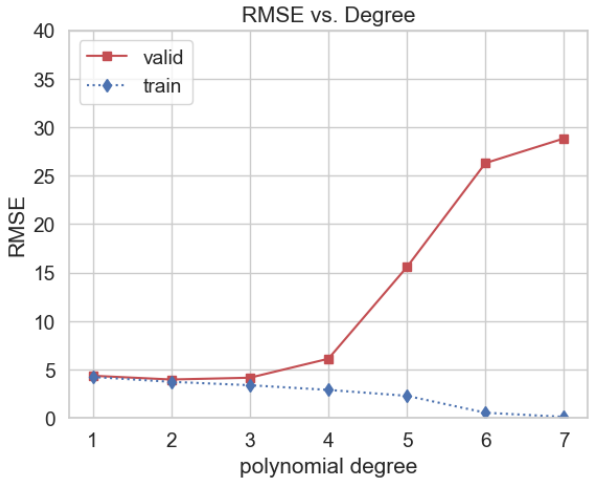
CS 135 HW 1 Regression, Cross-Validation, and Regularization

Liam Strand

Problem 1: Polynomial Regression - Selecting Degree on a Fixed Validation Set

Figure 1:



This plot aligns with what I would expect based on the course concepts. The high-degree polynomial overfits on the training data and is unable to make meaningful predictions on the validation data. I would recommend degree 2 or 3 based on this plot.

Short Answer 1a:

Rescaling is necessary because different features are in very different units. They are scaled differently, and some are discrete values between 4 and 8 as opposed to other continuous values. We rescale to standardize everything prior to attempting to fit the model.

If we remove this step, training error does not decrease as we would expect for a high-degree polynomial model. The validation error is reasonably low, but it does not spike on the higher degree polynomials.

Short Answer 1b:

```
-10.43 : x0          where x0 = horsepower
-18.23 : x1                x1 = weight
 -1.15 : x2                x2 = cylinders
  0.58 : x3                x3 = displacement
```

Increasing weight definitely decreases fuel efficiency, no doubt about that. And that makes sense! If there's more stuff to drag around, you need to burn more fuel to move it.

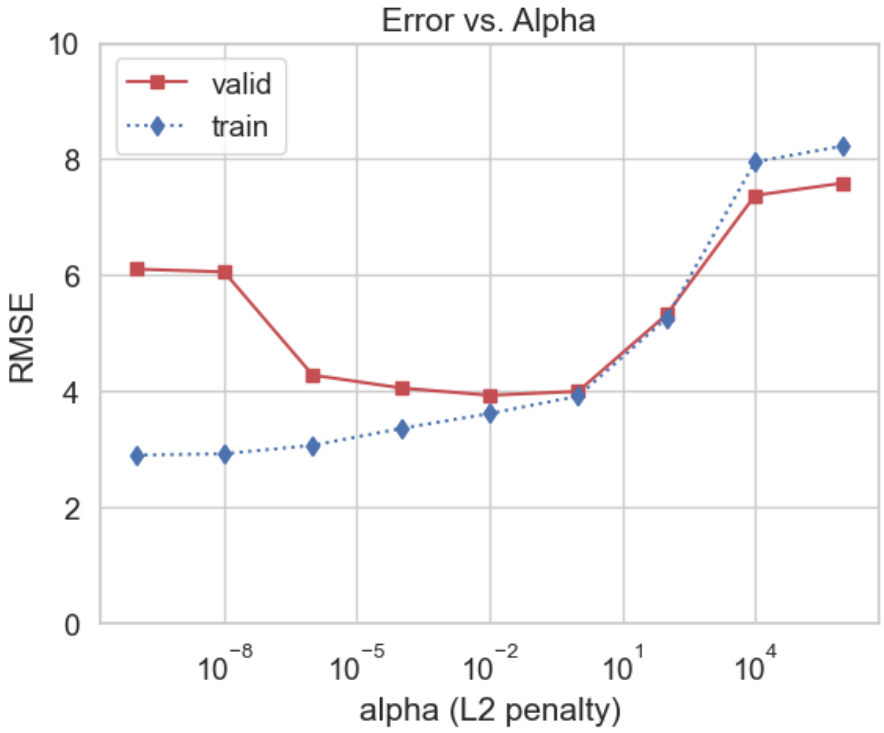
Increasing displacement seems to increase efficiency. That seems a little counterintuitive, but I suppose if I really think about it, a larger engine can make better use of the fuel it is given, because it can consume more air to burn that fuel more effectively.

Short Answer 1c:

The high-degree polynomial models have many parameters with *huge* weights. This is a big red flag for overfitting, which makes sense given that the training error continues to decrease while the validation error skyrockets. The lower-degree polynomial models have relatively small weights, which indicates that they are not overfit.

Problem 2: Penalized Polynomial Regression - Selecting Alpha on a Fixed Validation Set

Figure 2:



The figure does look like what I would expect. The low-alpha models appear overfit while the high-alpha models appear overfit. I would select $\alpha = 0.01$ based on this graph.

Short Answer 2a:

The magnitudes of the weights in part 2 are much much less than those without the added penalty.

Short Answer 2b:

You would essentially always pick $\alpha = 0$ because the best fit you could find would be high-degree low-alpha. You need the validation set to make sure that the model generalizes to new data.

Problem 3: Penalized Polynomial Regression + Model Selection with Cross-Validation

Regressor	Test RMSE
Baseline	7.104
Best Poly+Linear	3.991
Best Poly+Ridge (degree 4)	3.877
Best Poly+Ridge (any degree)	3.816

The model selected by using 10-fold cross validation performed the best. The rankings do match what I would expect given the course concepts. The more thoroughly we can evaluate and select hyperparameters, the better our resulting model can be.