Learning on Ridge Regression, LASSO and Elastic Net

Ruiqi Yin <ruiqi.yin@wisc.edu> 9071563093

Wei Hao <whao8@wisc.edu> 9078076974

Abstract/Executive Summary:

When performing linear regression or classification, there are some regularization terms to choose. Using what we have learned from Ridge Regression and LASSO in class, this project will introduce a hybrid of Ridge and LASSO, which is called '***Elastic Net***'. Empirical studies have suggested that the Elastic Net can outperform LASSO and Ridge Regression on data with highly correlated features. In other words, it can be more practical than merely using Ridge Regression or LASSO. This project will explore the advantages of Elastic Net by comparing the performance of these regularizations on a leukemia dataset[1]. Through this project, you are expected to learn about:

1. Applying different regularizations on the dataset and analyze the effectiveness of them by comparing their performance (Mean squared error, Weights' sparsity, etc.).
2. Understanding Elastic Net and how it is related to Ridge and LASSO
3. Understanding the advantages of Ridge/ LASSO/ Elastic Net on datasets with different characteristics and to which situations each method can be applied.

Background

There are two aspects that are important when evaluating the quality of a model: accuracy of the prediction on testing data and interpretability of the model[2]. As we learned in class, the accuracy of the model with Ridge and LASSO both outperform the ordinary least square solution. Additionally, models using LASSO have better interpretability, since LASSO tends to filter out irrelevant features and extract the important ones, while Ridge spreads out weights to all features. Nevertheless, there are some drawbacks of using only LASSO. One of them is that LASSO introduces randomness when there is one or more groups of highly correlated features (features that have strong linear relationships). In this case, LASSO would randomly pick out one feature from the highly correlated feature group.

0. Correlation

Correlation is a statistical relationship indicating the linear relationship between two random variables (in this case, it measures the relationship between two features). The correlation coefficient is a number in the range of [-1, 1] which indicates how strongly two random variables are correlated. The random variables can be positively correlated or negatively correlated, in other words, the randomness can fluctuate in an increasing or decreasing direction. The higher the absolute value of the correlation coefficient is, the higher the correlation they have. When the correlation coefficient is close to 0, there is no correlation between these two variables. If you are interested in learning more of how correlation coefficient is determined, here is a good start [3].

1. LASSO

Suppose that there are two highly correlated features (i.e. feature 1 as f1 and feature 2 as f2, where f1 = f2). The solution space of weights for these two features are any lines that satisfy $w_1 + w_2 = k \in R$, and the boundary of the violet region is the solution for LASSO (Figure 1.). As we can see, there is a overlapping between LASSO and weights solution space, and LASSO would randomly pick any point on the overlapping region. In real life, a subtle turbulence of the solution space caused by noise in the data set will let LASSO pick one feature and ignore the other which makes the solution highly unstable. This situation can hinder researchers from extracting the features which reflect the causation of other correlated features or analyzing a group of highly correlated features[4].

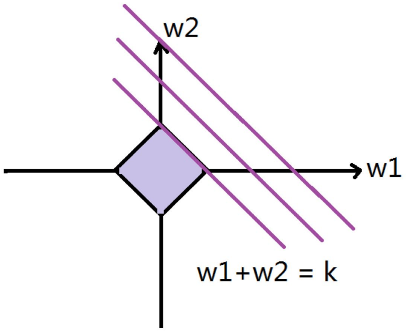
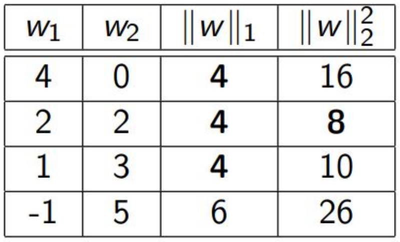


Figure 1.



w_1	w_2	$\|w\|_1$	$\|w\|_2^2$
4	0	**4**	16
2	2	**4**	**8**
1	3	**4**	10
-1	5	6	26

Figure 2.

2. LASSO compared with Ridge Regression

On the other hand, the solution of Ridge Regression doesn't have the uncertainty when some features are highly correlated. Continuing with the previous example, let $w_1 + w_2 = 4$ be the solution space. As Figure 2 shows, LASSO would not discriminate as long as the weights have the same sign[4]. However, Ridge will evenly spread out the weight among two features, which means researchers can determine a group of highly correlated features by observing the same weights. Empirically speaking, when noise is introduced, a group of highly correlated features will have roughly the same weights.

3. Elastic Net

Hence, data scientists incorporate both Ridge Regression and LASSO in the 2003 [2], so-called “Elastic Net”. The formula is:

$$\hat{\mathrm{w}} = \underset{w}{\arg\min} \|y - Xw\| + \lambda_1 \|w\| + \lambda_2 \|w\|^2$$

Geometrically, as Elastic Net assigns different emphasis on the two regularization terms based on their lambdas, the solution space becomes an arc (red lines in Figure 3.). In addition to the abilities of LASSO, Elastic Net can also handle any possible subtle turbulence in the solution space by spreading weights within the group of features that are highly correlated because of the introduced Ridge Regression. This ‘grouping effect’ is presented when there are multiple features as well[4]. In Figure 4., there are two groups of 3 highly correlated. LASSO assigns weights and filter features almost arbitrarily, while Elastic Net groups the features with high correlation and assign weights relatively evenly within each group.

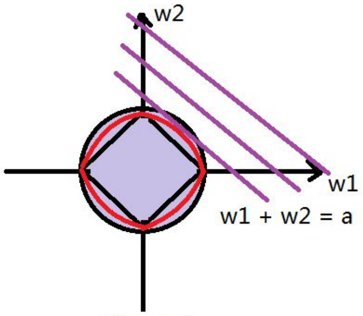


Figure 3.

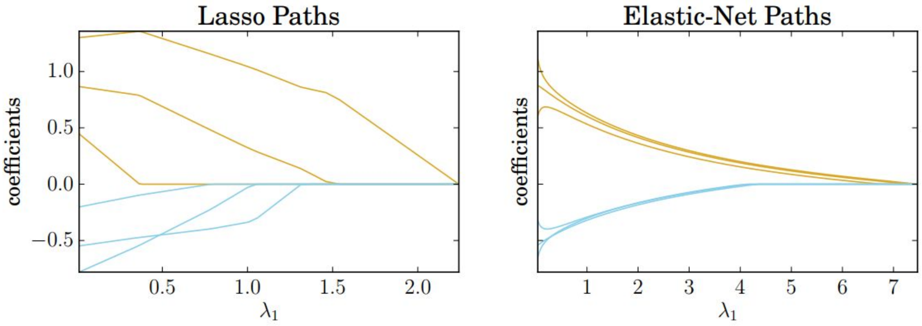


Figure 4.

Warm Up

Matlab has a built-in method for LASSO. This method actually can be used as LASSO, Ridge Regression or Elastic Net by adjusting different parameters:

[B,FitInfo] = lasso(X,y, 'Alpha',__,'Lambda',__)

1. '*Alpha*' represents the relationship between λ_1 and λ_2 . According to Matlab, "the value $Alpha = 1$ represents lasso regression, Alpha close to 0 approaches ridge regression, and other values represent Elastic Net optimization" [5].

For the following questions, we use $Alpha = 10e - 10$ to represent Ridge regression.

2. '*Lambda*' is the regularization coefficient. Both regularization terms will share the same λ , but with different ratios. Thus the formula of Elastic Net is rewritten in this way in Matlab:

$$w = \underset{w}{\operatorname{argmin}} \|y - Xw\| + \alpha\lambda\|w\|_1 + \frac{1-\alpha}{2}\lambda\|w\|_2^2$$

3. B – stores a matrix of weights where each column corresponds to a specific λ

4. $FitInfo$ - stores Fit information of models (e.g. FitInfo.MSE stores a vector of mean squared errors where each element corresponds to a specific λ ; FitInfo.DF stores a

vector of number of nonzero coefficients in B where each element corresponds to a specific λ ;FitInfo.Lambda stores a vector of Lambda parameters in ascending order)

Warm Up Questions

To recall in which situation LASSO and Ridge regression can be applied for better prediction, we developed some warm up questions for you:

Given a randomly generated dataset X∈R(1000X5000) which contains 1000 samples, the true model is y=Xw+ϵ where w∈R(5000x1) and ϵ~N_n(0,I) in R(5000x1). The first 900 samples are for training and the last 100 samples are for testing (skeleton code is provided).

1. If w'=($w_1, ...w_{15}, 0, ...0$) where $w_1 - w_{15} = 1$, and the rest w's are all zeros:
    a. In order to acquire a better model, which one between the two regularizations would you choose? Why?
    b. Implement both regularization with a desirable regularization coefficient. Compare their performance on the test data using mean squared error. (Hint: Very short code is needed if you use [B,FitInfo] = lasso(X,y, 'Alpha',__); And use [M,I] = min() to find the smallest MSE and its corresponding index)

2. Same setting for the data set. If w'=($0, ..., \ 0, \ w_{501}, ...w_{5000}$) where w_{501}~w_{5000} ~N_n(0,1) , and the rest w's are all zeros:
    a. In order to acquire a better model, which one between the two regularizations would you choose? Why?

b. Implement both regularization with a desirable regularization coefficient. Compare their performance on the test data using mean squared error. (The error for this question is huge, don’t panic).

3. If you are given a large dataset with a huge amount of features that assumptions of relations between these features can hardly be made,why using Elastic Net is a better idea compared to using a single regularization? Explain the intuition using the background knowledge and patterns in the warm up activities you observe.

Main Activity

The goal of this activity is to derive a model with Elastic Net on a specific data set and explore how Elastic Net can outperform Ridge and LASSO. The data set we use is stored in Leukemia.mat (see the journal paper posted on Canvas for more details about the dataset)[1]. This data set collects 3571 genes measurements from 72 Leukemia patients and classifies them into two classes (represented using ‘1’ and ‘-1’). For simplicity, we compare the performance of each model based on the mean squared error on the whole dataset.

1. Use LASSO to get a model with a desirable regularization coefficient.

(Hint: use [M,I] = min() to find the smallest MSE and its corresponding index)

   a) What is the MSE for your model?

   b) What is the λ_1 in your model (Notice λ and λ_1 are different)?

   c) Find the number of non-zero weights in your model.

2. Use Ridge regression to get a model with a desirable regularization coefficient.

   a) What is the MSE for your model?

   b) What is the λ_2 in your model (Notice λ and λ_2 are different)?

   c) Find the number of non-zero weights in your model.

3. Compare the performance of Ridge and LASSO. What assumption can you make about the dataset?

4. Express ‘Alpha’ in terms of λ_1 and λ_2 in the Lasso() function. And make a guess of a reasonable ‘Alpha’ if we use Elastic Net.

5. Use Elastic Net to get a model with desirable regularization coefficients (Try different 'Alpha's). Then,

a) What is the MSE for your model? Is it smaller than the MSE's from the Ridge and LASSO models?

b) Find the number of non-zero weights in your model.

c) Notice that when n<p, lasso can only 'extract' at most n features while Elastic Net can 'extract' more than n features, why is this?

d) Comparing with your assumption from question 3, what assumption can you make now about the dataset by looking at the number of non-zero weights in your model?

6. We choose 'Alpha' to be 0.001 for better illustration of the grouping effect. We find the correlation coefficient matrix in Matlab. Each element indicates the coefficient of the ith roe feature and jth column feature. We have observed that two groups of highly correlated features. The first group has features 6, 7, and 8, and the second group has features 9, 10, 11, and 12. Check the weights of these features in your Elastic Net model, can you think of the reasons why are the weights like so?

Reference

[1]Leukemia data:

https://web.stanford.edu/~hastie/CASI_files/DATA/leukemia.html

[2] Regularization and variable selection via the Elastic Net:

https://web.stanford.edu/~hastie/Papers/B67.2%20(2005)%20301-320%20Zou%20&%20Hastie.pdf

[3]Correlation: https://whatis.techtarget.com/definition/correlation

[4]LASSO, Ridge, and Elastic Net

https://davidrosenberg.github.io/mlcourse/Archive/2017/Lectures/3a.elastic-net.pdf

[5] LASSO and Elastic Net function

https://www.mathworks.com/help/stats/lasso.html#d120e453625