Housing Values in Boston Suburbs

Samba Njie Jr

1/2/2017

Housing Values in Suburbs of Boston

Dataset Description: `housing.csv` concerns housing values in suburbs of Boston. The dataset was created by Harrison, D. and Rubinfeld, D.L. and analyzed in *'Hedonic prices and the demand for clean air', J. Environment Economics and Management, vol. 5, 81 - 102, 1978.* There are 506 observaitons and 12 continuous attributes including the response variable `MEDV`.

Attribute Information:

- `CRIM`: per capita crime rate by town
- `ZN`: proportion of residential land zoned for lots over 25,000 sq. ft.
- `INDUS`: proportion of non-retail business acres per town
- `CHAS`: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- `NOX`: nitric oxides concentration (parts per 10 million)
- `RM`: average number of rooms per dwelling
- `AGE`: proportion of owner-occupied units built prior to 1940
- `DIS`: weighted distances to five Boston employment centres
- `RAD`: index of accessibility to radial highways
- `TAX`: full-value property-tax rate per $10,000
- `PTRATIO`: pupil-teacher ratio by town
- `LSTAT`: % lower status of the population
- `MEDV`: Median value of owner-occupied homes in $1000's.

Tasks:

1. Create randomly sampled training and test sets from the dataset using 90% of the observations for training and 10% for testing. Put aside your test set and only use it for the last task.
2. Plot the correlation matrix of all attributes. Which attributes you deem more predictive of the housing prices?
3. Implement Algorithm 6.1 and report the best model under C_p, BIC, adjusted R^2 and Cross-Validation (k-fold, k of your choice).
4. Implement Algorithm 6.2 and report the best model under C_p, BIC, adjusted R^2 and Cross-Validation (k-fold, k of your choice).
5. Find the best model under LASSO and Ridge-regularized LS. Use cross-validation to choose the best penalty. You may use `glmnet` or any other library for this task.
6. Use your 3 best models chose from the last 3 tasks to predict the housing values in your test set and compute the predicted MSE for each. Interpret your results.