PROJECT REPORT: C S 487/519 Applied Machine Learning I

Project 4: Compare regression methods

Written by

Omid Jafari

Department of Computer Science

New Mexico State University

October 2018

Preprocessing and notes:

For both datasets, we are using 70% of the data as the training set and the rest as the test set.

We are standardizing both datasets and show the results for both the original and standardized versions.

Regarding the California Renewable Production dataset, we are removing the timestamp column and columns containing missing values since the data should be complete for the regressors to work properly.

We will use Mean Squared Error and running time as the evaluation metrics.

The MSE value is reported for both training and testing sets to be able to detect overfitting.

We are using the following parameters for the regressors:

- max_trials=100
- min_samples=50
- residual_threshold=5.0
- alpha=1.0
- max_depth=3
- random_state=1

Results:

MSE of Housing dataset				
	Original data		Standardized data	
	Train set	Test set	Train set	Test set
Linear Regression	22.390	21.382	0.265	0.253
RANSAC	30.947	29.939	0.265	0.253
Ridge	22.645	21.412	0.265	0.253
Lasso	27.661	28.114	1.023	0.952
Normal Equation	22.390	21.382	0.265	0.253
Decision Tree Regressor	13.964	24.045	0.165	0.257

MSE of California Renewable Production dataset				
	Original data		Standardized data	
	Train set	Test set	Train set	Test set
Linear Regression	980413.183	984987.518	0.910	0.915
RANSAC	1056284.777	1052900.668	0.910	0.915
Ridge	980413.183	984987.519	0.910	0.915
Lasso	980413.207	984989.894	0.998	1.004
Decision Tree Regressor	948410.783	951481.160	0.881	0.884

Running time of Housing dataset (ms)						
	Original data			Standardized data		
	Training	Predicting		Training	Predicting	
		Train set	Test set		Train set	Test set
Linear Regression	11	0	0	1	0	0
RANSAC	47	0	1	2	0	0
Ridge	1	0	0	1	0	0
Lasso	1	0	0	1	0	0
Normal Equation	0	0	0	0	0	0
Decision Tree Regressor	1	0	1	1	0	0

Running time of California Renewable Production dataset (ms)						
	Original data			Standardized data		
	Training	Predicting		Training	Predicting	
		Train set	Test set		Train set	Test set
Linear Regression	7	0	0	8	1	0
RANSAC	94	1	0	14	1	1
Ridge	4	0	0	5	0	1
Lasso	5	1	0	3	0	0
Decision Tree Regressor	44	2	1	35	2	1

Analysis:

The MSE values of non-standardized data show that the datasets are not standardized by default. Therefore, we only focus on the standardized results.

By comparing the MSE values of train and test sets, we can infer that there is no overfitting in any of the regressors.

Looking at the MSE values of test set, we can see that all regressors except Lasso have almost the same values and are behaving the same. Lasso is having a worse MSE in this case.

Since the Housing dataset is small, the running times of it are not reliable. As a result, we use the running times of California Renewable Production dataset to compare the regressors in terms of running time. By comparing these values, we can see that the training time of Lasso is better than the rest. The predicting times of the regressors are almost like each other.

Performance Improvement:

We will use the feature selection library of SKlearn called LassoCV to extract the most important attributes, and then run regressors.

MSE of California Renewable Production dataset		
	Standardized data	
	Train set	Test set
Linear Regression	0.933	0.939
RANSAC	0.933	0.939
Ridge	0.933	0.939
Lasso	0.998	1.004
Decision Tree Regressor	0.895	0.906

Running time of California Renewable Production dataset (ms)			
	Standardized data		
	Training	Predicting	
		Train set	Test set
Linear Regression	4	0	0
RANSAC	8	1	1
Ridge	4	0	0
Lasso	3	1	0
Decision Tree Regressor	22	2	0

From the above tables, we can see there is a little increase in MSE values (except Lasso which has the same value as before). However, the running time of all of them is improved. This improvement can be drastically for larger datasets.