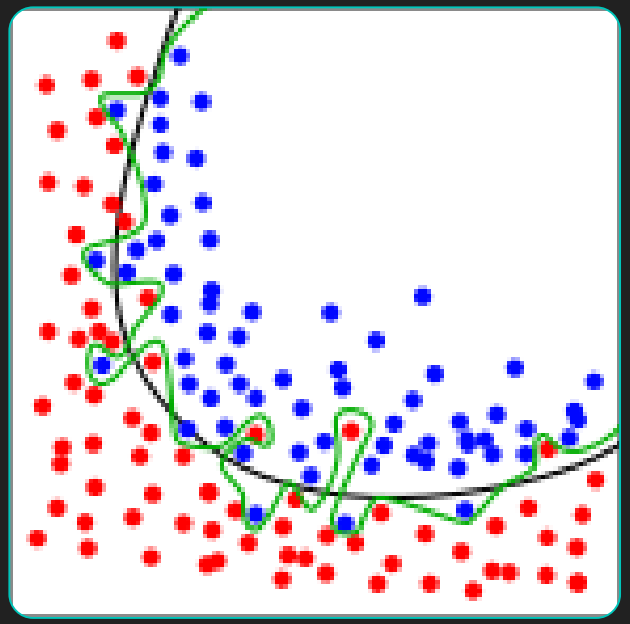
Data Science for Business
Chapter 5. Overfitting and Its Avoidance

One of the most important fundamental notions in Data Science.

Overfitting

- Overfitting occurs when a model is excessively complex, such as having too many parameters relative to the number of observations.



Generalization – Unseen cases

- It is the property a model or modeling process whereby the model applies to data that were not use to build the model.

Example: Population of phone customers' contract about to expire within six months

How to recognized overfitting in a data set?

- Analytic Tool : Fitting graph tool; shows the accurate of the model as a function of complexity.
- To examine overfitting – *Holdout* Data

The accuracy of the model depends on how complex we allow it to be.

Overfitting Examined

- **Holdout Data and Fitting Graphs -**

A **fitting graph** shows the accuracy of a model as a function of complexity .

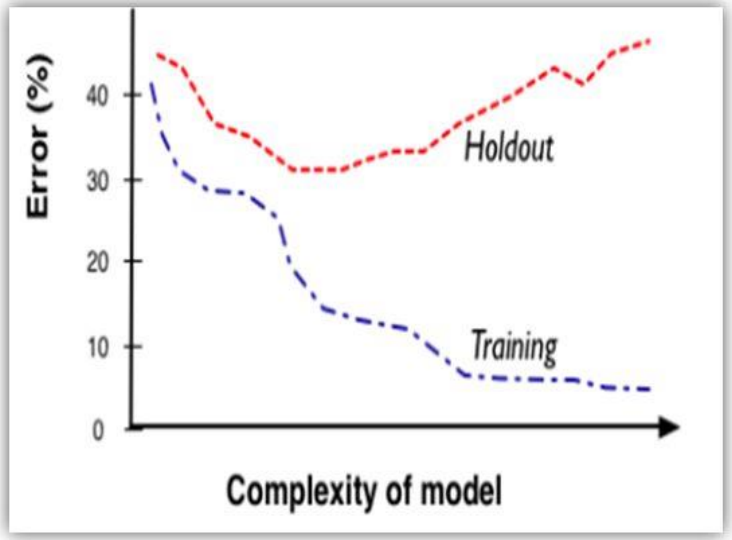


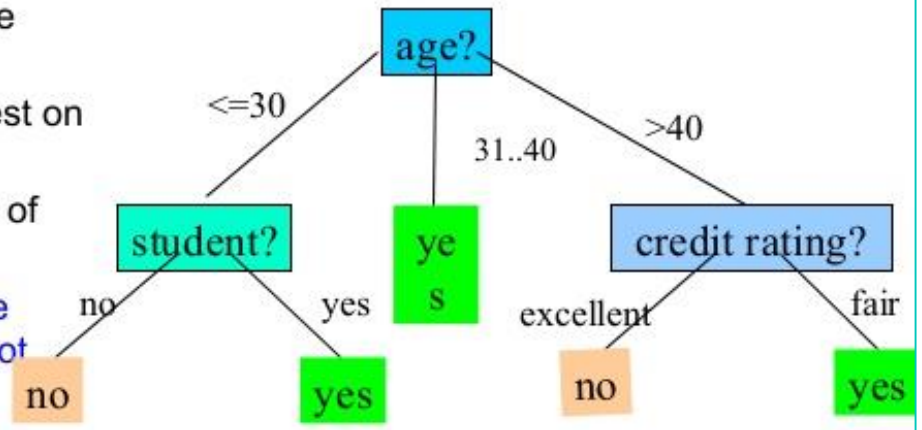
Figure 1. A typical fitting graph.

Overfitting in Tree induction

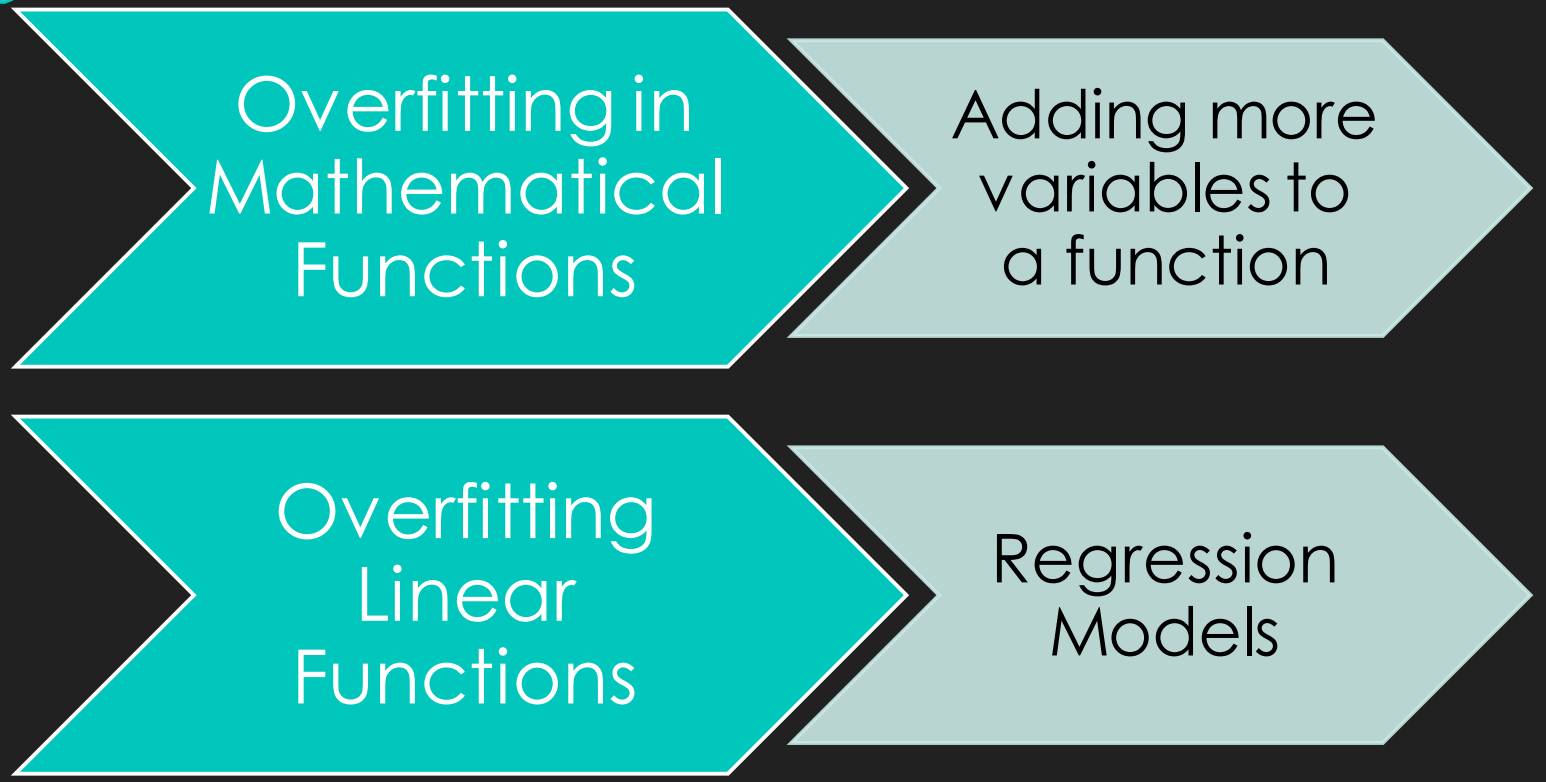
- Overfitting is a significant practical difficulty for decision tree models and many other predictive models.

Decision Tree Induction

- Decision tree
  - Flow chart like tree structure
  - Internal nodes - test on an attribute
  - Branch - outcome of the test
  - To classify sample trace path from root

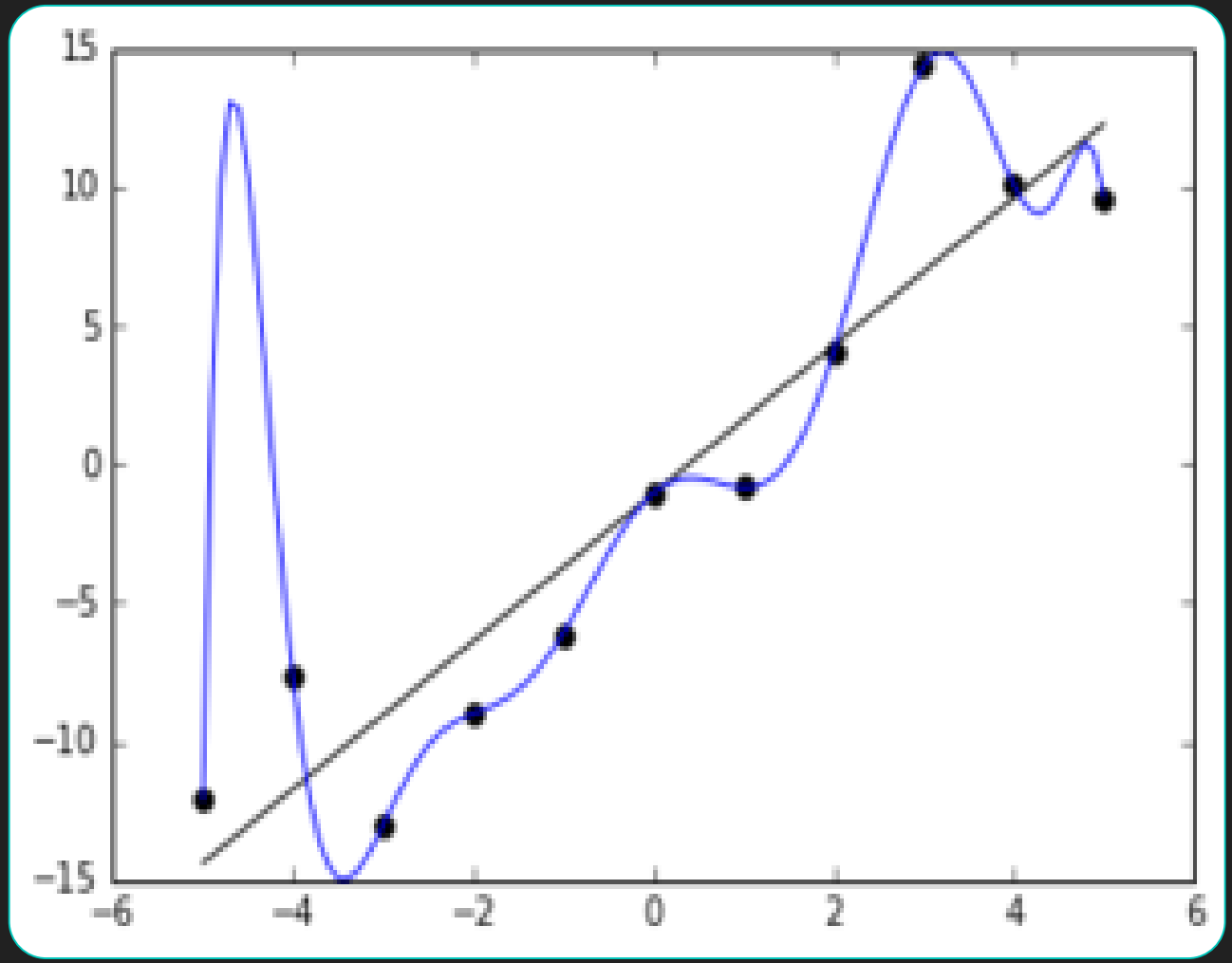


Overfitting Avoidance and Complexity Control



Overfitting Linear Functions

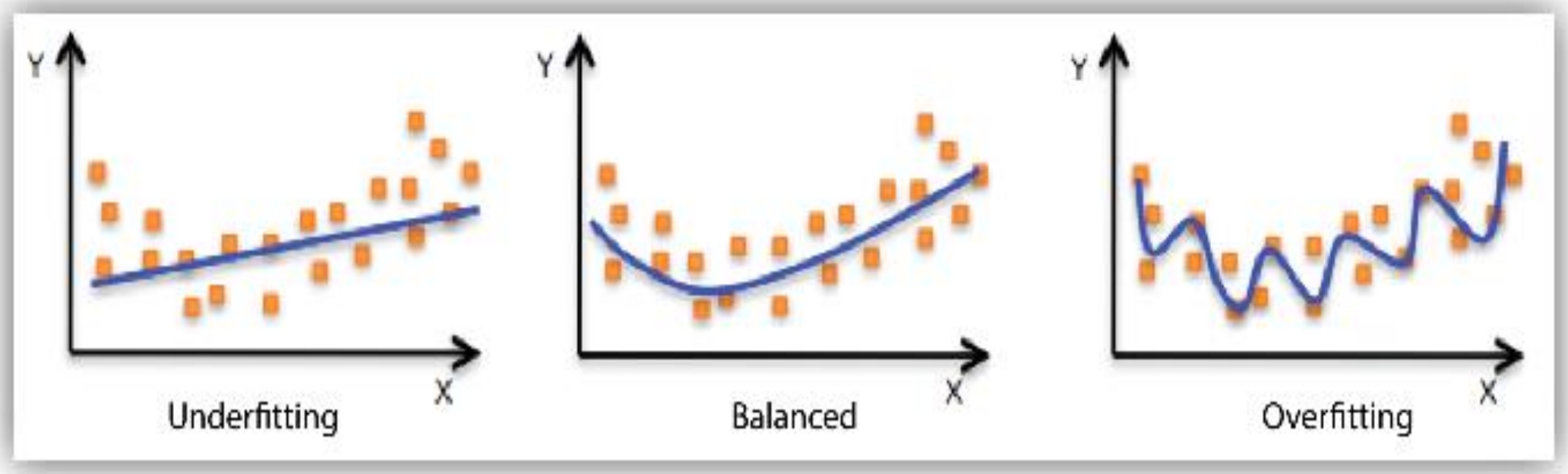
- The blue line represents an overfitted model and the black line represents a regularized model.
- Noisy (roughly linear) data is fitted to both linear and polynomial functions.



Overfitting code example in R

http://davide.eynard.it/teaching/2017_ML/overfitting.R

https://www.r-bloggers.com/cross-validation-for-predictive-analytics-using-r/



Summary :Amazon Learning Machine :Model Fit : Underfitting vs Overfitting.

http://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

Why Is Overfitting Bad?

- **Overfitting** is empirically **bad**. Suppose you have a data set which you split in two, test and training. ... An **overfitted** model uses more of the noise, which increases its performance in the case of known noise (training data) and decreases its performance in the case of novel noise.
- Overfitting essentially means taking too much information from your data and using it in a model. To see why this is bad, suppose you split a data set into two sets, test and training. In this case, to say a model is overfitted model means that the model performs significantly worse on the test data set than the training data set.