Some Prediction of White Wine Grading

Chung, K.I. and Xiao, B.J.
Supervised by: Prof. Lo, M.N.

January 13, 2017

- Introduction
    - Overview of the Data
    - Correlation
    - PCA
- Multiple Linear Regressions
    - Procedure
    - Regular Model Selection Method
    - Best Subset
    - Lasso
    - Principle Component Regression
- Classifiers
    - Decision Tree
    - Random Forest
- Conclusion
- Reference

Overview of the Data

- From the UC Irvine Machine Learning Repository Website
- 4898 Observations and 12 variables
- 4408 training data and 490 test data

Response

The quality of the the wine is graded from 3 to 9.

Grade	3	4	5	6	7	8	9
Frequency	20	163	1457	2198	880	175	5

Predictors

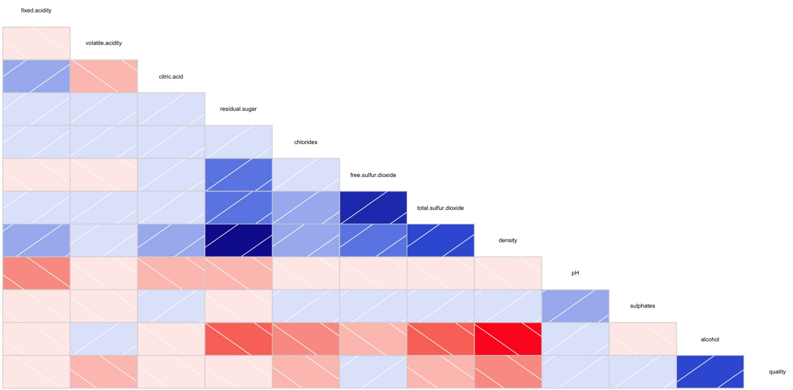
- Fixed acidity (g/l)
- Volatile Acidity (g/l)
- Citric Acid (g/l)
- Residual Sugar (g/l)
- Chlorides (g/l)
- Free Sulfur Dioxide (mg/l)
- Total Sulfur Dioxide (mg/l)
- Density (g/ml)
- pH
- Sulphate (g/l)
- Alcohol (%)

Target

- Find the relation between the variables
- Predict the quality of the wine
- Reduce the error rate

Table: Coefficient Matrix

1	-0.02	0.28	0.09	0.02	-0.05	0.09	0.26	-0.42	-0.02	-0.11	-0.11
-0.02	1	-0.14	0.06	0.07	-0.1	0.09	0.03	-0.04	-0.03	0.07	-0.19
0.28	-0.14	1	0.09	0.11	0.09	0.11	0.15	-0.16	0.07	-0.07	-0.01
0.09	0.06	0.09	1	0.09	0.29	0.40	0.84	-0.19	-0.03	-0.45	-0.10
0.02	0.07	0.11	0.09	1	0.11	0.2	0.26	-0.09	0.02	-0.36	-0.21
-0.05	-0.10	0.09	0.29	0.11	1	0.61	0.29	0.00	0.06	-0.25	0.01
0.09	0.09	0.11	0.40	0.20	0.61	1	0.53	0.00	0.14	-0.45	-0.17
0.26	0.03	0.15	0.84	0.26	0.29	0.53	1	-0.09	0.08	-0.78	-0.30
-0.42	-0.04	-0.16	-0.19	-0.09	0.00	0.00	-0.09	1	0.16	0.12	0.10
-0.02	-0.03	0.07	-0.03	0.02	0.06	0.14	0.08	0.16	1	-0.03	0.05
-0.11	0.07	-0.07	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.03	1	0.43
-0.11	-0.19	-0.01	-0.10	-0.21	0.01	-0.17	-0.30	0.10	0.05	0.43	1



The Correlation Plot

Table: PCA output

	PC1	PC2	PC3	PC4	PC5	PC6	h2	u2
fixed.acidity	0.28	-0.74	0.13	0.02	0.25	-0.10	0.71	0.287
volatile.acidity	0.01	0.06	-0.65	0.28	0.63	0.12	0.92	0.077
citric.acid	0.26	-0.43	0.56	0.15	0.05	0.13	0.61	0.393
residual.sugar	0.77	0.01	-0.24	-0.28	0.01	-0.28	0.80	0.200
chlorides	0.38	-0.01	-0.11	0.72	-0.32	0.38	0.92	0.076
free.sulfur.dioxide	0.54	0.36	0.31	-0.31	0.17	0.48	0.87	0.126
total.sulfur.dioxide	0.73	0.31	0.14	-0.06	0.29	0.27	0.80	0.195
density	0.92	0.01	-0.14	-0.02	-0.08	-0.32	0.97	0.028
pH	-0.23	0.73	0.14	0.10	-0.12	-0.19	0.66	0.336
sulphates	0.08	0.28	0.48	0.45	0.40	-0.47	0.89	0.114
alcohol	-0.78	-0.04	0.12	-0.14	0.33	0.13	0.78	0.219
SS loadings	3.22	1.58	1.22	1.02	0.97	0.94		
Proportion Var	0.29	0.14	0.11	0.09	0.09	0.09		
Cumulative Var	0.29	0.44	0.55	0.64	0.73	0.81		
Proportion Explained	0.36	0.18	0.14	0.11	0.11	0.10		
Cumulative Proportion	0.36	0.54	0.67	0.79	0.90	1.00		

Procedure

1. Fit a multiple linear regression with training data
2. Predict the response of the test data
3. Calculate the test MSE
4. Round off the predictions to integers range from 3 to 9
5. Calculate the test error rate

Full Model

Table: Coefficients of the Full Model

(Intercept)	fixed.acidity	volatile.acidity	citric.acid
137.5716	0.0569	-1.8633	0.0315
residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide
0.0771	-0.3096	0.0039	-0.0003
density	pH	sulphates	alcohol
-137.4853	0.6397	0.5982	0.2074

Collinearity Elimination

The following table shows that the predictors residual.sugar, density and alcohol as relatively high collinearity.

Table: VIF of the Full Model

fixed.acidity	volatile.acidity	citric.acid	residual.sugar
2.6444	1.1449	1.1562	12.5508
chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density
1.2390	1.7795	2.2262	27.8343
pH	sulphates	alcohol	
2.1600	1.1422	7.5554	

To eliminate the the effect of the collinearity, we will perform the "leave-one-in" model selection which provide us the least AIC model with one of three collinear predictors.

Table: AIC of the Three Candidates

residual.sugar	density	alcohol
10941.64	10722.53	10195.19

Insignificant Predictor Detection

Table: P-value of the Model without Collinearity

(Intercept)	fixed.acidity	volatile.acidity	citric.acid
0.00	0.00	0.00	0.70
chlorides	free.sulfur.dioxide	total.sulfur.dioxide	pH
0.00	0.00	0.47	0.75
sulphates	alcohol		
0.00	0.00		

The Final Model

Table: Coefficients of the Final Model

(Intercept)	fixed.acidity	volatile.acidity	citric.acid
3.1023	-0.0588	-1.8977	0.0221
free.sulfur.dioxide	total.sulfur.dioxide	alcohol	
0.0055	-0.0001	0.3345	

The Test MSE and the Test Error Rate

Table: Error Table

	Full	Without Collinearity	Final
MSE	0.705	0.724	0.727
Err. Rate	0.465	0.469	0.473

				tru			
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	1	2	0	0	0	0
5	0	8	61	27	0	0	0
6	0	6	79	184	69	9	1
7	0	0	0	21	15	7	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

				tru			
prd	3	4	5	6	7	8	9
4	1	1	0	0	0	0	0
5	0	7	57	30	1	1	0
6	0	8	82	185	66	8	1
7	0	0	1	17	17	7	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

				tru			
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	1	1	0	0	0	0	0
5	0	7	59	32	1	1	0
6	0	8	80	183	68	8	1
7	0	0	1	17	15	7	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Best Subset

(Intercept)	fixed.acidity	volatile.acidity	residual.sugar	free.sulfur.dioxide
141.8781	0.0602	-1.8935	0.0786	0.0035
density	pH	sulphates	alcohol	
-141.9033	0.6495	0.5957	0.2073	

- $MSE = 0.4970$
- $ErrorRate = 0.4653$

	tru						
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	1	2	0	0	0	0	0
5	0	8	59	26	0	0	0
6	0	6	81	186	69	9	1
7	0	0	0	20	15	7	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Lasso

(Intercept)	volatile.acidity	chlorides	pH	alcohol
2.5585	-1.9358	-1.3864	0.2088	0.3095

- $MSE = 0.5358$
- $ErrorRate = 0.4776$

				tru			
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	7	57	31	2	1	0
6	0	9	83	184	67	8	1
7	0	0	0	17	15	7	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

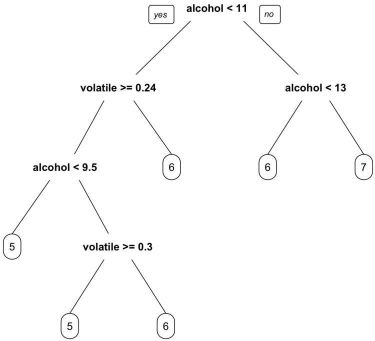
Principle Component Regression

- $MSE = 0.5853$
- $ErrorRate = 0.5122$

	tru						
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	5	27	18	4	1	0
6	0	11	112	209	77	13	1
7	0	0	1	5	3	2	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Decision Tree

$$ErrorRate = 0.4367$$



	tru						
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
5	1	11	82	44	4	0	0
6	0	4	58	178	64	11	1
7	0	1	0	10	16	5	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0

Random Forest

$$ErrorRate = 0.3041$$

	tru						
prd	3	4	5	6	7	8	9
3	0	0	0	0	0	0	0
4	0	3	1	0	0	0	0
5	1	10	97	24	3	1	0
6	0	3	42	194	38	6	1
7	0	0	0	14	42	4	0
8	0	0	0	0	1	5	0
9	0	0	0	0	0	0	0

Conclusion

- The data is too unbalanced
- The random forest performs the best
- It is not bad while error rate $= 0.5$

	MSE	Error Rate
Full	0.7050	0.4650
Non-collinearity	0.7240	0.4690
Final	0.7270	0.4730
Best Subset	0.4970	0.4653
Lasso	0.5358	0.4776
PCR	0.5853	0.5122
Decision Tree	NA	0.4367
Random Forest	NA	0.3041

Reference

1 James, G., Witten, D., Hastie, T. and Tibshirani, R., *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013

2 Johnson, R. and Wichern, D. *Applied Multivariate Statistical Analysis, 6th Edition*, Pearson, London, 2014

3 Kabacoff, R.I., *R in Action Data, analysis and graphics with R*, Manning, New York, 2014

4 Matloff, N., *The Art of R Programming, A Tour of Statistical Software Design*, No Starch Press, San Francisco, 2011

5 Montgomery, D.C., Peck, E.A. and Vining, G.G., *Introduction to Linear Regression Analysis, 4 Edition*, Wiley, 2006