Assignment 5: Regression and ML

Due April 15 11:59 pm

Part 1

Load the data from 'bb_agg.csv'. This data is an easy tabular format, everyone row correspond to aggregate statistics for a team in a given year. The columns are as follows:

(1) yearID, teamID (self explanatory, I hope)

(2) batting statistics: hits (H_bat), doubles (2B_bat), home runs (HR_bat), stolen bases (SB_bat), caught stolen (CS_bat), ground into double plays (GIDP_bat),

(3) pitching statistics: home runs allowed (HR_ptch), walks allowed (BB_ptch), strikeouts (SO_ptch), hits allowed (H_ptch)

(4) team statistics: wins (W), games played (G), league ID (lgID)

Part 2

Treat yearID and lgID as categorical variables. Make your y variable the total wins (W).

Treat the batting statistics and pitching statistics as continuous variables. Make a pipeline (using a column transformer) that does the following:

(1) One Hot encodes yearID and lgID, and then uses a polynomial transformer to get interaction terms between them (use degree=2 and interaction_only=True) at this stage.

(2) Does a second order polynomial transform on the continuous variables, and then standard scales them.

(Note: do not include a bias term in either of the polynomial transformations).

(3) Drops any additional variables.

(4) Ends with a ridge regression model (regularization constant unspecified).

Part 3

Use GridSearchCV to cross validate this pipeline. You can use 5 fold cross-validation, and use np.logspace(-1,2,20) as your parameter grid for alpha in the ridge regression model. How well does the model perform on the test set?

Part 4

Do the same thing as part 2/3, except fit a random forest model. Perform cross-validation to optimize the hyperparameters ’max_tree_depth’ and ’min_samples_split’.

Extra Credit

If I wanted to include in the cross validation some of the parameters in the preprocessing part of the pipeline, how would I do that? i.e. - if I wanted to cross_validate the degree of the polynomial transformation for the continuous variables?