A Study Of Ensemble Methods In Machine Learning

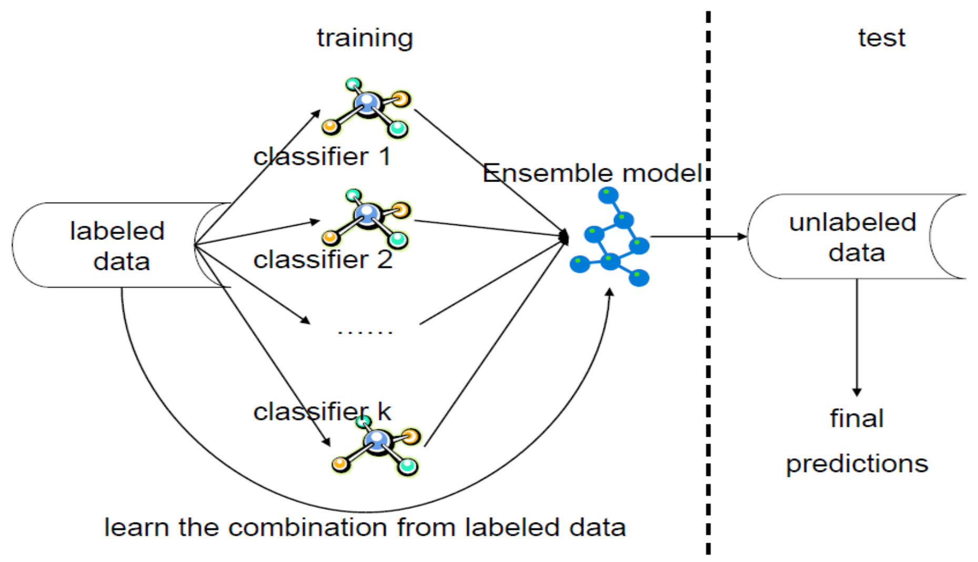
Kwhangho Kim, Jeha Yang

Department of Statistics, Stanford University

Motivation

Power of Ensemble Method

- Ensemble model improves accuracy and robustness over single model methods



- We discuss novel methods of model combination schemes which are very simple and efficient to produce the ensemble models
- We apply our ensemble algorithms to real data and compare the results to the case of single classification methods

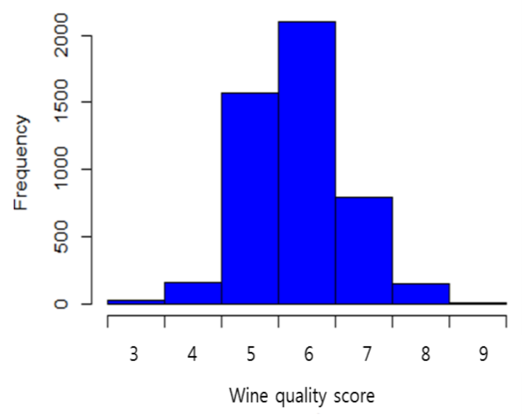
Data Description

Napa Valley Wine Quality Prediction

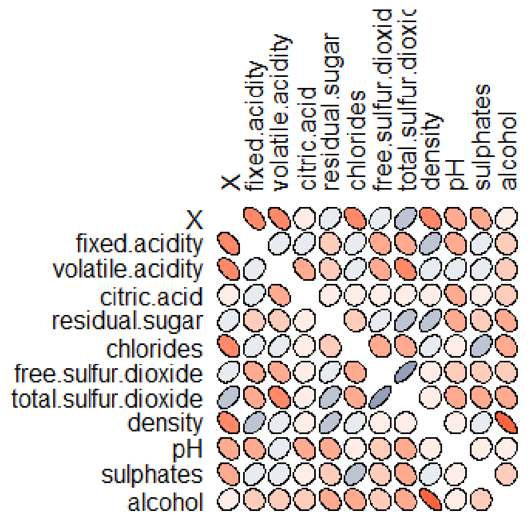
- The wine quality is graded after the wines have been released to the market
- It directly affects the price and demand of the wine
- We want to predict the wine quality score with given soil conditions of the year in which the wine was produced

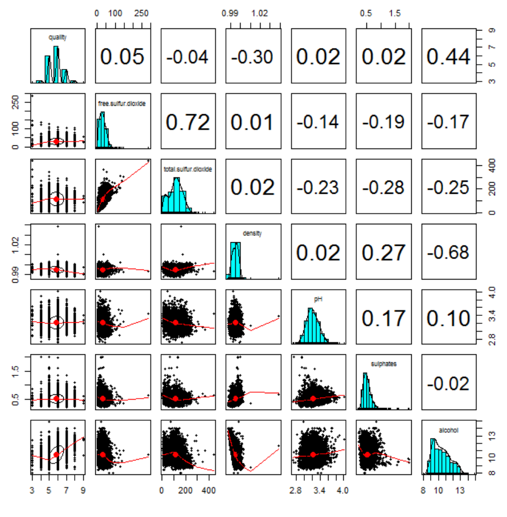
Data Description

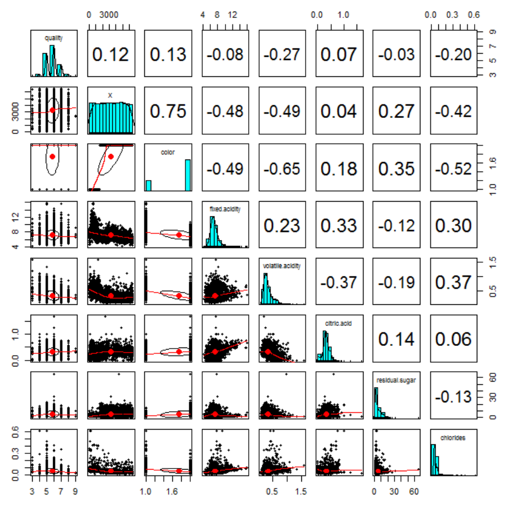
- We have 4800 observations
- There are 13 covariates which contain information about soil conditions
- The wine quality ranges from score 3 to score 9



- Covariates Information







Methods

Prediction Methodologies

(1) L1 Regression (backfitting + WLS) (L1R) (2) Elastic Net (ElaNet) Regression (3) Principle Component Regression (PCR)
(4) Natural Cubic Splines (LCS) with Lasso (5) Local Polynomial Regression (LPR) with variables from the best subset selection
(6) Support Vector Machine (SVM) with Radial Kernel (7) Linear Discriminant Analysis (LDA) (8) Logistic Regression (LOG)

1. Heuristic Ensemble

- Misclassification Errors Per Class

Score	L1R	ElaNet	PCR	NCS	LPR	SVM	LDA	LOG
3	100	100	100	95.5	100	100	**77.3**	100
4	98.8	98.8	98.8	100	98.2	94.5	**87.7**	98.8
5	46.8	47.6	47.9	43.7	44.6	**42.2**	**41.5**	48.7
6	25.0	23.6	23.7	28.1	26.1	**20.3**	31.2	35.1
7	74.3	75.2	75.3	75.5	75.5	**55.0**	69.7	78.3
8	100	100	100	100	100	**73.1**	99.3	91.4
9	100	100	100	100	100	100	100	100

- Combine SVM, LDA, and ElaNet (which is included because it is a regression type classifier so it may increase the model diversification, although being dominated by SVM) in the way that strengthens the strengths and makes up for the weaknesses ; that is, put the weights decreasing in the order of SVM, LDA, and ElaNet.

$$\Rightarrow 0.050\, ElaNet + 0.700\, SVM + 0.250\, LDA$$

2. Ensemble based on l_2 Minimization of CV Errors

- $y^{(k)}$ and $\hat{y}^{(-k,m)}$ denote the realized value of classes for the k^{th} CV fold and the predicted value of classes for the k^{th} CV fold based on the learning method L_m trained on the rest of the data, i.e. trained on the data points taken out for the k^{th} CV fold, respectively.
- $X^{(k)} := (\hat{y}^{(-k,1)}, \hat{y}^{(-k,2)}, \cdots, \hat{y}^{(-k,M)})$
- Weights (by minimizing the CV SSE)

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \mathbb{R}^M}{\operatorname{argmin}} \sum_{k=1}^{K} \|y^{(k)} - X^{(k)}\boldsymbol{\omega}\|_2^2 \;\; s.t. \;\; \mathbf{1}^{\mathbf{T}}\boldsymbol{\omega} = 1, \boldsymbol{\omega} \geq 0$$

$$\Rightarrow 0.092\, NCS + 0.170\, LPR + 0.507\, SVM + 0.231\, LDA$$

3. Simple SSEPC Ensemble

- Sum of Squared Errors Per Class (SSEPC) from the k^{th} CV fold

$$s_{cm}^{(k)} := \sqrt{\sum_{i: y_i^{(-k)} = c} (\hat{y}_i^{(-k,m)} - y_i^{(-k)})^2}, \quad S^{(k)} := (s_{cm}^{(k)})_{1 \leq c \leq C, 1 \leq m \leq M}$$

- Weights (by minimizing an upper bound of the CV SSE)

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \mathbb{R}^M}{\operatorname{argmin}} \sum_{k=1}^{K} \|S^{(k)}\boldsymbol{\omega}\|_2^2 \;\; s.t. \;\; \mathbf{1}^{\mathbf{T}}\boldsymbol{\omega} = 1, \boldsymbol{\omega} \geq 0$$

$$\Rightarrow 0.050\, LPR + 0.950\, SVM$$

4. Forward Stepwise Ensemble (using SSEPC)

- The smaller the ratio of the SSE of a class between a single model and the current model is, the better the single model works in the corresponding class. Therefore, summing up those ratios over all classes could be considered as a measure of how much a single model complements the current ensemble model, treating all classes equally. Furthermore, if those ratios are greater than 1 for all classes, than we don't need to include the corresponding single model to the ensemble.
- Let R be the $C \times M$ matrix of (CV) SSE per class.

(1) Define the best single model as the current model, and update R by removing the corresponding column.
(2) For each column of R, divide each element by the corresponding (CV) SSE per class of the current model, to obtain the matrix of those ratios.
(3) If all of those ratios are larger than 1, stop ; otherwise, find the column with the smallest sum of ratios, include the corresponding single model to the current set of models, and update the model by the least square ensemble of the new set of models and R by removing the corresponding column.
(4) Repeat 2 and 3 until convergence (in terms of the supremum of the change in weight over all single models).

$$\Rightarrow 0.107\, L1 + 0.520\, SVM + 0.373\, LDA$$

Results

We divide original data into 4500 training samples and 300 validation samples

CV RMSE for single classifier

L1R	ElaNet	PCR	NCS	LPR	SVM	LDA	LOG
0.731	0.731	0.769	0.792	0.724	0.717	0.726	1.501

CV RMSE for Ensemble of Classifiers

Heuristic	L2CV	SSEPC	FWD SSEPC
0.679	0.669	0.709	0.673

RMSE for single classifier

L1R	ElaNet	PCR	NCS	LPR	SVM	LDA	LOG
0.802	0.792	0.856	0.936	0.806	0.723	0.781	1.343

RMSE for Ensemble of Classifiers

Heuristic	L2CV	SSEPC	FWD SSEPC
0.683	0.687	0.718	0.687

1. All four ensemble methods perform better than any of single classifiers in terms of RMSE
2. Ensemble methods show their strengths particularly when they combine the complementary predictive powers of multiple models
3. It turns out that the background ideas of FWD SSEPC worked, in that LDA was picked following SVM

Conclusion

- We propose novel simple and efficient ensemble algorithms to systematically determine weights to optimally combine multiple models
- Base models are combined by learning from labeled data or by their consensus
- With the wine quality data application, we confirm that combining independent, diversified models improves both accuracy and predictive power

Future Work

- To improve FWD SSEPC, we can try to find a better measure for the third step as a function of the ratios
- One can examine whether simple SSEPC always guarantees sparse ensemble models
- We can compare L2CV with FWD SSEPC in terms of their performance