NYU Center for Data Science

DS-UA 112
Introduction to Data Science

Week 12: Lecture 1
Regularization – Training and Testing

How can we make accurate predictions about a population based on data in a sample?

DS-UA 112
Introduction to Data Science

Week 12: Lecture 1
Regularization – Training and Testing

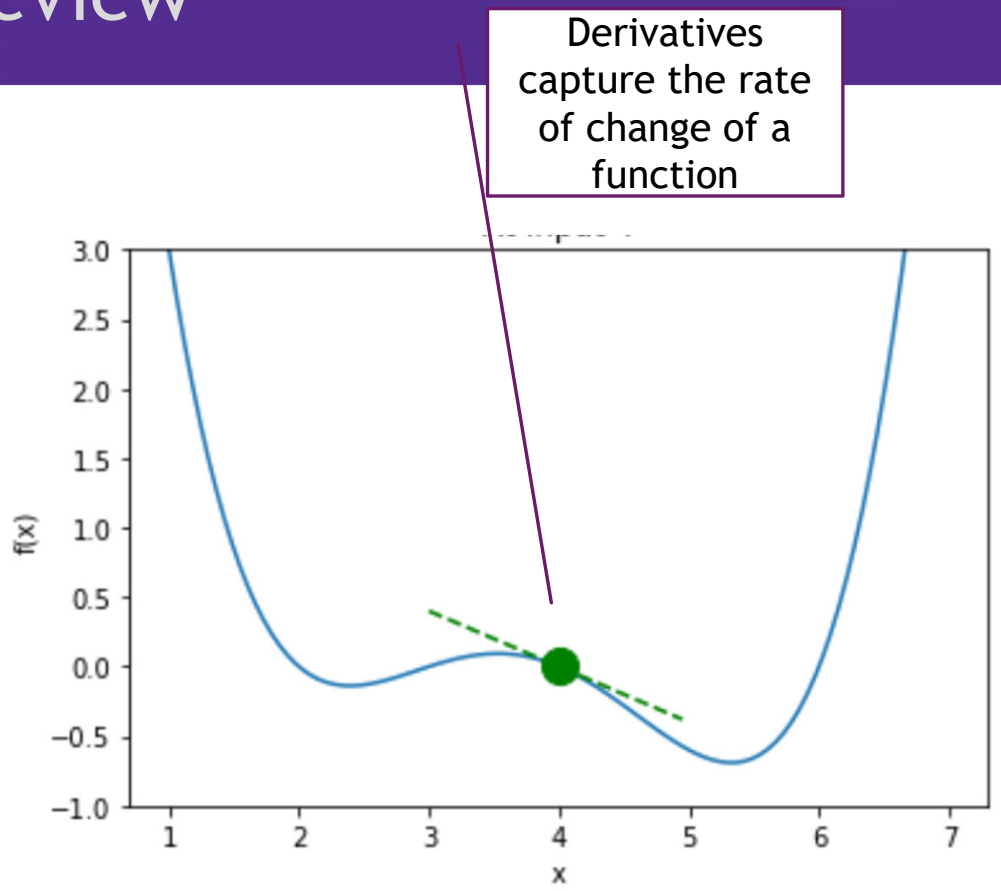
Adapted from Nolan, Speed, Gonzalez, Lau

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

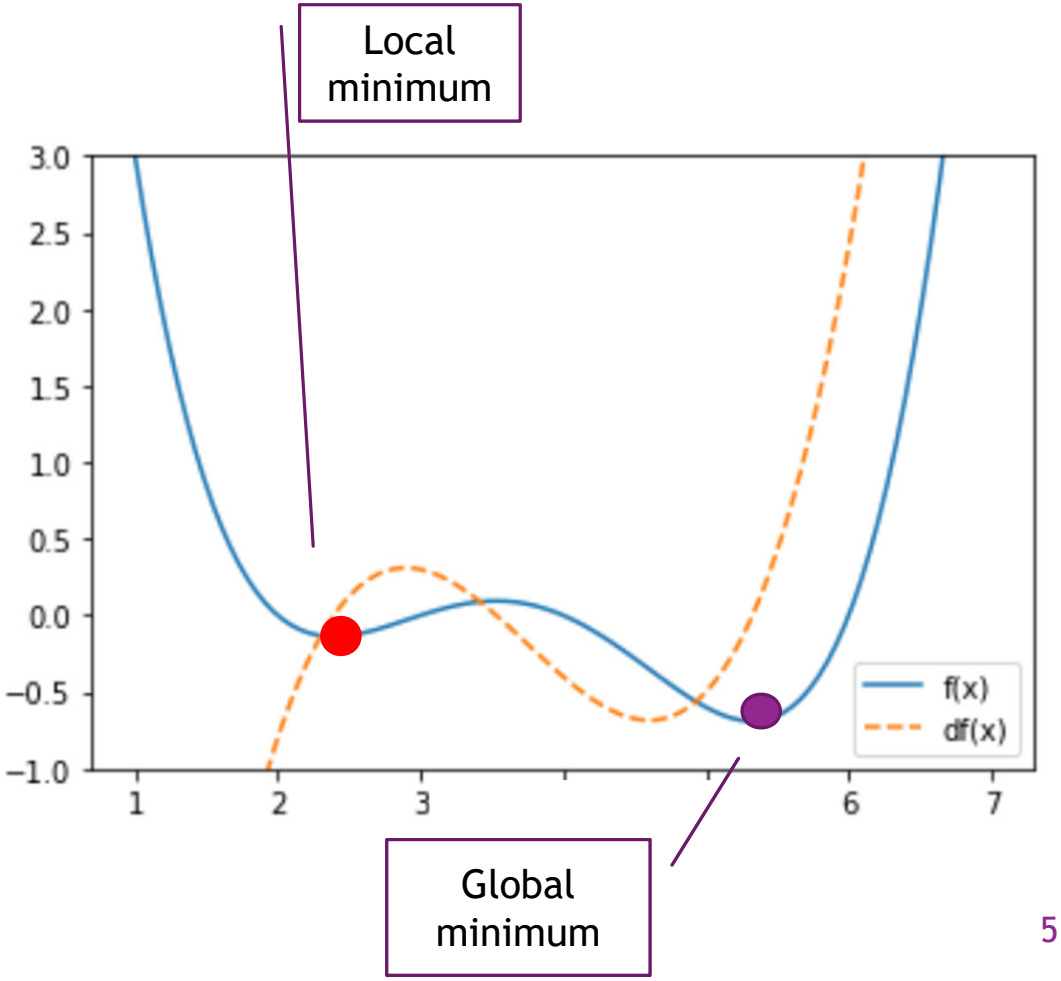
Review

- By minimizing the average loss, we obtain parameters that fit the model to the data.
- We should not guess inputs and check outputs to minimize the function because that approach is inefficient and inaccurate.
- Instead we will just make one guess and use the derivative to update the guess. We call the approach gradient descent.



Review

- We want to find the minimum output for the function among all inputs. We nickname the value the global minimum.
- However gradient descent can get stuck at a local minimum. We need to be careful about the implementation of gradient descent for minimizing functions with local minima

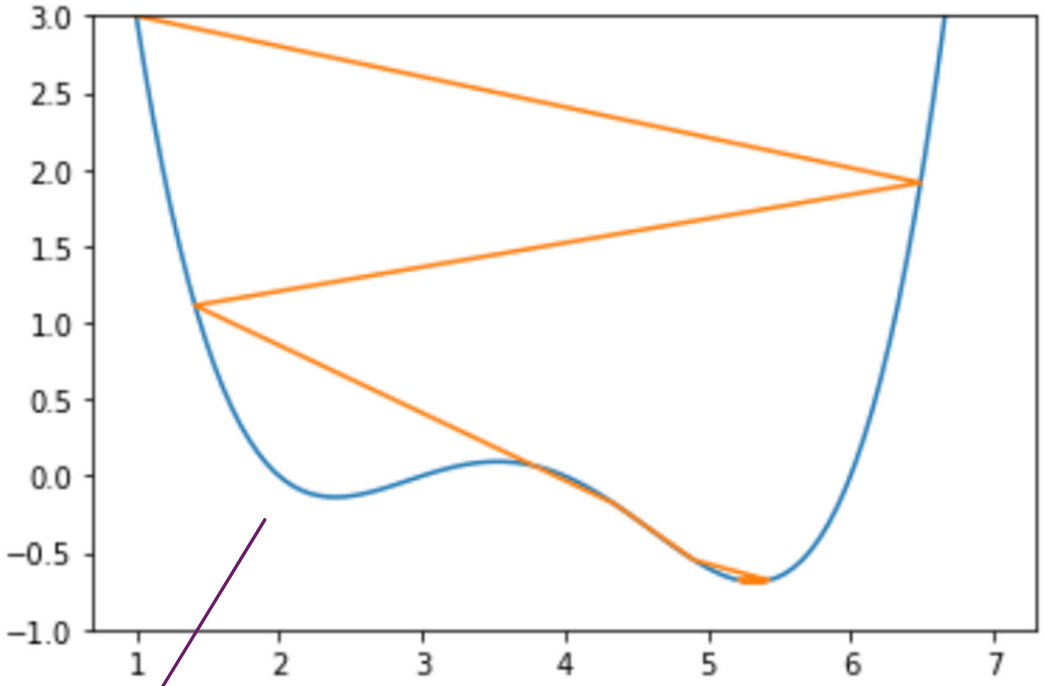


Review

- Starting from an initial guess $x^{(0)}$ we update the guess with the formula

$$x^{(t+1)} = x^{(t)} - \alpha \frac{d}{dx} f(x)$$

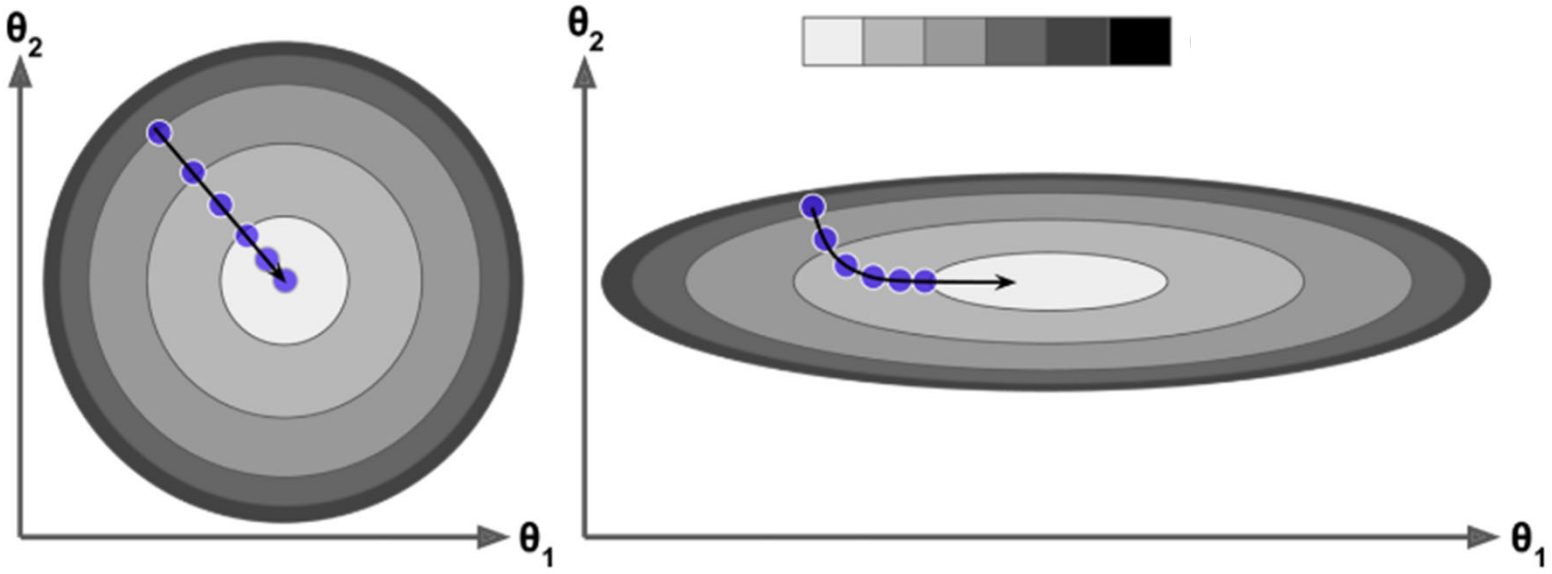
- Here α denotes the learning rate. If α is large, then guesses can change a lot between iterations. If α is small, then guesses can change a little between iterations



If the learning rate is too large then gradient descent might diverge from the minimum

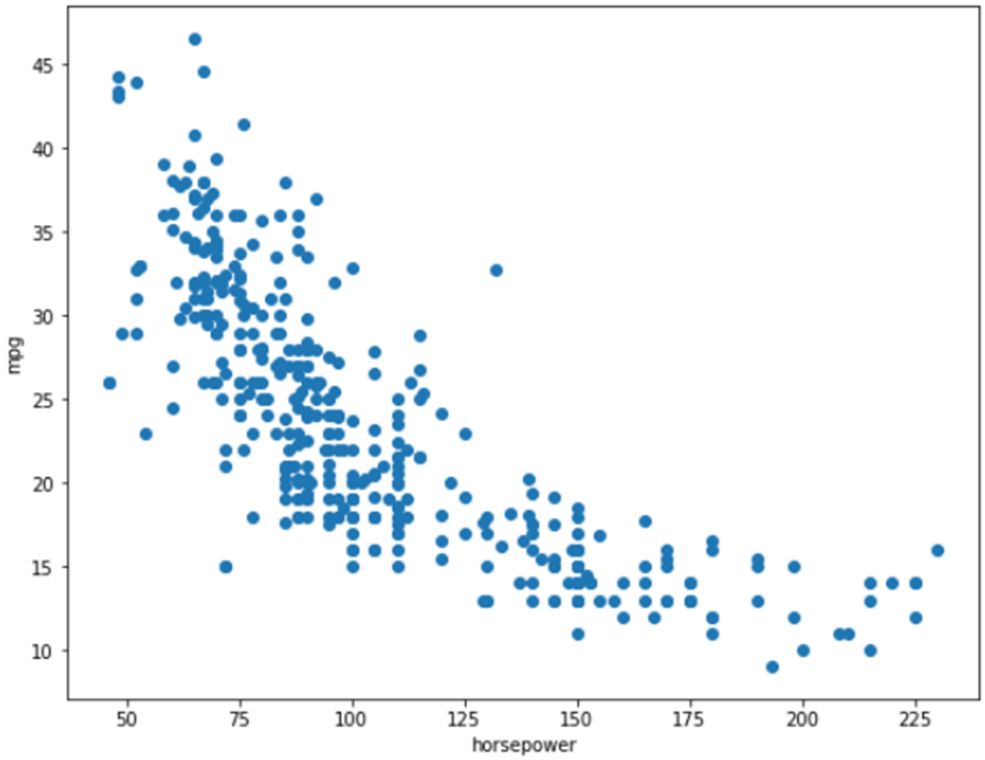
Review

- If the variables have different scales, then gradient descent might diverge from the minimum. So we could transform each column of the table to have mean 0 and standard deviation 1.



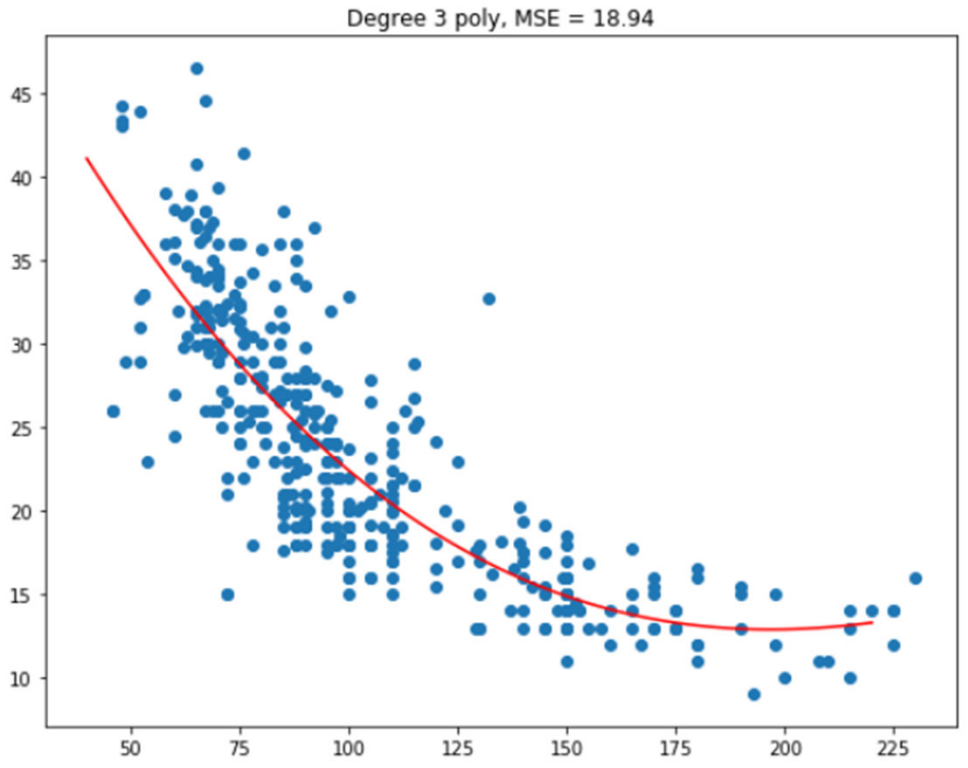
Agenda

- Validation
  - Training Set
  - Testing Set
- Number of Features
  - Underfitting
  - Overfitting
- Regularization



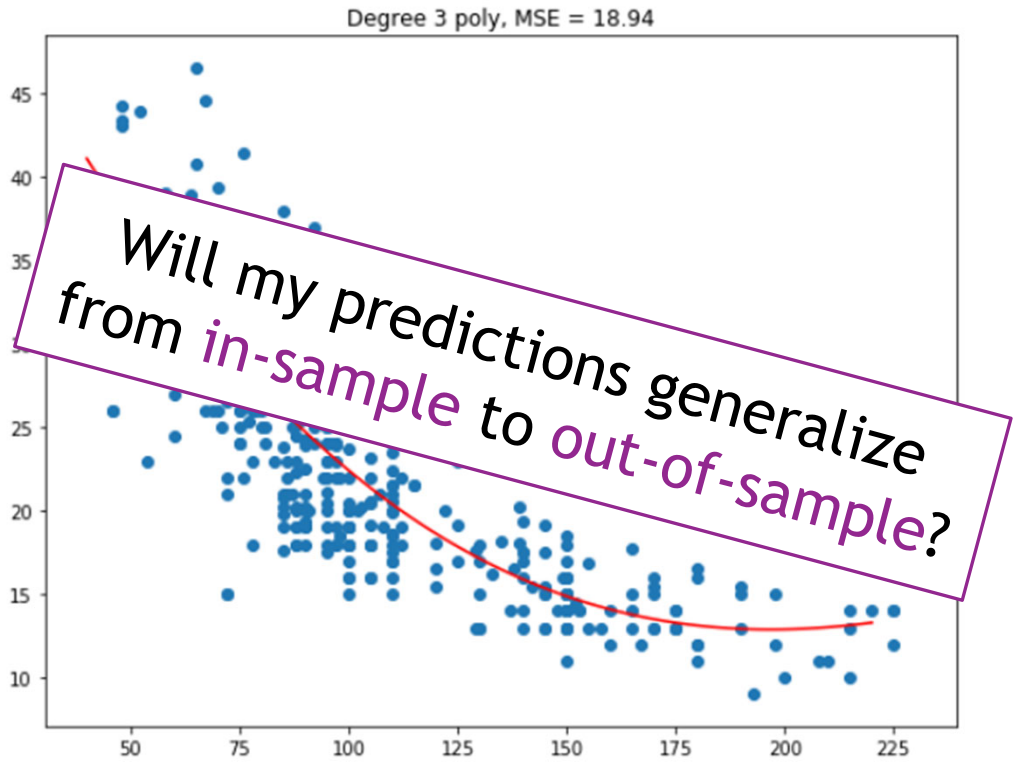
Fitting Data

- We want to extend linear models with additional features.
  - One-Hot Encoding
  - Polynomial Transformation
  - Logarithmic Transformation
- While adding explanatory variables improves the accuracy of the predictions in the sample, we need to look outside of the sample to the population. Here adding explanatory variables might worsen the accuracy of the predictions.



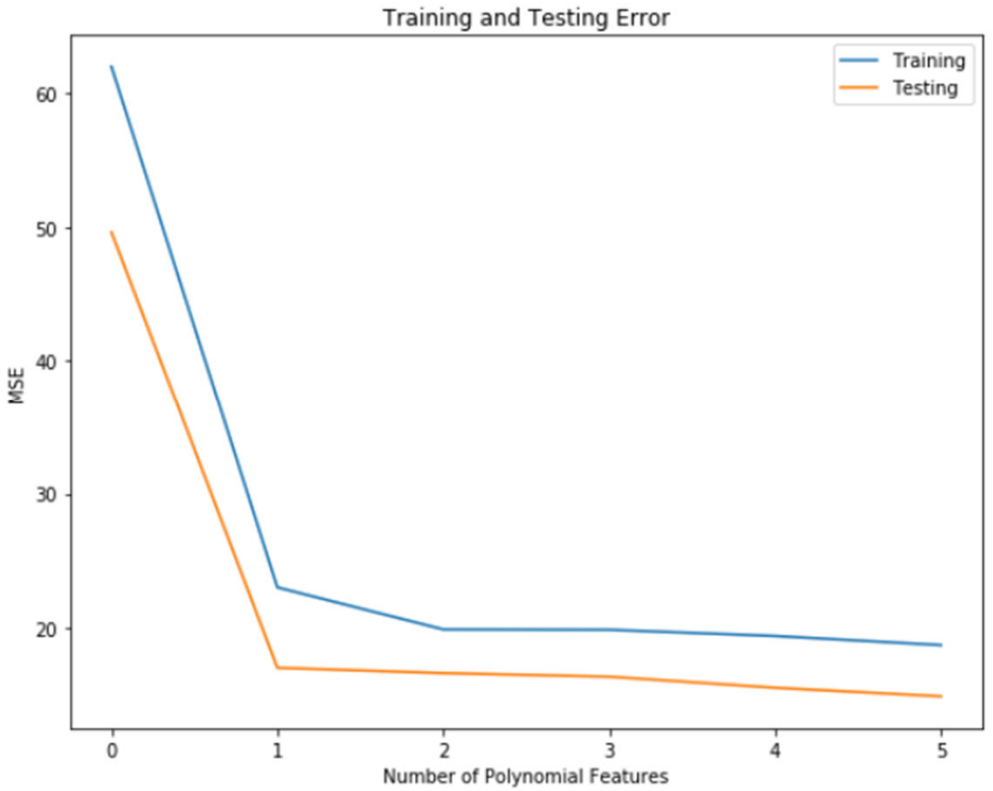
Generalization

- We want to extend linear models with additional features.
  - One-Hot Encoding
  - Polynomial Transformation
  - Logarithmic Transformation
- While adding explanatory variables improves the accuracy of the predictions in the sample, we need to look outside of the sample to the population. Here adding explanatory variables might worsen the accuracy of the predictions.



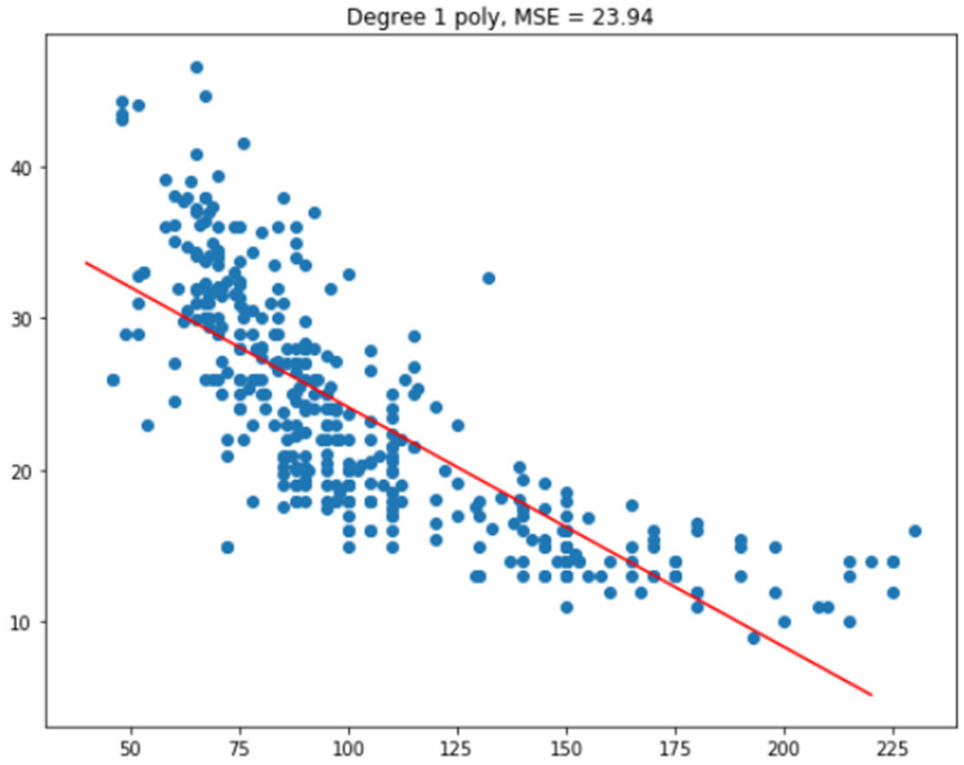
Training and Testing

- Instead of studying one sample, we need to study two samples
  - Training Set
  - 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.
- The testing set substitutes for the population helping us to access the model.



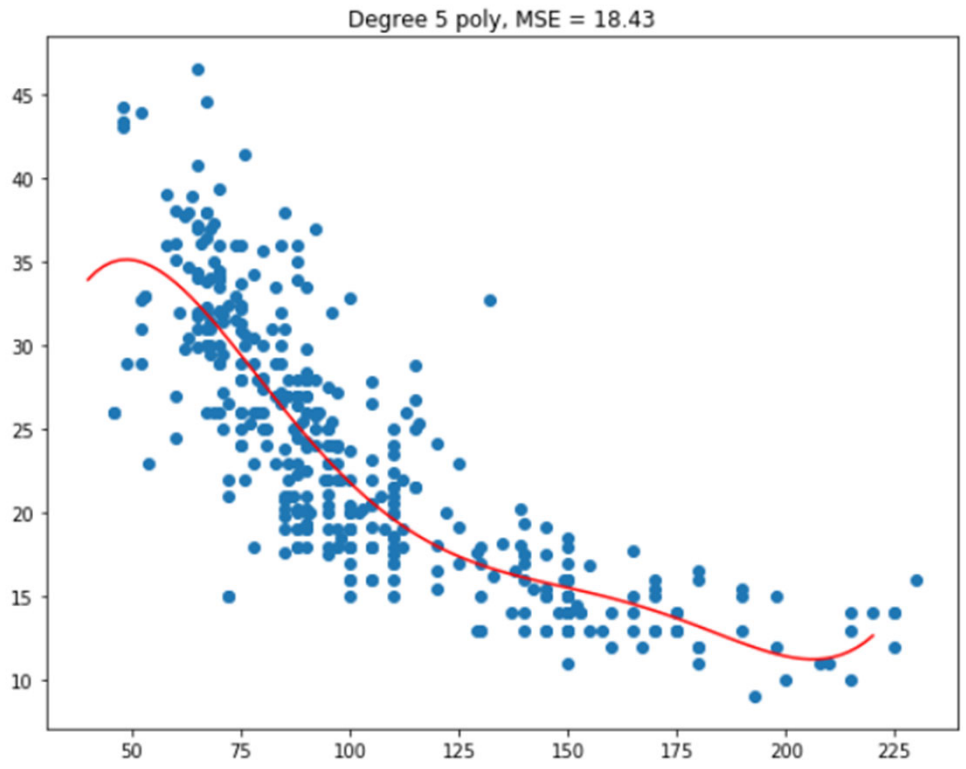
Underfitting

- If we lack features, then we may not be able to explain the response variables. We obtain inaccurate prediction on both the training set and testing set.
- With fewer features, we have few parameters. The lack of parameters means that related inputs to the prediction have related outputs without much variability



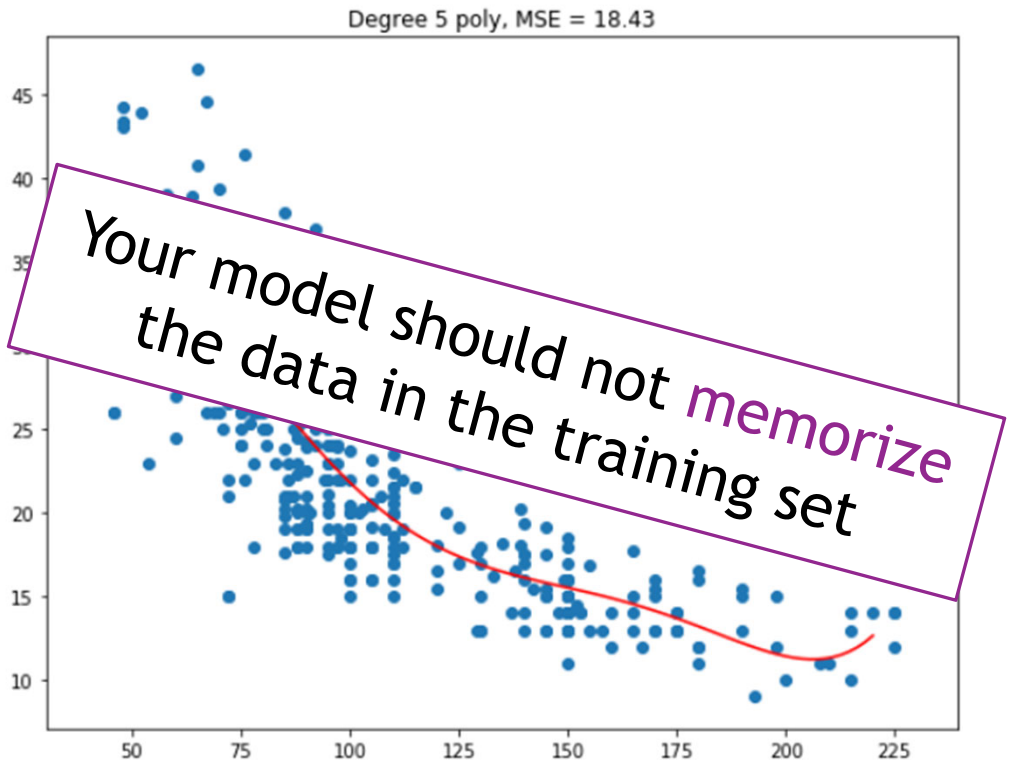
Overfitting

- If we add too many features. We obtain accurate prediction on the training set but inaccurate predictions on the testing set.
- With more features, we have more parameters. The amount of parameters means that related inputs to the prediction have unrelated outputs with a lot of variability



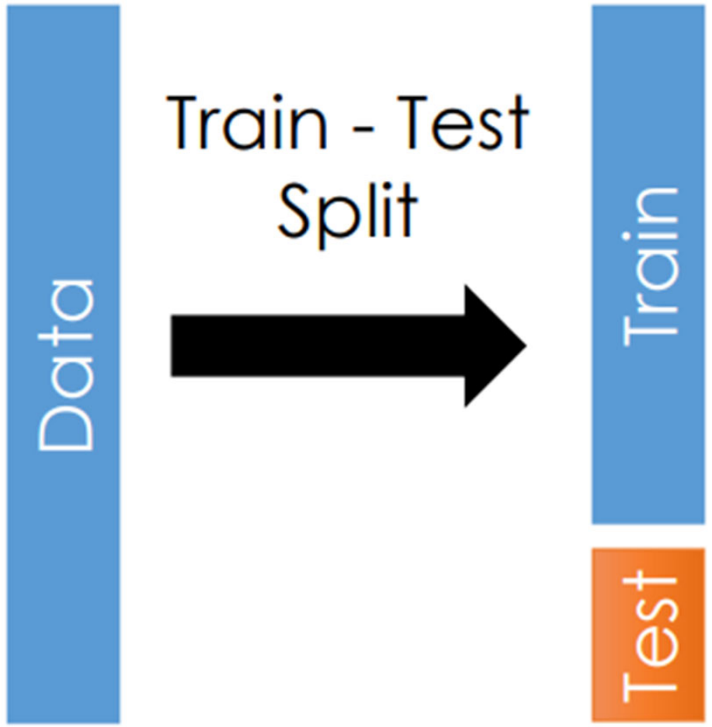
Overfitting

- If we add too many features. We obtain accurate prediction on the training set but inaccurate predictions on the testing set.
- With more features, we have more parameters. The amount of parameters means that related inputs to the prediction have unrelated outputs with a lot of variability



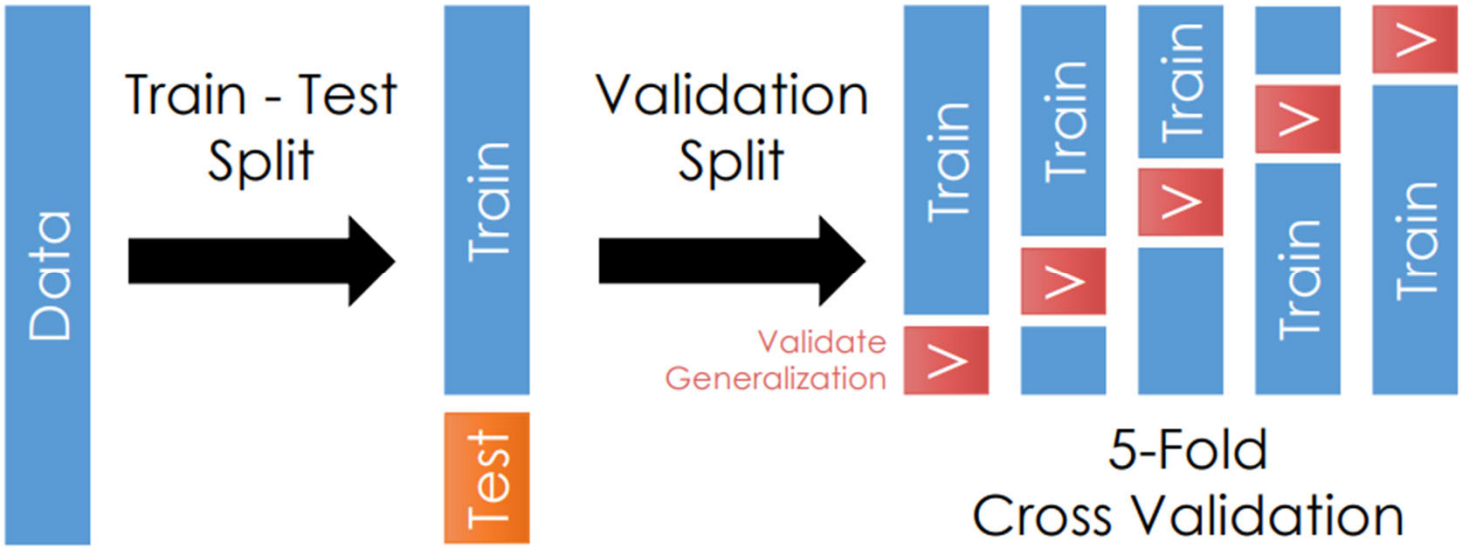
Validation

- We might not be able to generate two samples at random from a population. Instead we can split one sample into two samples
- Commonly we take 80% of the data for the training set and 20% of the data for testing set.
- We need to split a random to avoid bias in the study. So we shuffle the rows of the table before we split into two tables.



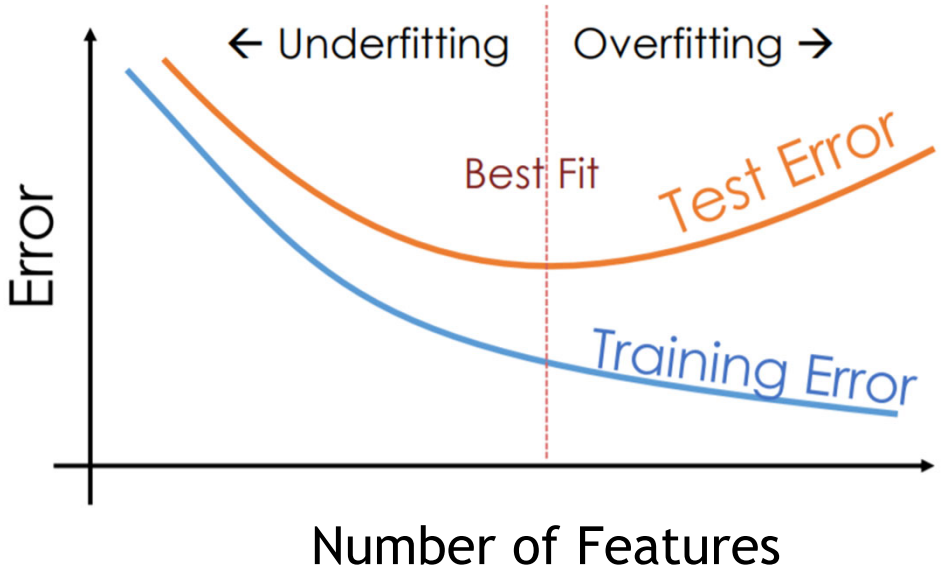
Cross Validation

- We could try to repeat validation with different folds. Each fold consists of a training set and testing set. Here we take different blocks of rows for the split.
- We split one sample into two sample to generate the testing set. Next we split one sample into k folds. Each fold has a training set and a validation set



Regularization

- The number of features equals the number of parameters. With few parameters we have a simple model. With many parameters we have a complicated model.
- We want to include many features to avoid underfitting. However, we need to limit the number of features to avoid overfitting.
- With regularization we somehow add the number of features to the average loss in the search for parameters



We add an extra parameter to trade-off between underfitting and overfitting

Summary

- Validation
  - Training Set
  - Testing Set
- Number of Features
  - Underfitting
  - Overfitting
- Regularization

Goals

- Randomly split a sample into training set and testing set
- Understand the difference between overfitting and underfitting
- Compare models with different combinations of features

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