

Project 1: Linear Regression

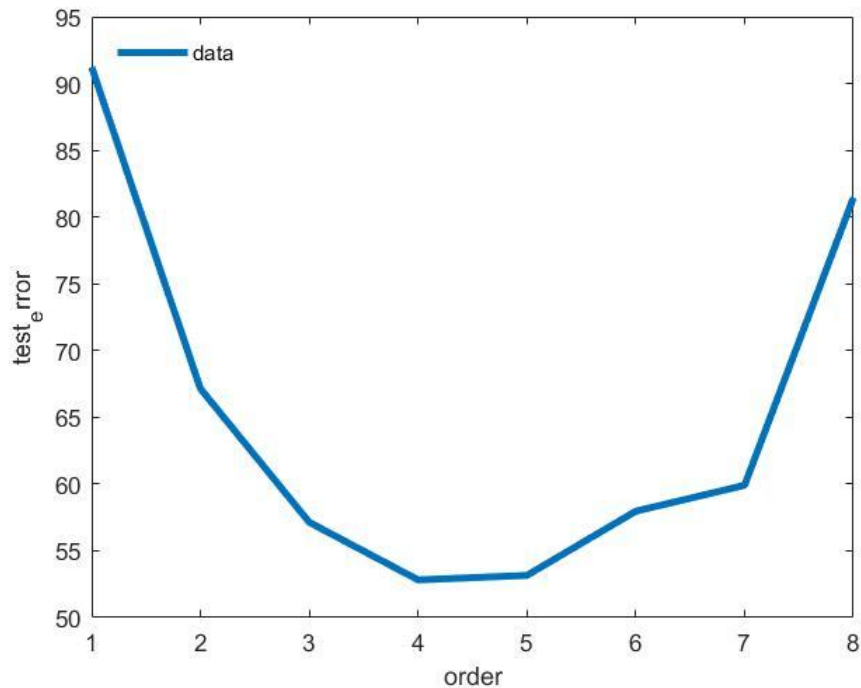
Aravind Vicinthangal Prathivaathi, Dingrong Wang

First part: Basic Linear Regression

- Using Polynomial Fit to implement basic linear regression, the training error is 31 and the test error is about 52. Empirically, the best polynomial order is 4 or 8.
- From the relatively large coefficient and the difference between train error and test error, the model was still a little overfit.
- Every feature has to be the same pattern in terms of the polynomial expansion.



First part: Basic Linear Regression



Second Part: Customized Linear Regression

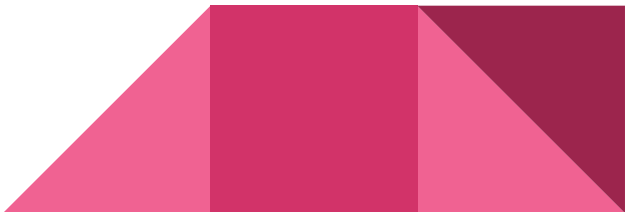
- First decide an order sequence, specifying the p order for each feature.
- Do the polynomial expansion for every feature according to their order in the previous sequence, then add a constant term
- After generating the Z matrix, compute the weight using linear regression and the test error using cross validation technique
- Choose the best order sequence using minimum test error and evaluate the model.

```
>> customized_linear_regression
```

```
minTR:39.7773
```

```
minTesterror:49.1175
```

```
averageError:40.1755
```



Third Part: Lasso Regression

Why we use Lasso to complete this project?

From the two steps talked above, we found the main reason for low accuracy is that there exist redundant features and overfitting problems, while Lasso can both do penalization and feature selection, it's a perfect choice. The other choice was ridge regression.

How do you implement it? Does it work?

I use a matlab built-in function to implement this function, with my own code to select lambda and clear irrelevant features. It works well, the test error has been reduced to 41, comparing to 52 in the first and second method.

Fourth Part: Summary

What features I use in the three methods

In the first method, every feature takes 8 polynomial order to achieve the best performance, the number of features is 65. ($8*8+1$)

In the second method, I loop every possible order for each feature from 1 to 3, the number of features is fluctuating between 9 ($1*8+1$) and 25 ($3*8+1$).

In the third method, every feature takes 5 polynomial order to achieve the best performance and I use lasso technique to conduct feature selection and compute the weight, the number of features is 15. (Some irrelevant features are removed.)

Fourth Part: Summary

How well do you estimate your team's work in terms of least-squares error?

I think it's pretty good. We have tried three methods in all. Finally, using the lasso technique, the test MSE error has been reduced to 16 or so, which is a big improvement comparing to the other two methods.

test_error	43.1349
testNum	277
traindata	926x9 double



End

Thank you for watching:)