Statistical Learning of Wine Quality

Group 8

Nai-Tan Chang, Yang Fu, Han Liu, Zinan Yue, Wei Dai

Outline

- → Overall Introduction
- → Data Handling
  - ◆ Data description
  - ◆ Group Quality into Class
  - ◆ Split data into training and testing sets
- → Classification
  - ◆ Multinomial Logistic Regression
  - ◆ Linear Discriminant Analysis
  - ◆ Quadratic Discriminant Analysis
- → Decision Tree and Random Forest
- → Support Vector Machines
- → Principal Component Analysis
- → Conclusions

http://discovermagazine.com/2016/sept/20-things-you-didnt-know-about--wine

Introduction

- Wine quality is measured throughout the production of wine, and is critical for both wine producers and consumers.

- Wine quality assessments usually include:
  - **Physicochemical tests**: wine density, pH, residual sugar, etc..
  - **Sensory tests**: rely on human experts to score the quality of wine.

- **Goal:** Predict the subjective wine Quality score using the eleven physicochemical properties of wine.

Data handling

Data Description

Original wine data set: 1,599 for Red Wine, 4,898 for White Wine.

Predictors	Response
11 numeric physicochemical characteristics of wine, including **fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.**	**Quality (default).** The subjective score graded by tasters in sensory tests; an integer scalar ranging from 1 to 10. **Class. The response we create by grouping Quality; letter A, B, and C.**

Data handling

- Create a second response Class by grouping Quality scores
  - Reduce the number of responses while retaining the order
- Save 25% data as the final test dataset

Class	**Class A**			**Class B**		**Class C**	
	1,277			4,974		246	
Quality	**Quality 9**	**Quality 8**	**Quality 7**	**Quality 6**	**Quality 5**	**Quality 4**	**Quality 3**
	5	193	1,079	2,836	2,138	216	30

Classification

- Multinomial logistic regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

Classification - methodology

Step 1: from the training data file, splitting data into training and test data

Training data (50%)	Test data(50%)

Step 2: training, testing, and reporting the training and test errors

Step 3: applying trained model to fit the data in the final test file and evaluating model performance as prediction error rate and model accuracy.

Step 4: applying step 2 and 3 to responses: Quality and Class

Classification - Quality as the response

A

Pred // Truth	01	02	03	04	05	06	07	
3 01								01
4 02		1	1					02
5 03	5	38	329	141	17	1		03
6 04		15	234	498	199	25		04
7 05			5	40	58	15	2	05
8 06								06
9 07				1				07
	01	02	03	04	05	06	07	

B

	01	02	03	04	05	06	07	
3 01	1	4	4	2				01
4 02		3	1	1				02
5 03	4	32	325	137	17	1		03
6 04		15	234	490	183	20		04
7 05			5	50	74	19	2	05
8 06								06
9 07						1		07
	01	02	03	04	05	06	07	

A: multinomial logistic regression

B: LDA

Classification - Class as the response

A. Multinomial logistic regression

B. LDA

C. QDA

Classification methods performance

	Multinomial Logistic regression		LDA		QDA
	Quality	**Class**	**Quality**	**Class**	**Class**
Training error	0.464	0.218	0.468	0.224	0.406
Test error	0.459	0.227	0.469	0.23	0.285
Final test error	0.455	0.212	0.451	0.214	0.251
Model accuracy (%)	54.52%	78.77%	54.95%	78.65%	74.95%

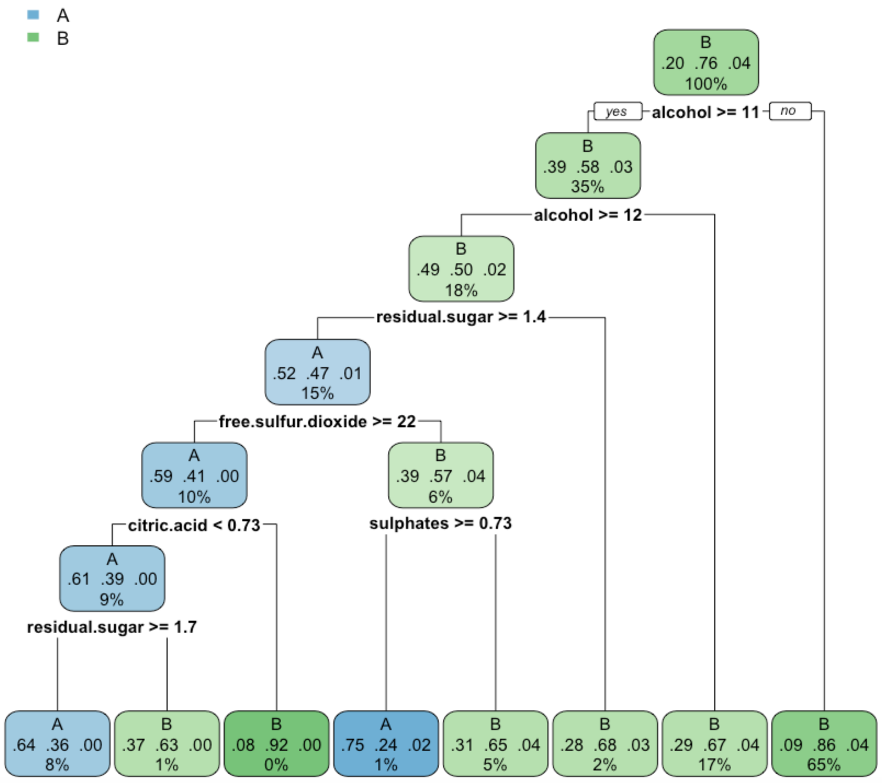
Decision Trees

→ Package: rpart

→ Input: all 11 variables

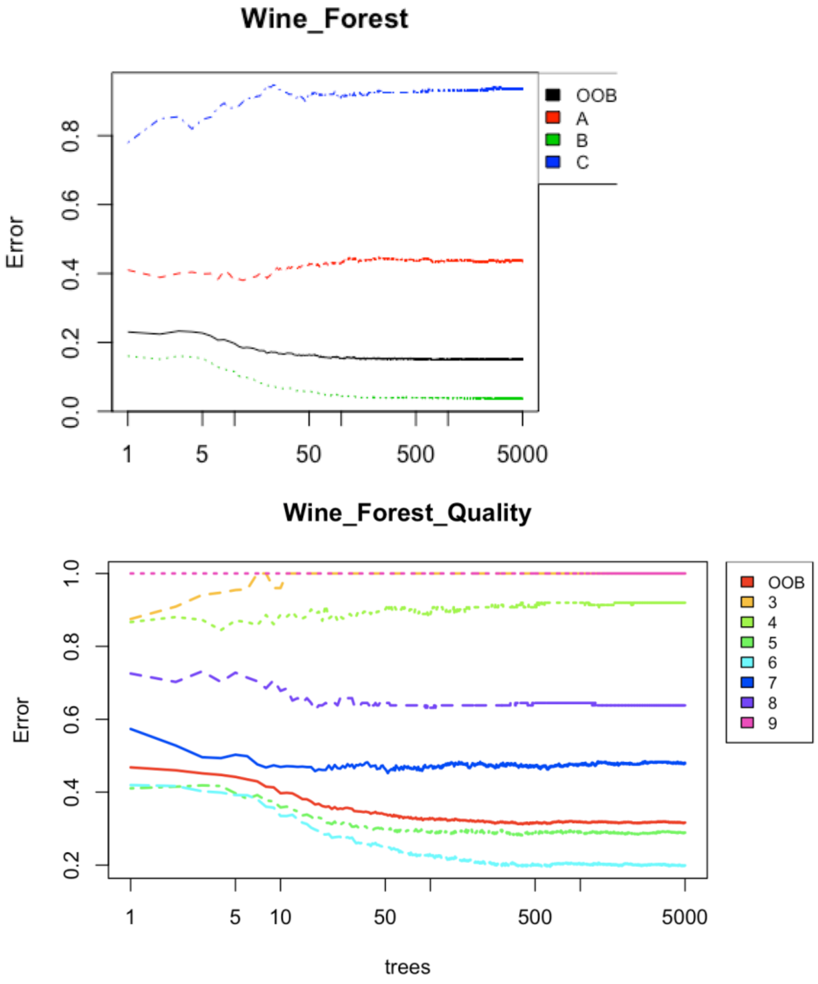
→ Methods: Tree, Prune

→ Predicts: Class, Quality



Random Forests

- ➔ Package: RandomForest
- ➔ Input: all 11 variables
- ➔ Predicts: Class, Quality



Model performance

Class	Decision Tree (pruned)	Random Forest (ntree=5000)
Training error	0.235	
Final test error	0.209	0.143 (OOB = 0.1502)
Model accuracy %	79.14	85.85

Quality	Decision Tree (pruned)	Random Forest (ntree=5000)
Training error	0.472	
Final test error	0.462	0.302
Model accuracy (%)	53.8	69.8

Support vector machines (SVM)

- → Quality prediction and Class prediction
- → Four kernels:
  - ◆ Linear
  - ◆ Polynomial
  - ◆ Radial
  - ◆ Sigmoid
- → Tune by cross validations
- → Evaluate performance using the final test dataset

Model performance on Quality prediction

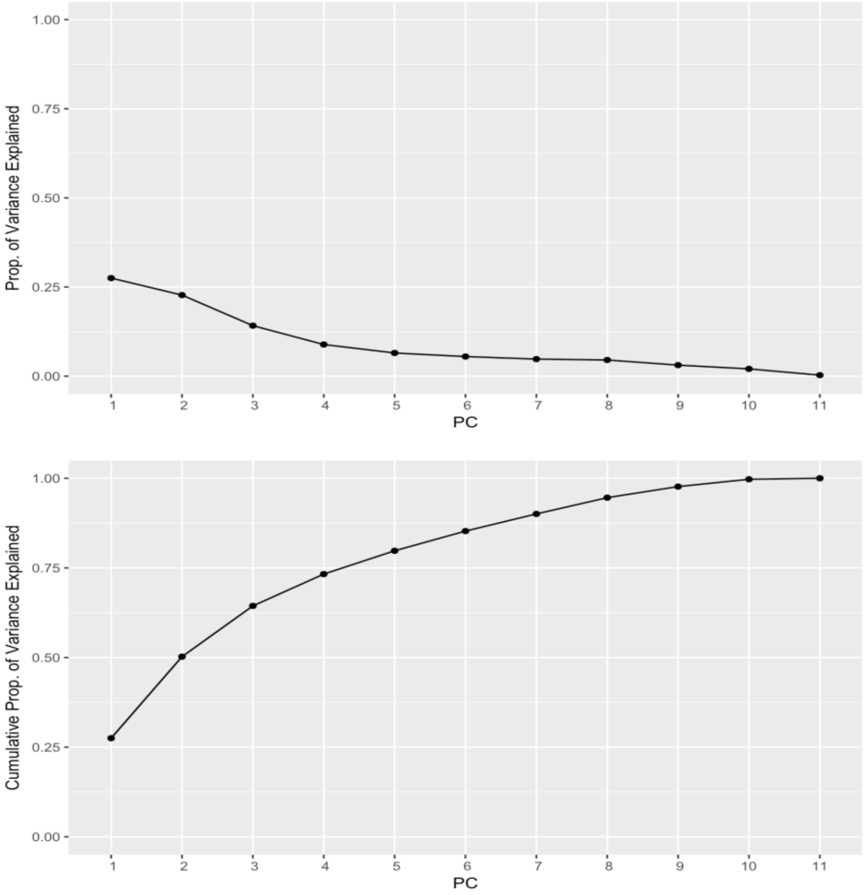
Kernel	Parameters	Training error	Test error	Final test error	Model accuracy (%)
Linear	cost = 0.1	0.471	0.461	0.455	54.5
Polynomial	degree = 5 coef0 = 0.5	0.411	0.479	0.449	55.1
Radial	**gamma = 0.5**	**0.427**	**0.472**	**0.446**	**55.4**
Sigmoid	cost = 0.1	0.514	0.486	0.480	52.0

Model performance on Class Prediction

Kernel	Parameters	Training error	Test error	Final test error	Model accuracy (%)
Linear	cost = 1	0.234	0.237	0.231	76.9
Polynomial	degree = 4 coef0 = 0.5	0.182	0.215	0.202	79.8
Radial	**cost = 10 gamma = 2**	**0.184**	**0.193**	**0.183**	**81.7**
Sigmoid	cost = 0.1	0.239	0.243	0.236	76.4

Process data with PCA...

- ➔ Scale = TRUE
- ➔ Relatively flat scree plot
- ➔ Pick PC4 (70%) and PC7 (90%)



Followed by Logistic regression and SVM

- ➔ Class prediction
- ➔ No big difference between PC4 and PC7
- ➔ SVM radial slightly outperformed
- ➔ PCA did not help...

	SVM radial		**Multinomial logistic regression**	
	PC4	**PC7**	**PC4**	**PC7**
Parameters	cost = 1 gamma = 5	cost = 10 gamma = 5		
Training error	0.209	0.187	0.232	0.220
Test error	0.218	0.193	0.241	0.232
Final test error	0.241	0.232	0.258	0.299
Model accuracy (%)	75.9	76.8	74.2	70.1

Conclusions and future directions

Performance of best models from each category of methods

Model accuracy (%)	Random Forest	SVM with a radial kernel	Multinomial logistic regression
Quality prediction	69.8	55.4	54.5
Class prediction	85.9	81.7	78.8

Conclusions and future directions cont.

- Predicting wine Quality based on physicochemical properties is challenging.
- Wine Class is slightly easier to work with.
- Random Forest, SVM (radial), and multinomial logistic regression are relatively good models.
- Quality 5 and 6 & Class B have the most true positive predictions
- PCA did not help, unfortunately...
- Future studies could be feature engineering.