The Comparison of Several Regression Algorithms

PINGFAN TANG
University of Utah, Salt Lake City, UT, USA

1 Problem and Data

In this project, we compared several regression algorithms by using them to explore "Communities and Crime Data Set" (http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime). We tried four basic algorithms: linear least squares regression, ridge regression, principal component analysis regression and Lasso regression, and used Lasso regression to find some important attributes which have a close relationship to the crime rate in a community. Moreover, we made a little extension to Theil-Sen estimator, and use it to carry out regression in the two dimensional space.

There are 1994 instances in the data set, and we use the first 1690 instances to build the model and use the remaining 304 instances to test the model. There are 122 predictive attributes in the original data, and we delete 23 attributes in which the data have missing values. So, in the processed data there are 99 predictive attributes, and one goal attribute which is the "total number of violent crimes per 100000 population".

Suppose the 1994 instances are in a set $X = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_{1994}, y_{1994})\}$ where $\mathbf{x}_i$ is the vector containing the value of 99 predictive attributes of $i-th$ instance, and y_i is the value of the goal attribute of $i-th$ instance. We define

$$\begin{aligned} X_1 &= \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_{338}, y_{338})\}, \\ X_2 &= \{(\mathbf{x}_{339}, y_{339}), \cdots, (\mathbf{x}_{676}, y_{676})\}, \\ X_3 &= \{(\mathbf{x}_{677}, y_{677}), \cdots, (\mathbf{x}_{1014}, y_{1014})\}, \\ X_4 &= \{(\mathbf{x}_{1015}, y_{1015}), \cdots, (\mathbf{x}_{1352}, y_{1352})\}, \\ X_5 &= \{(\mathbf{x}_{1353}, y_{1353}), \cdots, (\mathbf{x}_{1690}, y_{1690})\}, \\ X_6 &= \{(\mathbf{x}_{1691}, y_{1691}), \cdots, (\mathbf{x}_{1994}, y_{1994})\}. \end{aligned} \tag{1}$$

where $|X_1| = |X_2| = |X_3| = |X_4| = |X_5| = 338$ and $|X_6| = 304$.

2 Comparison of Four Regression Algorithms

For ridge regression, the goal is to minimize

$$\sum_{\mathbf{x}_i} (\mathbf{x}_i^T \mathbf{a} - y_i)^2 + s\|\mathbf{a}\|_2^2. \tag{2}$$

For Lasso regression, the goal is to minimize

$$\sum_{\mathbf{x}_i} (\mathbf{x}_i^T \mathbf{a} - y_i)^2 \quad \text{such that } \|\mathbf{a}\|_1 \le t. \tag{3}$$

We use cross validation to choose parameters s and t in (2) and (3). For fixed s or t, we use $X - X_6 - X_i$ ($i \in \{1, 2, 3, 4, 5\}$) to carry out the regression, and then use X_i to compute the Error_square e_i. For example, we use $X - X_6 - X_1$ to do regression, and then use the result and X_1 to compute the estimate value of goal attribute $\{\hat{y}_1, \cdots, \hat{y}_{338}\}$, and $e_1 = \sum_{i=1}^{338} (\hat{y}_i - y_i)^2$. Thus, for each s or t we can obtain an error $e = \frac{1}{5}\sum_{i=1}^{5} e_i$. We choose s and t that can minimize this error.

For ridge regression, the relationship between e and s is shown in Figure 1 (a), and we choose regularization parameter $s = 1.1$, because when $s = 1.1$ the error e achieves its minimum value. For Lasso regression, the relationship between e and t is shown in Figure 1 (b), and we choose the parameter $t = 5.1$, because when $t = 5.1$ the error e achieves its minimum value. We will also consider the case $t = 1.5$, because when $t = 1.5$ the error is close to the minimum value and the solution is sparse, so we can know which attributes are important to the goal attribute.

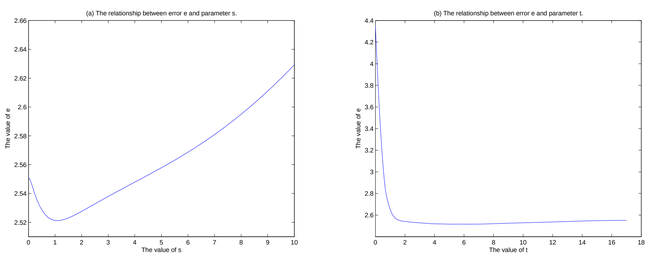

Figure 1. The relationship between error e and regularization parameter.

After fixing the parameter s and t, we use $X - X_6$ as the data to carry out regression and use X_6 to compute the error for different algorithms. Suppose for those instances in X_6, the estimated value of goal attribute is $\{\hat{y}_{1691}, \hat{y}_{1692}, \cdots, \hat{y}_{1994}\}$. We define the following three errors to compare the performance of different algorithms.

$$\text{Error\_max} = \max_{1691 \le i \le 1994} \{|y_i - \hat{y}_i|\}, \quad \text{Error\_square} = \Big(\sum_{i=1691}^{1994} (y_i - \hat{y}_i)^2\Big)^{\frac{1}{2}}, \quad \text{Error\_mean} = \frac{1}{304}\sum_{i=1691}^{1994} |y_i - \hat{y}_i|.$$

The running result of three algorithms is shown in Table 1.

Table 1. The error of different algorithms.

	Error_max	Error_square	Error_mean
Least Square Regression	0.5986	2.2880	0.0923
Ridge Regression (s=1.1)	0.5869	2.2373	0.0899
Lasso Regression (t=5.1)	0.5698	2.2358	0.0896
Lasso Regression (t=1.5)	0.5372	2.2055	0.0884
PCA Regression	17.7278	80.1834	3.6403

From Table 1, we can see Lasso regression works better than ridge regression and least square regression, and PCA regression is not applicable for this question, because it tries to minimizes the perpendicular distances from the data to the fitted model (a hyperplane), which means the error in the goal attribute may be very large.

3 Something Interesting Found in Data

In Table 1, when $t = 1.5$ the Lasso regression brings us the best result. Suppose when $t = 1.5$ the model built by Lasso regression is $y = \mathbf{x}^T\mathbf{a} + b$, then only 25 elements in vector a are nonzero. Obviously, these nonzero elements are the coefficients of those important attributes. We choose 6 attributes with the largest absolute coefficients, and show them in Table 2. When the coefficient is greater than zero, it means this attribute is positively correlated to the crime rate in a community; when the coefficient is less than zero, it means this attribute is negatively correlated to the crime rate in a community. Therefore, Table 2 indicates that the stability and harmony of family play an important role in reducing the crime rate of a community.

Table 2. Important Attributes.

Attribute	Coefficient
percentage of kids in family housing with two parents	-0.2095
percentage of kids born to never married	0.1870
percentage of population that is African American	0.1753
percentage of males who are divorced	0.1507
percent of persons in dense housing (more than 1 person per room)	0.1324
number of vacant households	0.1019

4 Theil-Sen Estimator in Two Dimensional Space

Given a data set $S = \{p_i \mid p_i = (x_i^1, x_i^2, y_i) \in \mathbb{R}^3, i = 1, 2, \cdots, n\}$, we try to find a plane $y = a_1x^1 + a_2x^2 + b$ to fit these data. Like Theil-Sen estimator in $\mathbb{R}^1$, for any set of three points $\{p_i, p_j, p_k\} \subset S$, if p_i, p_j, p_k are not on the same line, we compute the normal vector of the plane passing these three points, and then take the L_1 median of these normal vectors, and then compute the offset b. The detail of this process is given in the following algorithm.

Input: $S = \{p_i \mid p_i = (x_i^1, x_i^2, y_i) \in \mathbb{R}^3, i = 1, 2, \cdots, n\}$
Set $A = \emptyset$, $\widetilde{S} = \{\{p_i, p_j, p_k\} \subset S \mid p_i, p_j, p_k$ are not on the same line$\}$
for each $\{p_i, p_j, p_k\} \in \widetilde{S}$ **do**
  $\gamma_1 = (x_i^2 - x_k^2)(y_i - y_j) - (x_i^2 - x_j^2)(y_i - y_k)$
  $\gamma_2 = (x_i^1 - x_j^1)(y_i - y_k) - (x_i^1 - x_k^1)(y_i - y_j)$
  $\gamma_3 = (x_i^1 - x_k^1)(x_i^2 - x_j^2) - (x_i^1 - x_j^1)(x_i^2 - x_k^2)$
  if $\gamma_3 \neq 0$ **then**
    $a_1 = \frac{\gamma_1}{\gamma_3}$, $a_2 = \frac{\gamma_2}{\gamma_3}$
    $A = A \cup \{(a_1, a_2)\}$
$(a_1, a_2) = L_1$ median of A
$b = \text{median}\{y_i + a_1x_i^1 + a_2x_i^2 | (x_i^1, x_i^2, y_i) \in S\}$
return $y = -a_1x^1 - a_2x^2 + b$

To test Theil-Sen estimator in the two dimensional space, we randomly generate a set of 150 points $D = \{(x_i^1, x_i^2) \mid i = 1, 2, \cdots, 150\}$ in the region $[0, 10] \times [0, 10] \subset \mathbb{R}^2$ according to uniform distribution, and then for each point $(x_i^1, x_i^2) \in X$ we compute

$$y_i = 0.5x_i^1 + 1.6x_i^2 + 6 + r_i \tag{4}$$

where $r_i \in [-2, 2]$ is a random number which obeys uniform distribution.

To verify the robustness of Theil-Sen estimator, we introduce three outliers $y_1^\alpha = y_1 + \alpha$, $y_2^\alpha = y_2 + \alpha$ and $y_3^\alpha = y_3 + \alpha$. We use $D_1 = \{(x_i^1, x_i^2) | i = 1, 2, \cdots, 100\}$, $\mathbf{y}_1^\alpha = \{y_1^\alpha, y_2^\alpha, y_3^\alpha, y_4, y_5, y_6, \cdots, y_{100}\}$ to carry out regression, and use $D_2 = \{(x_i^1, x_i^2) | i = 101, 102, \cdots, 150\}$, $\mathbf{y}_2 = \{y_{101}, y_{102}, \cdots, y_{150}\}$ to test the result. For least square regression and Theil-Sen estimator, the relationship between Error_square on test data and the value of α is shown in Figure 2. From this figure,we can see as α increases the Error_square of the result obtained from least square regression increases dramatically, but the Error_square of the result obtained from Theil-Sen estimator is almost a constant. This implies the Theil-Sen estimator is more robust than the least square regression.

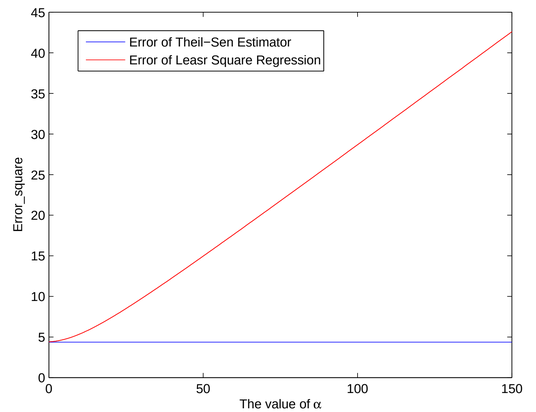

Figure 2. The relationship between Error_square and $\alpha = 100$.

Figure 3 gives a pictorial description of the results obtained from these two methods. When $\alpha = 100$, the plane obtained from Theil-Sen estimator is $y = 0.4704x^1 + 1.6217x^2 + 5.0561$ which is close to (4). The plane $y = 0.4704x^1 + 1.6217x^2 + 5.0561$ and training data X_1,$\mathbf{y}_1^{100}$ are shown in Figure 3 (a), where the red points are outliers. For $\alpha = 100$, the plane obtained from least square regression is $y = 0.4967x^1 + 0.6634x^2 + 12.5094$, which is shown in Figure 3 (b) with training data. From this figure, we can clearly see, due to the influence of three red outliers, the plane apparently deviates from the blue points.

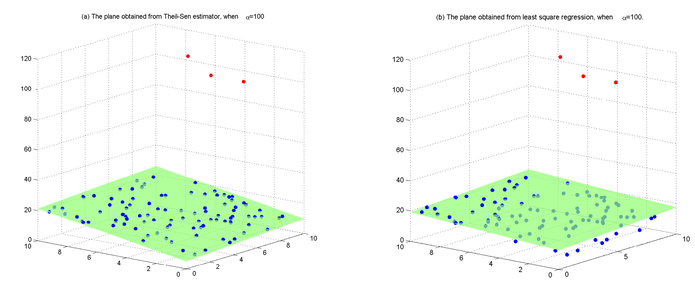

Figure 3. The plane obtained from two algorithms.

5 Conclusion

(1) The performance of a regression algorithm on the training data may be very well, but the model obtained from this algorithm perhaps can not work well on the test data. This is the phenomenon of over fitting. Actually, in "Communities and Crime Data Set", many attributes have no much relationship with the goal attribute "crime rate". Least square regression considers all these irrelevant attributes, so it cannot give a good prediction on the test data. On the contrary, ridge regression and Lasso regression restrict the norm of the solution, so sometimes they can work better. Moreover, we can use Lasso regression to find which attributes have the most important influence on the goal attribute.

(2) Compared with least square regression which is unstable to outliers, Theil-Sen estimator is more robust. It works well in two dimensional space. Theoretically, it can be generalized to high dimensional space, but for n instances in $\mathbb{R}^d$ we need to compute the normal vectors of $\binom{n}{d+1}$ hyperplanes, which is impractical for large n and d. Perhaps, we can find a method to only choose those representative hyperplanes, and then compute their normal vetoers to obtain a robust and practical regression algorithm in high dimensional space.