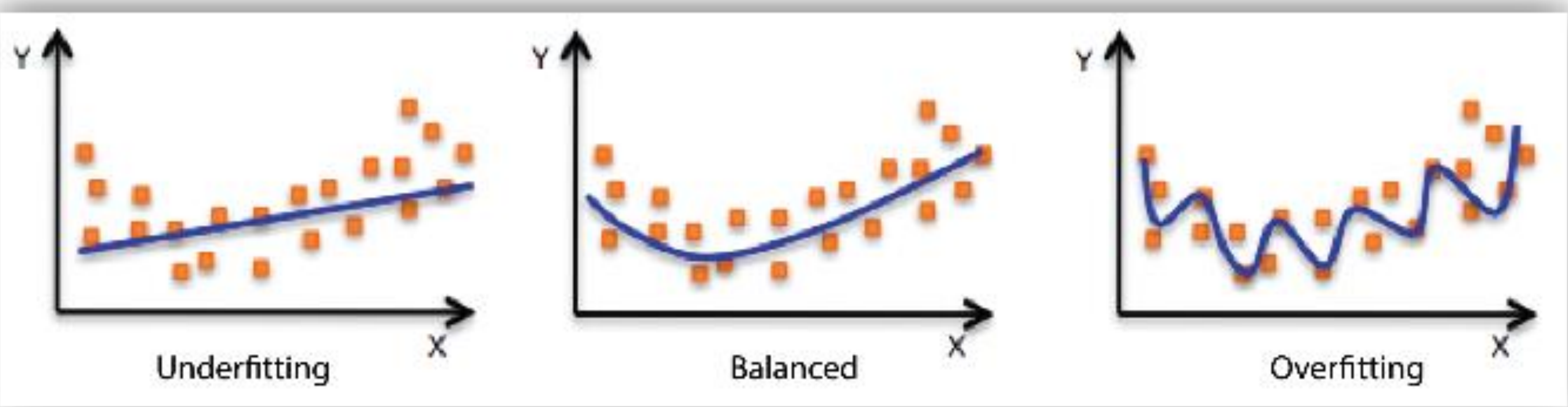
Model fitting challenges

Understanding model fit

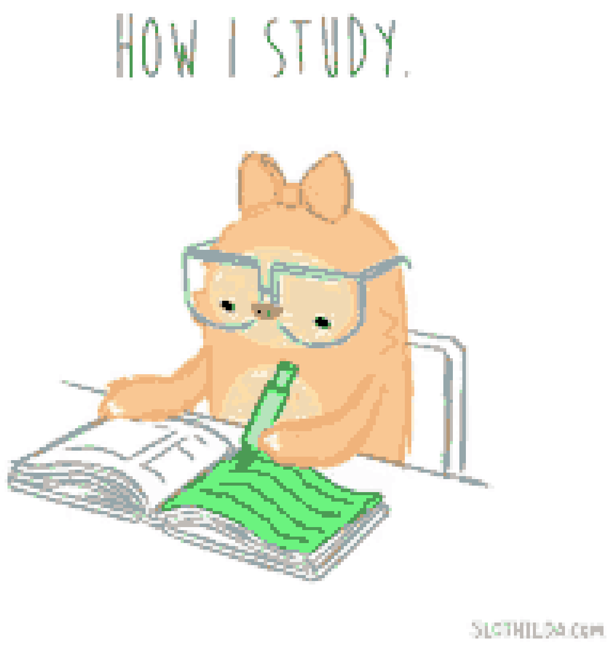


Underfitting/Overfitting is one of the main causes for poor model accuracy

Underfitting

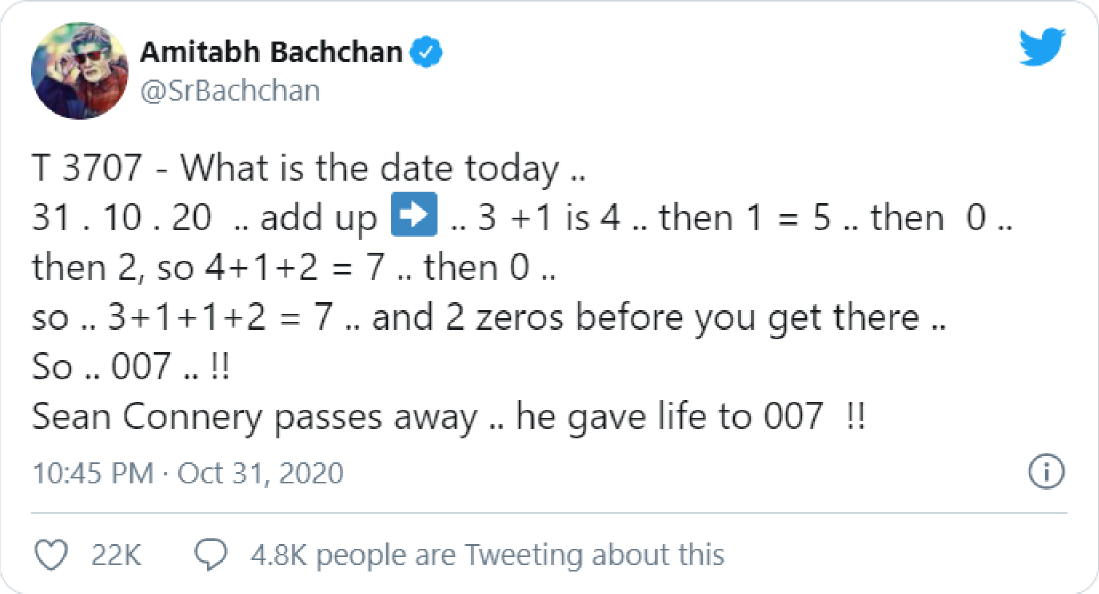
Not modelling the data well enough!

- Inadequate learning while training for the exam
- Consequently, bad performance in the exam



Overfitting

Fitting data at all costs!



Amitabh Bachchan
@SrBachchan

T 3707 - What is the date today ..
31 . 10 . 20 .. add up ➡ .. 3 +1 is 4 .. then 1 = 5 .. then 0 ..
then 2, so 4+1+2 = 7 .. then 0 ..
so .. 3+1+1+2 = 7 .. and 2 zeros before you get there ..
So .. 007 .. !!
Sean Connery passes away .. he gave life to 007 !!

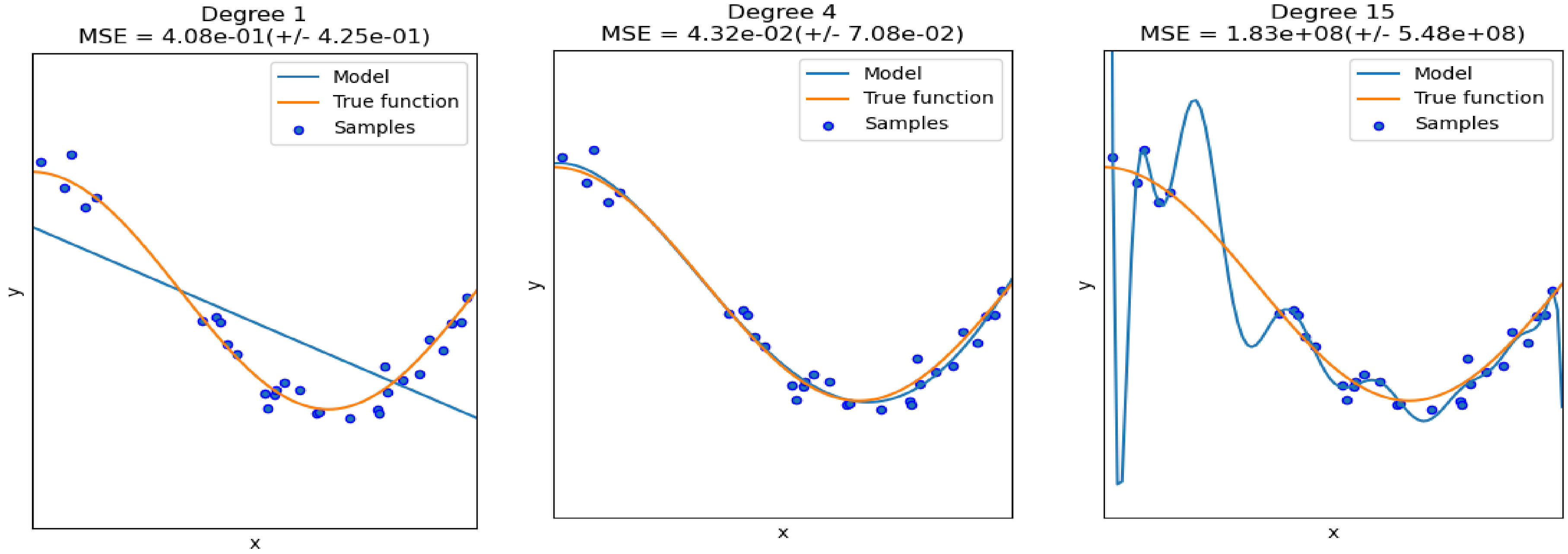
10:45 PM · Oct 31, 2020

22K 4.8K people are Tweeting about this

Overfitting

$$y = \beta_0 + \beta_1 x$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n .$$



Overfitting

linear regression (polynomial degree = 1)

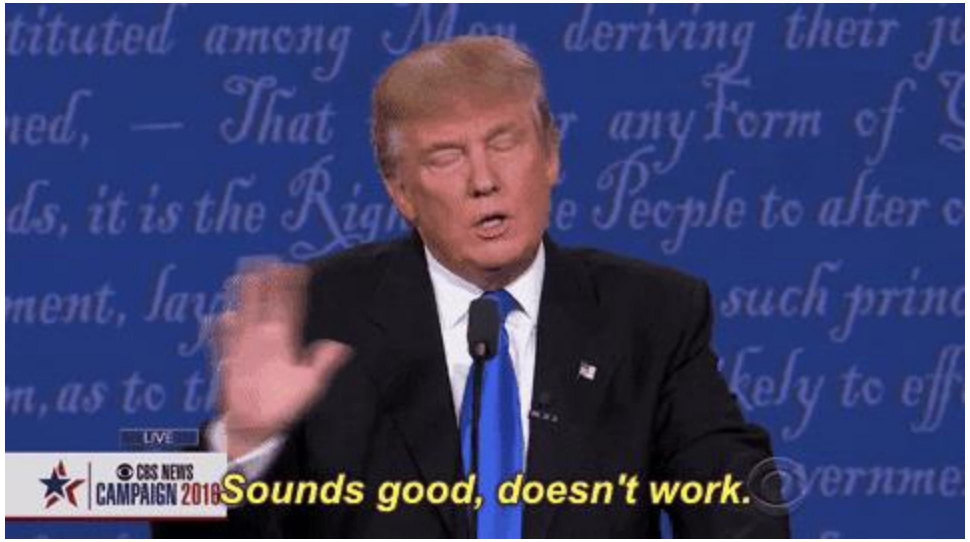
linear regression with higher degree polynomial features to approximate nonlinear functions

Detecting Overfitting

The key indicator

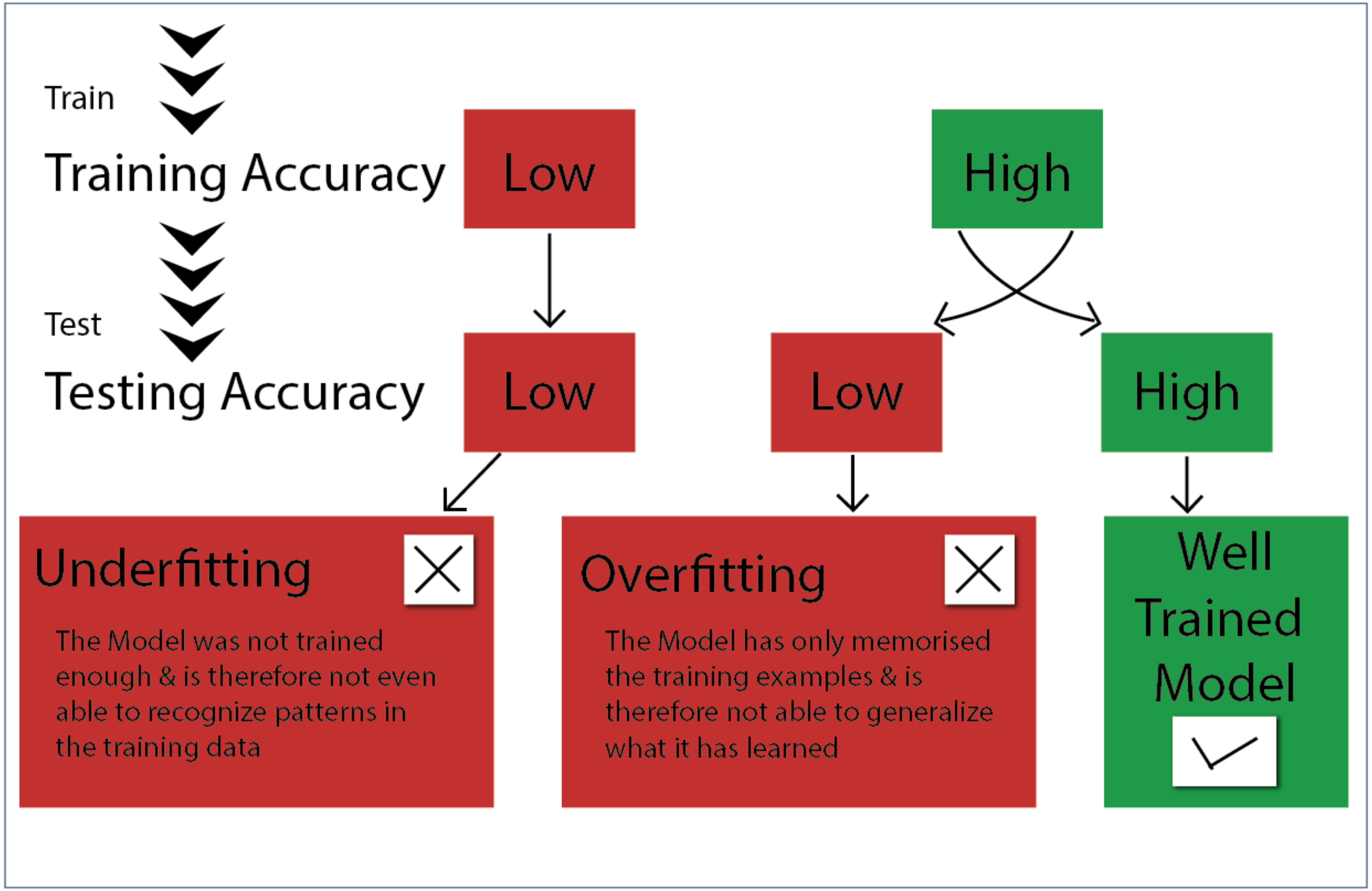
TRAINING ACCURACY

TESTING ACCURACY



Detecting Underfitting/Overfitting

Summary



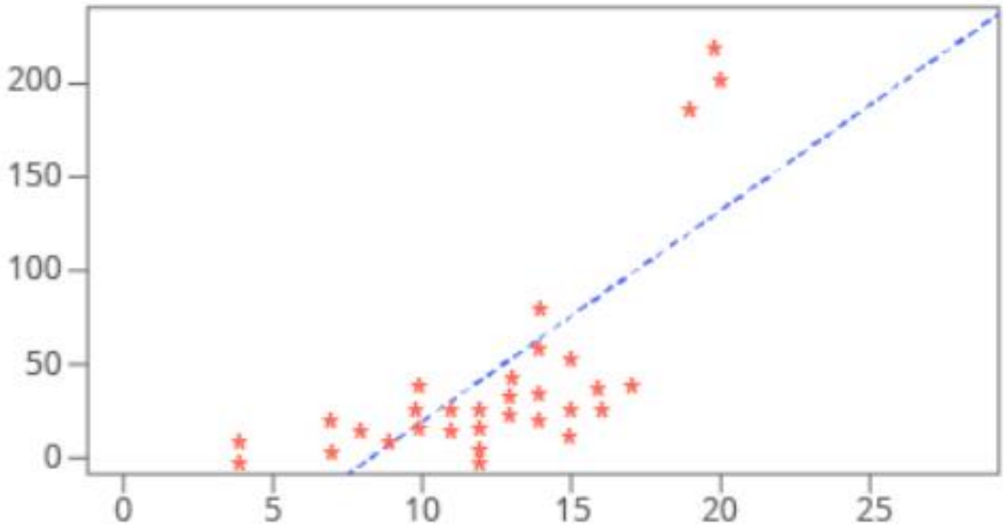
Reasons for Underfitting

- Short training
- Simplistic model
- Lack of enough features

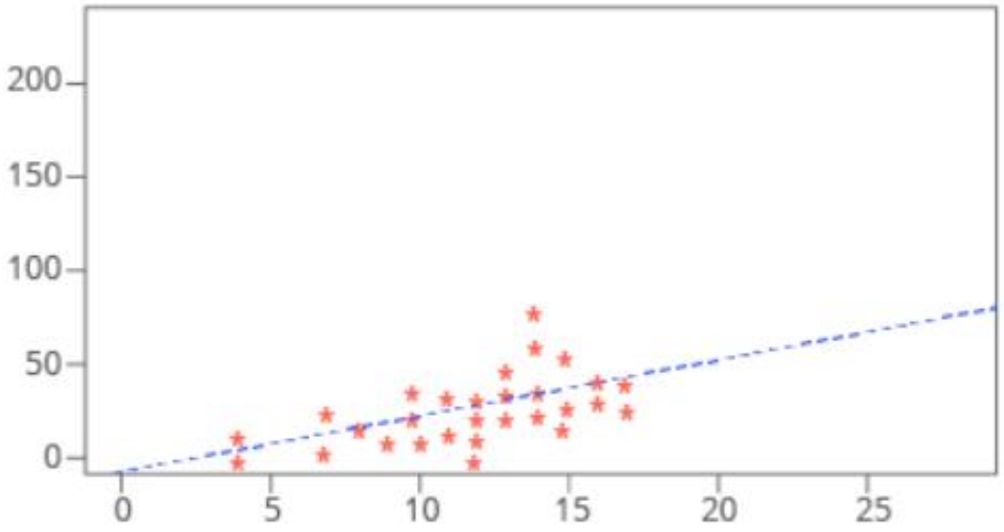
Reasons for Overfitting

- Learning noise in the data
- Using a complex model for a simple task
- Inappropriate model hyperparameters

Reasons for Overfitting
Learning noise in the data



With outliers



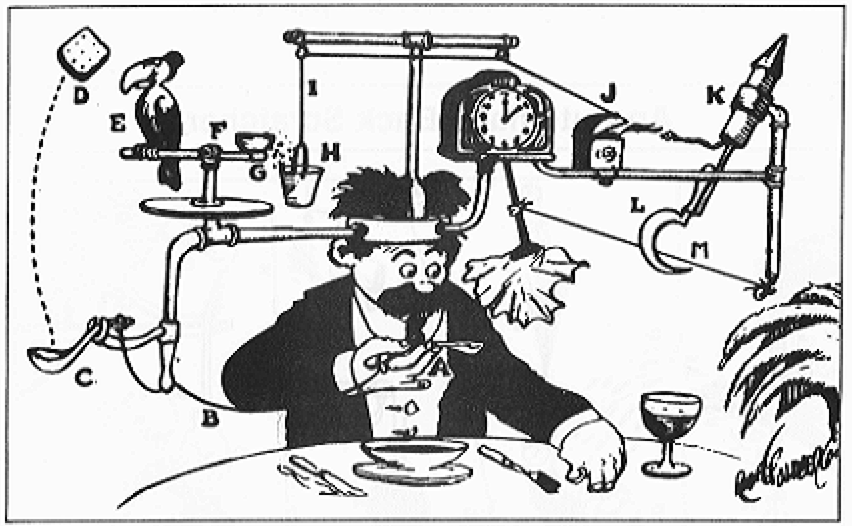
Without outliers
A much better fit

https://www.knowledgehut.com/blog/data-science/overfitting-and-underfitting-in-machine-learning

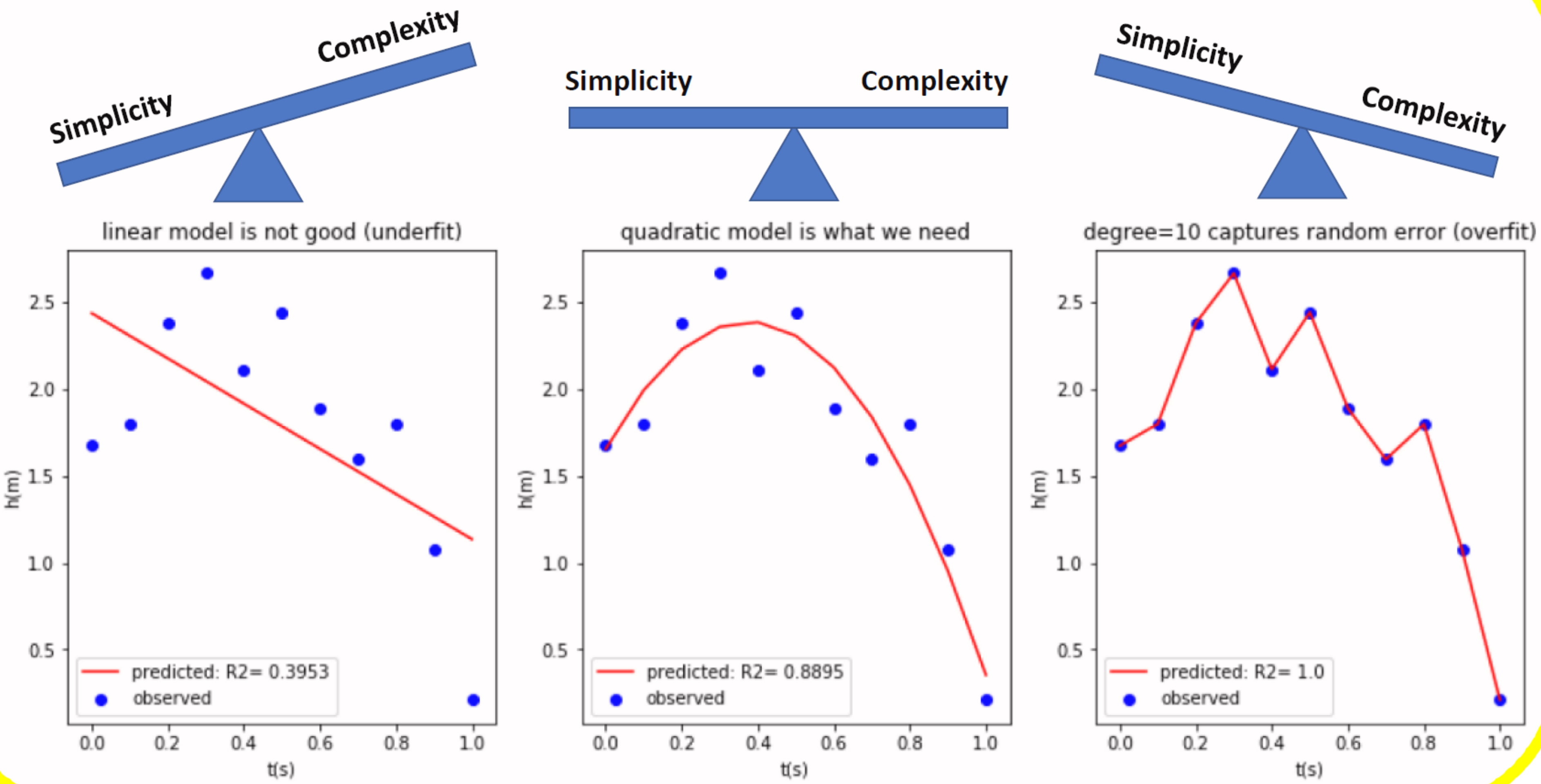
Reasons for Overfitting

Using a complex algorithm for a simple task

Self-Operating Napkin



A **Rube Goldberg machine**, named after American cartoonist Rube Goldberg, is a machine intentionally designed to perform a simple task in an indirect and overly complicated way.



https://towardsdatascience.com/simplicity-vs-complexity-in-machine-learning

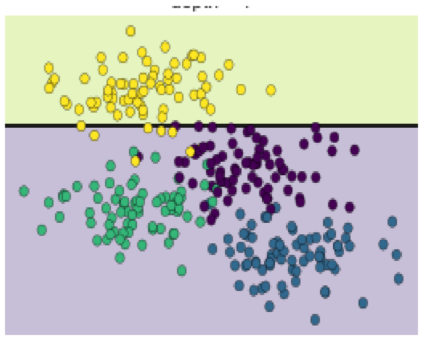
Reasons for Overfitting

Inappropriate model hyperparameter(s)

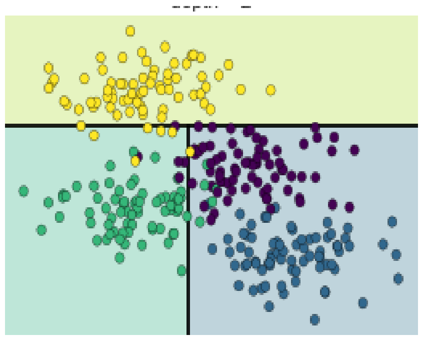
A simple decision tree built on this data will iteratively split the data along one or the other axis according to some quantitative criterion

With default max_depth = None, the nodes are expanded until all leaves are pure

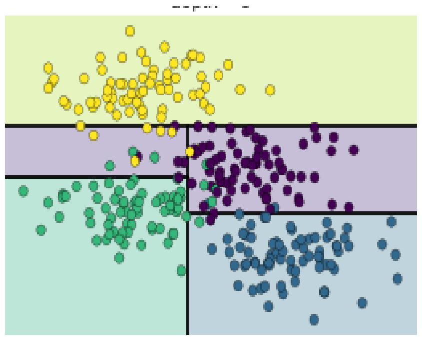
Depth = 1



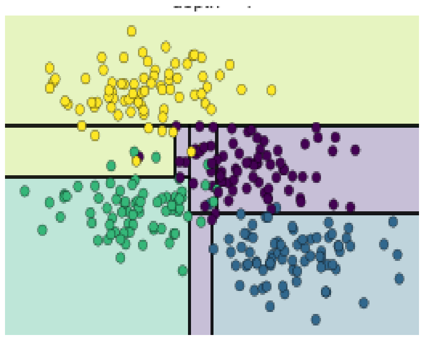
Depth = 2



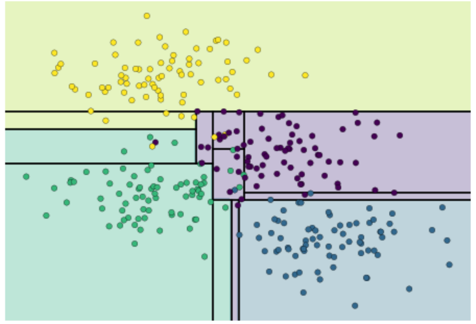
Depth = 3



Depth =4



Depth = 5



……

Python Data Science Handbook by Jake VanderPlas

Overfitting Example

Hyperparameter tuning – Decision tree

- **Demo_Overfitting_Diabetes**
  - Dataset: Diabetes dataset
  - Objective: To understand the influence of tuning hyperparameters to avoid Overfitting

Addressing Overfitting

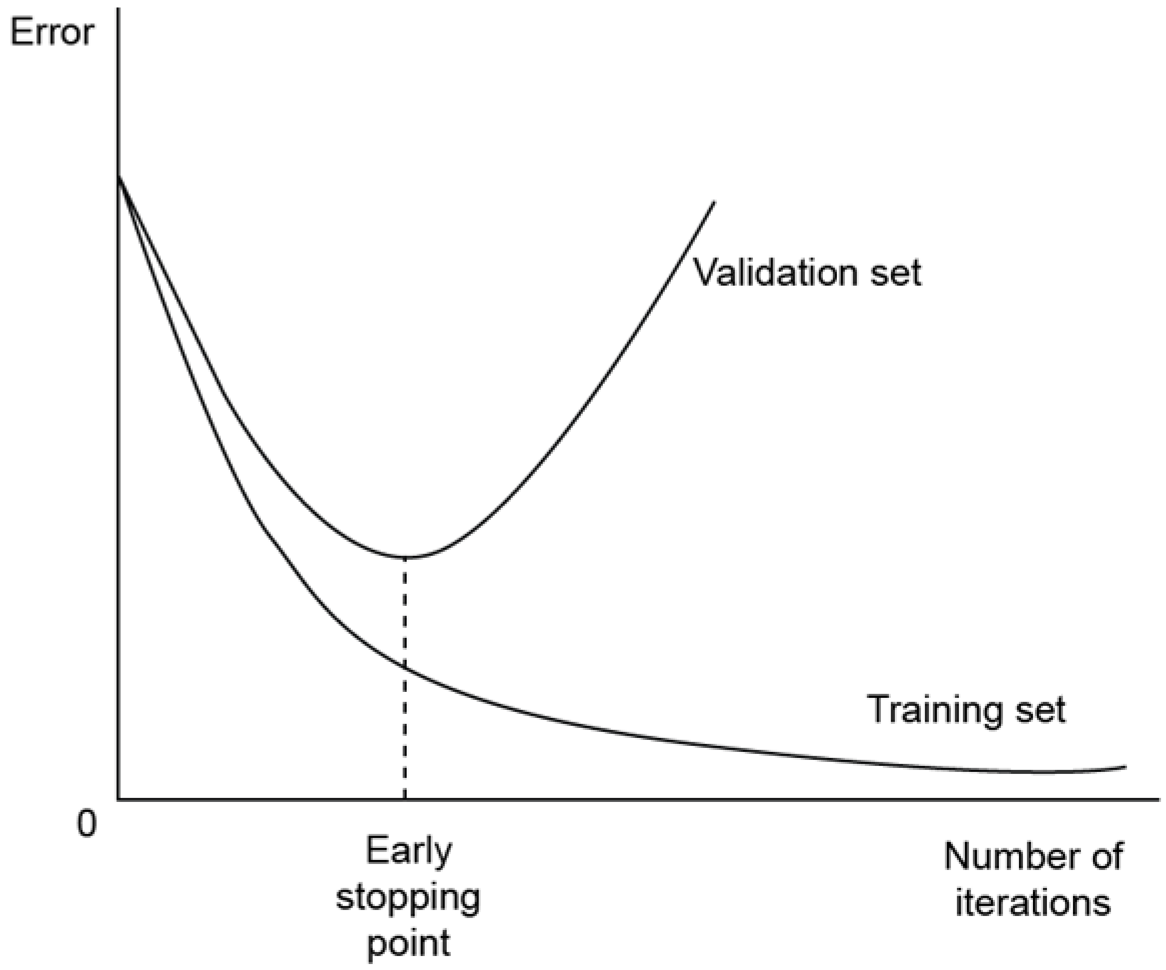
- Use More Data → Can learn more parameters

- Reduce the number of features or dimensions
  - Select which features to keep
  - Reduce dimension
  - Select which model to use

How to prevent/reduce overfitting?

- Train with more data
- Reduce/Remove features
- **Early Stopping**
- **Regularization**
- Ensembling

Early Stopping



Regularization

Regularization is a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.

What to Minimize? (regularization)

- We were minimizing:

$$\varepsilon = \sum_i (z_i - y_i)^2 = \sum_i (z_i - f(x_i))^2$$

- We can minimize

  Fit error + some measure of complexity of the model

  ε + Sum of squares of Weights

  ε + Number of Non-zero Weights

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

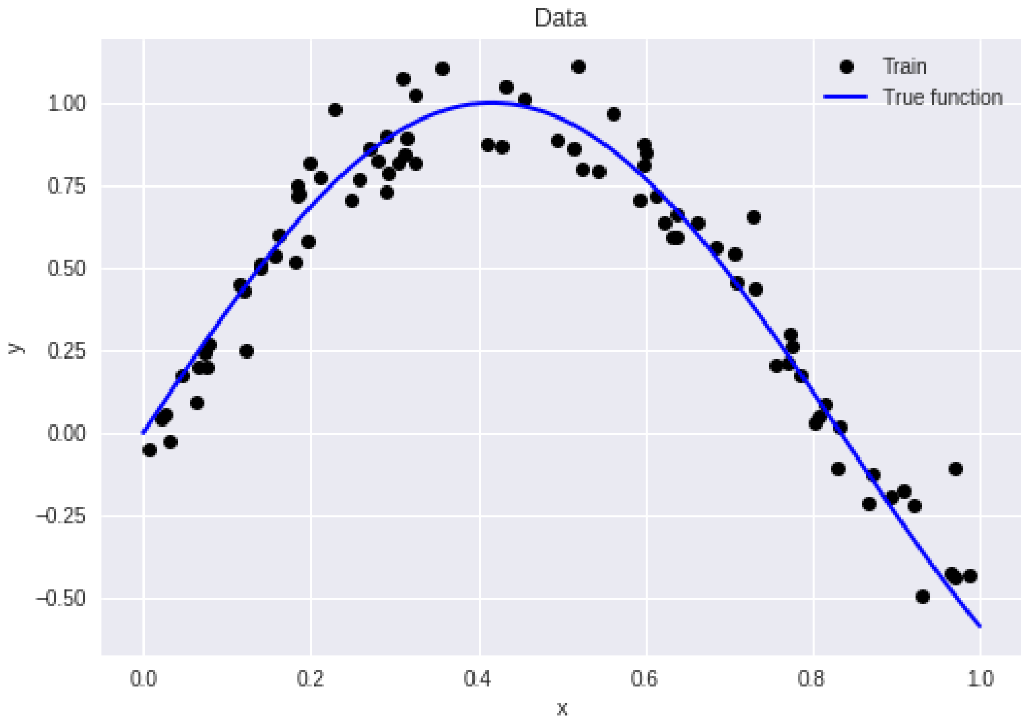
Ridge Regression (L2-Norm), Lasso Regression (L1-norm) - Least Absolute Shrinkage and Selection Operator

Effect of Regularization

Fit a curve to the data

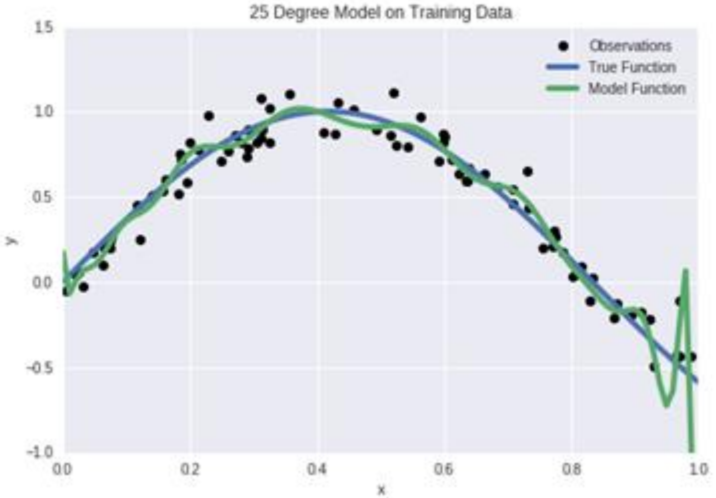
- Learn W to fit the following polyn regression

$$y = w_{25}.x^{25} + w_{24}.x^{24} + .. + w_1.x + w_0$$

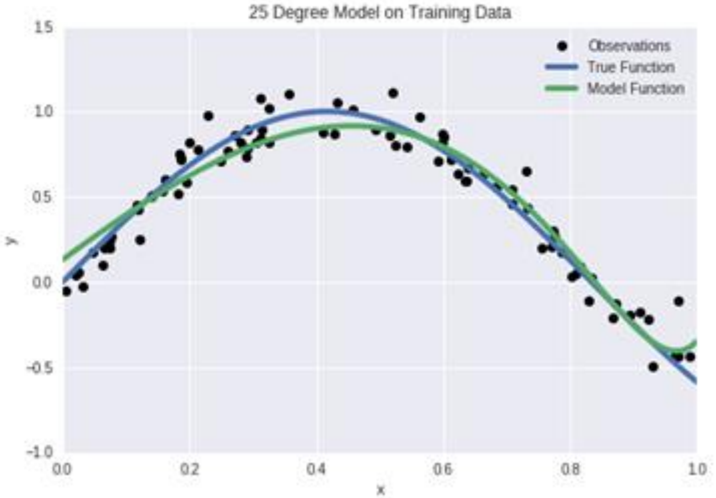


Blue curve represents the true function

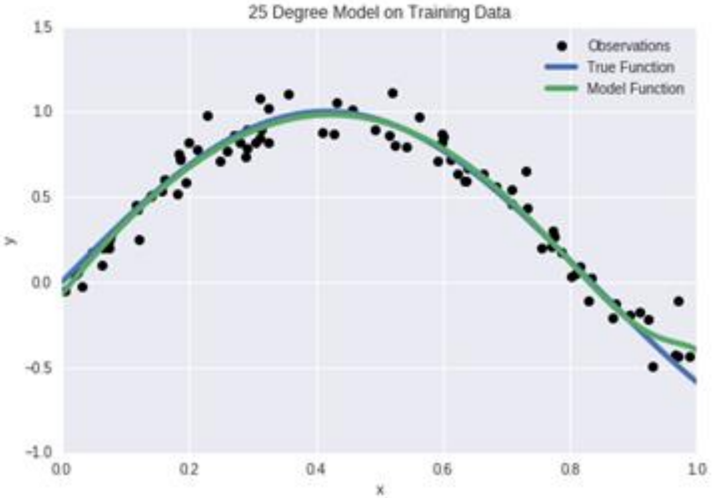
Curve fitting with different regularizations



No Regularisation



L1 Regularisation
$\lambda = 0.0125$



L2 Regularisation
$\lambda = 0.0125$

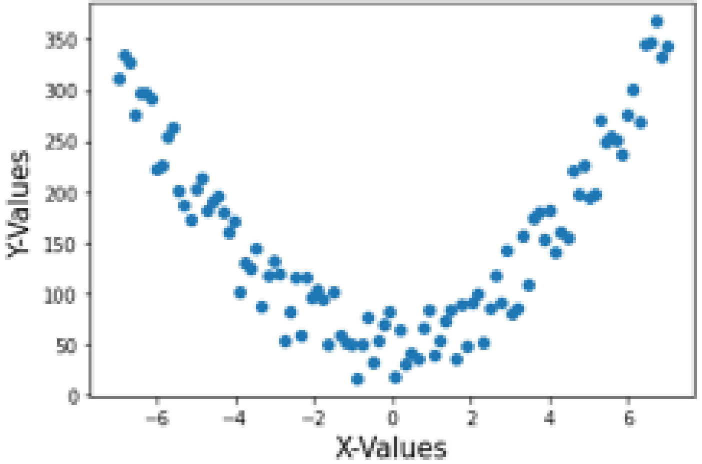
Lambda=Regularization Parameter (alpha)

Polynomial Regression

Third Degree Polynomial Feature Transformation

Second Degree Polynomial Feature Transformation

X	Y
-7	311
-6.85859	335.524
-6.71717	328.288
-6.57576	275.292
-6.43434	296.536
-6.29293	297.0199
-6.15152	292.7438
-6.0101	222.7077
-5.86869	225.9115
-5.72727	254.3554
-5.58586	264.0392
-5.44444	201.963
-5.30303	186.1267
-5.16162	173.5305
-5.0202	203.1742
-4.87879	214.0579
-4.73737	182.1815
-4.59596	190.5451
-4.45455	196.1488
-4.31313	179.9923
-4.17172	160.0759
-4.0303	170.3994
-3.88889	101.963



From the plot, it is seen that there is no linear relationship between X and Y.

$(1+X)^2=1+2x+X^2 = (1,X,X^2)$			
f1	f2	f3	Y
1	-7	49	311
1	-6.85859	47.0402	335.524
1	-6.71717	45.1204	328.288
1	-6.57576	43.24059	275.292
1	-6.43434	41.40078	296.536
1	-6.29293	39.60096	297.0199
1	-6.15152	37.84114	292.7438
1	-6.0101	36.12131	222.7077
1	-5.86869	34.44149	225.9115
1	-5.72727	32.80165	254.3554
1	-5.58586	31.20182	264.0392
1	-5.44444	29.64198	201.963
1	-5.30303	28.12213	186.1267
1	-5.16162	26.64228	173.5305
1	-5.0202	25.20243	203.1742
1	-4.87879	23.80257	214.0579
1	-4.73737	22.44271	182.1815
1	-4.59596	21.12284	190.5451
1	-4.45455	19.84298	196.1488

$(1+X)^3=1+3X+3X^2+X^3 = (1, X, X^2, X^3)$				
f1	f2	f3	f4	Y
1	-7	49	-343	311
1	-6.85859	47.0402	-322.629	335.524
1	-6.71717	45.1204	-303.081	328.288
1	-6.57576	43.24059	-284.34	275.292
1	-6.43434	41.40078	-266.387	296.536
1	-6.29293	39.60096	-249.206	297.0199
1	-6.15152	37.84114	-232.78	292.7438
1	-6.0101	36.12131	-217.093	222.7077
1	-5.86869	34.44149	-202.126	225.9115
1	-5.72727	32.80165	-187.864	254.3554
1	-5.58586	31.20182	-174.289	264.0392
1	-5.44444	29.64198	-161.384	201.963
1	-5.30303	28.12213	-149.133	186.1267
1	-5.16162	26.64228	-137.517	173.5305
1	-5.0202	25.20243	-126.521	203.1742
1	-4.87879	23.80257	-116.128	214.0579
1	-4.73737	22.44271	-106.32	182.1815
1	-4.59596	21.12284	-97.0797	190.5451
1	-4.45455	19.84298	-88.3914	196.1488

Now, instead of using ‘X ‘ as input features, We will use f1,f2,f3,… as features for modeling .
So it becomes simply a multivariable regression problem.

Summary

- Find model that is simple and fitting the data.
- Problem of overfitting.
  - How to detect? What to observe during learning?
  - How to avoid.
- Regularize the solution by adding extra term that also minimize the “complexity”.
- Right way to design solution with
  - Train, Val and Test splits.

Thanks!!

Questions?