Wine Rating Prediction

Ke Xu (kexu@), Xixi Wang (xixiwang@)

Predicting

- We want to predict rating points of wines based on historical reviews from experts.
- We used the price, wine variety, and winery information as the training signals, and output the predicted rating for a wine.
- We explored several linear regression models and one neural network model.

Data & Features

Source: ~30k wine review data points scraped from WineEnthusiast.

Input features:

- Price: the cost for a bottle in float
- Variety: the type of grapes used to make the wine (i.e. Pinot Noir) in string
- Winery as string
- Location: derived signal by combining country, province and region of the wine as a string

Label:
Original rating points. Ranges from 80 to 100.

Pre-processing:
Treat string input features as categorical (using one-hot encoding) features. Removed duplicates, any empty features and rarely seen values (less than 10 occurrence in the whole data set).

Models

Linear Regression
We began with basic linear regression, and observed large coefficients and outliers. To improve that, we added 3 types of regularization: Lasso (L1), Ridge (L2) and Elastic Net (L1 and L2).
The formula for Elastic Net:

$$\hat{\theta} = \arg\min_{\theta}\{(y-\theta X)^2 + \lambda_1\|\theta\|_1 + \lambda_2\|\theta\|_2^2\}$$

Neural Network
Data processing: normalized price values into [0, 1]
NN-Structure:

- 2 hidden layers with 100 neurons in each
- ReLU is used as the activation function
- Optimizes the squared-loss using L-BFGS, which is an optimizer in the family of quasi-Newton methods
- 200 iterations in total

Results

We defined three metrics to evaluate the model performance, R^2 score, mean square error and median absolute error. From these metrics, we can see the best performance model is the linear regression model using Ridge regularization.

	Training			Testing		
	R^2 score	Mean Square Error	Median Absolute Error	R^2 score	Mean Square Error	Median Absolute Error
Basic Linear Regression	0.537	5.035	1.468	-1.38E+17	1.48E+18	1.602
Linear Regression w Lasso	0.225	8.428	2.086	0.239	8.163	2.044
Linear Regression w Ridge	0.535	5.060	1.481	0.468	5.711	1.593
Linear Regression w Elastic Net	0.225	8.427	2.078	0.240	8.159	2.046
Neural Network	0.522	5.144	1.488	0.466	5.860	1.610

Discussion and Future Work

We explored several different models and tuned with different parameters for each model, and had one common observation: for all the models, the performance for both training and testing data sets is not good as expected. This indicates that the review points can't be perfectly predicted by current features.

In the future, we would like to add features such as acidity, alcohol by volume, the age of the wine, to improve the performance of the models.

References

[1] Scikit-learn: Machine Learning in Python, Pedregosa etal., JMLR 12, pp. 2825-2830, 2011.
[2]https://en.wikipedia.org/wiki/Limited-memory_BFGS