Machine Learning and Computational Statistics, Spring 2017
Homework 2: Lasso Regression

Due: Monday, February 13, 2017, at 10pm (Submit via Gradescope)

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or MathJax via iPython), though if you need to you may scan handwritten work. You may find the minted package convenient for including source code in your LaTeX document.

1 Introduction

In this homework you will investigate regression with ℓ_1 regularization, both implementation techniques and theoretical properties. On the methods side, you'll work on coordinate descent (the "shooting algorithm"), homotopy methods, and [optionally] projected SGD. On the theory side you'll deriver the explicit solution to the coordinate minimizers used in coordinate descent, you'll derive the largest ℓ_1 regularization parameter you'll ever need to try, you'll investigate what happens with ridge and lasso regression when you have two copies of the same feature, and you'll [optionally] work out the details of the classic picture that "explains" why ℓ_1 regularization leads to sparsity.

For the experiments, we constructed a small dataset with known properties. The file `generate_data.py` contains the code used to generate the data. Section 1.1 describes the steps followed to construct the dataset.

1.1 Dataset construction

Start by creating a design matrix for regression with $m = 150$ examples, each of dimension $d = 75$. We will choose a true weight vector $\boldsymbol{\theta}$ that has only a few non-zero components:

1. Let $X \in \mathbf{R}^{m \times d}$ be the "design matrix," where the i'th row of X is $x_i \in \mathbf{R}^d$. Construct a random design matrix X using `numpy.random.rand()` function.

2. Construct a true weight vector $\boldsymbol{\theta} \in \mathbf{R}^{d \times 1}$ as follows: Set the first 10 components of $\boldsymbol{\theta}$ to 10 or -10 arbitrarily, and all the other components to zero.

3. Let $y = (y_1, \ldots, y_m)^T \in \mathbf{R}^{m \times 1}$ be the response. Construct the vector $y = X\boldsymbol{\theta} + \epsilon$, where ϵ is an $m \times 1$ random noise vector where each entry is sampled i.i.d. from a normal distribution with mean 0 and standard deviation 0.1. You can use `numpy.random.randn()` to generate ϵ in Python.

4. Split the dataset by taking the first 80 points for training, the next 20 points for validation, and the last 50 points for testing.

Note that we are not adding an extra feature for the bias in this problem. By construction, the true model does not have a bias term.

2 Ridge Regression

By construction, we know that our dataset admits a sparse solution. Here, we want to evaluate the performance of ridge regression (i.e. ℓ_2-regularized linear regression) on this dataset.

1. Run ridge regression on this dataset. Choose the λ that minimizes the square loss on the validation set. For the chosen λ, examine the model coefficients. Report on how many components with true value 0 have been estimated to be non-zero, and vice-versa (don't worry if they are all nonzero). Now choose a small threshold (say 10^{-3} or smaller), count anything with magnitude smaller than the threshold as zero, and repeat the report. (For running ridge regression, you may either use your code from HW1, or you may use `scipy.optimize.minimize` (see the demo code provided for guidance). For debugging purposes, you are welcome, even encouraged, to compare your results to what you get from `sklearn.linear_model.Ridge`.)

3 Coordinate Descent for Lasso (a.k.a. The Shooting algorithm)

The Lasso optimization problem can be formulated as

$$\hat{w} = \arg\min_{w \in \mathbf{R}^d} \sum_{i=1}^{m} (h_w(x_i) - y_i)^2 + \lambda \|w\|_1,$$

where $h_w(x) = w^T x$, and $\|w\|_1 = \sum_{i=1}^{d} |w_i|$. Since the ℓ_1-regularization term in the objective function is non-differentiable, it's not clear how gradient descent or SGD could be used to solve this optimization problem. (In fact, as we'll see in the next homework on SVMs, we can use "subgradient" methods when the objective function is not differentiable.)

Another approach to solving optimization problems is coordinate descent, in which at each step we optimize over one component of the unknown parameter vector, fixing all other components. The descent path so obtained is a sequence of steps each of which is parallel to a coordinate axis in $\mathbf{R}^d$, hence the name. It turns out that for the Lasso optimization problem, we can find a closed form solution for optimization over a single component fixing all other components. This gives us the following algorithm, known as the **shooting algorithm**:

Algorithm 13.1: Coordinate descent for lasso (aka shooting algorithm)

```
Initialize $\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$;
repeat
    for $j = 1, \ldots, D$ do
        $a_j = 2\sum_{i=1}^{n} x_{ij}^2$;
        $c_j = 2\sum_{i=1}^{n} x_{ij}(y_i - \mathbf{w}^T\mathbf{x}_i + w_j x_{ij})$ ;
        $w_j = \text{soft}(\frac{c_j}{a_j}, \frac{\lambda}{a_j})$;
until converged;
```

(Source: Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.)

The "soft thresholding" function is defined as

$$\text{soft}\,(a, \delta) = \text{sign}(a)\,(|a| - \delta)_+\,.$$

NOTE: Algorithm 13.1 does not account for the case that $a_j = c_j = 0$, which occurs when the jth column of X is identically 0. One can either eliminate the column (as it cannot possibly help the solution), or you can set $w_j = 0$ in that case since it is, as you can easily verify, the coordinate minimizer. Note also that Murphy is suggesting to initialize the optimization with the ridge regession solution. Although theoretically this is not necessary (with exact computations and enough time, coordinate descent will converge for lasso from any starting point), in practice it's helpful to start as close to the solution as we're able.

There are a few tricks that can make selecting the hyperparameter λ easier and faster. First, you can show that for any $\lambda \geq 2\|X^T(y - \bar{y})\|_\infty$, the estimated weight vector $\hat{w}$ is entirely zero, where $\bar{y}$ is the mean of values in the vector y, and $\|\cdot\|_\infty$ is the infinity norm (or supremum norm), which is the maximum over the absolute values of the components of a vector. Thus we need to search for an optimal λ in $[0, \lambda_{\max}]$, where $\lambda_{\max} = 2\|X^T(y - \bar{y})\|_\infty$. (Note: This expression for $\lambda_{\max}$ assumes we have an unregularized bias term in our model. That is, our decision functions are $h_{w,b}(x) = w^T x + b$. For the experiments, you can exclude the bias term, in which case $\lambda_{\max} = 2\|X^T y\|_\infty$.)

The second trick is to use the fact that when λ and λ' are close, the corresponding solutions $\hat{w}(\lambda)$ and $\hat{w}(\lambda')$ are also close. Start with $\lambda = \lambda_{\max}$, for which we know $\hat{w}(\lambda_{\max}) = 0$. You can run the optimization anyway, and initialize the optimization at $w = 0$. Next, λ is reduced (e.g. by a constant factor), and the optimization problem is solved using the previous optimal point as the starting point. This is called **warm starting** the optimization. The technique of computing a set of solutions for a chain of nearby λ's is called a **continuation** or **homotopy method**. The resulting set of parameter values $\hat{w}(\lambda)$ as λ ranges over $[0, \lambda_{\max}\,]$ is known as a **regularization path**.

3.1 Experiments with the Shooting Algorithm

1. Write a function that computes the Lasso solution for a given λ using the shooting algorithm described above. This function should take a starting point for the optimization as a parameter. Run it on the dataset constructed in (1.1), and select the λ that minimizes the square error on the validation set. Report the optimal value of λ found, and the corresponding test error. Plot the validation error vs λ. [Don't use the homotopy method in this part, as we

want to measure the speed improvement of homotopy methods in question 3. Also, no need to vectorize the calculations until question 4, where again we'll compare the speedup. In any case, having two different implementations of the same thing is a good way to check your work.]

2. Analyze the sparsity[1] of your solution, reporting how many components with true value zero have been estimated to be non-zero, and vice-versa.

3. Implement the homotopy method described above. Compare the runtime for computing the full regularization path (for the same set of λ's you tried in the first question above) using the homotopy method compared to the basic shooting algorithm.

4. The algorithm as described above is not ready for a large dataset (at least if it has been implemented in basic Python) because of the implied loop over the dataset (i.e. where we sum over the training set). By using matrix and vector operations, we can eliminate the loops. This is called "vectorization" and can lead to dramatic speedup in languages such as Python, Matlab, and R. Derive matrix expressions for computing a_j and c_j. (Hint: A matlab version of this vectorized method can be found in the assignment zip file.) Implement the matrix expressions and measure the speedup in computing the regularization path.

3.2 Deriving the Coordinate Minimizer for Lasso

This problem is to derive the expressions for the coordinate minimizers used in the Shooting algorithm. This is often derived using subgradients (slide 15), but here we will take a bare hands approach (which is essentially equivalent).

In each step of the shooting algorithm, we would like to find the w_j minimizing

$$\begin{aligned} f(w_j) &= \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda \left|w\right|_1 \\ &= \sum_{i=1}^{n} \left[w_j x_{ij} + \sum_{k \neq j} w_k x_{ik} - y_i \right]^2 + \lambda \left|w_j\right| + \lambda \sum_{k \neq j} \left|w_k\right|, \end{aligned}$$

where we've written x_{ij} for the jth entry of the vector x_i. This function is strictly convex in w_j, and thus it has a unique minimum. The only thing keeping f from being differentiable is the term with $|w_j|$. So f is differentiable everywhere except $w_j = 0$. We'll break this problem into 3 cases: $w_j > 0$, $w_j < 0$, and $w_j = 0$. In the first two cases, we can simply differentiate f w.r.t. w_j to get optimality conditions. For the last case, we'll use the fact that since $f : \mathbf{R} \to \mathbf{R}$ is convex, 0 is a minimizer of f iff

$$\lim_{\varepsilon \downarrow 0} \frac{f(\varepsilon) - f(0)}{\varepsilon} \geq 0 \quad \text{and} \quad \lim_{\varepsilon \downarrow 0} \frac{f(-\varepsilon) - f(0)}{\varepsilon} \geq 0.$$

[1] One might hope that the solution will a sparsity pattern that is similar to the ground truth. Estimators that preserve the sparsity pattern (with enough training data) are said to be **"sparsistent"** (sparse + consistent). Formally, an estimator $\hat{\beta}$ of parameter β is said to be consistent if the estimator $\hat{\beta}$ converges to the true value β in probability as our sample size goes to infinity. Analogously, if we define the support of a vector β as the indices with non-zero components, i.e. $\text{Supp}(\beta) = \{j \mid \beta_j \neq 0\}$, then an estimator $\hat{\beta}$ is said to be sparsistent if as the number of samples becomes large, the support of $\hat{\beta}$ converges to the support of β, or $\lim_{m \to \infty} P[\text{Supp}(\hat{\beta}_m) = \text{Supp}(\beta)] = 1$.

This is a special case of the optimality conditions described in slide 12 here, where now the "direction" v is simply taken to be the scalars 1 and -1, respectively.

1. First let's get a trivial case out of the way. If $x_{ij} = 0$ for $i = 1, \dots, n$, what is the coordinate minimizer w_j? In the remaining questions below, you may assume that $\sum_{i=1}^{n} x_{ij}^2 > 0$.

2. Give an expression for the derivative $f(w_j)$ for $w_j \neq 0$. It will be convenient to write your expression in terms of the following definitions:

$$\begin{aligned} \operatorname{sign}(w_j) &:= \begin{cases} 1 & w_j > 0 \\ 0 & w_j = 0 \\ -1 & w_j < 0 \end{cases} \\ a_j &:= 2\sum_{i=1}^{n} x_{ij}^2 \\ c_j &:= 2\sum_{i=1}^{n} x_{ij}\left(y_i - \sum_{k\neq j} w_k x_{ik}\right). \end{aligned}$$

3. If $w_j > 0$ and minimizes f, show that $w_j = \frac{1}{a_j}(c_j - \lambda)$. Similarly, if $w_j < 0$ and minimizes f, show that $w_j = \frac{1}{a_j}(c_j + \lambda)$. Give conditions on c_j that imply that a minimizer w_j is positive and conditions for which a minimizer w_j is negative.

4. Derive expressions for the two one-sided derivatives at $f(0)$, and show that $c_j \in [-\lambda, \lambda]$ implies that $w_j = 0$ is a minimizer.

5. Putting together the preceding results, we conclude the following:

$$w_j = \begin{cases} \frac{1}{a_j}(c_j - \lambda) & c_j > \lambda \\ 0 & c_j \in [-\lambda, \lambda] \\ \frac{1}{a_j}(c_j + \lambda) & c_j < -\lambda \end{cases}$$

   Show that this is equivalent to the expression given in 3.

4 Lasso Properties

4.1 Deriving $\lambda_{\max}$

In this problem we will derive an expression for $\lambda_{\max}$. For the first three parts, use the Lasso objective function excluding the bias term i.e, $L(w) = \|Xw - y\|_2^2 + \lambda\|w\|_1$. Show that for any $\lambda \geq 2\|X^T y\|_\infty$, the estimated weight vector $\hat{w}$ is entirely zero, where $\|\cdot\|_\infty$ is the infinity norm (or supremum norm), which is the maximum absolute value of any component of the vector.

1. The one-sided directional derivative of $f(x)$ at x in the direction v is defined as:

$$f'(x; v) = \lim_{h \downarrow 0} \frac{f(x + hv) - f(x)}{h}$$

Compute $L'(0; v)$. That is, compute the one-sided directional derivative of $L(w)$ at $w = 0$ in the direction v. [Hint: the result should be in terms of $X, y, \lambda,$ and v.]

2. Since the Lasso objective is convex, for w^* to be a minimizer of $L(w)$ we must have that the directional derivative $L'(w^*; v) \geq 0$ for all v. Starting from the condition $L'(0; v) \geq 0$, rearrange terms to get a lower bounds on λ. [Hint: this should be in terms of $X, y,$ and v.]

3. In the previous problem, we get a different lower bound on λ for each choice of v. Compute the maximum lower bound of λ by maximizing the expression over v. Show that this expression is equivalent to $\lambda_{\max} = 2\|X^T y\|_\infty$.

4. [Optional] Show that for $L(w, b) = \|Xw + b\mathbf{1} - y\|_2^2 + \lambda \|w\|_1$, $\lambda_{\max} = 2\|X^T(y - \bar{y})\|_\infty$where $\bar{y}$ is the mean of values in the vector y, and $\mathbf{1} \in \mathbf{R}^n$ is a column vector of 1's .

4.2 Feature Correlation

In this problem, we will examine and compare the behavior of the Lasso and ridge regression in the case of an exactly repeated feature. That is, consider the design matrix $X \in \mathbf{R}^{m \times d}$, where $X_{\cdot i} = X_{\cdot j}$ for some i and j, where $X_{\cdot i}$ is the i^{th} column of X. We will see that ridge regression divides the weight equally among identical features, while Lasso divides the weight arbitrarily. In an optional part to this problem, we will consider what changes when $X_{\cdot i}$ and $X_{\cdot j}$ are highly correlated (e.g. exactly the same except for some small random noise) rather than exactly the same.

1. Without a loss of generality, assume the first two colums of X are our repeated features. Partition X and θ as follows:

$$X = \begin{pmatrix} x_1 & x_2 & X_r \end{pmatrix}, \theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_r \end{pmatrix}$$

We can write the Lasso objective function as:

$$\begin{aligned} L(\theta) &= \|X\theta - y\|_2^2 + \lambda \|\theta\|_1 \\ &= \|x_1\theta_1 + x_2\theta_2 + X_r\theta_r - y\|_2^2 + \lambda|\theta_1| + \lambda|\theta_2| + \lambda \|\theta_r\|_1 \end{aligned}$$

With repeated features, there will be multiple minimizers of $L(\theta)$. Suppose that

$$\hat{\theta} = \begin{pmatrix} a \\ b \\ r \end{pmatrix}$$

is a minimizer of $L(\theta)$. Give conditions on c and d such that $\left(c, d, r^T\right)^T$ is also a minimizer of $L(\theta)$. [Hint: First show that a and b must have the same sign, or at least one of them is zero. Then, using this result, rewrite the optimization problem to derive a relation between a and b.]

2. Using the same notation as the previous problem, suppose

$$\hat{\theta} = \begin{pmatrix} a \\ b \\ r \end{pmatrix}$$

minimizes the ridge regression objective function. What is the relationship between a and b, and why?

3. [Optional] What do you think would happen with Lasso and ridge when $X_{\cdot i}$ and $X_{\cdot j}$ are highly correlated, but not exactly the same. You may investigate this experimentally or theoretically.

5 [Optional] The Ellipsoids in the ℓ_1/ℓ_2 regularization picture

Recall the famous picture purporting to explain why ℓ_1 regularization leads to sparsity, while ℓ_2 regularization does not. Here's the instance from Hastie et al's *The Elements of Statistical Learning:*

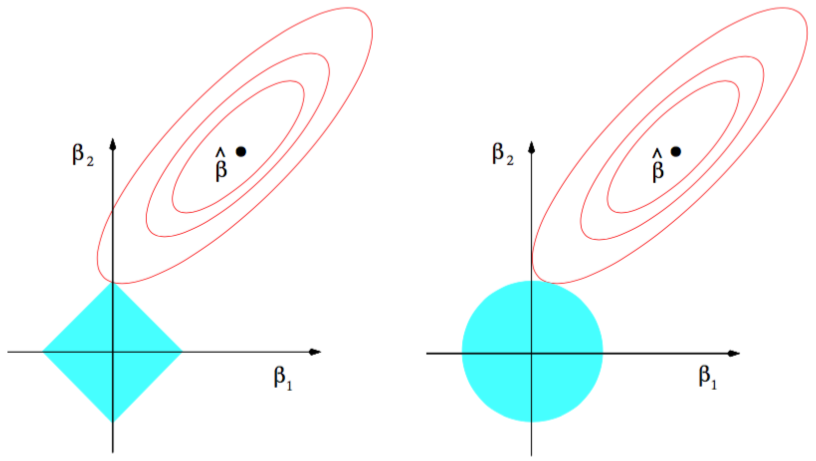

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions $|\beta_1| + |\beta_2| \le t$ and $\beta_1^2 + \beta_2^2 \le t^2$, respectively, while the red ellipses are the contours of the least squares error function.*

(While Hastie et al. use β for the parameters, we'll continue to use w.)

In this problem we'll show that the level sets of the empirical risk are indeed ellipsoids centered at the empirical risk minimizer $\hat{w}$.

Consider linear prediction functions of the form $x \mapsto w^T x$. Then the empirical risk for $f(x) = w^T x$ under the square loss is

$$\begin{aligned} \hat{R}_n(w) &= \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 \\ &= \frac{1}{n} (Xw - y)^T (Xw - y). \end{aligned}$$

1. [Optional] Let $\hat{w} = \left(X^T X\right)^{-1} X^T y$. Show that $\hat{w}$ has empirical risk given by

$$\hat{R}_n(\hat{w}) = \frac{1}{n}\left(-y^T X\hat{w} + y^T y\right)$$

2. [Optional] Show that for any w we have

$$\hat{R}_n(w) = \frac{1}{n}\left(w - \hat{w}\right)^T X^T X\left(w - \hat{w}\right) + \hat{R}_n(\hat{w}).$$

   Note that the RHS (i.e. “right hand side”) has one term that’s quadratic in w and one term that’s independent of w. In particular, the RHS does not have any term that’s linear in w. On the LHS (i.e. “left hand side”), we have $\hat{R}_n(w) = \frac{1}{n}\left(Xw - y\right)^T\left(Xw - y\right)$. After expanding this out, you’ll have terms that are quadratic, linear, and constant in w. Completing the square is the tool for rearranging an expression to get rid of the linear terms. The following “completing the square” identity is easy to verify just by multiplying out the expressions on the RHS:

$$x^T Mx - 2b^T x \quad = \quad \left(x - M^{-1}b\right)^T M(x - M^{-1}b) - b^T M^{-1}b$$

3. [Optional] Using the expression derived for $\hat{R}_n(w)$ in 2, give a very short proof that $\hat{w} = \left(X^T X\right)^{-1} X^T y$ is the empirical risk minimizer. That is:

$$\hat{w} = \arg\min_w \hat{R}_n(w).$$

   Hint: Note that $X^T X$ is positive semidefinite and, by definition, a symmetric matrix M is positive semidefinite iff for all $x \in \mathbf{R}^d$, $x^T Mx \geq 0$.

4. [Optional] Give an expression for the set of w for which the empirical risk exceeds the minimum empirical risk $\hat{R}_n(\hat{w})$ by an amount $c > 0$. If X is full rank, then $X^T X$ is positive definite, and this set is an ellipse – what is its center?

6 [Optional] Projected SGD via Variable Splitting

In this question, we consider another general technique that can be used on the Lasso problem. We first use the variable splitting method to transform the Lasso problem to a smooth problem with linear inequality constraints, and then we can apply a variant of SGD.

Representing the unknown vector θ as a difference of two non-negative vectors θ^+ and θ^-, the ℓ_1-norm of θ is given by $\sum_{i=1}^{d} \theta_i^+ + \sum_{i=1}^{d} \theta_i^-$. Thus, the optimization problem can be written as

$$(\hat{\theta}^+, \hat{\theta}^-) = \underset{\theta^+, \theta^- \in \mathbf{R}^d}{\arg\min} \sum_{i=1}^{m} (h_{\theta^+,\theta^-}(x_i) - y_i)^2 + \lambda \sum_{i=1}^{d} \theta_i^+ + \lambda \sum_{i=1}^{d} \theta_i^-$$
$$\text{such that } \theta^+ \geq 0 \text{ and } \theta^- \geq 0,$$

where $h_{\theta^+,\theta^-}(x) = (\theta^+ - \theta^-)^T x$. The original parameter θ can then be estimated as $\hat{\theta} = (\hat{\theta}^+ - \hat{\theta}^-)$.

This is a convex optimization problem with a differentiable objective and linear inequality constraints. We can approach this problem using projected stochastic gradient descent, as discussed in lecture. Here, after taking our stochastic gradient step, we project the result back into the feasible set by setting any negative components of θ^+ and θ^- to zero.

1. [Optional] Implement projected SGD to solve the above optimization problem for the same λ's as used with the shooting algorithm. Since the two optimization algorithms should find essentially the same solutions, you can check the algorithms against each other. Report the differences in validation loss for each λ between the two optimization methods. (You can make a table or plot the differences.)

2. [Optional] Choose the λ that gives the best performance on the validation set. Describe the solution $\hat{w}$ in term of its sparsity. How does the sparsity compare to the solution from the shooting algorithm?