From Food to Wine

Justin Meier

CS229 Final Project

1. Introduction

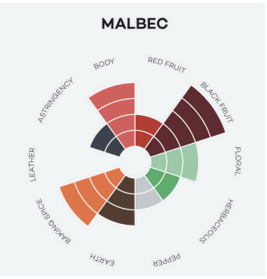
The art of food and wine pairing has dated back centuries and since its first practice, has honed its rules and regulations on what pairs well together. However, within these stringent rules, nuances and differing opinions exist that make selecting the perfect wine for a food more complicated than it may seem. Even beyond this, the complexities involved when selecting a wine for an entire dinner are even larger. The overall goal of the following presented algorithms will be to use these complexities and see if the nuances of wine pairing can be simplified. More specifically, the goal of the Naïve Bayes and K-Means algorithms will be to predict either the color of the wine (red vs. white) or a specific type of wine (Chianti, Pinot Grigio, etc.) to pair with a dinner.

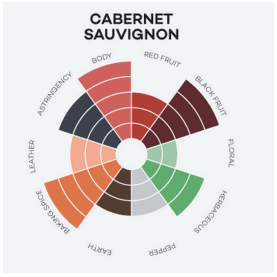
2. Dataset

The total number of datasets used in the algorithms is three; however note that certain algorithms do not use all three datasets in order to predict a wine. The first dataset contains self-reviewed wine and food pairings in which a specific type of food is compared to multiple types of wine. Each pairing is associated with a rating, ranging from 1 – 5, that determines the quality of the pairing. A depiction of this dataset is given below:

Food		Type	Varietal	Rating
Tomatoes Provencale	Paired with:	Rosé	Côtes de Provence, Rosé AOC	
Fried Green Tomatoes	Paired with:	White Wine	Sauvignon Blanc	
Fried Green Tomatoes	Paired with:	White Wine	Pinot Grigio	
Tomatoes	Paired with:	Red Wine	Barbera DOC	
Tomatoes	Paired with:	Red Wine	Chianti DOCG	
Tomatoes	Paired with:	Red Wine	Sangiovese	

The second dataset maps wines to a flavor profile. Each flavor profile contains ten elements which ranged from a scale of 0 – 5 in which 0 represents no presence of the taste in the wine and 5 represents a strong presence of the taste in the wine. An example of this is provided below:





The third dataset was far more difficult to find and the data was only obtained after numerous emails to the company associated with the site. The final dataset is a pairing of a specific wine to a specific dinner in which the ingredients of the dinner come as a list of foods. This dataset is ultimately used as my testing dataset with which I test whether my algorithms produced the correct color or type of wine. An example of this dataset is given below:

Steamed Clams and Mussels with Grilled Bread

Wine:
Color: White, **Varietal:** Sauvignon Blanc, **Vineyard:** Beringer,
Year: 2006, **Region:** Napa Valley, California

Seafood - Shell Fish

Users rating: view details... →

Finally, for better clarity of the results presented later, listed below are some general dataset statistics that help to put the results in a better context.

Total Pairings = 2521
Unique Foods = 477
Unique Wines = 131
Dinners = 495

3. Features & Preprocessing

3.1 Raw Input Data

For the first dataset, there are a total of three original features associated with the raw-input data: a specific type of wine, whether the wine is red or white, and a user rating. All of these features are mapped to a given food. The original features associated with the raw-input data for the second dataset are simply the listed 10 flavors and their numeric value. The specific flavors are body, red fruit, black fruit, floral, herbaceous, pepper, earth, baking spice, leather, and astringency. The third datasets features are a list of foods that composed a dinner and this list was mapped to a particular wine.

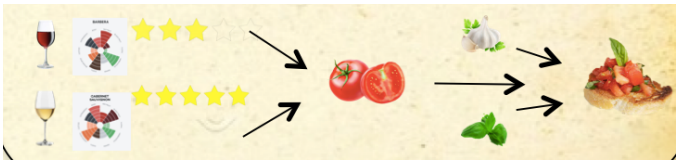
3.2 Derived Features:

The derived data comes from a combination of all three datasets. The first step involves calculating an average wine flavor profile for each food found within

the first dataset. The averaged wine flavor profile results from taking the flavor profiles for each wine, multiplying them by their rating, adding all of these flavor profiles together, and dividing by the sum of the ratings found for each pairing of the given food to its wine. This process ensures that the flavor profile is an average of all the flavor profiles of the wines associated with the food and also ensures that wines that have a higher ratings received more input in determining the flavor profile for the food.

A similar process is used for determining the average flavor profile for a given dinner. That is, the average flavor profile for a dinner is determined by averaging together the wine flavor profiles of the ingredients found within the dinner.

After completing both of these transformations, a database exists that contains the averaged flavor profiles for each individual food and also each individual wine. An illustration of this transformation can be pictured below, so as to better explain how the process worked:



$$RateSum(food) = \sum_{i \in wines(food)} rating(food)$$

$$FlavProf(food) = \frac{1}{RateSum(food)} \sum flavorProf(wine) * rating$$

$$FlavProf(dinner) = \frac{1}{numOfIngredients} \sum flavorProf(food)$$

4. Models & Results

The two types of models used are a Naïve Bayes algorithm and a K-Means algorithm. The goal of both algorithms is to either predict a specific wine or to predict whether one should drink a red or white wine. The input is the list of ingredients found within the third database, which uses the raw-input variables and the derived variables to predict a specific type of wine. The accuracy of prediction is given by whether the predicted wine was the color or the type of wine given by the third database of wines paired to dinners. The specifics of the algorithms are listed below:

4.1 Naïve Bayes:

The Naïve Bayes algorithm uses the generalized Naïve Bayes formula with multiples variables, given below:

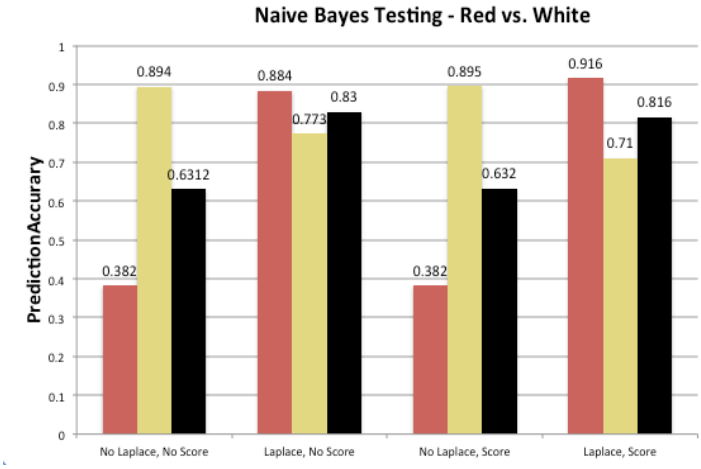
$$P(x|f_1 \dots f_n) = \frac{P(x)\prod_{i=1}^{n} P(f_i|x)}{P(f_1 \dots f_n)}$$

$$\arg\max P(x|f_1 \dots f_n) = \arg\max_x P(x) \prod_{i=1}^{n} P(f_i|x)$$

That is, the best prediction (whether it be a the color of the wine or a specific wine), will be the label that produces the highest product when considering the probability that we have a specific ingredient given our label. Below are the summaries of the two types of predictions that were made using Naïve Bayes:

4.1.1 Red vs. White Prediction:

For the red and white prediction, I use a combination of score consideration and Laplace smoothing. If I do not consider the ranking (no score consideration), then I simply calculate the probability of a food given a specific color based on the occurrence of the food and wine color together. If I consider the score, then I use the ranking in determining the probability that a given food occurred given the color of the wine. That is, wine pairings with higher rankings increase this probability, thereby giving preference to wines that are paired well with a food. Laplace smoothing works normally when there is no score consideration. However, when there is score consideration, the smoothing parameter is given as the average of all the ratings of food and wine. The accuracy of these four types of Naïve Bayes classification can be seen below:



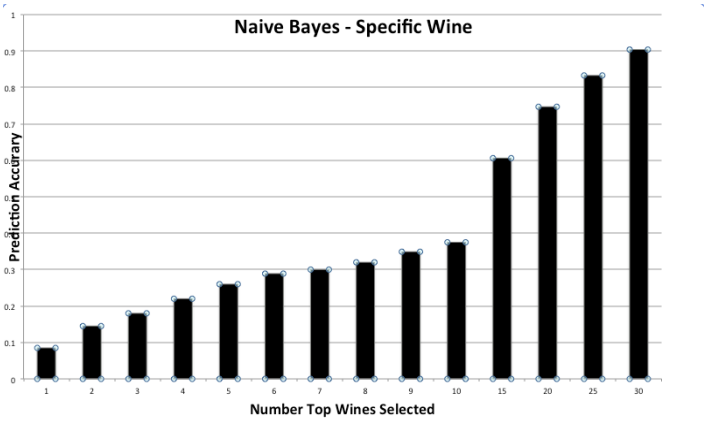
Red vs. White			
	Red Pred. Accuracy	White Pred. Accuracy	Both Pred. Accuracy
No Laplace No Score	38.2%	89.4%	63.1%
Laplace No Score	88.4%	77.3%	83.0%
No Laplace Score	38.2%	89.5%	63.2%
Laplace	91.6%	73.0%	81.6%

Score			

4.1.2 Specific Wine Prediction:

When calculating the probabilities for finding a specific wine, the algorithm uses the model with the maximum total accuracy from the red/white prediction problem. In this case, the probabilities involve no score consideration but do involve Laplace smoothing.

In order to best interpret the accuracy of the algorithm, I decided to not only keep track of the top wine predicted by the algorithm by varying top wine predictions. Then, the accuracy of the algorithm is determined by whether the actual wine paired with the dinner is found within the top N wines that were kept track of. The results of this are demonstrated below:



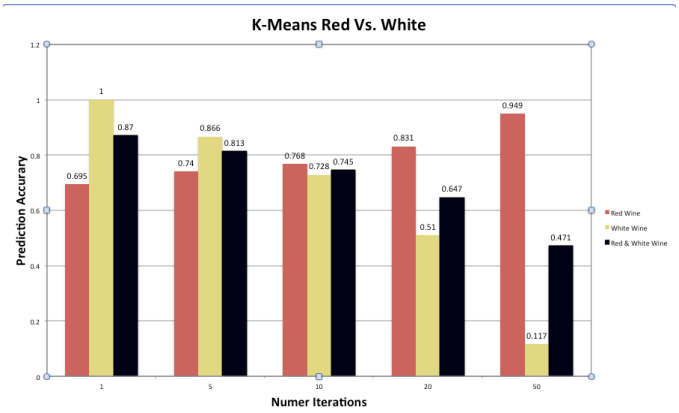
Specific Wine (with top N wines)	
N = 1	8.5%
N = 3	18.0%
N = 5	26.0%
N = 10	37.5%
N = 20	74.7%
N = 30	90.4%

4.2 K-Means

The K-Means algorithm functions as a normal K-Means algorithm with the variables used to determine distance being the values associated with the flavor profile, from 0 to5. The points of the algorithm are the flavor profiles of the dinners and the centroids are the flavor profiles of the wine, given by the original data from the second database. As each point converges closer to a centroid, the algorithm predicts the wine represented by the centroid for the dinner represented by the point. These same ideas are used to calculate both the color of the wine as well as a specific type of wine. The specifics and results of each of these is given below:

4.2.1 Red vs. White Wine Prediction

This algorithm can be approached in two separate ways. The first follows the exact procedure detailed above and then the prediction for the color of the wine is given based on the color of the predicted wine. Ultimately, my k-means algorithm would not converge in a measurable and feasible time range and despite numerous attempts to resolve this issue, it still remains. I tried a multitude of things, including decreasing the side of the variables associated with calculating the Euclidean distance between the points and the centroids. Instead, I decided to change the number of iterations The results of this are shown below:



Red vs. White from Specific Wine			
	Red Pred. Accuracy	White Pred. Accuracy	Both Pred. Accuracy
Iters = 1	69.5%	100%	87.0%
Iters = 5	74.0%	86.6%	81.3%
Iters = 10	76.8%	72.8%	74.5%
Iters = 20	83.1%	51.0%	64.7%
Iters = 50	94.9%	11.7%	47.1%

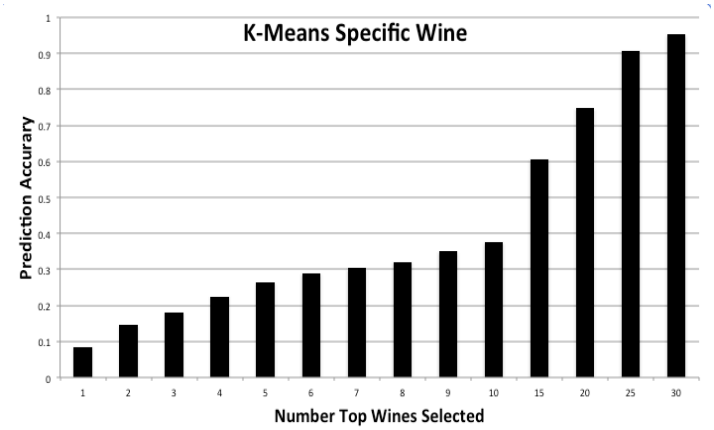
The second approach involves computing an average flavor profile for a red wine and for a white wine by averaging together all the red wine flavor profiles and all the white wine flavor profiles. From here, there are a total of two centroids representing all of the white wines and all the red wines. Therefore, as a dinner converges on a centroid, the algorithm would associate it with the color that centroid represented. The results of both of these algorithms are given below.

Red vs. White Prediction Accuracy with Averaged Red/White Flavor Profiles		
Red Pred. Accuracy	White Pred. Accuracy	Both Pred. Accuracy
61.7%	80.0%	70.9%

4.2.2 Specific Wine Prediction:

For finding a specific wine, I use the number of iterations that produced the highest accuracy when predicting the color of the wine. This occurrs when

there was only one iteration. Much like the Naïve Bayes equivalent to solve this algorithm, the algorithm keeps track of the top N predicted wines rather than simply the top predicted wine so as to better show the accuracy of the algorithm. In order to calculate the top N wines given that the wine is grouped to a specific centroid and, therefore, only one specific wine, the algorithm calculates the Euclidian distance from the point to all other centroids and kept track of the top N centroids that have the smallest distance. The results of this algorithm can be seen below:



Specific Wine (with top N wines)	
N = 1	8.6%
N = 3	18.0%
N = 5	26.3%
N = 10	37.6%
N = 20	75.1%
N = 30	95.2%

5. Discussion

5.1 Naïve Bayes

Overall, the Naïve Bayes algorithm functioned quite well in predicting both the color of the wine and a specific type of wine. In terms of score consideration and Laplace, it was evident that the algorithm performed best when Laplace smoothing was applied and had a negligible difference with score consideration as opposed to no score consideration.

The increasing in the accuracy with Laplace smoothing makes sense. Going through the database, I noticed a few foods that occurred with white wines with low ranks and then only a few red wines with high ranks. Ultimately, the Lagrange smoothing more evenly distributes the probability given a red or white wine and thereby increases the overall accuracy of the equation. This is further demonstrated by an increase in the accuracy of red wine prediction with a decrease in the accuracy of white wine prediction.

In terms of predicting a specific wine, the results far exceed what would be expected. With a total of 131 wines, the accuracy of randomly selecting a single correct wine from this set would be 0.76% but the algorithm is able to predict the top wine with 8.0% accuracy, more than 10 fold better than guessing. As well, the algorithm sad the correct wine in the top 30 guesses 90% of the time, demonstrating that the algorithm is able to recognize the correct wine as a strong contender, if even it doesn't select the wine as the number one choice.

5.2 K-Means

Overall, the K-Means model has a high accuracy rate but, based on consistency, functions worse than the Naïve-Bayes model, as demonstrated by the K-Means attempt to predict the color of wine to pair with the dinner. In addition to its inability to converge, the K-Means model also varies drastically on the accuracy for red wines and the accuracy for white wines. That is, with fewer iterations, the accuracy for white wine prediction is much higher, reaching a peak of 100%, in fact, with a single iteration. However, as the number of iterations increased, the accuracy of white wine drops to a low of 11.7% with 50 iterations. This effect is reversed for the accuracy of a red wine prediction.

One theory I have behind this phenomenon is that quite a few of the dinners that are extremely strongly associated with a white wine. That is, there are a few white-wine-paired dinners that have only highly white wine weighted ingredients such as any type of fish, clam, mussels, lemon, lime, any number of vegetables, and olive oil. Essentially, any type of seafood dish. However, there are a variety of white-wine-paired dishes and nearly all red-wine-paired dishes that contain both ingredients strongly associated with red wine and ingredients associated with white wines. Therefore, at the beginning, the white wine prediction is high. However, as the number of iterations increases, the white wine centroids become more similar to the extreme white-wine-paired dishes. As this occurs, more dishes will be paired with a red wine because only the few cases of extreme white-wine paired dishes will be clustered with a white wine centroid, thereby decreasing the accuracy of predicting a white wine and increasing the accuracy of predicting a red wine.

For predicting a specific type of wine, the K-Mean models accuracy is extremely close to the accuracy achieved in the Naïve Bayes model. That is, with only selecting the one predicted wine, the model accurately predicts the correct wine 8.1% of the time. And with selecting the top 30 predicted wines, the model accurately predicted the correct wine in these predictions 95.2% of the time.

6. Conclusion

Based on the results of the two algorithms, the naïve Bayes is more consistent at predicting the color of the wine. In terms of predicting a specific type of wine, both algorithms perform well with negligible accuracy difference between the two.

In terms of the broad conclusion draw from the results, it seems that there exists a strong set of rules that dominate how wine and food pair together. Our original features and derived features originate from the first two databases whereas our method for checking accuracy came from an entirely different database. With the accuracy demonstrated by training and testing on completely separate databases, it seems that all three databases have an underlying theme of wine and food pairing that tied them together. That is, if any of the databases did not accurately reflect the wine/food pairing rules demonstrated in the others, we would likely not see this high level of prediction accuracy. Therefore, using my models, it seems that there is a strict set of rules that apply to wine/food pairing.

Finally, in terms of progressing from wine to food pairing into wine to dinner pairing, it seems that many of the same rules are still applicable. That is, using the rules and nuances of wine/food pairing, one can predict with a high level of accuracy a wine that pairs with a grouping of food as well. Therefore, it seems that the rules for wine pairing rules are specific enough to be represented accurately across a variety of databases and with a variety of models, but are also generalized enough to accommodate for having multiple foods.

7. Future

I think the one thing that would drastically increase the accuracy of the algorithm would be considering the importance that each food plays in the meal. That is, meat is usually paired with a red wine whereas most vegetables are usually paired with a white wine. However, in a case where we have a meat as the main component of the dish and the vegetable as the side component, it would seem that the meat should play a larger role in determining both what color wine to have and what specific type of wine to have. Although difficult, this could be done by obtaining the estimated or averaged weight of each ingredient in the meal and using this as a weight to weight the influence of each ingredient on predicting a wine. I would also do more analysis of the K=Means algorithm in an attempt to better understand why it predicts white wines with such high accuracy when it has so few iterations but with such low accuracy when it has a larger number of iterations. Finally, I think have more databases to use would be extremely helping, especially in bolstering the idea that strict wine pairing rules exist across a multitude of databases. Although this appeared to be the case for the three databases used in the models, having more databases to test this would be ideal.

I'd love to continue working on this project in hopes of making it a practical application that a variety of people could use to help pair a wine with their meal. In fact, while at the poster presentation, I was confronted by a man working on an iPhone application that paired a wine with a list of ingredients and then went a step further and showed a specific vintage and also the best place to buy the wine in order to obtain the cheapest price. Hopefully I am able to continue working on my models and help contribute to this application.

References

What To Pair [Online]. Available: http://www.whattopair.com

M. Puckette. (2014, Sept. 22). *Flavor Profiles of Wines (Infographic)* [Online] Available: http://winefolly.com/review/red-wine-flavor-profiles/

Match My Wine [Online]. Available: http://www.matchmywine.com/