Lab Report

Machine Learning/CS3612/Homework2

Jiahao Zhao
Shanghai Jiao Tong University
School of Electronic Information and Electrical Engineering

April 6, 2023

Contents

Result (and analysis) **1**
Some analysis . . . 1

Visualization of parameters (and analysis) **1**
$\beta - Linear\ Rgression$: . . . 1
$\beta - Ridge\ Rgression$: . . . 2
$c - RBFKernel\ Regression$: . . . 3
$\beta - Lasso\ Regression$: . . . 4
Some analysis . . . 4

Discussions (and conclusions) **4**

Acknowledgement **5**

References **5**

Results (and analysis)

Attention: In the lab, the data I used is the normalized one.

Mode	σ	λ	lr	epochs	**prediction errors (SSE)**	method
Linear Regression	/	/	$2e-5$	$1e6$	190.08	gradient descent
Ridge Regression	/	1	$3.3e-05$	$1e6$	211.25	gradient descent
	/	10	$3.3e-05$	$1e6$	216.23	gradient descent
	/	100	$3.3e-05$	$1e5$	**191.62**	gradient descent
	/	1000	$3.3e-05$	$1e5$	238.39	gradient descent
RBF Kernel Regression	50	1	/	/	158.92	analytic expression
	50	10	/	/	156.97	analytic expression
	50	100	/	/	167.47	analytic expression
	50	1000	/	/	224.35	analytic expression
	1	10	/	/	255.66	analytic expression
	500	10	/	/	156.30	analytic expression
	5000	10	/	/	**156.29**	analytic expression
Lasso Regression	/	1	$2e-5$	$1e6$	223.06	gradient descent
	/	10	$2e-5$	$1e6$	181.43	gradient descent
	/	100	$2e-5$	$1e6$	170.43	gradient descent
	/	1000	$2e-5$	$1e6$	275.76	gradient descent
	/	$10^{1.7} \to 10^{0.01}$	/	/	**166.18**	coordinate descent

Table 1: Prediction errors results

For ridge regression, I use gradient descent instead of analytic expression because the latter seems to overfit on training set and it gives a bad result. For lasso regression, I obtain a better result by using coordinate descent. Hyperparameters for each experiments are shown in the table and I highlight each best result of ridge, kernel and lasso regression.

Some analysis

1. If we use analytic expression to obtain the Linear Regression Model, the SSE of it will be over 250. I think that is caused by overfitting on training set, since analytic expression means that model has a best performance on the training set.

2. Ridge regression performs worse than pure linear regression. The reason may be that the data distribution can not be fitted well by linear model. That is to say there is not strong linear correlation between arguments and dependent variable. That is also proved by RBF kernel regression. By using RBF kernel, I get a relatively good result, and kernel regression is to some extent non-linear model.

3. However, lasso regression seems to achieve a good performance. It seems that there may be some dimensions of X(input) linearly correlated with Y. Thus, I calculate correlation coefficient between each dimension of X and dependent variable Y. I also compute the determination coefficient[1]. The results is shown blow.

4. I find that none of arguments has a strong linear correlation with dependent variable while there are some dimensions correlate with dependent variable more than others. That explains linear model performs badly while lasso regression could give a relatively good result. And further,

X	Y	1th	day	FFMC	DMC	DC
0.075	0.071	0.151	0.006	0.052	0.11	0.108

ISI	temp	RH	wind	rain	DeCo (analytic)	DeCo (GD)
−0.039	0.019	−0.062	0.046	−0.046	0.067	0.065

Table 2: Correlation coefficient of each argument and determination coefficient(DeCo)

Visualization of parameters (and analysis)

$\beta - Linear\ Rgression$:

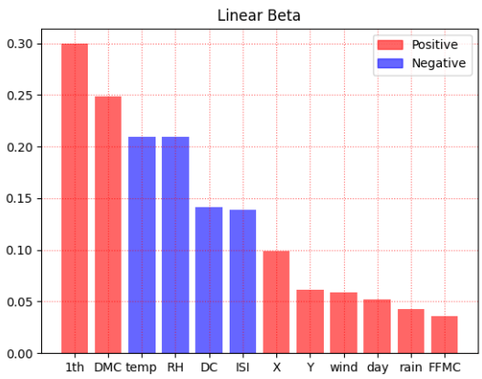
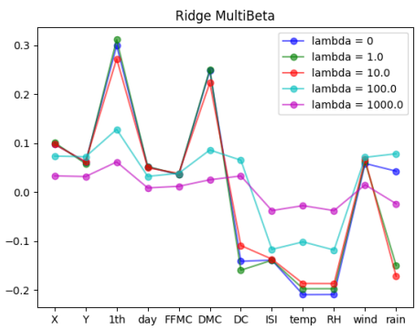


Figure 1: Pure linear model

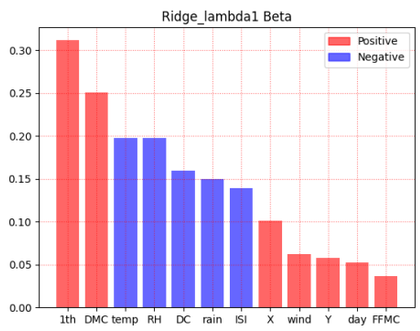
In this section, blue bar means negative value while red bar means positive value. For each β or c, I plot a bar chart, and for different parameters of a same model, I also plot a line chart for comparing them.

As is shown in Figure 1, values of β show that linear regression learned a almost right model. For example, RH and ISI are negatively correlated with areas (dependent variable) and their values are negative. However, there are some arguments like rain not consistent with their correlation with areas. It might be because of weak correlation.

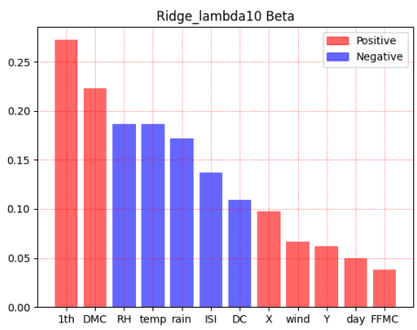
$\beta - Ridge\ Regression$:



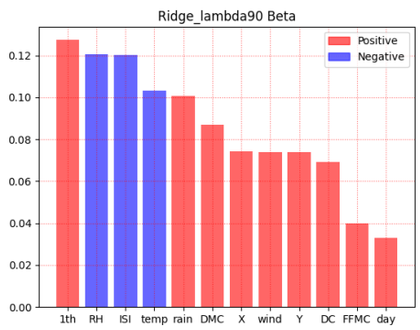
(a) β with different λ



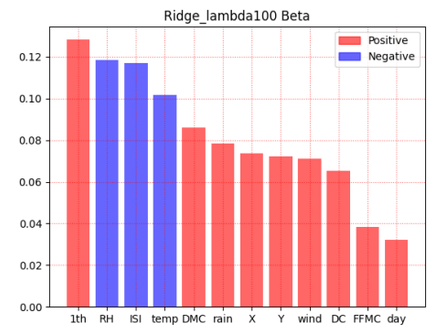
(b) $\lambda = 1$



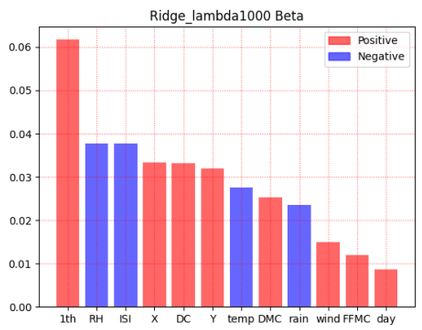
(c) $\lambda = 10$



(d) $\lambda = 90$



(e) $\lambda = 100$

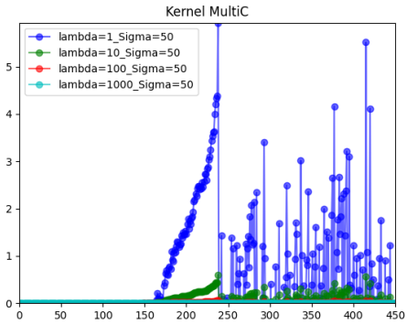


(f) $\lambda = 1000$

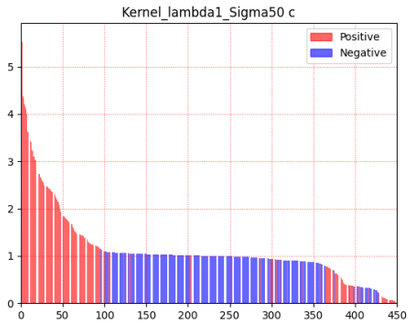
Figure 2: Ridge Regression

From Figure 2, we can see the power of L2-norm punishment. As λ increases, absolute value of each dimension of β decreases. The line of β tends to smooth as we can see in (a). And we can see that with $\lambda = 1000$, the model seems to learn values consistent with the correlation, although its prediction error is relatively high.

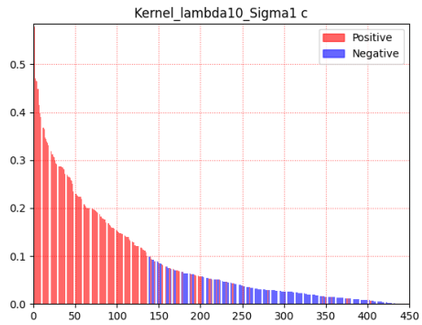
c − RBFKernel Regression :



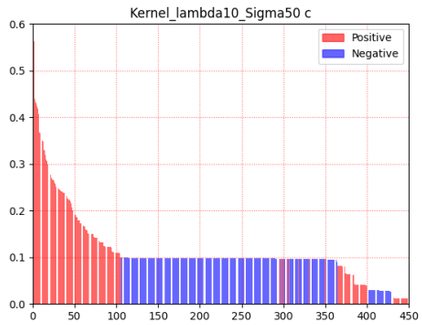
(a) c with different λ.



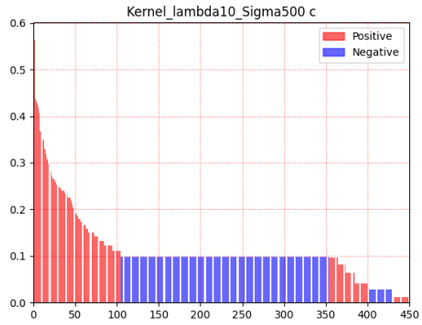
(b) $\lambda = 1$, $\sigma = 50$



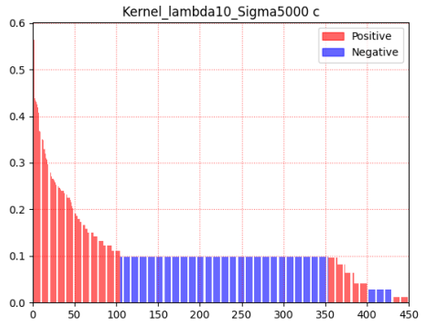
(c) $\lambda = 10$, $\sigma = 1$



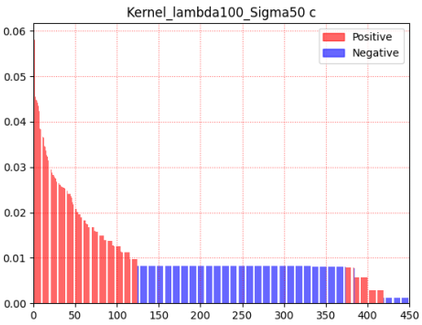
(d) $\lambda = 10$, $\sigma = 50$



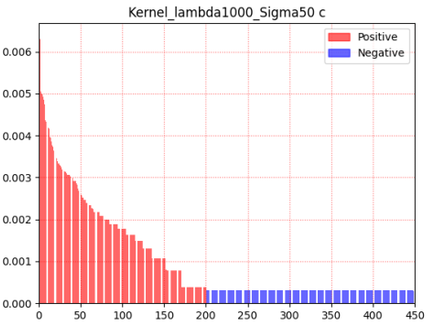
(e) $\lambda = 10$, $\sigma = 500$



(f) $\lambda = 10$, $\sigma = 5000$



(g) $\lambda = 100$, $\sigma = 50$



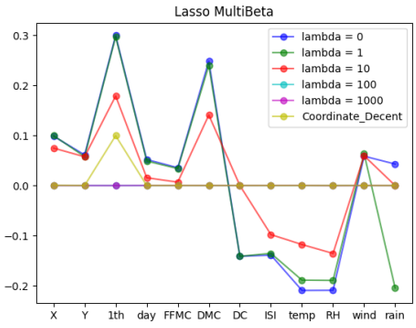
(h) $\lambda = 1000$, $\sigma = 50$

Figure 3: RBFKernel Regression

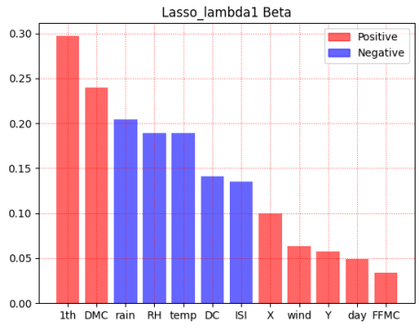
From Figure 3 we can see different functions of λ and σ in RBF Kernel regression. As λ increases, absolute value of each dimension of c decreases. As σ increases, absolute value of each dimension of c tends to be the same. That is because large σ means that data follows an almost even distribution, so that the model learned almost equivalent value for each feature.

From table 1 we can see that with large σ, the prediction error is low. That also proves that the data foes not follows linear correlation.

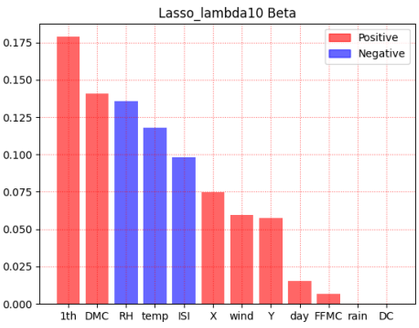
$\beta - Lasso\ Regression$:



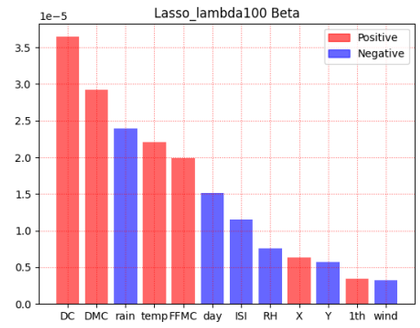
(a) β with different λ



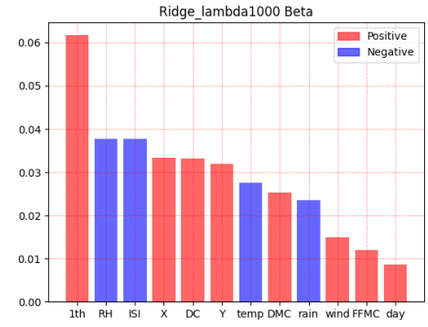
(b) $\lambda = 1$



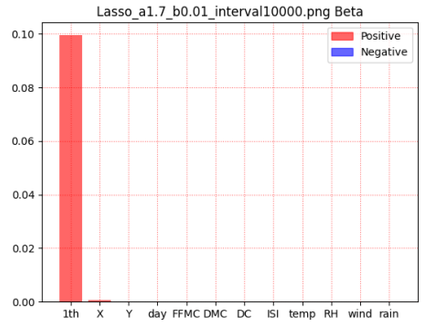
(c) $\lambda = 10$



(d) $\lambda = 100$



(e) $\lambda = 1000$



(f) Coordinate Decent

Figure 4: Lasso Regression

From Figure 4 we can see the power of L1-norm punishment. It tends to make β sparse, which means only some dominate dimensions have relatively large value. The effect is obvious in (f) in which I used coordinate decent for optimization. In gradient decent for lasso regression, I used sign function for derivative and that might be poor-performance in practice.

Some analysis

1. Ridge regression limits the absolute value of each dimension of β by punish L2-norm of β. It contributes to learning a right model.

2. In RBF kernel regression, λ limits the absolute value of each dimension of c while σ limits the absolute-value difference among different dimensions. Large σ obtains good result, which to some extend indicate that the data follows an almost even distribution.

3. Lasso regression makes β sparse by punish L1-norm of β. In practice, gradient decent for lasso regression is hard to converge since I use sign function for the derivative of L1-norm and it is not continuous and smooth. However, we can see an obvious effect by employing coordinate decent.

Conclusions (and discussions)

In this lab, I have implemented linear regression, ridge regression, RBF kernel regression and lasso regression and tried using all these four models to make predictions by different optimization method. I have learned the power of L1-norm and L2-norm punishment, the function of RBF kernel and the practical value of coordinate decent. And since some results are out of expectation, I also measured the correlation in the training data set. I believe that linear model can not fit well in this data set in fact. And from this lab, I am also aware of the following things:

1. Linear model is of great limitation while kernel regression is of great power (by adding non-linear kernel).

2. Analytic expression is to some extend overfitting, so gradient decent can achieve a better performance often.
3. Normalization sometimes might greatly affects the convergence of learning model (since the scale of different dimensions differs greatly).
4. And finally, I strongly suggest that using a better data set (which means there are relatively stronger linear correlation between arguments and dependent variable) while testing linear model.

(Other detailed analysis is in the previous sections.)

Acknowledgement

During the process of finishing my homework, I have discussed the hyperparameters and the data set with Yao Xu and Xiaotong Huang, I appreciate it.

References

[1] "Correlation and determination coefficient in linear regression". In: URL: `https://zhuanlan.zhihu.com/p/32335608`.