Task: Regularized linear regression

(a) Join the House Prices: Advanced Regression Techniques competition on Kaggle. Download the training and test data. The competition page describes how these files are formatted.

(b) Tell us about the data. How many samples are there in the training set? How many features?

Which features are categorical?

(c) What variables seem to be important? Which seem to correlate with the sale price? Plot the relationship between sale price and year of sale, garage area, lot area, and other variables of your choice. Choose 7 variables and, along with the response variable, make a scatterplot matrix (hint: look at pandas.plotting.scatter_matrix or seaborn.pairplot). Explain what you see.

(d) Using statsmodels (as done on last slide of lecture 8), run ordinary least squares on all the features and report which features have a 95% confidence interval that contains 0 and which do not. Comment on what this means.

(e) Split the training data into a training (80%) and test set (20%). Try to run a variety of regression methods using sklearn methods:

• Ordinary least squares

• k-Nearest Neighbors with 10-fold cross validation to choose k

• Ridge regression with 10-fold cross validation to chose λ

• LASSO with 10-fold cross validation to chose λ

• Backward stepwise (linear) regression with 10-fold cross validation to choose k (number of features)

• Forward stepwise (linear) regression with 10-fold cross validation to choose k (number of features)

For each, give a brief description of how well it works and why that might be and report the test accuracy (note: this refers to split off test set above; not the Kaggle test set).

(Hint: sklearn has methods called RidgeCV and LassoCV.)

(f) Repeat the above, except for OLS, after adding all the quadratic features:
$X_{ij}X_{ik}$ for all $i, j = 1, \ldots, p$ (this includes X_{ij}^2)

(g) Which variables are being retained by LASSO and the stepwise regression models and which are regularized away? Do these variables match your intuitions about which variables are important and which are not? Compare this to (d).