the metrics of

mean squared error, residual sum of squares, and r-squared in order to compare across the different regression models to then look towards the coefficients of that model.

Table 4: Comparison of Classifiers

Regression Model	Mean Squared Error	Residual Sum of Squares	r-squared
Ridge Regression	6.2980	152,379.097	0.3132
ElasticNet Regression	8.0851	195,617.9242	0.1183
Linear Regression	6.3456	153,532.1311	0.380

We see that the best performer in terms of mean squared error and RSS was Ridge regression, with Linear regression being a close second. ElasticNet regression had the worst performance in all aspects. It seems to be the case that a combination of l_1 and l_2 regularization does not work well with this data set, as can be seen with ElasticNet's performance, while either of the two regularization methods works well, as can be seen with Ridge and Linear. Linear regression has the highest r-squared value and we look towards this model to analyze the coefficients. This is due to the relatively high performance of linear regression and the simplicity of its model.

3.3.2 Positive Features Analysis

We analyze the features of linear regression associated the most with higher ratings in order to better understand what goes into a highly rated wine. We look towards the top 20 features with the largest, positive coefficient after fitting the linear regression model. This selection is made after a k-best selection based upon f1-score to have 250 components.

Table 5: Top 20 Features of Linear Regression by Coefficient

Feature	Coefficient
country: Argentina	4.118747e+09
country: Peru	3.886923e+09
province: England	8.453873e+08
country: Ukraine	1.868244e+08
taster name: Anne Krebiehl MW	2.764144e+00
province: Wachau	2.416535e+00
province: Puente Alto	2.284224e+00
province: Kumeu	2.269647e+00
province: Kamptal	2.219636e+00
taster name: Matt Kettmann	2.178286e+00
province: Traisental	2.139456e+00
variety: Muscadelle	2.078293e+00
variety: Picolit	2.061431e+00
province: Madeira	1.936232e+00
province: Leithaberg	1.802688e+00
province: Kremstal	1.773534e+00
province: Washington	1.721713e+00
province: California	1.688816e+00
variety: Sangiovese Grosso	1.636819e+00
variety: Sagrantino	1.618324e+00

We see that certain countries of production, provinces of production, and varieties have the largest positive impacts on the rating of a given wine. It is surprising to see that South American countries, such as Argentina and Peru, have a larger impact on the rating of the wine than other countries of production. This is most likely due to the much smaller sample of wines from those countries compared to a country like the United States. It does seem to be the case, rather surprisingly that English wines are highly rated, which we would not have assumed considering England is not particularly known for its outstanding terroir. It is interesting to note the bias towards higher scores that certain tasters have, and this allows us to account for the upward bias of their ratings.

It is perhaps not too surprising that varieties that hail traditionally from Italy and France are associated with higher ratings, as is the case with Muscadelle, Picolit, and Sangiovese Grosso. It is also interesting to see that the wines from Washington and California perform well in the ratings. This is rather in line with the general sentiment of American wines rising in quality, especially those of Napa Valley. We also note that many of the coefficients do not provide much valuable information, as they are derived from a very small subsample of wines in some cases, such as the case of wines from Ukraine. It is interesting to note that certain varieties of wine are associated with higher ratings. This points towards either higher perceptions of certain varieties, or certain varieties being produced under a limited set of conditions that would yield a better product. If the review system were to be impartial, it should be the case that a great example of one variety should be rated similarly to a great example of another.

3.3.3 Negative Features Analysis

Table 6: Bottom 20 Features of Linear Regression by Coefficient

Feature	Coefficient
province: Other	-4.118747e+09
province: Mendoza Province	-4.118747e+09
province: Ica	-3.886923e+09
country: England	-8.453873e+08
province: Ukraine	-1.868244e+08
variety: Tempranillo Blanco	-4.281550e+00
variety: Brachetto	-4.115134e+00
variety: Airen	-4.078972e+00
province: Molina	-3.743532e+00
variety: Portuguese Ros	-3.453592e+00
province: Rio Claro	-3.359180e+00
province: Greece	-3.034868e+00
variety: Viura-Chardonnay	-2.956762e+00
variety: Garnacha-Syrah	-2.738010e+00
country: Brazil	-2.629350e+00
province: Bulgaria	-2.545243e+00
province: Central Valley	-2.494183e+00
variety: Pinot Noir-Gamay	-2.465141e+00
variety: Prieto Picudo	-2.421631e+00
variety: Inzolia	-2.369286e+00

We look also towards the top 20 factors that lead to a lower rating in the different wines through looking at the coefficients of linear regression. It is not surprising to see that wines from miscellaneous provinces performed poorly, as they would not have famous terroir nor recognition for the

wineries present. It is interesting to see that Mendoza province is associated with lower scores, considering that it is within Argentina, which we previously found is associated with higher scores. This is most likely due to a small sample of wines produced within that specific province.

It is surprising to see that wines produced in the country of England have lower ratings, while wines produced in the province of England have higher ratings. This is most likely due to reviewers misclassifying England as a province for a small selection of wines that happened to be great, leading to misleading coefficients. It does make sense that English wines, which are not traditionally recognized as some of the best, would have lower scores. We see that wines of hybrid varieties that are more niche tend to have lowered scores.

4 Discussion and Conclusion

We first looked at the efficacy of Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) in topic modeling with this data set of 130,000 wine reviews. We saw that NMF seemed to generate the most coherent topics and also had a more even distribution of topic weightings. Through the latent topics found that deal with aging, notes, brightness, scent, and blend of wines, we are able to obtain a better idea of the buckets of criteria that a virtual sommelier should look towards in evaluating a wine. It may actually be the case that LDA will be able to outperform NMF, but due to computational constraints, we would use NMF in subsequent analyses.

We then looked at the application of classifiers to the task of classifying wines to above or below average, looking at a series of factors surrounding the production of a wine. We saw that out of Bernoulli Naive Bayes, Linear SVM, and Logistic Regression, the Linear Support Vector Machine had the best performance in terms of precision, recall, and f1-score. This may be not too surprising considering its previous use in the literature for a similar task, as mentioned above [4]. We opted to evaluate wines on a binary scale in order to have cleaner results due to having much fewer classifications than a 100-point scale, but we sacrificed granularity. We instead rely upon regression models to arrive at a numeric score and use the classifiers as a check for the ratings, but there could be potential in converting classification to a points scale given more reviews.

Regressions were able to shed light upon which factors add to and subtract from wine ratings, with ridge regression and Linear regression performing the best. Our findings that the terroir of the region of production and the desirability of the variety impacting the score are not too surprising. Through the combination of these three different types of machine learning tasks, we are able to model the rating of a wine with information regarding its production. We have also extracted the criteria that sommeliers use in evaluating a wine, allowing us to have a set of criteria for a virtual sommelier. Nonetheless, we still maintain a healthy dose of skepticism regarding much of the factors surrounding a rating of a wine, as much of the ratings is dependent upon personal preferences, even between sommeliers. However, machine learnt sommelier should be able to account for such inconsistencies through even more training data.

5 Extensions

One possible extension is to assimilate these three types of models into one virtual sommelier. This would entail the creation of a user interface that would have inputs regarding factors surrounding the production of the wine, the type of wine, and the taste of the wine. There may also be better methods of preprocessing the reviews, as not all of the words within our bag-of-words representation were particularly meaningful. We were also limited by computational power and were not able to run as many iterations as could have been optimal.

Another form of analysis would be to cut out any human interaction with the wine in our analysis by having sensors to detect tannin levels, acidity, alcohol levels, oxidation, and other factors of wines. With the application of machine learning methods such as classification with these readings, we would be able to generate ratings and reviews of wines. This would be an exciting direction towards evaluation of wines in that much of the irrational human factor, such as certain reviewers giving higher ratings than others, would be cut out.

Acknowledgments

References

[1] Berry Bros. & Rudd.

[2] What Makes Great Wine?

[3] Teodoro Aguilera, Jess Lozano, Jos A. Paredes, Fernando J. lvarez, and Jos I. Surez. Electronic Nose Based on Independent Component Analysis Combined with Partial Least Squares and Artificial Neural Networks for Wine Prediction. *Sensors*, 12(6):8055–8072, June 2012.

[4] Paulo Cortez, Antnio Cerdeira, Fernando Almeida, Telmo Matos, and Jos Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, November 2009.

[5] Claudia Gonzalez Viejo, Sigfredo Fuentes, Damir D. Torrico, Kate Howell, and Frank R. Dunshea. Assessment of Beer Quality Based on a Robotic Pourer, Computer Vision, and Machine Learning Algorithms Using Commercial Beers. *Journal of Food Science*, 83(5):1381–1388, 2018.