ML INTRO

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Cross validation

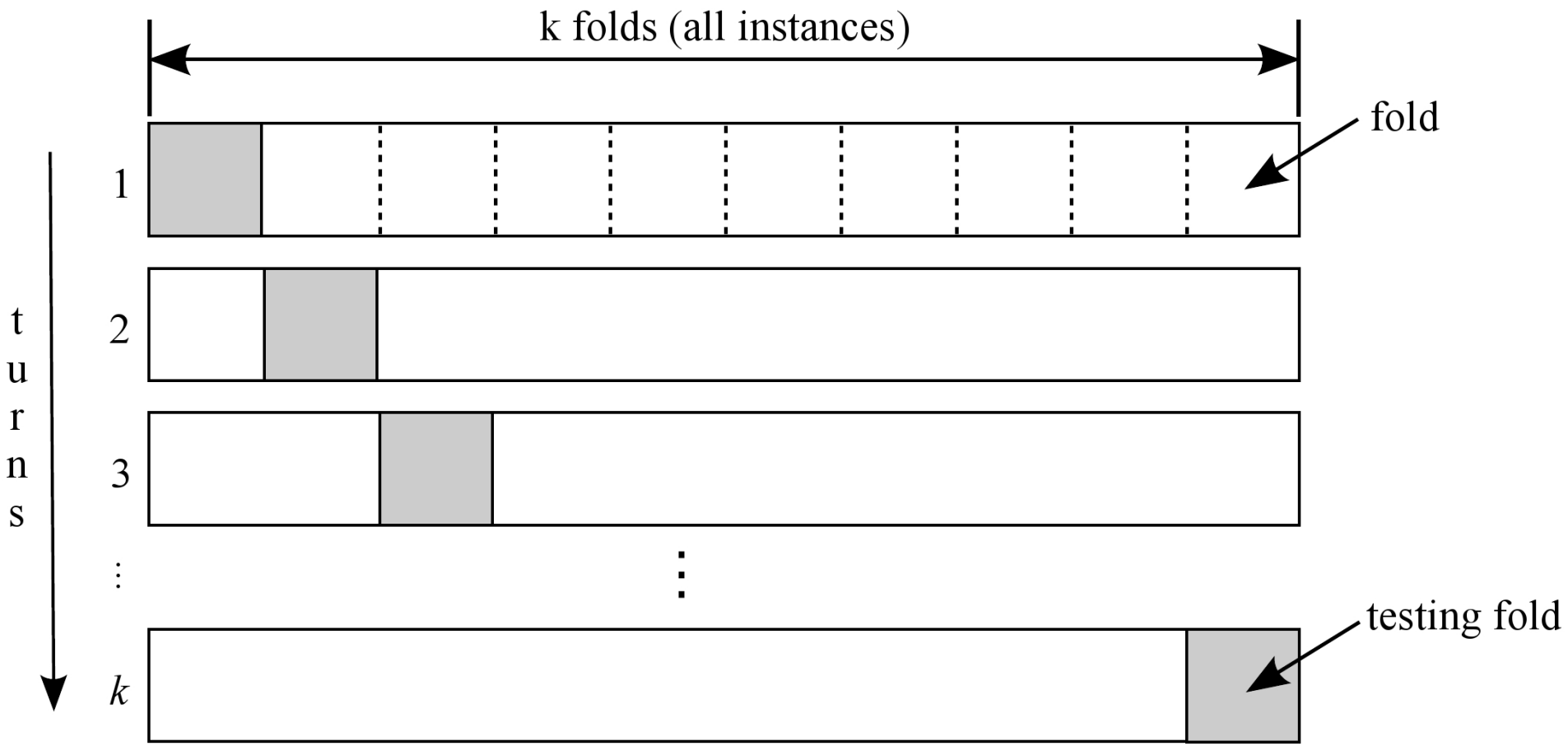
CROSS VALIDATION

Often used approach for validating ML models, as it allows for an effective use of the available data

This is how it works:

- The training set is divided into K subsets or *folds*
  - most often K is set to 10 (known as *10 fold cross validation*)
- Then, we perform K iterations of training and validating the ML model:
  - in each iteration, 1 data fold is used for validation, while the rest of the data (K-1 folds) are used for training the model
  - in each iteration, a different data fold is used for validating the model

CROSS VALIDATION



Source: http://goo.gl/BLIKRv

Cross Validation

In each iteration, the model's performance is measured

At the end, after the K^{th} iteration is finished, we compute the model's average performance on all K iterations

Performance measures computed in this way provide more realistic picture of the model's performance

If the performance measures computed in all (K) iterations do not vary much, the evaluation of the model's performance can be considered reliable

Over-fitting (variance)

vs.

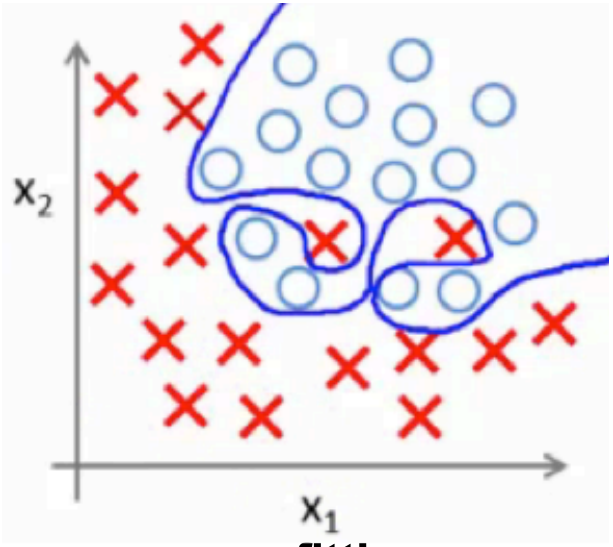
Under-fitting (bias)

Bias / Variance Trade-off

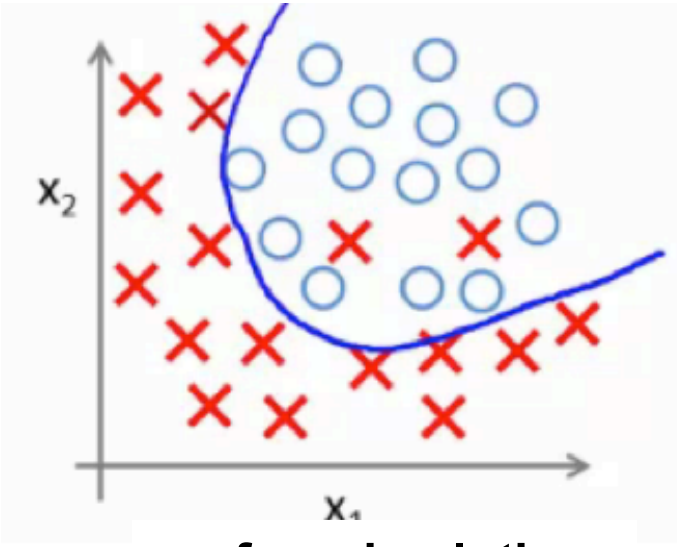
Over-fitting (Variance) and *Under-fitting* (*Bias*) are two very important things that we need to take into account when creating a ML model

OVER-FITTING

It refers to a situation when a ML model learns to work (almost) perfectly with the data from the training data set, but has poor performance on the test data, or any data that even minimally differ from the data used for training purposes



over fitting



preferred solution

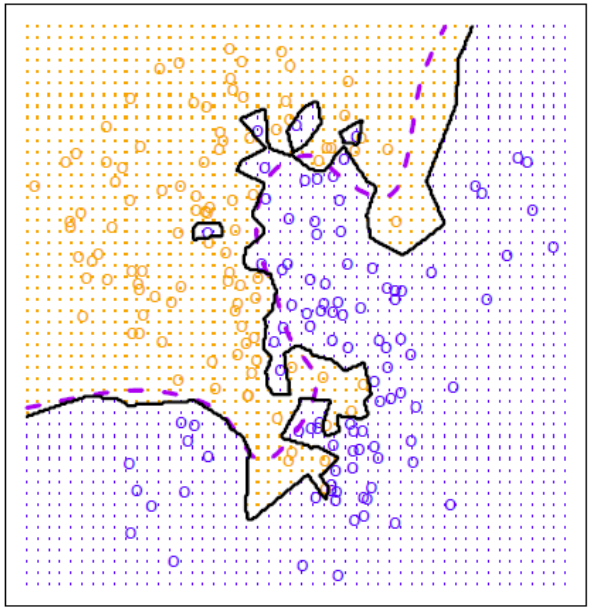
Source: https://www.coursera.org/course/ml

OVER-FITTING

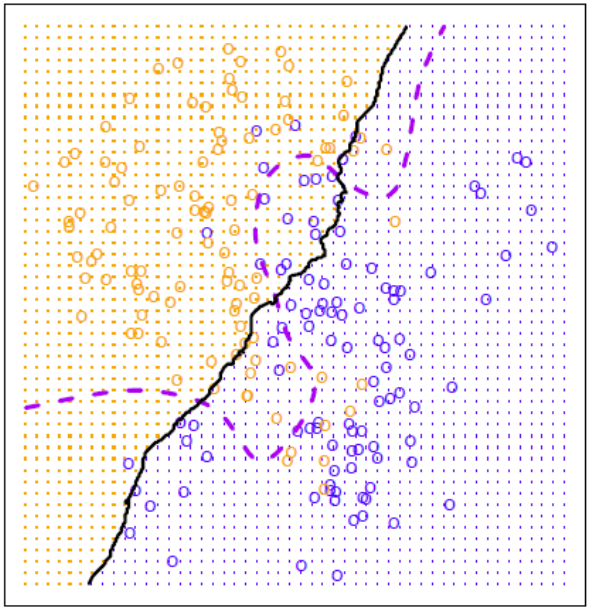
This problem is closely related to the *high variance* of the applied ML method

VARIANCE OF ML METHODS

- ***Variance*** measures the extent of change in a ML model if it is (re-)built with somewhat different training set (i.e., if the data values in the training set change)
- In general, the higher the complexity / flexibility of a ML method, the higher its *variance* will be



High variance

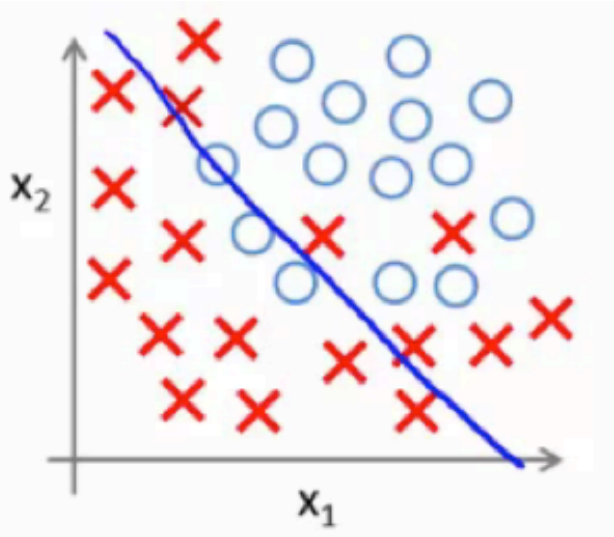


Low variance

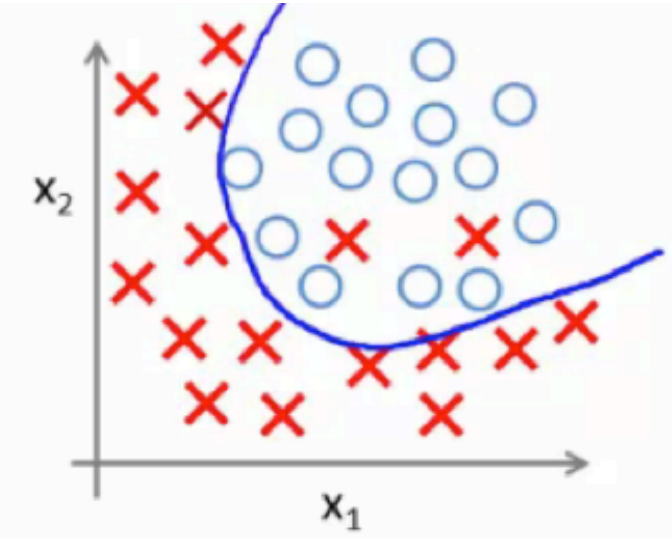
Source:
http://www.alsharif.info/#!iom530/c21o7

UNDER-FITTING

Under-fitting refers to the situation when a ML model, often due to its simplicity, fails to approximate the data from the training set, so it has poor performance even on the training data



under fitting



desired solution

Source: https://www.coursera.org/course/ml

UNDER-FITTING

This problem is closely tied to the *high bias* of the applied ML method

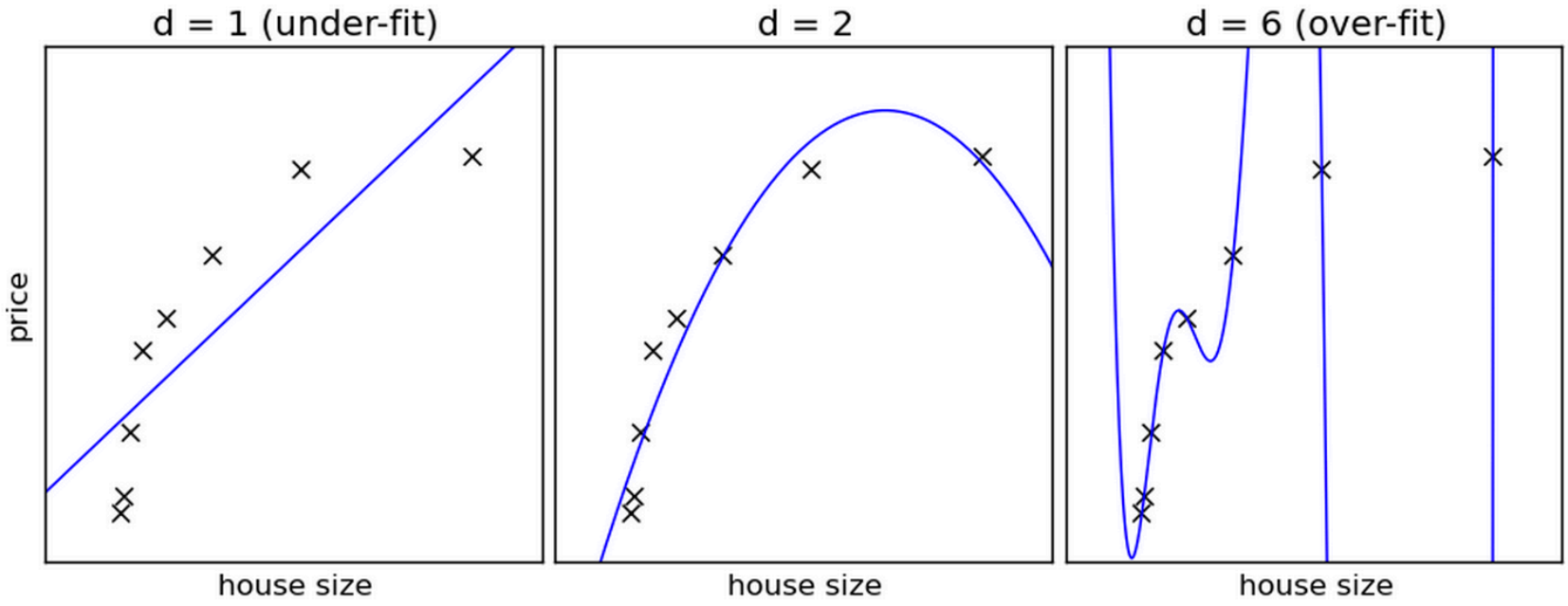
BIAS OF ML METHODS

- ***Bias*** refers to the low performance of ML models caused by using an overly simple ML method for solving a complex problem
- For instance:
  - linear regression assumes the existence of linear relationship between a dependent and an independent variable;
  - however, in reality, relations between variables are often non-linear;
  - this is why high bias is often associated with ML models that rely on linear regression
- In general, the higher the complexity / flexibility of a ML method, the lower its *bias* will be

BIAS / VARIANCE: EXAMPLE

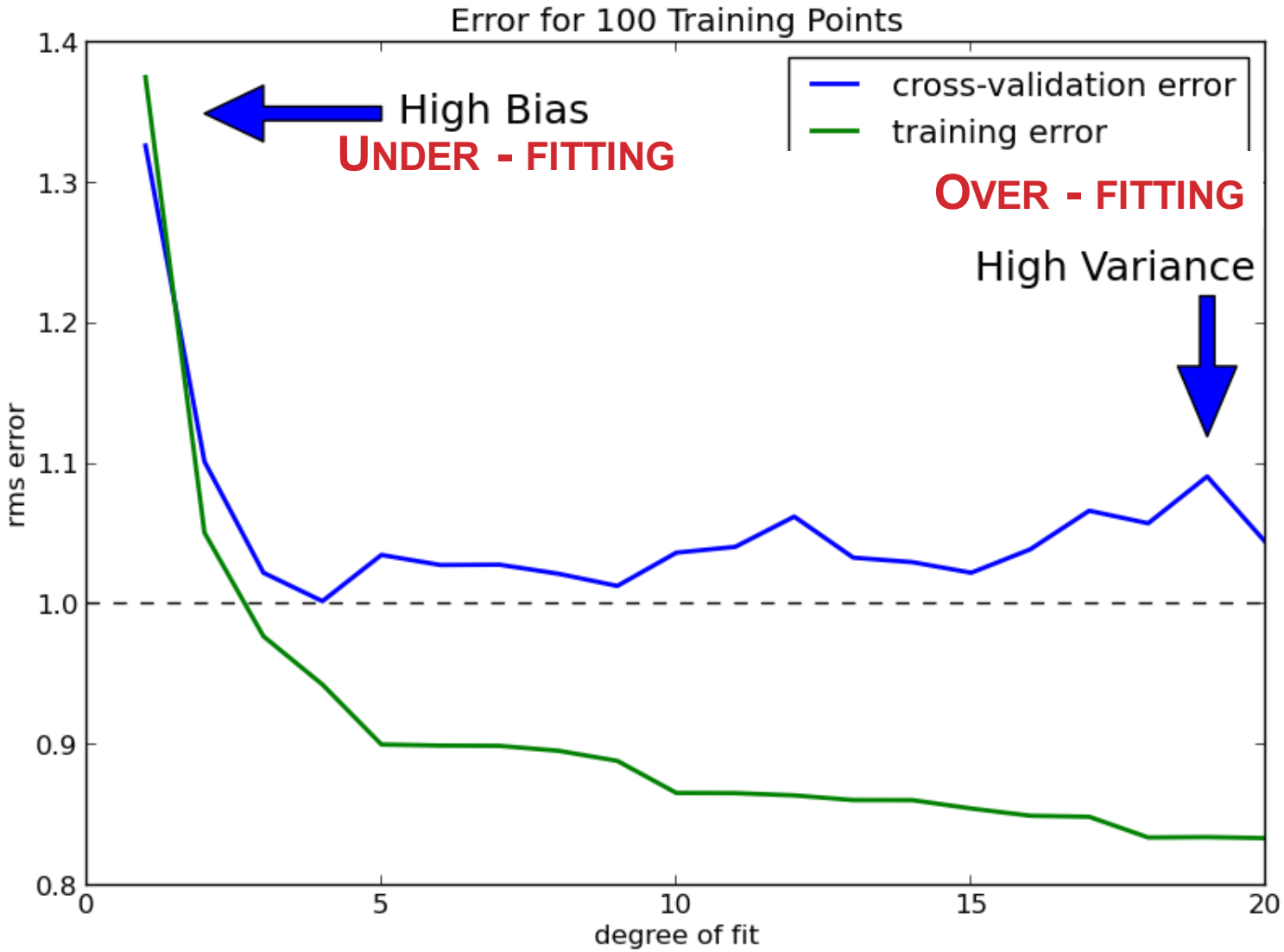
Bias/variance of a regression model for predicting the price of a house (dependent variable) based on its size (predictor)

d is the degree of the polynomial of the applied regression method



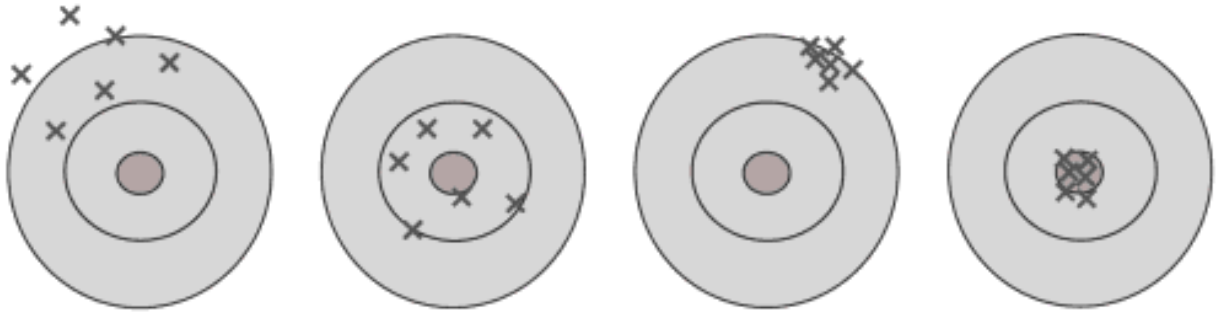
Source:
http://www.astroml.org/sklearn_tutorial/practical.html

OVER-FITTING VS. UNDER-FITTING



Source: http://www.astroml.org/sklearn_tutorial/practical.html

BIAS - VARIANCE DILEMMA / TRADE-OFF



Bias Variance Decomposition. Figure 1. The bias-variance decomposition is like trying to hit the bullseye on a dart-board. Each dart is thrown after training our "dart-throwing" model in a slightly different manner. If the darts vary wildly, the learner is *high variance*. If they are far from the bullseye, the learner is *high bias*. The ideal is clearly to have both low bias and low variance; however this is often difficult, giving an alternative terminology as the bias-variance "dilemma" (*Dartboard analogy*, Moore & McCabe (2002))

Source: http://www.springer.com/us/book/9780387307688