Predicting wine varieties from corpus of text descriptions

Yuan Wang[1]

[1] Princeton University, Department of Computer Science, Princeton, NJ

yuanwang@princeton.edu

ABSTRACT

- The blind-tasting portion of the Master Sommelier's exam requires a candidate to use "deductive tasting" to infer a wine's vintage, varietal, and origin within 25 minutes.
  - But, such process requires highly-trained skills and is laborious.
- We are interesting in whether we can model the deductive tasting technique in a data-driven way.
  - We trained models that take the input of wine descriptions and predict the label of the corresponding grape varietal(s).
  - We use the dataset scraped by Zack Thoutt and train four types of multi-class classifiers: naive Bayes, logistic regression, XGBoost, and bidirectional long short-term memory (LSTM) neural network [1].
  - We evaluated model performance with area under ROC/Precision-Recall curve, F-1 score, accuracy, and log-loss. Overall, we found that logistic regression and XGBoost with TF-IDF embeddings produced competing performance.

BACKGROUND

country	description	designation	points	price	province	region_1	region_2	variety	winery
US	This tremendous 100% varietal wine hails from Oakville and was aged over three years in oak. Juicy red-cherry fruit and a compelling hint of caramel greet the palate, framed by elegant, fine tannins and a subtle minty tone in the background. Balanced and rewarding from start to finish, it has years ahead of it to develop further nuance. Enjoy 2022–2030.	Martha's Vineyard	96	235.0	California	Napa Valley	Napa	Cabernet Sauvignon	Heitz
Spain	Ripe aromas of fig, blackberry and cassis are softened and sweetened by a slathering of oaky chocolate and vanilla. This is full, layered, intense and cushioned on the palate, with rich flavors of chocolaty black fruits and baking spices. A toasty, everlasting finish is heady but ideally balanced. Drink through 2023.	Carodorum Selección Especial Reserva	96	110.0	Northern Spain	Toro	NaN	Tinta de Toro	Bodega Carmen Rodríguez
US	Mac Watson honors the memory of a wine once made by his mother in this tremendously delicious, balanced and complex botrytised white. Dark gold in color, it layers toasted hazelnut, pear compote and orange peel flavors, reveling in the succulence of its 122 g/L of residual sugar.	Special Selected Late Harvest	96	90.0	California	Knights Valley	Sonoma	Sauvignon Blanc	Macauley

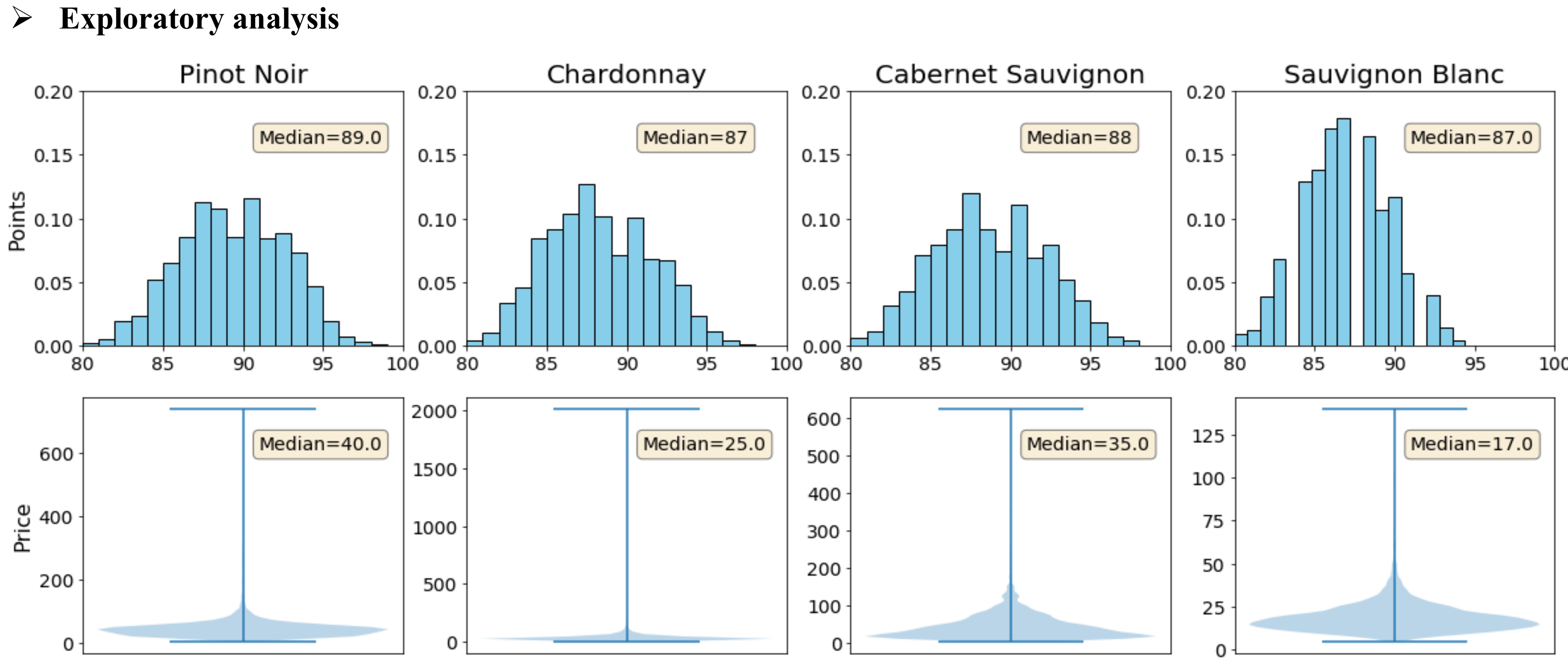
150,930 × 11

Data Processing

- Removed duplicated rows and unused columns
- Selected top 20 most represented varieties

71,200 × 2

- **Exploratory analysis**



WORD EMBEDDINGS

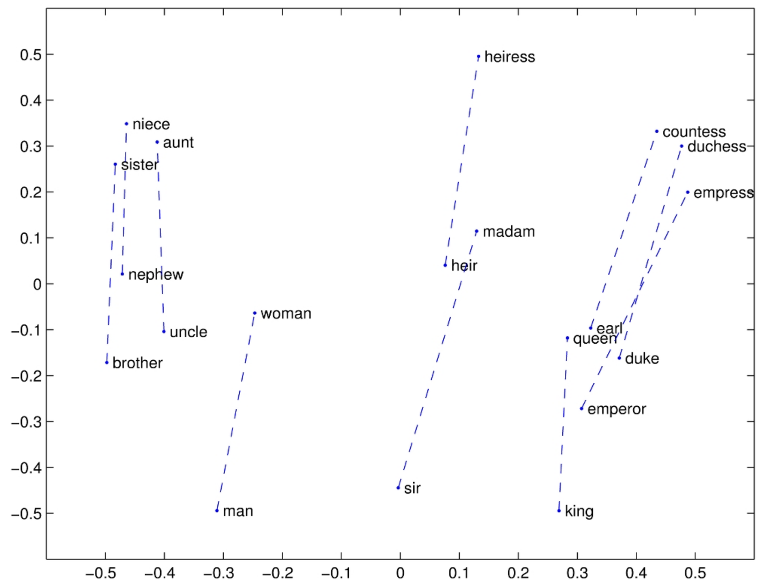
① Term Frequency-Inverse Document Frequency (TF-IDF)

- **Term frequency (TF)** gives us the frequency of the word in each document in the corpus. For word i and document j, the TF weight is calculated as
$$tf_{i,j} = n_{i,j} / \sum_{j \in N} n_{i,j}.$$
- **Inverse document frequency (IDF)** measures how much weight a word carries. It is given as
$$idf(i) = \log(N/df_i)$$
- Then, the TF-IDF equals
$$tfidf = tf_{i,j} \times idf(i)$$

df_i = # of documents containing i
N = total number of documents

- Overall, with the TF-IDF embedding, words that are more rare but actually help distinguishing between the data will carry more weight, while words that appear too frequently will be penalized with less weight.
- Vocab = 20,000

② Global Vectors for Word Representation (GloVe)



Linear structure

- GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence matrix from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space [2].
- In our case, we used the version trained on Wikipedia 2014 + Gigaword5 (6B tokens, 400K vocab, 300D vectors)

MULTICLASS CLASSIFIERS

Input:
- **TF-IDF**
  20,000 words
  71,200 documents
- **GloVe**
  300D embedding
  71,200 documents

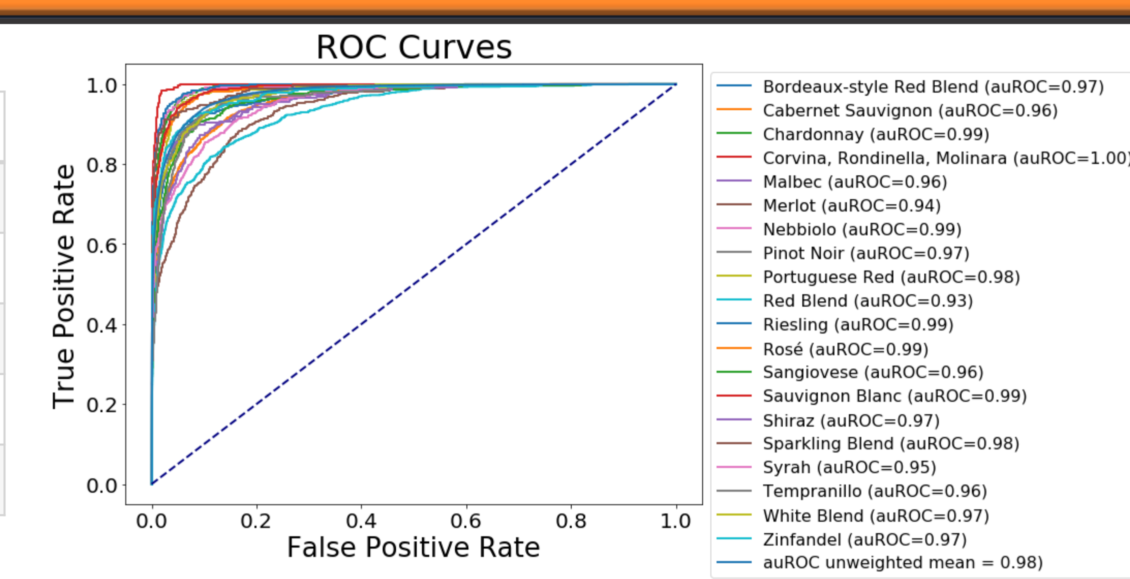
Multi-class Classifiers
- Naïve Bayes
- Logistic regression
- XGBoost (gbtree, 5-fold CV)
- Bidirectional LSTM [3]
  - Loss = categorical crossentropy
  - Optimizer = Adam

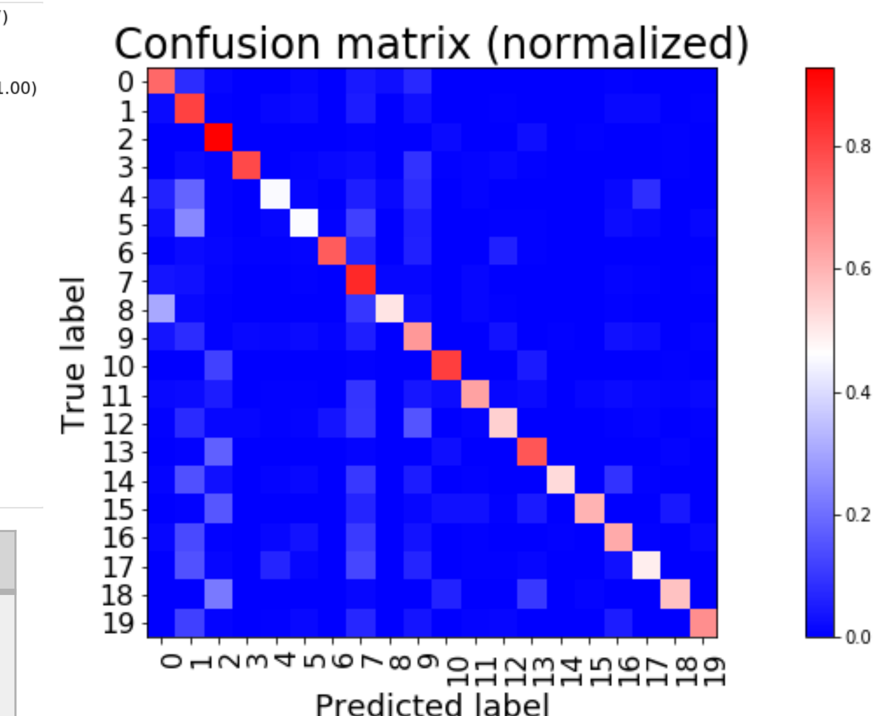
Output:
Predicted 20 labels
(80/20 train/test)

300D GloVe word embedding
→ Bidirectional LSTM layer with 100 hidden units
→ Dense layer with 100 units → ReLU activation
→ Dropout 0.3
→ Dense layer with 21 units → softmax activation

PERFORMANCE AND RESULTS

	TF-IDF			GloVe		
	Acc	F-1 (w)	LL	Acc	F-1 (w)	LL
NB	0.51	0.42	1.64	-	-	-
LR	0.73	0.72	0.94	0.61	0.60	1.25
XGB	0.72	0.65	1.08	0.57	0.56	1.43
LSTM	-	-	-	0.68	0.68	1.05





Cabernet Sauvignon	Chardonnay	Malbec	Merlot	Pinot Noir	Riesling	Sauvignon Blanc	Syrah	Zinfandel
blackberry, tannins	apple, Chablis	black, saturated	cherry, herb	silky, *Pommard*	honey, citrus	grapefruit, passion	pepper, *Hermitage*	Alcohol, blackberry, raspberries

FUTURE WORK

- Further fine-tune the hyperparameters of XGBoost with GloVe embeddings
- Use a better class labeling scheme
- Explore other deep neural networks such as CNNs

REFERENCES

[1] Kaggle: Wine reviews. https://www.kaggle.com/zynicide/wine-reviews. Accessed: 2019-04-22.
[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In IN EMNLP, 2014.
[3] im Aiken and Clara Meister. Applying Natural Language Processing to the World of Wine. Stanford University, CS230, 2018.