Predicting quality of wine based on chemical attributes

Amelia Lemionet, Yi Liu, Zhenxiang Zhou

Introduction

- **Dataset consists of eleven potential predictors and one response variable: "wine quality".**
- **We propose an additive logistic regression method and compare its performance with LR and kNN.**
- **Overall work includes:**
  - Exploratory analysis.
  - Theoretical foundations of additive logistic regression.
  - Implementation, test and comparison of the three methods.
  - Main conclusions and open discussion to other issues of interest.

Data

Prediction is based on 11 explanatory variables

- **Analysis focused in a Portuguese white wine database consisting of 4,898 observations.**
- **Prediction based on 11 chemical attributes:**
  - Fixed acidity
  - Volatile acidity
  - Citric acid
  - Residual sugar
  - Chlorides
  - Free sulfuric dioxide
  - Total sulfuric dioxide
  - Density
  - pH
  - Sulfates
  - Alcohol

Wine quality concentrated on "average" wines

Very low quality and very high quality wines are under-represented

Summary of variables

Feature	Min	1st. Q	Med	3rd. Q	Max	Mean
fixed.acidity	3.80	6.30	6.80	7.30	14.20	6.86
volatile.acidity	0.08	0.21	0.26	0.32	1.10	0.28
citric.acid	0.00	0.27	0.32	0.39	1.66	0.33
residual.sugar	0.60	1.70	5.20	9.90	65.80	6.39
chlorides	0.01	0.04	0.04	0.05	0.35	0.05
free.sulfur.dio	2.00	23.00	34.00	46.00	289.00	35.31
total.sulfur.dio	9.00	108.00	134.00	167.00	440.00	138.40
density	0.99	0.99	0.99	1.00	1.04	0.99
pH	2.72	3.09	3.18	3.28	3.82	3.19
sulphates	0.22	0.41	0.47	0.55	1.08	0.49
alcohol	8.00	9.50	10.40	11.40	14.20	10.51

Weighted linear regression

More weight was added to points with low density:

- W_{1i}: if test point is close to training point, point will gain more weight.
- W_{2i}: add more weight to the high quality scores.
- W_{3i}: add more weight to the low frequency quality scores.

Weights

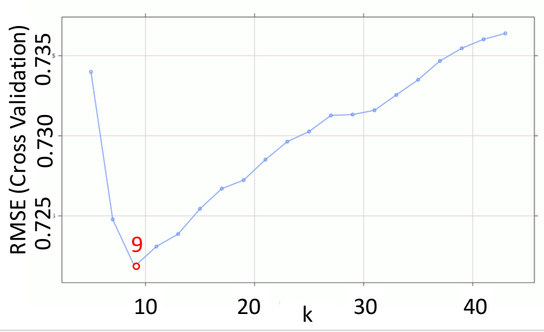
$$W_{1i} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

$$W_{2i} = \frac{\max(Y_i)}{Y_i}$$

$$W_{3i} = \frac{Mean(frequency of\ Y)}{frequency of\ Y_i}$$

K-nearest neighbors

- To make use of the ordinal structure of the data we used the mean of the kNN responses.
- CV to select the optimal value for based on RMSE.
  - 10 fold, 3 times



Additive Logistic Regression (ALR)

Ordinal variables $y \in \{1,2,3,\dots,n\}$

Lemma:

$$E(y \mid X) = \sum_{i=1} P(y \geq i \mid X)$$

Proof:

$$E(y \mid X) = \sum_{i=1}^{n} i \times P(y = i \mid X)$$

$$E(y \mid X) = \sum_{i=1}^{n} \sum_{j=1}^{i} P(y = j \mid X)$$

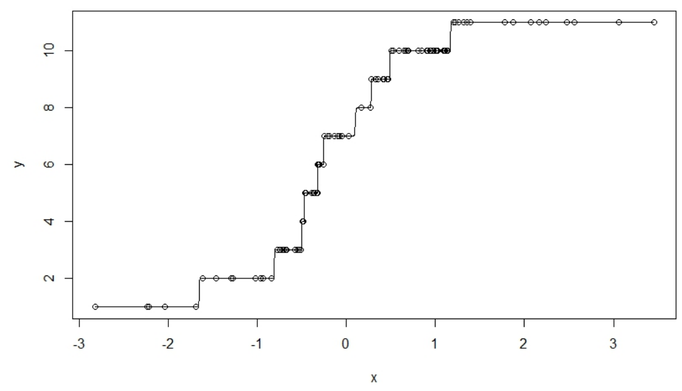
$$= \sum_{i=1}^{n} P(y \geq i \mid X)$$

$\Longrightarrow$ We can estimate $P(y \geq i \mid X)$ via a logistic regression method.

Algorithm 1 Additive Logistic Regression Algorithm

1: **for** j = 0 to n **do**
2: Divide the Training data into $y^{(i)} \leq j$ and $y^{(i)} > j$
3: Estimate $P(y^{(i)} \leq j \mid x^{(i)})$ using logistic regression
4: $\phi_j^{(i)} = \hat{P}(y^{(i)} \leq j \mid x^{(i)})$
5: **end for**
6: $\hat{y}^{(i)} = \sum_j^m \phi_j^{(i)}$

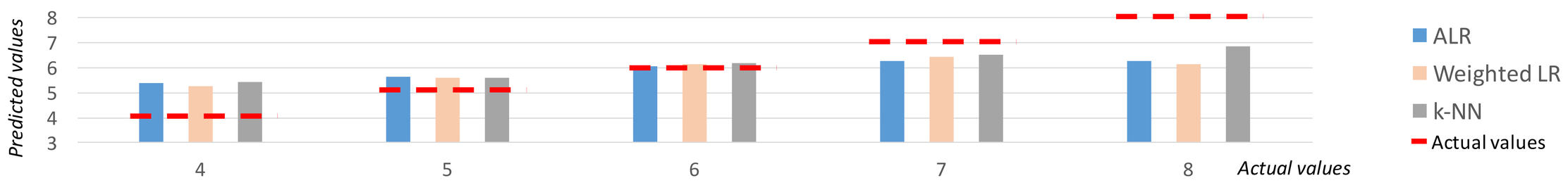
Illustrative example with $x \in \mathbb{R}$



Results

Predictions are centered around average values for all three methods

ALR has the smallest error rate



	ALR	Weighted LR	K-NN
% Error	44.65%	47.55%	47.33%
% overestimate	26.39%	31.63%	32.52%
% underestimate	18.15%	15.92%	14.70%
Additional advantages	• High interpretability in terms of the effect of each variable	• Capturing more information about predictors to improve prediction	• Mean considers ordinal nature of response variable

Conclusions

Although all models yielded relatively high error rates and most of the predicted values are in a compact range, some positive takeaways are to be considered:

- **Additive logistic regression proved to be an effective prediction method when compared to kNN and WLR:**
  - Lowest error rate
  - Inexpensive computation-wise
  - High interpretability regarding the effect of each variable in the model
- **kNN as well as weighted linear regression offer more freedom in model design since they both have tuning parameters that can be customized depending on the problem.**
  - Number of neighbors for kNN and weights for WLR

Potential improvements

- **Reduce the variance through bagging or an adapted version of random forest**
  - By choosing the different variables each time we estimate $P(y \geq i \mid X)$.
- **Look for ways to separate the response values so that the predicted values are not congested together.**
- **Consider possible interactions between the variables.**

References

Winemaker's Academy, understanding wine acidity, http://winemakersacademy.com/understanding-wine-acidity/
Max Kuhn, Classification and Regression Training, Package 'caret', https://cran.r-project.org/web/packages/caret/caret.pdf
Daria Alekseeva, Red and White Wine Quality, https://rpubs.com/Daria/57835
Steven Grubbs, Jargon: What is Residual Sugar?, http://drinks.seriouseats.com/2013/04/wine-jargon-what-is-residual-sugar-riesling-fermentation-steven-grubbs.html