NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 12: Lecture 2
Regularization – Ridge and Lasso

CC BY NC ND

How can we avoid both underfitting and overfitting?

DS-UA 112
Introduction to Data Science

Week 12: Lecture 2
Regularization – Ridge and Lasso

Adapted from Nolan, Speed, Gonzalez, Lau

cc BY NC ND

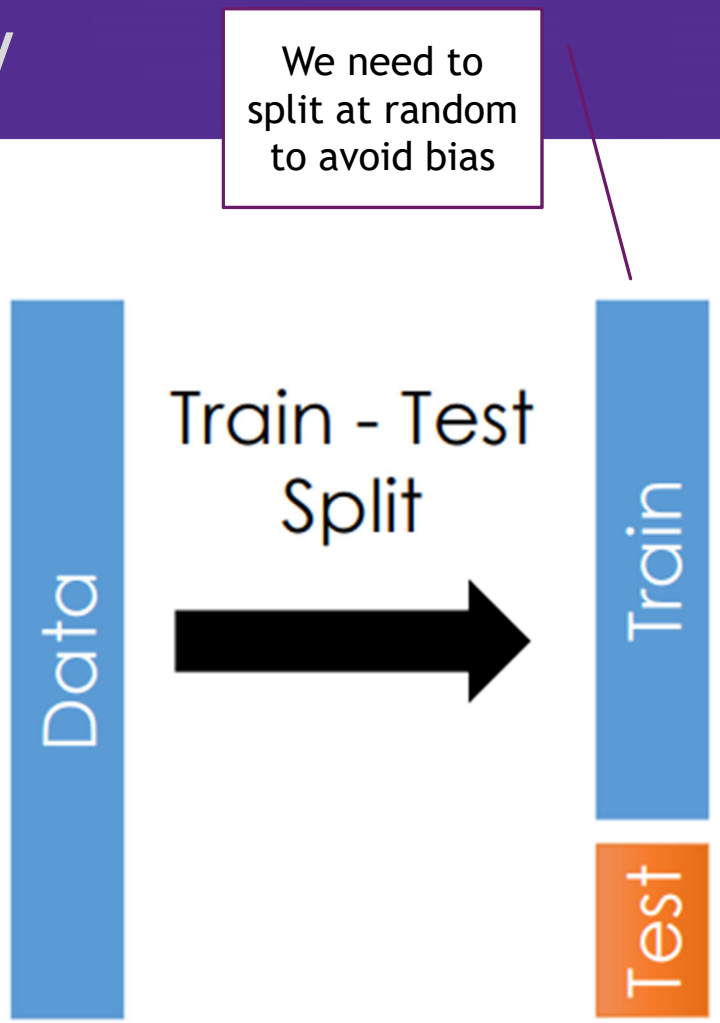
Announcements

- Please check Week 12 agenda on NYU Classes
  - Lab 12
    - Due on Friday April 24 at 11:59PM EST
  - Homework 4
    - Due on Saturday April 18 at 11:59PM EST
  - Homework 5
    - Released Thursday April 16

CC BY NC ND

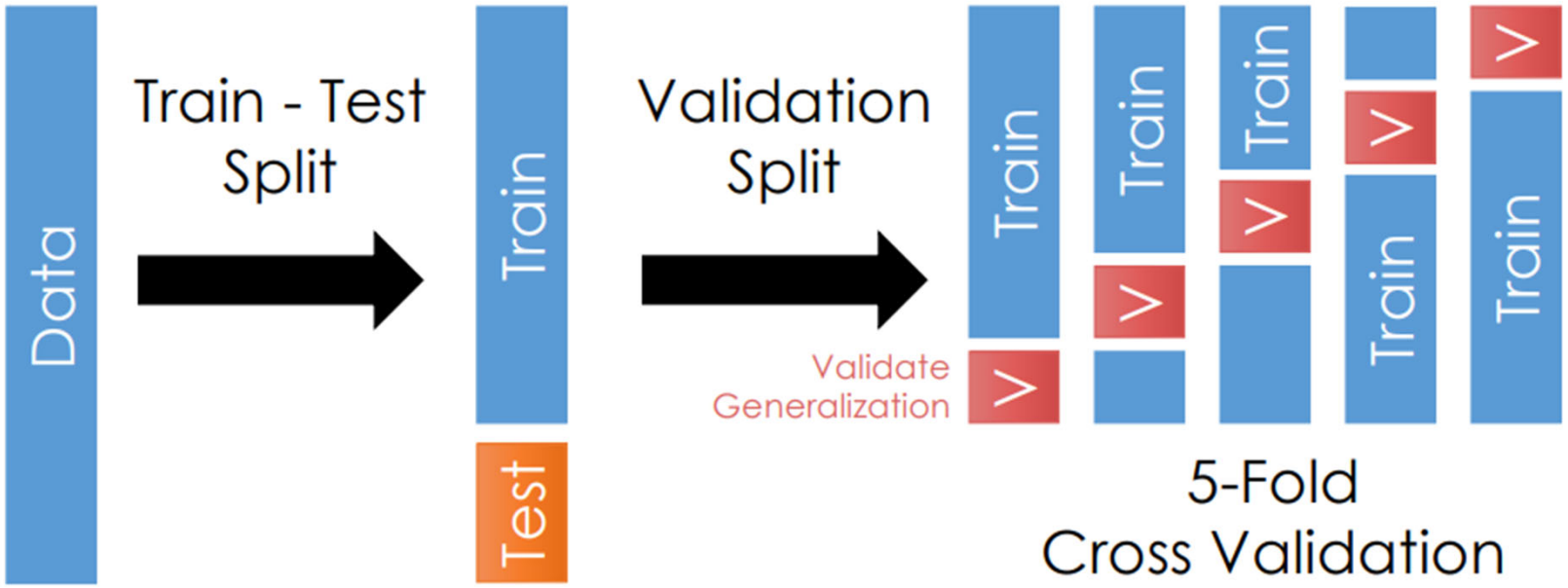
Review

- Instead of studying one sample, we need to study two samples the training set and testing set
- We will fit the model to the data in the training set. We will check the accuracy of the predictions on the testing set.
- Usually we take 80% of the data for the training set and 20% of the data for testing set.



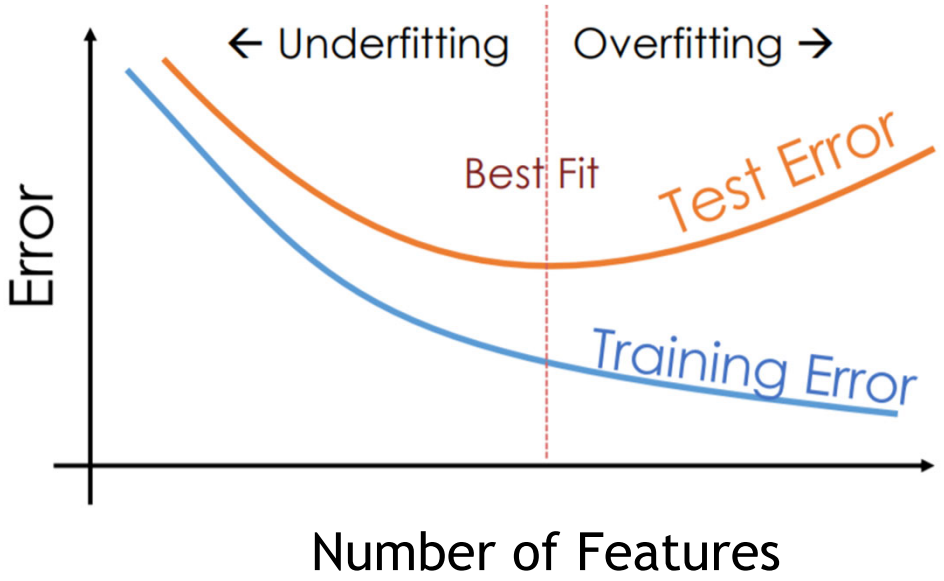
Review

- We split one sample into two sample to generate the training set and testing set. Next we split the training set into k folds. Each fold has a training set and a validation set



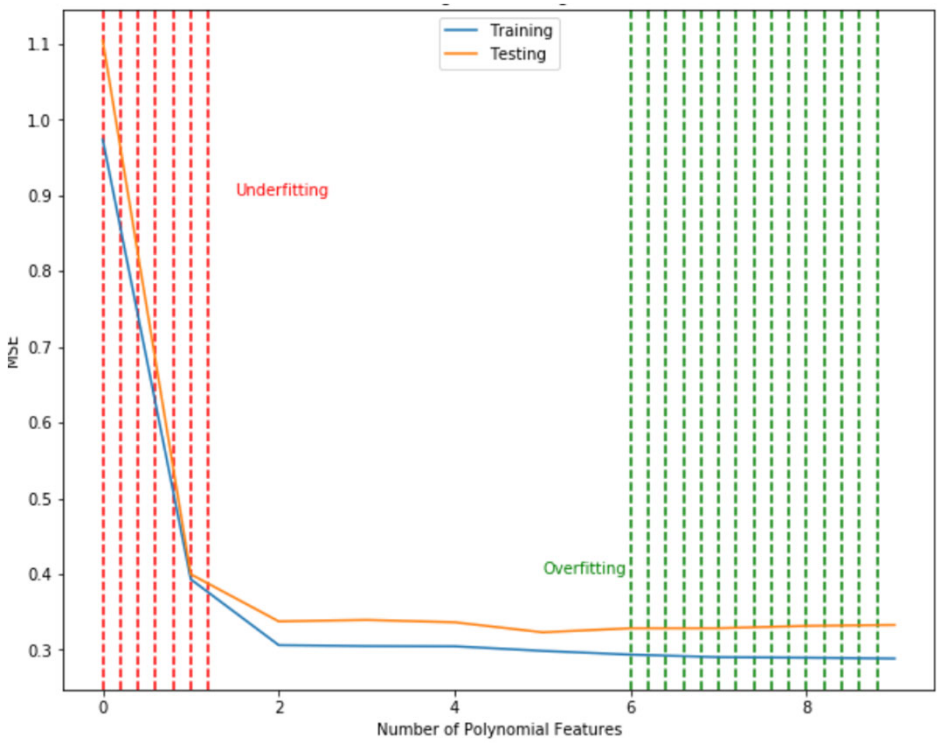
Review

- We want to include many features to avoid underfitting. However, we need to limit the number of features to avoid overfitting. Regularization means balancing between underfitting and overfitting
- We want to add an extra parameter to the model to help with regularization. For large values, it prevents overfitting. For small values it prevents underfitting.



Agenda

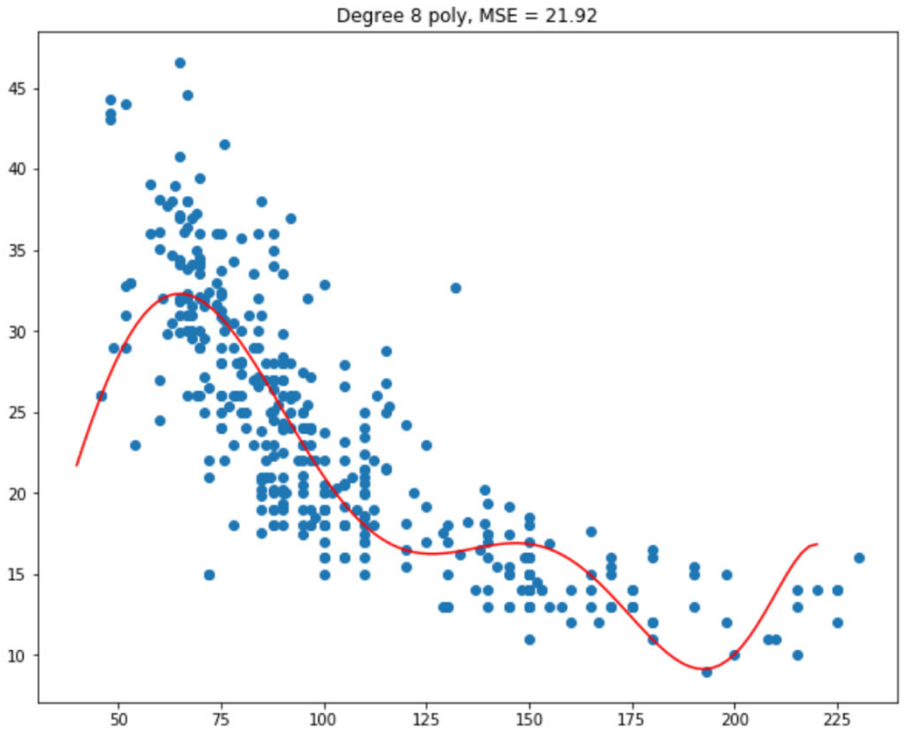
- Cross Validation
- Feature Selection
- Regularization



cc BY NC ND

Memorizing Data

- If we overfit the data, then the predictions might not generalize from sample to population. The model should not memorize the training set.
- Suppose we have explanatory variables $x_1,\ldots,x_n$ and response variables $y_1,\ldots,y_n$. For some x, the model could be
  - if $x = x_i$ for some i then predict $y = y_i$
  - Otherwise predict $y_i = 0$
- While we would have no errors on the training set, we would have large errors on the testing set.



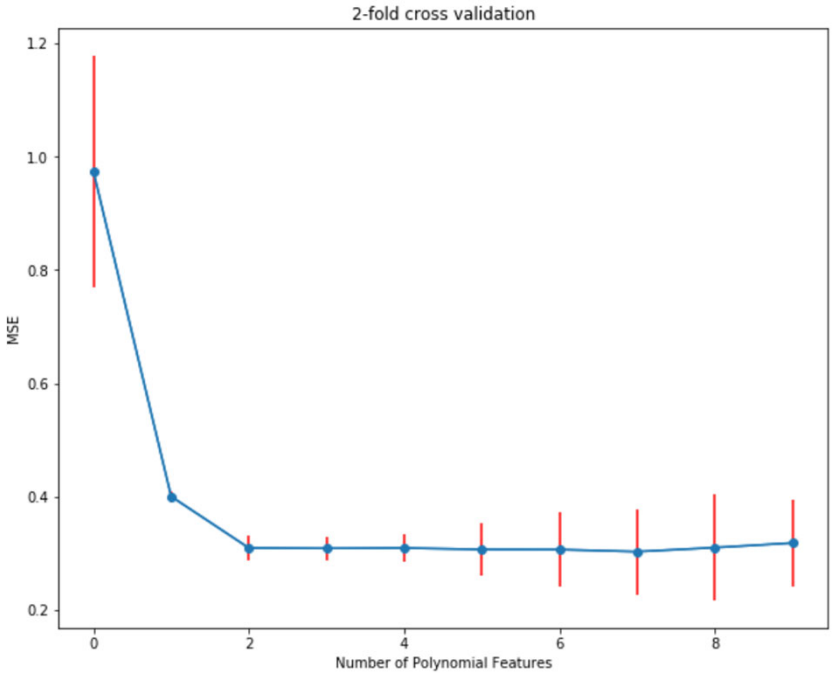
Standard Units

- If features have different scales then we might have trouble with linear regression
  - The aspect ratio in charts will distort the relationships between features
  - The large or small values of the derivatives might prevent gradient descent from converging to a minimum.
  - We will need the same scale for regularization.
- So we should convert to standard units

- For each column of the table compute
  - mean
  - standard deviation
- Subtract the mean
- Divide by the standard deviation
- Following the transformation each column should have mean 0 and standard deviation 1

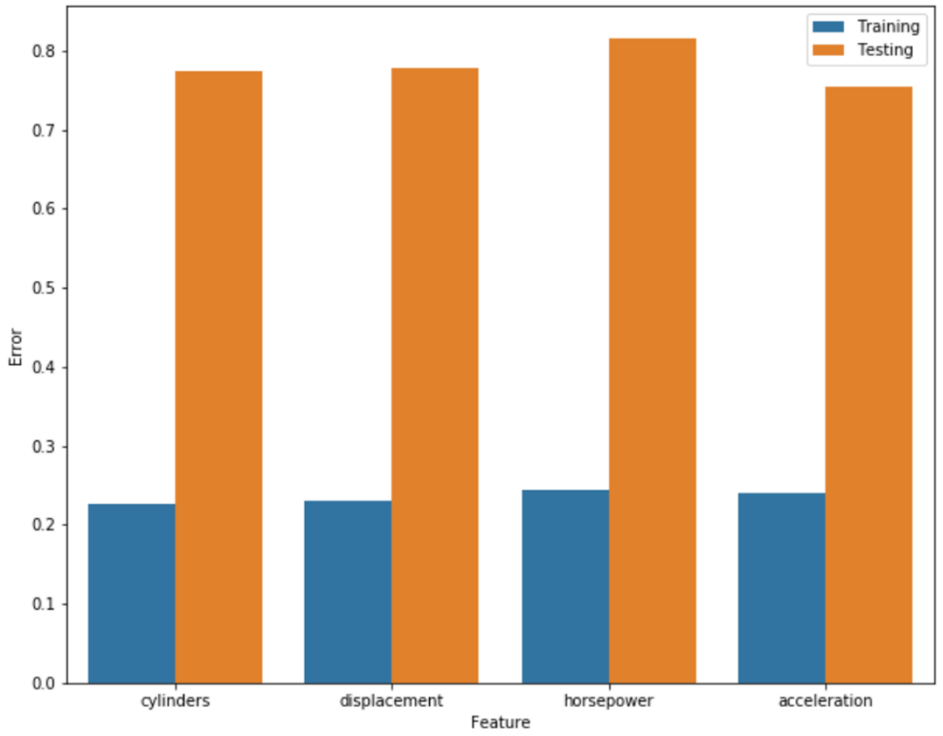
Cross Validation

- We want to choose models that are both accurate and consistent.
- With cross validation we measure the difference between predictions and observations on many datasets.
  - Small errors give us accuracy
  - Similarity between errors give us consistency
- We can visualize both the accuracy and consistency through a line chart with error bars



Feature Selection

- If we want to remove features to prevent against overfitting, then we could try to assess the effect of dropping combinations of features.
- In backward feature selection we
  - select a feature
  - drop it from the table
  - fit a model to the data
  - calculate average loss
- The feature that led to the smallest increase in loss should be excluded from predictions



Lasso Regression

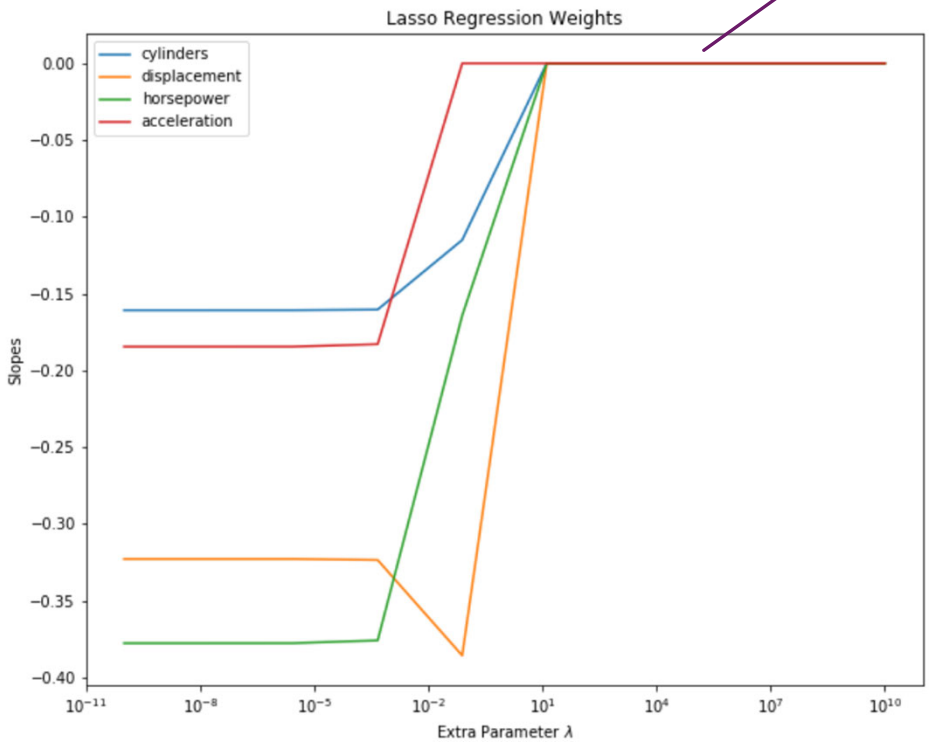
- Remember that for each feature we have a slope. If we set the slope to zero, then we have effectively removed the feature from the model.
- We can replace the average loss for linear regression

$$\frac{1}{n}\sum_{i=1}^{n}(a+bx_i-y_i)^2$$

with regularized average loss for lasso regression

$$\lambda(|a|+|b|)+\frac{1}{n}\sum_{i=1}^{n}(a+bx_i-y_i)^2$$

Fitting the model to the data will exclude less important features



Summary

- Cross Validation
- Feature Selection
- Regularization

Goals

- Transform data to standard unit to prevent different scales
- Use cross validation to assess the accuracy and consistency of model
- Apply Lasso regression to select features

CC BY NC ND

Questions

- Questions on Piazza?
  - Please provide your feedback along with questions
- Question for You!
  - How could overfitting lead to bias in predictions?

ProPublica

Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement

Facebook's ad delivery algorithm further skews the audience based on ... investigation of Facebook over discrimination in ads for housing and ...

right. As Facebook promised in the settlement, advertisers on the new portal can no longer explicitly target by age or gender. Nevertheless, the composition of audiences can still tilt toward demographic groups such as men or younger workers, according to a study published today by researchers at Northeastern University

cc BY NC ND