Learning the Language of Wine

Aaron Effron, Alyssa Ferris, and David Tagliamonti
{aeffron,acferris,dtag}@stanford.edu

Motivation

- Wine has been an integral element of human society for millennia.
- Interpreting wine speak can be intimidating.
- Can machine learning make wine more accessible?

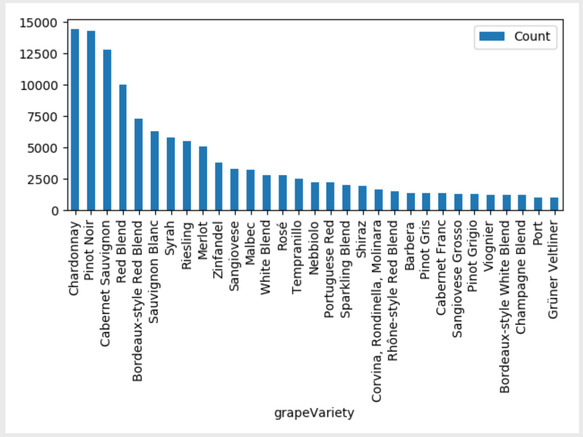
Task Definition

Use expert wine descriptions to:

1. Classify red vs. white
2. Classify grape varieties
3. Make recommendations

Data & Features

- 150k wine dataset from Kaggle with: red/white, grape variety, and ~25 word descriptions
- Features: word features, character features, and word embeddings (word2vec)



Challenges

- Many classes of grape varieties
- Many infrequently used words, which can lead to overfitting the data.

Models & Results

Task 1: Red vs White	Word Features (1)		Char Features (5)	
	Train	Dev	Train	Dev
Baseline	**0.617**	**0.657**	**0.617**	**0.657**
Naïve Bayes	0.880	0.791	N/A	
Logistic Regression	0.998	0.978	1.000	0.982
SVM	0.997	0.977	1.000	0.980
Decision Trees	1.000	0.964	1.000	0.961
Random Forest	0.999	0.967	0.999	0.971

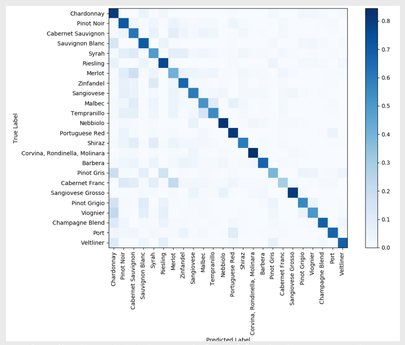
Task 2: Grape Variety	Word Features (1)		Char Features (5)	
	Train	Dev	Train	Dev
Baseline	**0.18**	**0.192**	**0.18**	**0.192**
Naïve Bayes	0.825	0.431	N/A	
Logistic Regression	0.951	0.662	1.000	0.670
SVM	0.936	0.655	1.000	0.657
Decision Trees	1.000	0.537	1.000	0.536
Random Forest	0.991	0.572	0.987	0.584

- The results are based on 25,000 examples with a 70/30 train/dev split.
- Models outperform random classification but drastically overfit the training data

Logistic regression

- Used L2 regularization
- Weighting emphasizes examples from underrepresented classes
- Weighted & regularized logistic (softmax) objective:

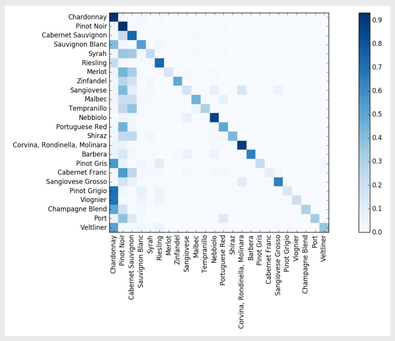
$$J(\theta) = -\sum_{i=1}^{m} w^{(i)} \sum_{k=1}^{K} y_k^{(i)} \log \hat{y}_k^{(i)} + \lambda \sum_{k=1}^{K} \|\theta_k\|_2^2$$



Confusion Matrix, weighted logistic regression

Random Forest

- Changing max tree depth, min number of samples per split, and/or increasing the number of trees has a minimal effect
- Most errors are misclassification of wines as the most abundant varieties



Confusion Matrix, Random Forest

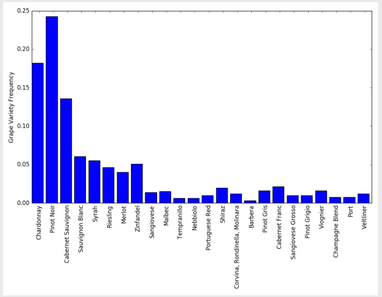
Discussion

- Results shed light on language usage in describing wines
- **Key Words for Predicting Red/White:**
  - *Red*: tannins, cherry, berry, blackberry, strawberry
  - *White*: yellow, tropical, pear, pineapple, apple
- **Key Words for Predicting Sauvignon Blanc:**
  - *Yes*: gooseberry, grass, herbaceous, herbaceousness, fig
  - *No*: tannins, cherry, berry, offdry, blackberry

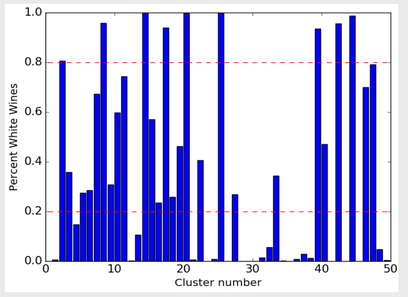
Wine Recommendations

Method 1

- Use k-means clustering to recommend similar wines
- Clustering metrics:
  - Mix of red or white within a cluster
  - Mix of grape varieties within a cluster



Distribution of grape varieties within one cluster



Distribution of white wines in different clusters

Method 2

- Use word2vec similarity to recommend similar wines
- Example recommendation pair:
  - Cabernet Franc: *Front-loaded, herbal, earthy and a bit dull. No fruit.*
  - Pinot Noir: *Thin and earthy, this is at its most herbal, with no fruit in sight.*

Future Work

- Use dimensionality reduction (e.g. PCA or word2vec) for regularization in classification
- An interesting extension of our project is to make wine recommendations using both textual similarity and a user profile of wine preferences

References

[1] Kaggle Featured Dataset. (2017). Wine Reviews [Data File]. Retrieved from: https://www.kaggle.com/zynicide/wine-reviews