CS 487 - HW 6

Cyrus Baker
April 12, 2020

Notes:

This assignment uses two datasets, the housing dataset and the California Renewable Production dataset. Six regressors will be evaluated by comparing the mean squared error of each regressor. The six regressors are:

- Linear Regression
- RANSAC Regression
- Ridge
- Lasso
- Normal Equation
- SVR

Two versions of each dataset will be tested, one with standardization and one without.

Results:

Mean Squared Error Housing (non-standardized)		
	Train	**Test**
Linear Regression	19.96	27.20
RANSAC	24.32	35.83
Ridge	20.13	27.72
Lasso	24.04	31.68
Normal Equation	19.96	27.20
SVR	72.32	82.41

Mean Squared Error Housing (standardized)		
	Train	**Test**
Linear Regression	0.236	0.322
RANSAC	0.236	0.322
Ridge	0.236	0.322
Lasso	1.00	0.99
Normal Equation	0.236	0.322
SVR	0.09	0.25

Mean Squared Error CRP (non-standardized)		
	Train	**Test**
Linear Regression	984468	978928
RANSAC	1272020	1261896
Ridge	984468	978928
Lasso	984468	978928
SVR	1122780	1118171

Mean Squared Error CRP (standardized)		
	Train	**Test**
Linear Regression	0.91	0.91
RANSAC	0.91	0.91
Ridge	0.91	0.91
Lasso	1.00	1.00
SVR	0.72	0.73

Running Time in ms (Housing non-standardized)			
	Fit	**y_pred_test**	**y_pred_train**
Linear Regression	6.23	0.06	0.06
RANSAC	50.20	0.08	0.08
Ridge	1.38	0.08	0.16
Lasso	0.66	0.14	0.23
Normal Equation	4.25	0.03	0.03
SVR	5.38	1.34	2.99

Running Time in ms (Housing standardized)			
	Fit	**y_pred_test**	**y_pred_train**
Linear Regression	0.43	0.07	0.07
RANSAC	3.48	0.07	0.06
Ridge	2.37	0.10	0.14
Lasso	0.68	0.08	0.05
Normal Equation	0.30	0.02	0.01
SVR	5.64	1.61	2.42

Running Time in ms (CRP non-standardized)			
	Fit	**y_pred_test**	**y_pred_train**
Linear Regression	10.29	0.25	0.29
RANSAC	110.97	0.23	0.27
Ridge	7.11	0.33	0.33
Lasso	5.50	1.00	0.66
SVR	38651	8742	20475

Running Time in ms (CRP standardized)			
	Fit	**y_pred_test**	**y_pred_train**
Linear Regression	3.58	1.55	0.64
RANSAC	12.53	0.31	0.30
Ridge	4.20	0.44	0.31
Lasso	3.94	0.25	0.33
SVR	54910	12684	29461

Analysis:

The mean squared errors are very large for the non standardized datasets. This means that the standardization is necessary to perform regression on these particular datasets. Because the results of the training set and the test set are very similar we can infer that the regressors are not overfitting.

I found that the mean squared error is lowest for the support vector regressor but it takes significantly longer to run compared to the other methods. All of the other methods, except lasso, gave almost the same mean squared error with lasso being slightly worse.

Of the regressors SVR is the best, although slow. If speed were a factor in choosing a regressor, then ridge, ransac, and linear regression would all be good candidates.