DATA SCIENCE: REGRESSION MODELS

André R. Brodtkorb

TODAYS TOPIC

- Interpolating data with polynomial interpolation
- Approximating data with regression models
- Training and test datasets

MOTIVATION

What is a (my) home worth today?

All other similar estimates

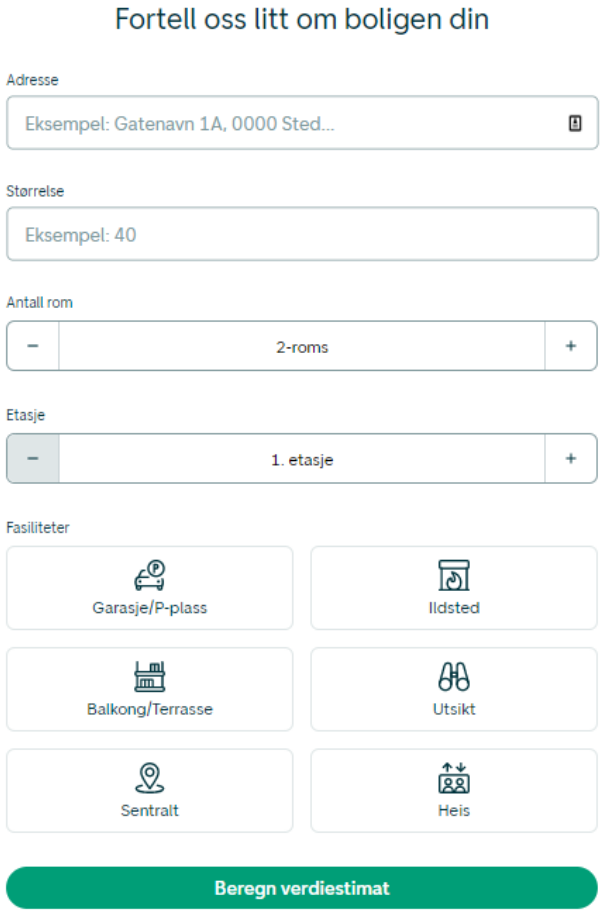
– electricity price

– How effective will a drug be for a patient?

– …

samsolgt

by DNB Eiendom



Fortell oss litt om boligen din

Adresse

Eksempel: Gatenavn 1A, 0000 Sted...

Størrelse

Eksempel: 40

Antall rom

− 2-roms +

Etasje

− 1. etasje +

Fasiliteter

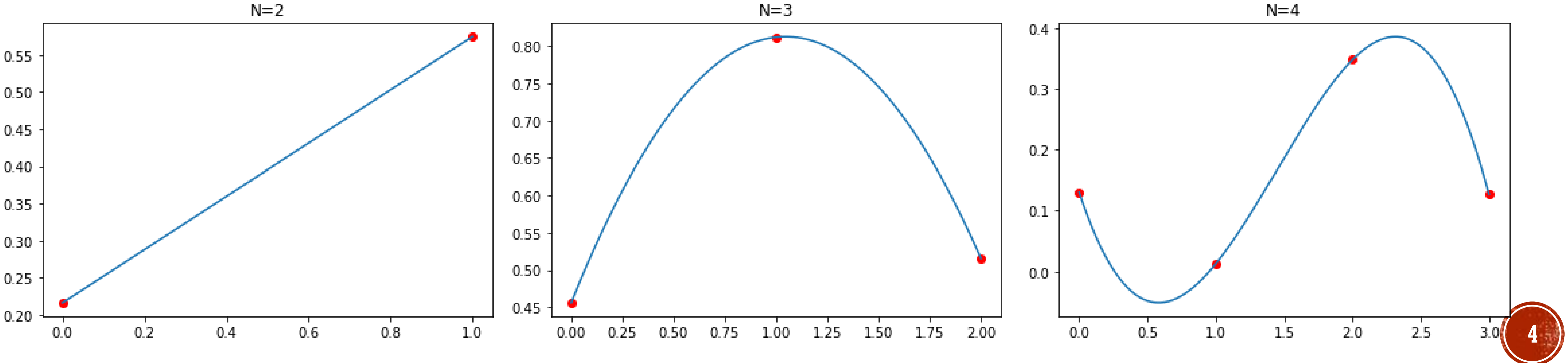
Garasje/P-plass	Ildsted
Balkong/Terrasse	Utsikt
Sentralt	Heis

Beregn verdiestimat

(figure from https://dnbeiendom.no/altombolig/samsolgt/se-hva-boligen-din-er-verdt--her-og-na)

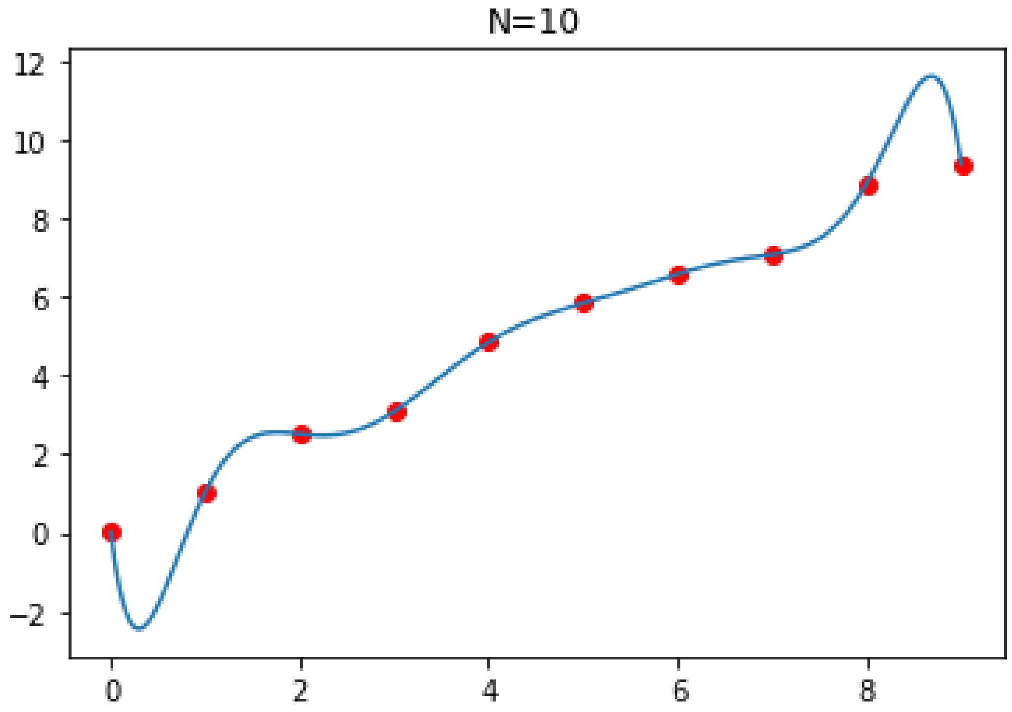
POLYNOMIAL INTERPOLATION

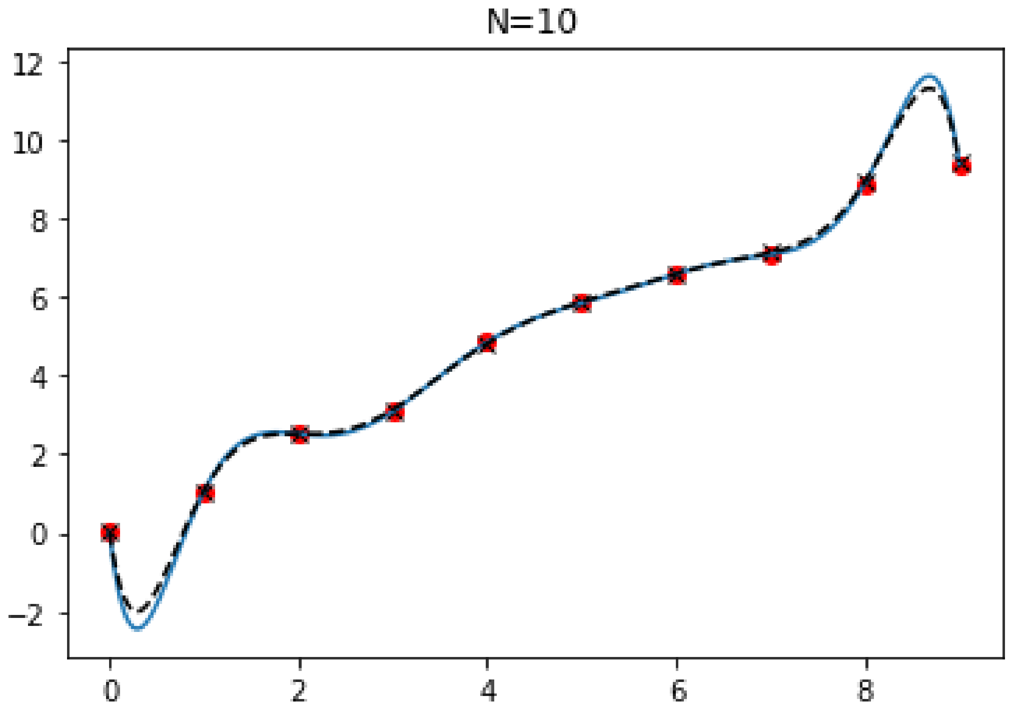
- For n data points, we can find a degree n-1 polynomial that interpolates all data points
  - Two points: line (f(x) = ax + b)
  - Three points: parabola (f(x) = ax^2 + bx + c)
  - Four points: cubic function (f(x) = ax^3 + bx^2 + cx + d)
  - …



PROBLEMS

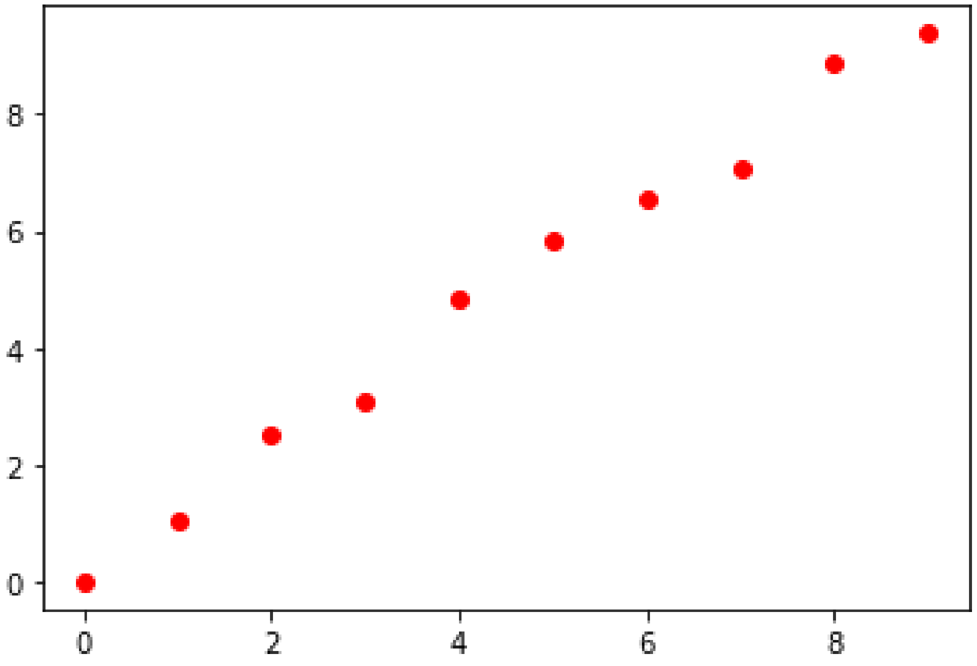
- Polynomial interpolation is unstable for large n
- Sensitive to noise

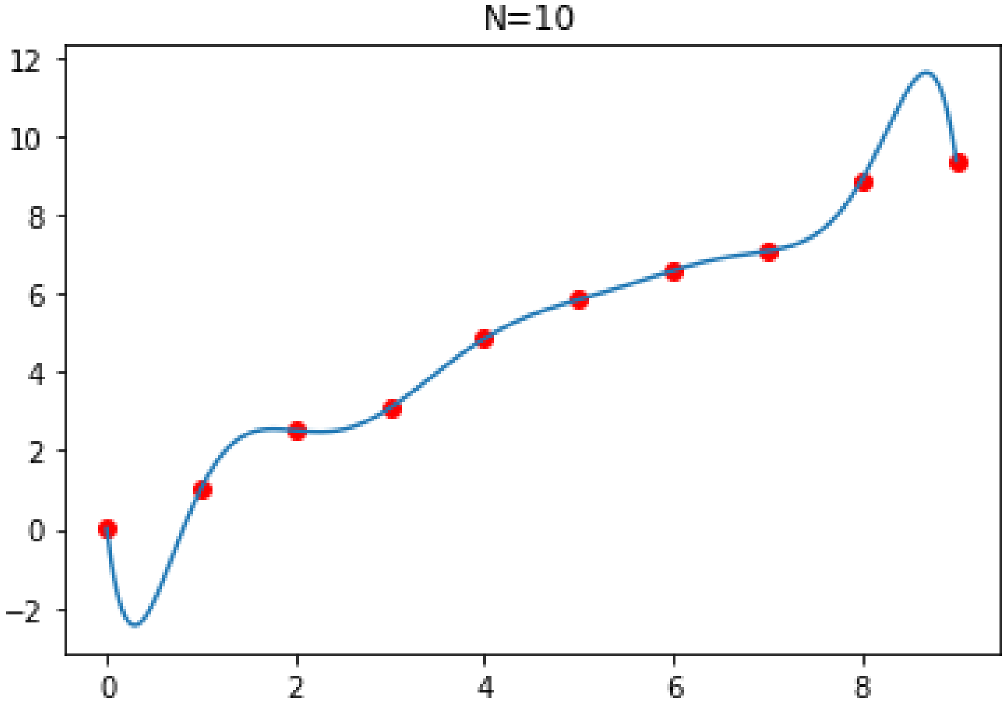




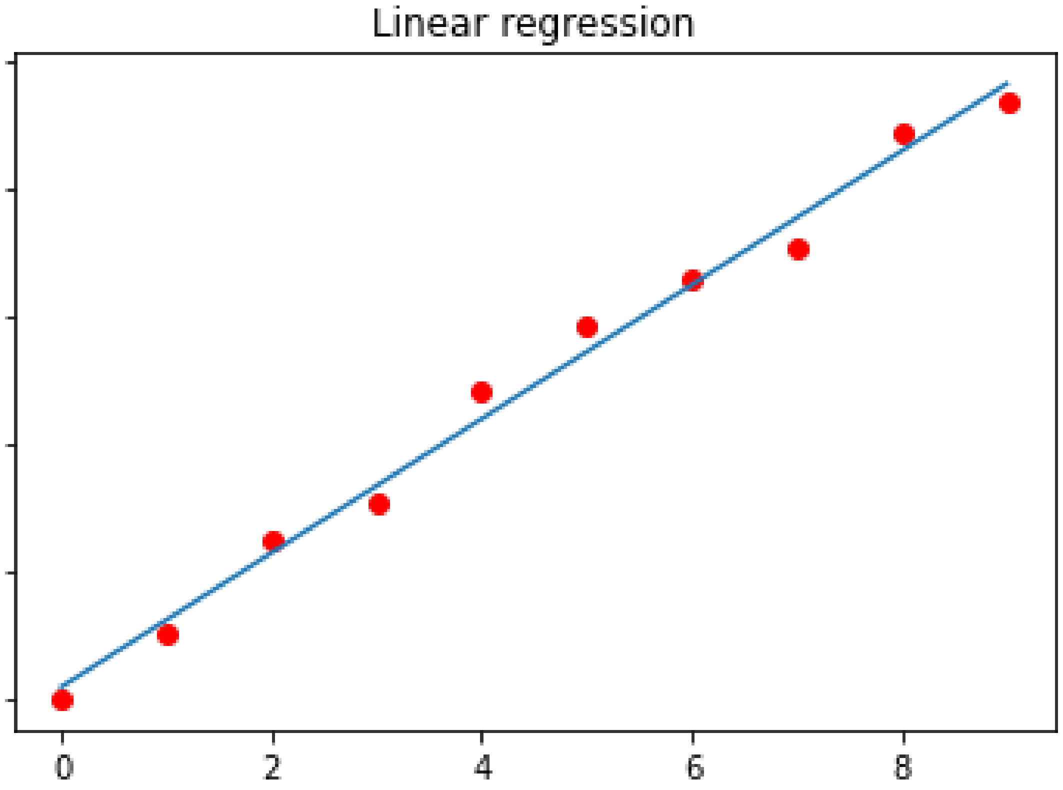
IS IT A GOOD REPRESENTATION?

- Is a degree 10 polynomial a good fit for our data?





APPROXIMATING DATA



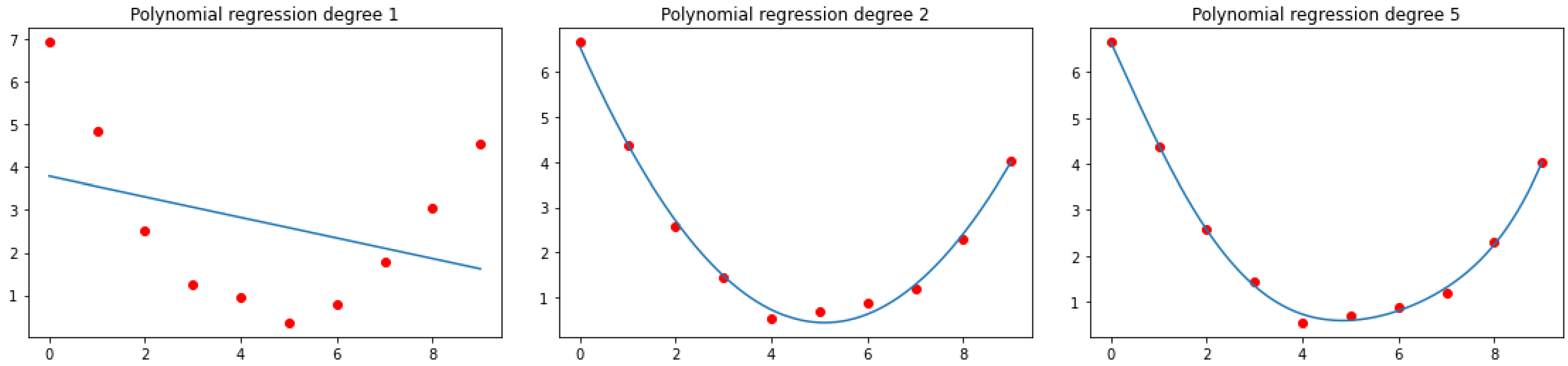
- Instead of interpolating all values, we can find a function that approximates our data (also called regression analysis)
- Our data "looks" linear, lets try linear regression

JUPYTER NOTEBOOK

- Linear regression in Python

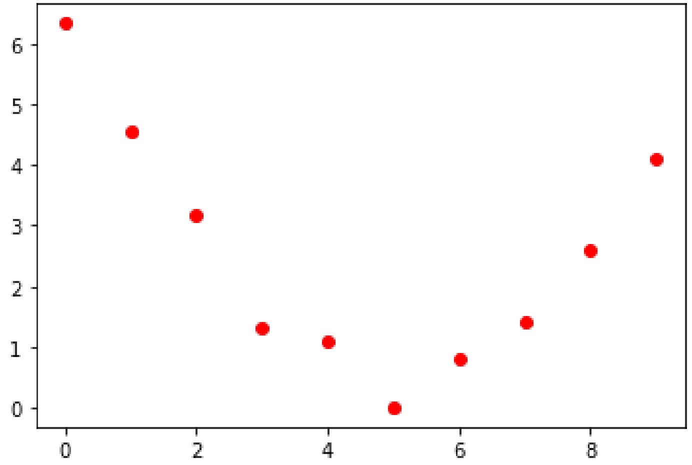
CHOOSING A GOOD MODEL FOR OUR DATA

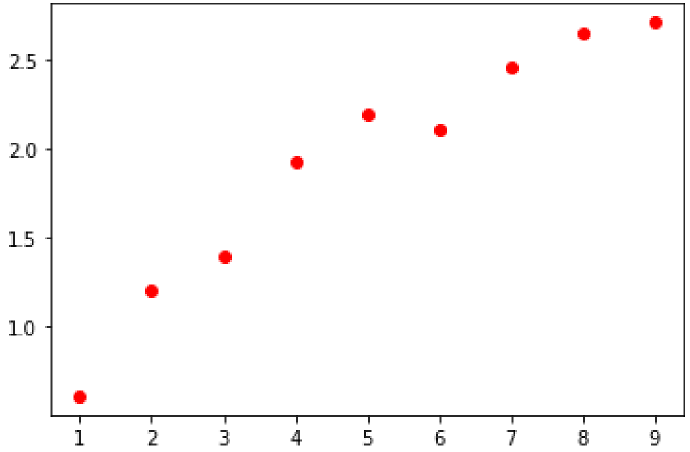
- Need to inspect data
- Need an educated guess on what type of model should fit our data "well"
- "Easy" for one-dimensional data, very difficult for 4D or higher.

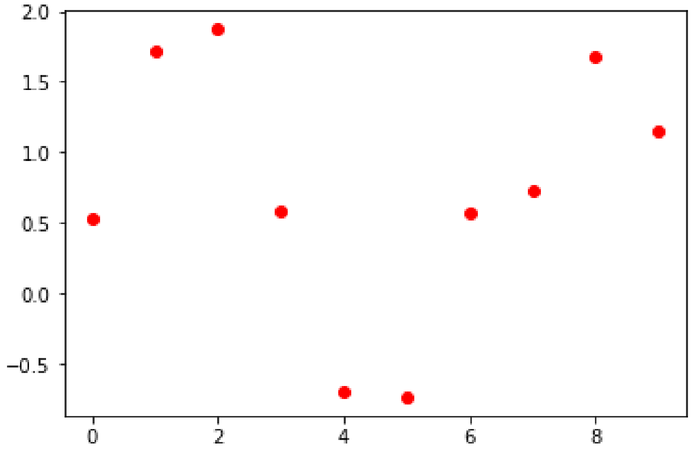


WHAT DOES OUR DATA LOOK LIKE?

- X^2, log(x), sin(x), …?

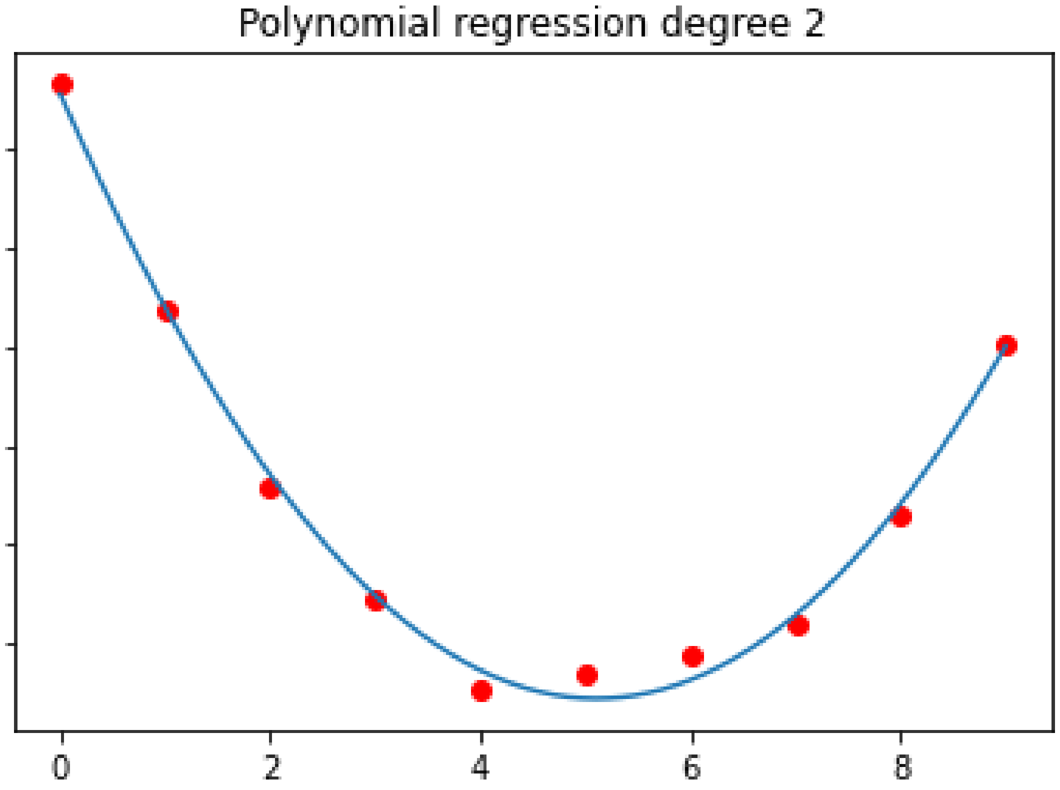






- Sometimes it is difficult to determine or unknown!

QUANTIFYING THE ERROR

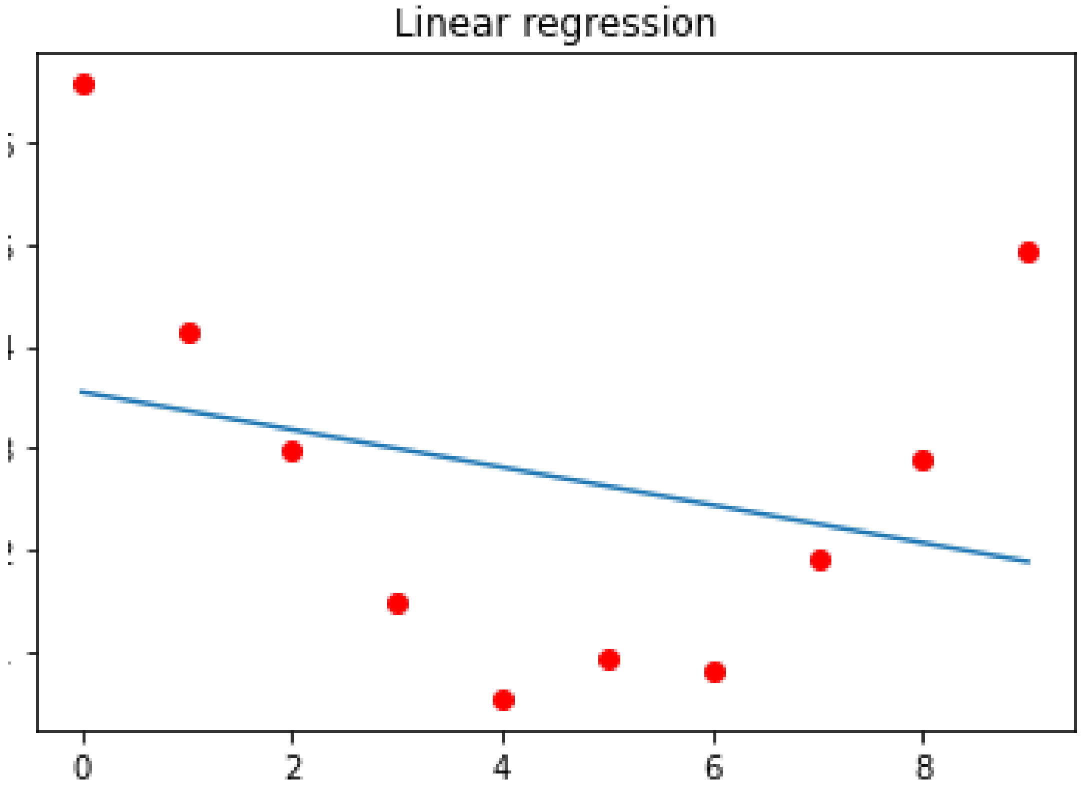


- Mean average error (MAE)
  - Average of absolute difference between prediction and observation
- Mean squared error (MSE)
  - Average of square of difference between prediction and observation
- Root mean squared error (RMSE)
  - Square root of mean squared error
- (more as well, see scikit.learn model evaluation for example)

JUPYTER NOTEBOOK

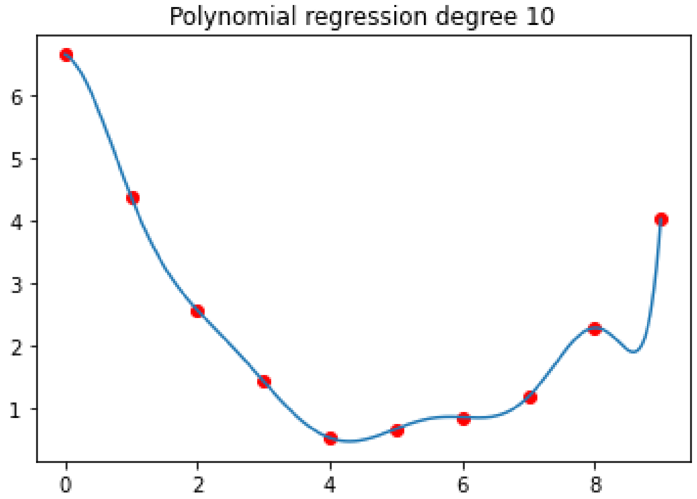
- Score of our linear regression example

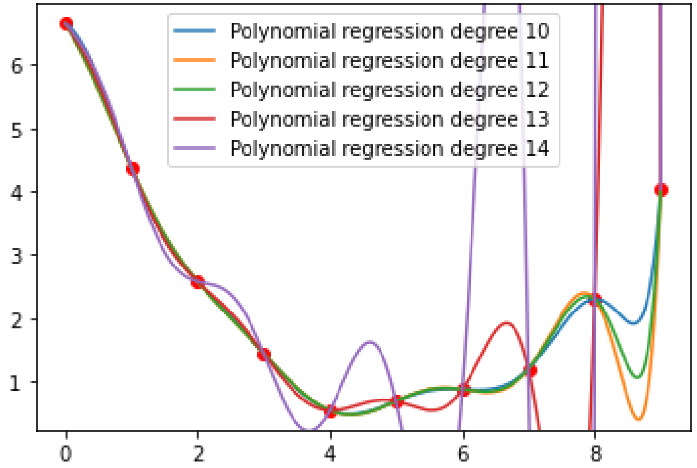
UNDERFITTING



- Underfitting happens when we have a too simple model
- Example: Using a linear model to predict nonlinear behaviour
- Symptoms: poor predictive skill, even on the data we try to fit

OVERFITTING

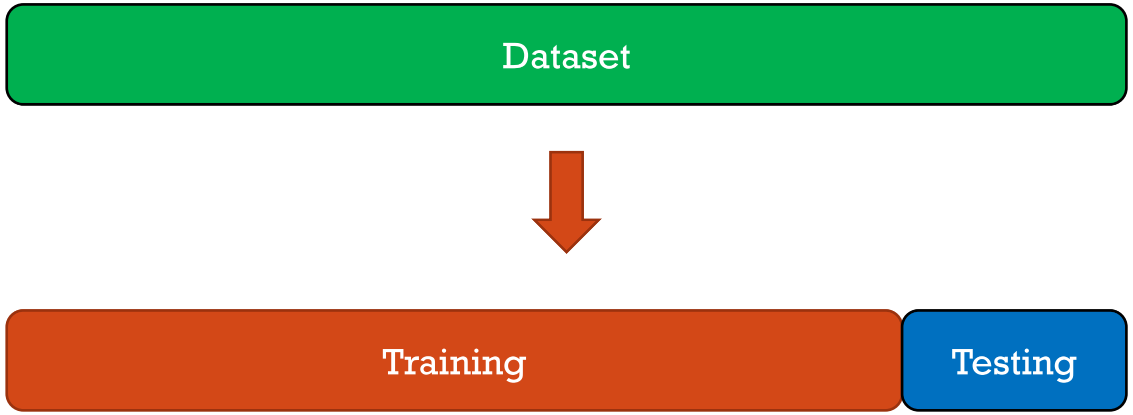


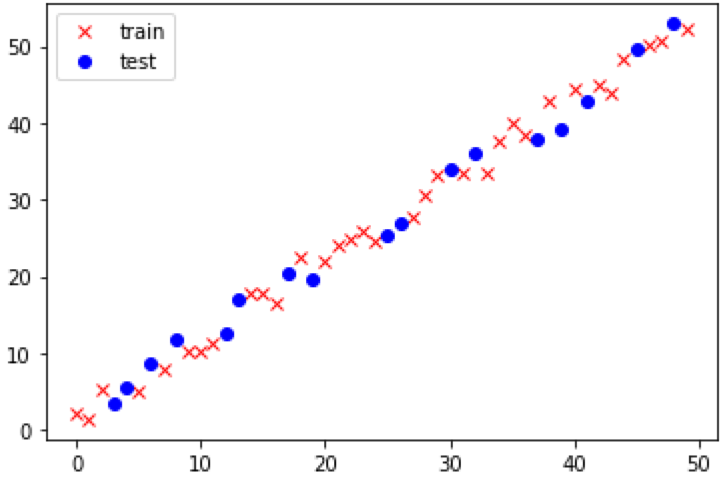


- Overfitting is when we have too much freedom in our model
- Example: Using a polynomial of degree n-1 for n data points (interpolation)
- Symptoms: Model is extremely good at predicting known data, but terrible at predicting new data

TESTING THE MODEL

- So far, we have tested the model on data that it's already "seen"
- This is not a very good way to quantify model performance
- In machine learning, the dataset is usually divided into train and test subsets



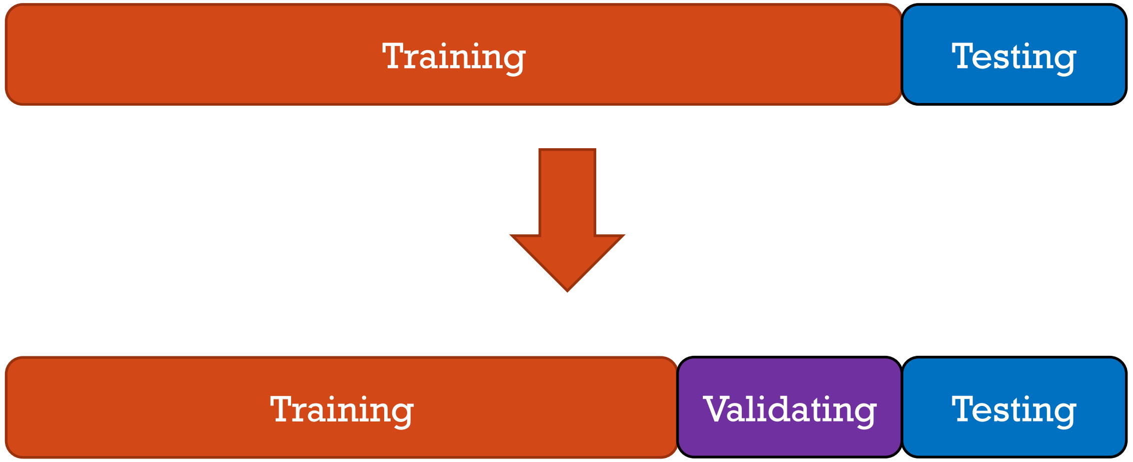


JUPYTER NOTEBOOK

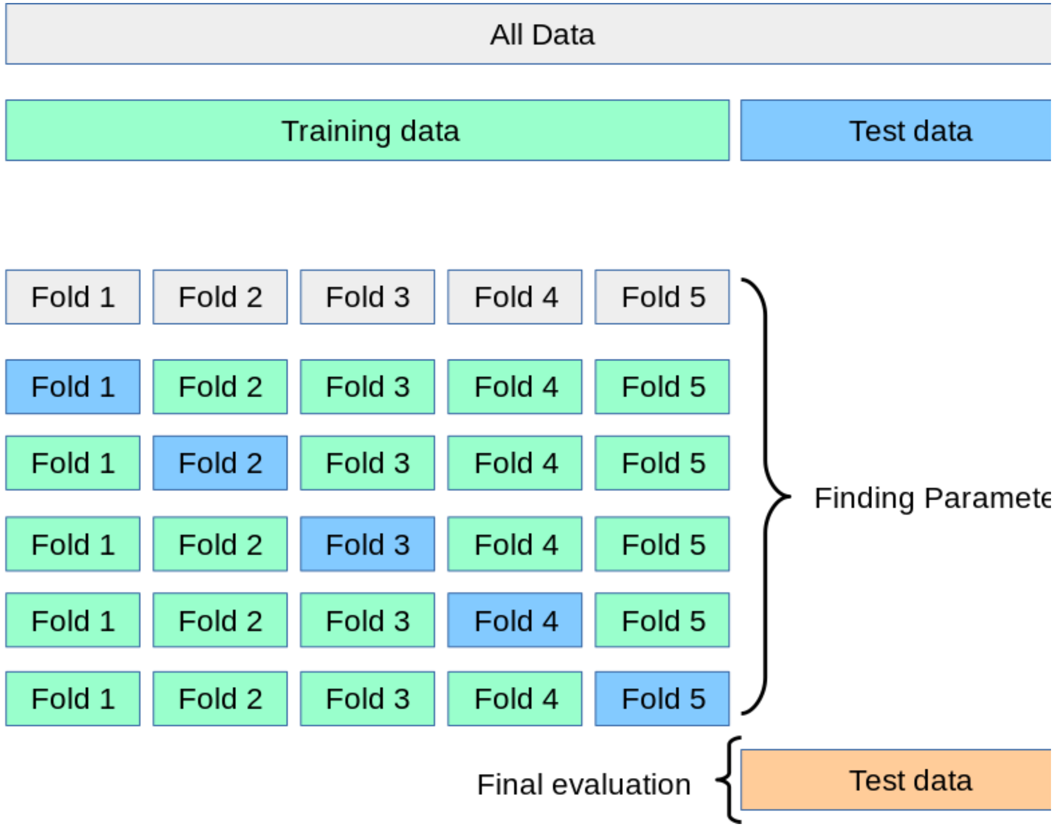
- Testing model performance on test dataset

VALIDATION

- Validation data is used to check model performance and set hyperparameters
- Model may "see" the validation data through performance feedback
- Testing data is still not part of training



K-FOLD CROSS-VALIDATION


All Data
Training data
Test data
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Fold 1 Fold 2 Fold 3 Fold 4 Fold 5
Finding Paramete
Final evaluation
Test data

- Divide data into k subsets
- Train k models, using a different subset as test data for each model
- Use the rest of the data for training
- Evaluate on separate test dataset

Image from https://scikit-learn.org/stable/modules/cross_validation.html

SUMMARY

- Polynomial interpolation does not scale
  - Sensitive to noise and high order

- Regression models approximate data
  - Check for underfitting and overfitting and find the sweetspot in between

- Testing and training datasets
  - K-fold cross-validation

- Source code on github: https://github.com/babrodtk/
- Slides on webpage: https://brodtkorb.org/

BONUS: BOOTSTRAPPING

- Assume you have population you want to model
- Create a "sample" (subset) of size n
- Pick n data points (with replacement) from your population to create a "bootstrap sample"
- Fit a model to each bootstrap sample
- Average models for prediction

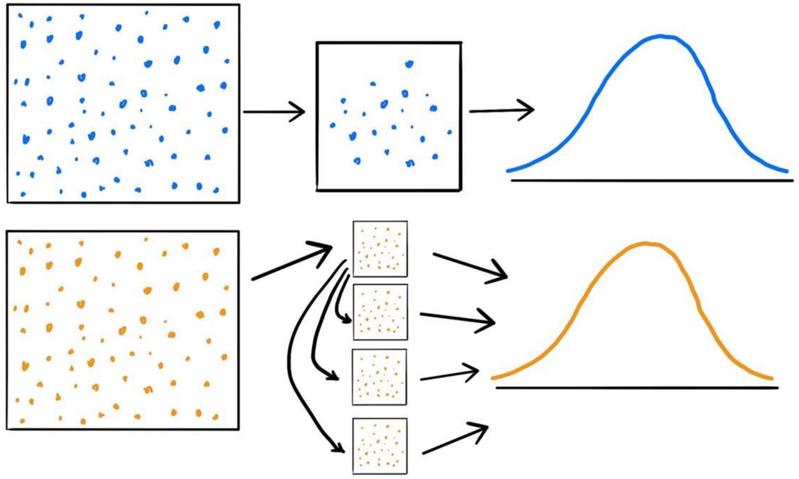


Image from https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307