Ian Dardik
Heeru Ahuja
CS6140
Exploring Regularization Techniques for Small Data Sets

1 Introduction

In this paper we focus on regularization techniques that improve linear regression modeling on small data sets. Ordinary Least Squares (OLS) is a popular solution for the weights of a linear model, however the method becomes numerically unstable as the number of samples decrease, and has no solution when the number of parameters exceed the sample size. Statistical regularization techniques stabilize our solution and reduce the variance which becomes increasingly important as the number of samples in our data set shrink. Unforunatley, statistical regularization introduces bias into our model so we must choose an appropriate balance. The models we consider in this paper control the degree of regularization through one or two tuning parameters that are completely suppressed at zero, and have more influence on the solution as they increase towards infinity. We can estimate the ideal tuning parameter values by using cross-validation to estimate the MSE for a small universe of tuning parameters, and select the parameters that minimize MSE. Each regularization technique that we consider is formulated as an optimization problem that minimizes an objective function J of the form:

$$\hat{\theta} = \arg\min_{\theta} J(\theta) = \arg\min_{\theta} \sum_{i=1}^{N} L(\theta) + \sum_{i=1}^{P} R(\theta, \lambda)$$

Where $L(\theta)$ is the loss function and $R(\theta, \lambda)$ is the regularization rule that depends on the tuning parameter λ, and sometimes an additional tuning parameter α. We often refer to the regularization rule as the *penalty*. We only consider squared loss in this paper, i.e. $L(\theta) = (Y_i - X_i'\theta)^2$, however we consider and discuss several options for choice of the regularization rule. The regularization techniques that we consider are:

- L1 penalty (LASSO)
- L2 penalty (Ridge)
- $L\frac{1}{2}$
- Elastic Net
- L2 with Custom Soft-Thresholding

Many of the regularization techniques we discuss do not have a closed analytical solution, so we will consider two algorithms for numerically arriving at a solution: Batch Gradient Descent and Pathwise Coordinate Descent. The following sections detail how to use each regularization rule with each algorithm.

For the remainder of this paper we assume that all data X and responses Y are standardized.

2 Gradient Descent

Gradient Descent algorithms work by moving in the direction that the gradient of θ points, with distance scaled by a learning rate parameter α. The algorithm continuously updates it value for the parameter vector θ until convergence. Gradients point towards local optima, so we clearly need to use convex objective functions to arrive at a gobally optimal solution in Gradient Descent. Fortunately, computing the gradient of objective functions listed above is straightforward which makes Gradient Descent fairly simple to implement. Notice that Gradient Descent works by computing the gradient; it does not itself solve an optimization problem per se; since the algorithm is only dependent on the gradient of the objective function, we can parameterize it based on the gradient. Therefore we end up with a core algorithm the form: gradient_descent($\nabla\theta, X, Y$)

3 Pathwise Coordinate Descent

While Gradient Descent operates on the entire parameter vector at once, Pathwise Coordinate Descent (PCD) concentrates on each parameter individually. Pathwise Coordinate Descent considers each parameter θ_j as its own optimization problem[1]:

$$\hat{\theta}_j = \arg\min_{\theta_j} J(\theta)$$

In each iteration of the PCD algorithm, we update each θ_j to be the solution to the individual optimization problem.

3.1 Parameterizing the PCD Algorithm

Pathwise Coordinate Descent is parameterized by the objective function J that it solves for each θ_j. However, for the universe of objective functions we consider in this paper, we can parameterize the PCD algorithm by just two nonnegative parameters we call A and B. For each objective function we consider below, we will show that the solution to optimizing each θ_j takes the form:

$$\theta_j = S\left(\sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k), B\right)/A$$

Where we define[1] $S(t, B) = \text{ sign}(t)(|t| - B)_+$, and $()_+$ is the soft-thresholding operator. It follows that each algorithm reduces to the following parameterized form PATHWISE_CD(A, B). In the following sections we will show how each of the objective functions we consider fit into this parameterization.

3.2 LASSO

LASSO uses the L1 norm for a penalty:

$$\hat{\theta}_j = \arg\min_{\theta_j} \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - \sum_{k=1}^{p} x_{ik}\theta_k\right)^2 + \lambda\sum_{k=1}^{p}|\theta_k|$$

Where $\lambda \geq 0$ is a tuning parameter. Let J be the objective function we want to optimize above. We can solve for θ_j by finding the stationary points of J:

$$\begin{aligned}\frac{\partial J}{\partial \theta_j} &= -\sum_{i=1}^{N} X_{ij}(Y_i - X_{ij}\theta_j - \sum_{k\neq j}^{p} X_{ik}\theta_k) + \lambda \text{ sign}(\theta_j) \\ &= -\sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k) + \sum_{i=1}^{N} X_{ij}^2\theta_j + \lambda \text{ sign}(\theta_j) \\ &= -\sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k) + (N-1)\theta_j + \lambda \text{ sign}(\theta_j)\end{aligned} \tag{1}$$

The last step follows from the fact that our data is standardized, and thus $\sum_{i=1}^{N} X_{ij}^2 = N - 1$.

Set $\frac{\partial J}{\partial \theta_j} = 0$ and, for convenience, let $Q = \sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k)$ and and we get:

$$(N-1)\theta_j + \lambda \text{ sign}(\theta_j) = Q$$

To solve for θ_j notice that when $\theta_j \geq 0$, $\theta_j = \frac{Q-\lambda}{N-1}$ so $Q \geq \lambda$. However when $\theta_j \leq 0$, $\theta_j = \frac{Q+\lambda}{N-1}$ so $Q \leq -\lambda$. Notice the following two properties:

(1) $|Q| \geq \lambda$

(2) $\theta_j \geq 0$ iff $Q \geq 0$ and $\theta_j \leq 0$ iff $Q \leq 0$

Therefore, we can write the solution for θ_j as:

$$\theta_j = \frac{\text{sign}(Q)(|Q| - \lambda)_+}{N-1}$$

Where the soft-thresholding operator guarantees property (1), while multiplying by $\text{sign}(Q)$ makes use of (and guarantees) property (2). Recall from Section 2 that $S(t, B) = \text{sign}(t)(|t| - B)_+$, then our final solution is:

$$\theta_j = S\left(\sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k), \lambda\right)/(N-1)$$

Notice this yields a solution nearly identical to Elements of Statistical Learning[1]; the difference is that we divide by $N-1$ and we believe this is likely a typo in the book.

Notice that the soft-thresholding parameter is λ; this explains why larger λ values causes LASSO to drive parameter values to zero. It is now clear that LASSO performs subset selection due to the choice of regularization penalty *and* algorithm.

Finally, it is clear that the parameterization for LASSO is PATHWISE_CD$(N-1, \lambda)$.

3.3 Elastic Net

The Elastic Net uses the tuning parameter $\alpha \in [0, 1]$ to compromise between Ridge and Lasso penalization. Ridge penalizes the L2 norm, so Elastic Net solves the following optimization problem:

$$\hat{\theta}_j = \arg\min_{\theta_j} \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - \sum_{k=1}^{p} x_{ik}\theta_k\right)^2 + \lambda\sum_{k=1}^{p}\left(\alpha\theta_k^2 + (1-\alpha)|\theta_k|\right)$$

This math is very similar to LASSO above so we have left the details to the appendix. Our final solution is:

$$\theta_j = S\left(\sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k), B\right)/A$$

Where $A = (N-1) + 2\lambda\alpha$, and $B = \lambda(1-\alpha)$. This equation is already in our desired form, and hence the parameterization is PATHWISE_CD$(N - 1 + 2\lambda\alpha, \lambda(1-\alpha))$.

3.4 Ridge

Ridge Regularization has a closed form analytical solution so it does not require an algorithm. Nevertheless, we will discuss Ridge in the context of PCD to realize why the L2 penalty does not perform subset selection while the L1 does. We can simply use our solution for the Elastic Net with $\alpha = 1$ for the Ridge solution. This implies that $A = (N-1) + 2\lambda$ and $B = 0$, yielding:

$$\theta_j = S\left(\sum_{i=1}^{N} X_{ij}(Y_i - \sum_{k\neq j}^{p} X_{ik}\theta_k), 0\right)/(N - 1 + 2\lambda)$$

With parameterization PATHWISE_CD$(N - 1 + 2\lambda, 0)$. Notice that the soft-thresholding parameter is zero; this implies that *no* soft-thresholding will be applied for Ridge, regardless of the parameter λ. This explains why Ridge tends to shrink parameter values but doesn't drive them towards zero.

3.5 Custom Shrinkage Methods

We have found above that LASSO uses λ exclusively for linear shrinkage and soft-thresholding, while Ridge uses λ exclusively for inverse shrinkage. We propose three "custom" ideas for changing and combining the shrinkage types in the context of PCD:

3.5.1 Ridge with Soft-Thresholding

This customization, inspired by Nearest Shrunken Centroids, applies soft-thresholding to the Ridge solution. This yields the parameterization PATHWISE_CD($N - 1 + 2\lambda, \gamma$), where the tuning parameter γ governs the degree of soft-thresholding. We found that cross-validation does not perform well for selecting a good λ, γ pair so we instead consider the case where $\gamma = \lambda$, yielding PATHWISE_CD($N - 1 + 2\lambda, \lambda$).

3.5.2 Ridge with Quadratic Soft-Thresholding

This idea is similar to the previous, except we vary the soft-thresholding quadratically with respect to the tuning parameter λ: PATHWISE_CD($N - 1 + 2\lambda, \lambda^2$).

3.5.3 Ridge with Exponential Soft-Thresholding

This idea is similar to the previous two, except we vary the soft-thresholding exponentially with respect to the tuning parameter λ: PATHWISE_CD($N - 1 + 2\lambda, e^\lambda - 1$).

4 Results

N<p Ortho

Algorithm	Test MSE	Parameters Selected	# Correct	# Incorrect	MSE Diff	# Correct Diff	# Incorrect Diff
LASSO	7.22	1, 4, 8	2	1	0.93	0	3
Ridge w/ST	7.08	1, 4, 8, 11	2	2	1.07	0	2
Ridge w/Quad ST	8.8	1, 8	2	0	-0.65	0	4
Ridge w/Exp ST	9.64	8	1	0	-1.49	1	4
Elastic Net	9.5	8	1	0	-1.35	1	4
LASSO w/BGD	10.36	All	4	34	-2.21	-2	-30
sklearn.linear_model	8.15	1, 4, 5, 8, 9, 11	2	4	0	0	0

N=p Ortho

Algorithm	Test MSE	Parameters Selected	# Correct	# Incorrect	MSE Diff	# Correct Diff	# Incorrect Diff
LASSO	8.2	8	1	0	-1.27	1	3
Ridge w/ST	7.84	1, 8	2	0	-0.91	0	3
Ridge w/Quad ST	7.85	1, 8	2	0	-0.92	0	3
Ridge w/Exp ST	7.99	1, 8	2	0	-1.06	0	3
Elastic Net	7.48	1, 8	2	0	-0.55	0	3
LASSO w/BGD	7.76	All	4	34	-0.83	-2	-31
sklearn.linear_model	6.93	1, 2, 6, 8, 14	2	3	0	0	0

N>p Ortho

Algorithm	Test MSE	Parameters Selected	# Correct	# Incorrect	MSE Diff	# Correct Diff	# Incorrect Diff
LASSO	7.35	1, 7, 8	3	0	0.2	0	3
Ridge w/ST	7.18	1, 7, 8	3	0	0.37	0	3
Ridge w/Quad ST	7.18	1, 7, 8	3	0	0.37	0	3
Ridge w/Exp ST	7.44	1, 7, 8	3	0	0.11	0	3
Elastic Net	7.78	1, 4, 7, 8	3	1	-0.23	0	2
LASSO w/BGD	7.95	All	4	34	-0.4	-1	-31
sklearn.linear_model	7.55	1, 4, 7, 8, 13, 20	3	3	0	0	0

N<p Corr

Algorithm	Test MSE	Parameters Selected	# Correct	# Incorrect	MSE Diff	# Correct Diff	# Incorrect Diff
LASSO	1.42	None	0	0	0	0	0
Ridge w/ST	1.42	None	0	0	0	0	0
Ridge w/Quad ST	1.42	None	0	0	0	0	0
Ridge w/Exp ST	1.42	None	0	0	0	0	0
Elastic Net	1.42	None	0	0	0	0	0
LASSO w/BGD	1.42	All	5	33	0	-5	-33
sklearn.linear_model	1.42	None	0	0	0	0	0

N=p Corr

Algorithm	Test MSE	Parameters Selected	# Correct	# Incorrect	MSE Diff	# Correct Diff	# Incorrect Diff
LASSO	1.53	14, 18, 21, 25	0	4	-0.11	0	-3
Ridge w/ST	1.5	14, 18, 21	0	3	-0.08	0	-2
Ridge w/Quad ST	1.43	14	0	1	-0.01	0	0
Ridge w/Exp ST	1.44	14	0	1	-0.02	0	0
Elastic Net	1.64	2, 7, 14, 18, 21, 24, 25	3	4	-0.22	-3	-3
LASSO w/BGD	1.42	All	5	33	0	-5	-32
sklearn.linear_model	1.42	14	0	1	0	0	0

N>p Corr

Algorithm	Test MSE	Parameters Selected	# Correct	# Incorrect	MSE Diff	# Correct Diff	# Incorrect Diff
LASSO	1.41	3	0	1	0.01	0	-1
Ridge w/ST	1.43	3, 14, 24	1	2	-0.01	-1	-2
Ridge w/Quad ST	1.47	3, 4, 8, 14, 24	1	4	-0.05	-1	-4
Ridge w/Exp ST	1.42	None	0	0	0	0	0
Elastic Net	1.42	3, 24	1	1	0	-1	-1
LASSO w/BGD	1.42	All	5	33	0	-5	-33
sklearn.linear_model	1.42	None	0	0	0	0	0

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, The Elements of Statistical Learning 2nd Edition