Authors: John Waldo and Adam Wolfson

HW 2 Write up

Part 1: Housing Regression

We found that it was difficult to work with a dataset with so many encoded variables. It really made it clear that in order to push our model to the next level we must have domain expertise and understand that the numeric features might have a nonlinear relationship to the target and by going through all our non numeric datapoints we could potentially extract a lot more information from the data than using one hot encoding. Regularization techniques help a lot when we throw a bunch of features at it but regularization alone cannot solve all our problems.

Part 2: Sentiment Analysis

Interestingly, the Bag of Words model has very good accuracy for more simple sentiment tasks. We found that a lot of the work for bag of words type models is in the preprocessing stage. Lining up similar words to the same token makes a huge difference not only in reinforcing certain trends but also in avoiding noise or excessive features.