

1 Pathwise Coordinate Descent

Pathwise coordinate descent is an effective algorithm for implementing regression algorithms such as LASSO and Elastic Net[1]. Pathwise coordinate descent works by holding tuning parameters (λ, α) constant and optimizing each parameter individually[1]. Given data \mathbf{X} , response variable \mathbf{Y} , and tuning parameters λ and α , we can find the solution to an optimization problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \{O(\theta, X, Y, \lambda, \alpha)\}$$

by considering each parameter j as an individual optimization problem where we hold all other parameters $k \neq j$ constant. We can find $\hat{\theta}_j$ by finding the stationary points of a convex O with respect to θ_j :

$$\hat{\theta}_j = \theta_j \text{ s.t. } \frac{\partial O}{\partial \theta_j} = 0$$

In Pathwise Coordinate Descent we cycle through each parameter and make the appropriate update. After each complete cycle, we check to see if the difference between the previous and the newly updated parameters is smaller than a given tolerance, and terminate if so.

2 Parameterizing the PCD Algorithm

Section 1 shows that Pathwise Coordinate Descent is parameterized by the optimization problem O that it solves for each θ_j . However, for the objective functions we consider in our project, 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] $S(t, B) = \operatorname{sign}(t)(|t| - B)_+$, and $(\cdot)_+$ is the soft-thresholding operator. It follows that each algorithm reduces to the following parameterized form: $\text{pathwise_cd}(A, B, X, Y)$.

3 Standardized Data

We will consider the LASSO and the Elastic Net in the following sections. In both regression problems we will assume that our data is standardized so that each data vector in our design matrix has mean 0 and variance 1. Thus we can omit the intercept term as well in our regression, since the solution will be equal to the mean[1].

4 LASSO

LASSO regression solves the following optimization problem for each θ_j [1]:

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

Where $\lambda \geq 0$ is a tuning parameter. Let O be the formula we are optimizing above. We can solve for θ_j by finding the stationary points of O :

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

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

Set $\frac{\partial O}{\partial \theta_j} = 0$ and, for convenience, let $Q = \sum_{i=1}^N X_{ij}(Y_i - \sum_{k \neq j}^p X_{ik}\theta_k)$ and 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 I'm not sure why they omit this in the book.

Finally, it is clear that we can parameterize `pathwise.cd` for the LASSO if we let $A = N-1$ and $B = \lambda$.

5 Elastic Net

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

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

This problem is very similar to LASSO above so I will write slightly more concisely here. Let O be the formula we are optimizing above. We can solve for θ_j by finding the stationary points of O :

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

Set $\frac{\partial Q}{\partial \theta_j} = 0$ and, for convenience, let $Q = \sum_{i=1}^N X_{ij}(Y_i - \sum_{k \neq j}^p X_{ik}\theta_k)$, $A = (N - 1) + 2\lambda\alpha$, and $B = \lambda(1 - \alpha)$ and we get:

$$A\theta_j + B \operatorname{sign}(\theta_j) = Q$$

Notice that since $A \geq 0$ and $B \geq 0$, we can use the same two case logic from the LASSO section, yielding the solution:

$$\theta_j = \frac{\operatorname{sign}(Q)(|Q| - B)_+}{A}$$

Our final solution is:

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

This is already in our desired form, and it is clear that we can parameterize `pathwise_cd` for the Elastic Net for our choices of $A = (N - 1) + 2\lambda\alpha$ and $B = \lambda(1 - \alpha)$.

References

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