Linear Model Regularization

Shashank Hebbar, Dhiraj Gharana, Lili Zhang
Department of Statistics and Analytical Sciences, Kennesaw State University

Kennesaw State University
College of Science and Mathematics
Department of Statistics and Analytical Sciences

Introduction

In the era of big data, there have been a lot of voices that more data and simple algorithms beat complex algorithms. [1] A family of popular simple analytical models are linear models. Because of its high interpretability, effectiveness, and simplicity, they have been widely used. But the standard linear regression models have restrict assumptions that may be violated. Linear models with penalized term (i.e. ridge regression) can yield better prediction accuracy.

Objectives

1. Explore linear models with reguliarization on the selected dataset.
2. Make recommendations for the practice.

Theory of Linear Model Regularization

As known, the standard linear regression is fitted by minimizing residual sum of squares (RSS). But when fitting ridge regression, a penalty is added in the objective.[2]

Linear regression:

$$\min RSS$$

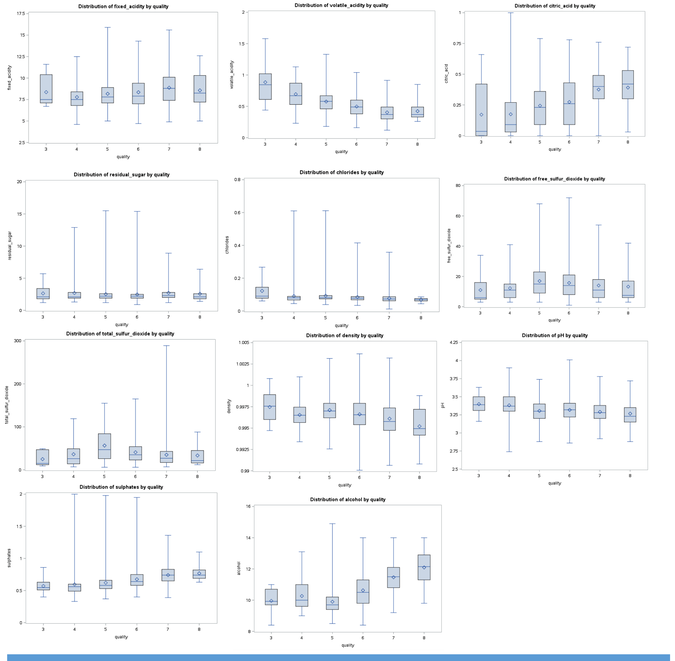
Ridge regression:

$$\min RSS + \lambda * (sum\ of\ square\ of\ coeffcients)$$
$$where\ \lambda\ is\ nonnegative\ constant$$

Why?

- Force coefficients of some features to zero where the features don't make contributions for predict the target.

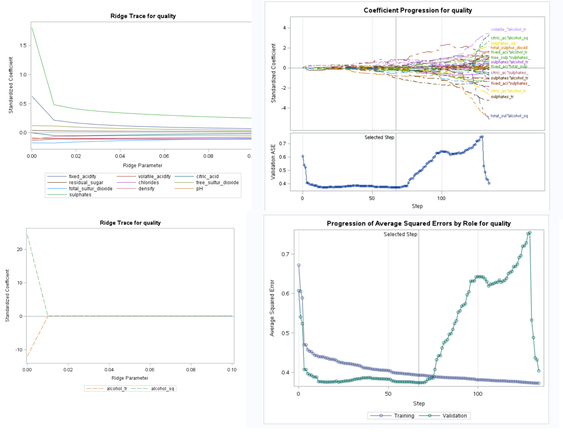
Data Exploratory Analysis



Model Construction

The order of predictor variables in the model:
fixed_acidity(3) volatile_acidity(2) citric_acid(3) residual_sugar(1) chlorides(1) free_sulfur_dioxide(2) total_sulfur_dioxide(2) density(1) pH(1) sulphates(3) alcohol(3)

Ridge Regression Analysis



Discussions

As shown in above figures, the coefficients shrinks as the parameter λ becomes larger. The less useful a coefficient is, the faster it shrinks.

The practice is to set a set of λ values, run the model on each λ, and find the best λ that gives best performance result.

Cross validation is used to select the optimal tuning parameter for lasso. The criteria used is the average square error of validation data sets.

The split ratio was (70-30). The RMSE for the best model is 0.637

Relevant Code

```
proc sort data = WORK.WINEQUALITY_RED;
   by quality;
run;

proc boxplot data=WORK.WINEQUALITY_RED;
   plot fixed_acidity*quality ;
run;

proc boxplot data=WORK.WINEQUALITY_RED;
   plot volatile_acidity*quality ;
run;

proc boxplot data=WORK.WINEQUALITY_RED;
   plot citric_acid*quality ;
run;
```

```
DATA WORK.TRANSFORMED_DATA;
   SET WORK.WINEQUALITY_RED;
   fixed_acidity_tr = fixed_acidity**3;
   fixed_acidity_sq = fixed_acidity**2;
   volatile_acidity_sq = volatile_acidity**2;
   citric_acid_tr = citric_acid**3;
   citric_acid_sq = citric_acid**2;
   free_sulfur_dioxide_sq = free_sulfur_dioxide**2;
   total_sulfur_dioxide_sq = total_sulfur_dioxide**2;
   sulphates_tr = sulphates**3;
   sulphates_sq = sulphates**2;
   alcohol_tr = alcohol**3;
   alcohol_sq = alcohol**2;
   RUN;
```

```
proc glmselect data=TRANSFORMED_DATA plots=all;
partition fraction(validate=.3);
model quality = fixed_acidity_tr|fixed_acidity_sq|volatile_acidity_sq|citric_acid_tr|citric_acid_sq|
free_sulphur_dioxide_sq|total_sulphur_dioxide_sq|sulphates_tr|sulphates_sq|alcohol_tr|alcohol_sq @2
/ selection=lasso(stop=none choose=validate);
run;
```

```
ods graphics on;
PROC REG DATA = WORK.TRANSFORMED_DATA PLOTS(ONLY)=RIDGE(unpack VIFaxis=log)
   OUTEST=RIDGE RIDGE=0 to 0.1 by .01;
   MODEL quality = fixed_acidity volatile_acidity citric_acid residual_sugar
                   chlorides   free_sulfur_dioxide  total_sulfur_dioxide density
                   pH sulphates alcohol
                   fixed_acidity_tr fixed_acidity_sq volatile_acidity_sq
                   citric_acid_tr citric_acid_sq free_sulfur_dioxide_sq
                   total_sulfur_dioxide_sq sulphates_tr sulphates_sq
                   alcohol_tr alcohol_sq
                   /
   OUTPUT  OUT=RIDGE_FORCAST PREDICTED=prediction RESIDUAL=residual;
   RUN;
PROC PRINT DATA=RIDGE;
   RUN;
```

Reference

[1] Mayer-Schönberger, Viktor, and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[2] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Vol. 6. New York: springer, 2013.