SDSC 2001 Python for Data Science
Machine Learning

Dr. Xinyue Li

Machine Learning

- What is machine learning?
  - Models: build mathematical models to help understand data
  - Learning: give these models tunable parameters that can be adapted to well describe observed data and relationships—learning from the data
  - Machine: machine-like learning to mimic the learning exhibited by the human brain—try to increase the similarity between model-based mathematical learning and human brain learning
  - Prediction: fitted models can be used to predict and understand aspects of newly observed data
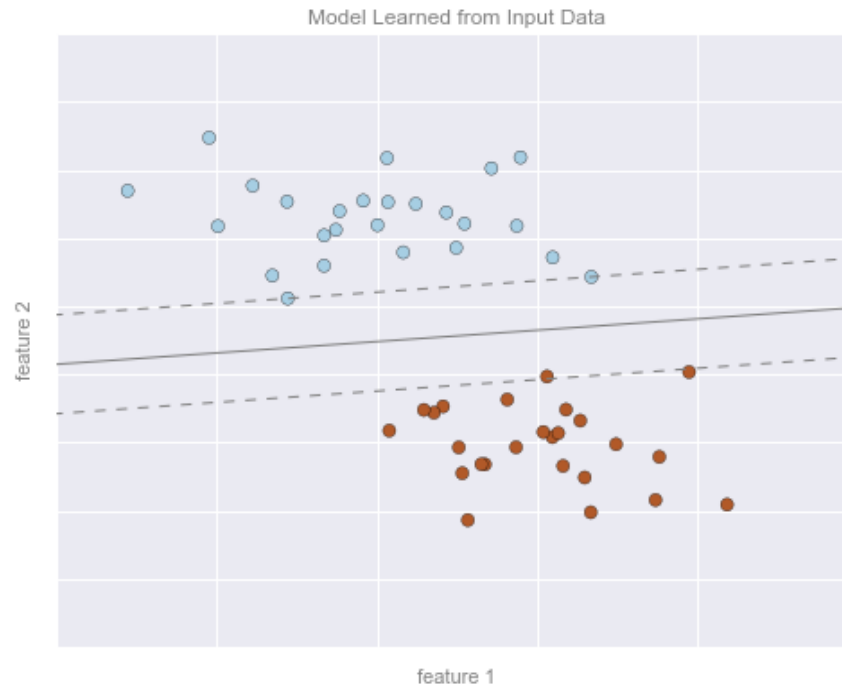
Supervised Learning

- Supervised learning involves somehow modeling the relationship between measured features of data and some <u>labeled</u> outcomes associated with the data

- Fitted model can be used to predict labels based on new, unknown feature data

- Classification: labels are discrete categories

- Regression: labels are continuous quantities

Unsupervised Learning

- Data-driven: modeling the features of a dataset <u>without</u> reference to any labeled outcomes
    - letting the dataset speak for itself

- Clustering: identify distinct groups of data

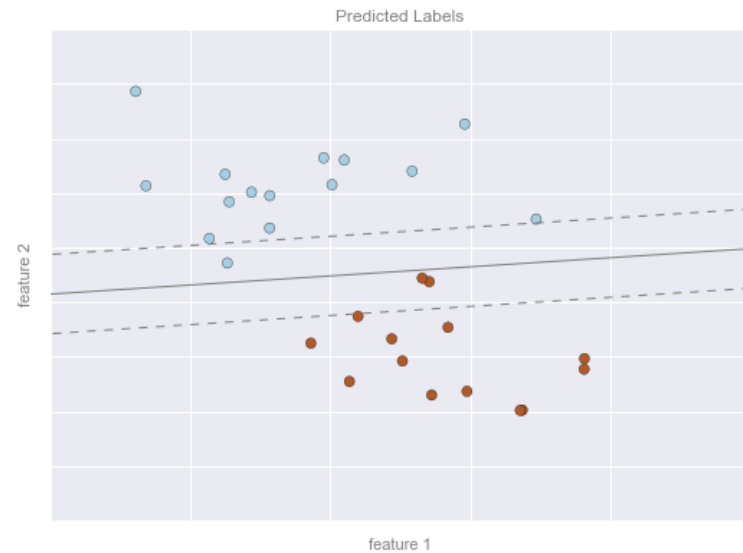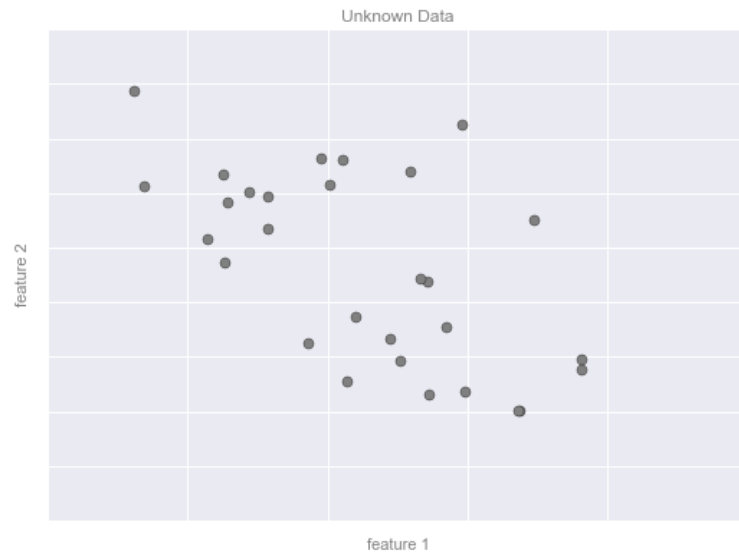- Dimension reduction: identify succinct representations of data

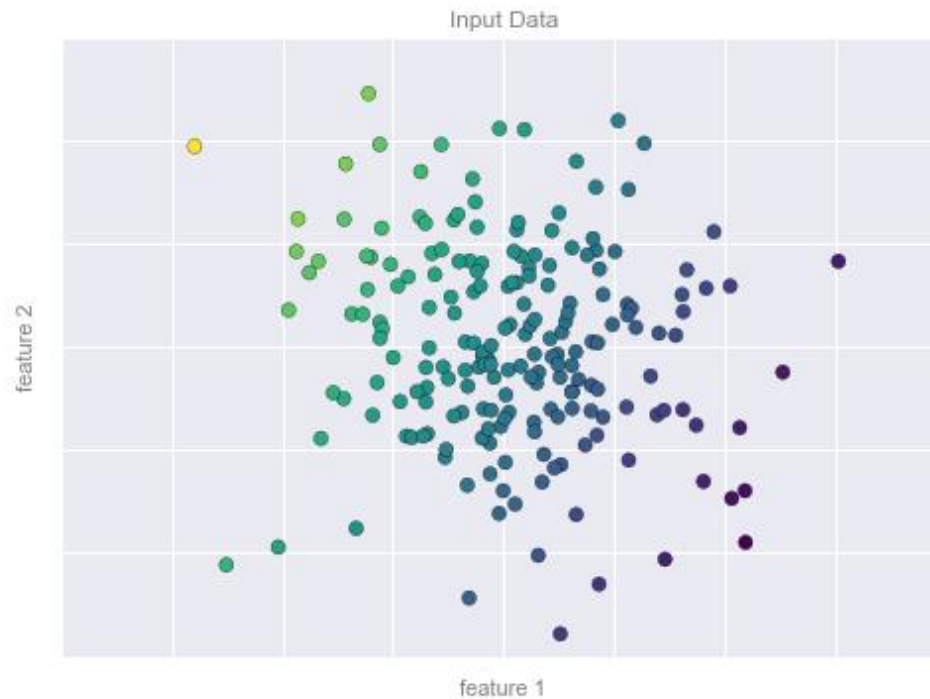# Supervised Learning

- Classification
  - Model fitting

Model Learned from Input Data

Supervised Learning

- Classification
  - Prediction

Supervised Learning

- Regression
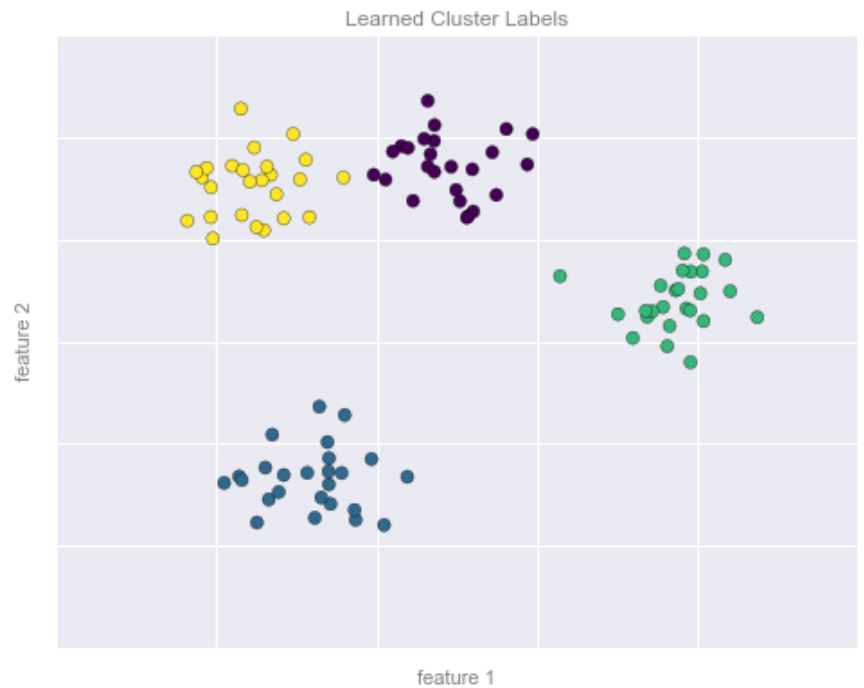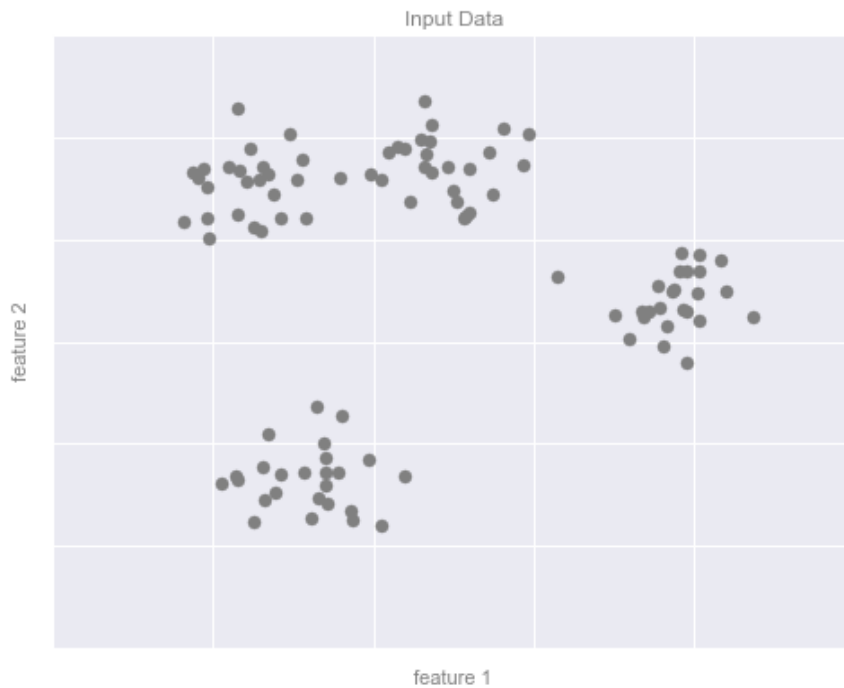 - Model fitting

Input Data

feature 2

feature 1

Supervised Learning
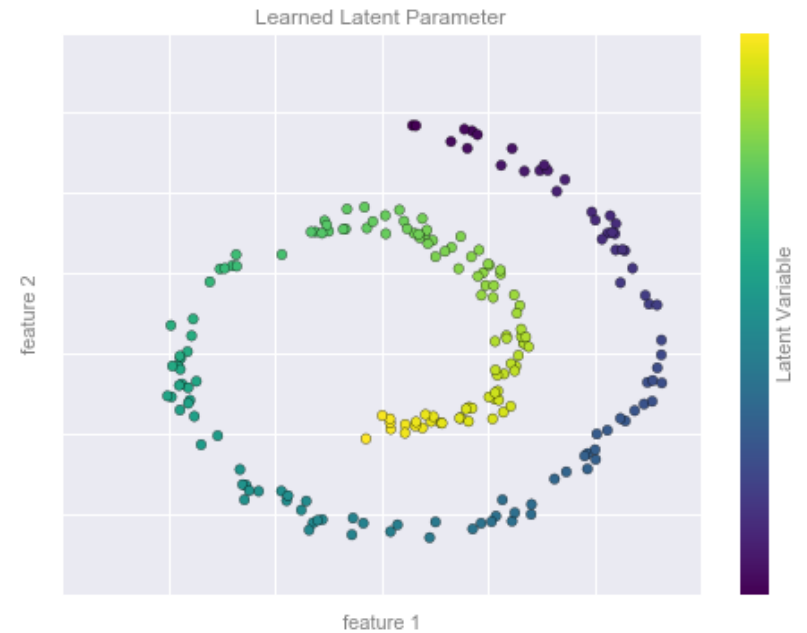
- Regression
  - Prediction
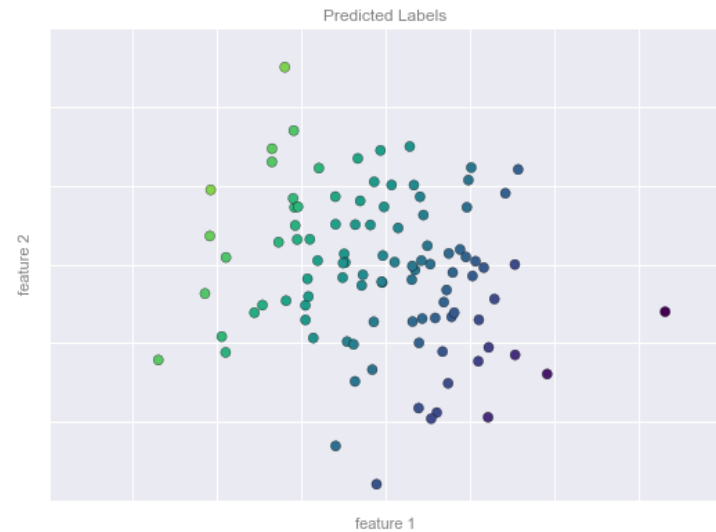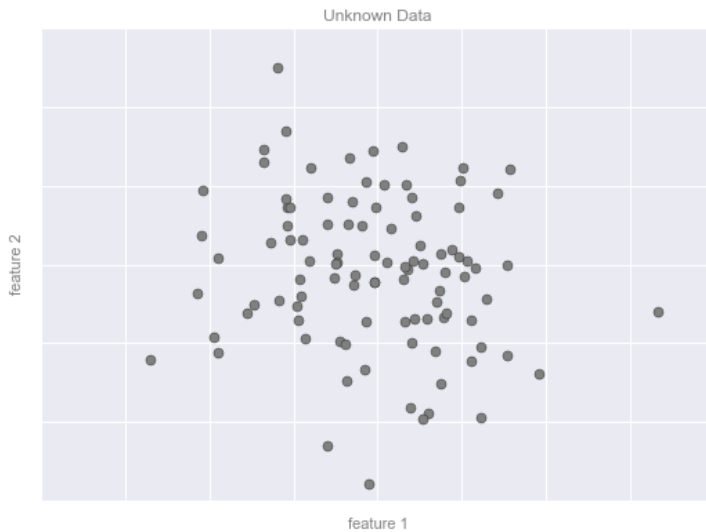
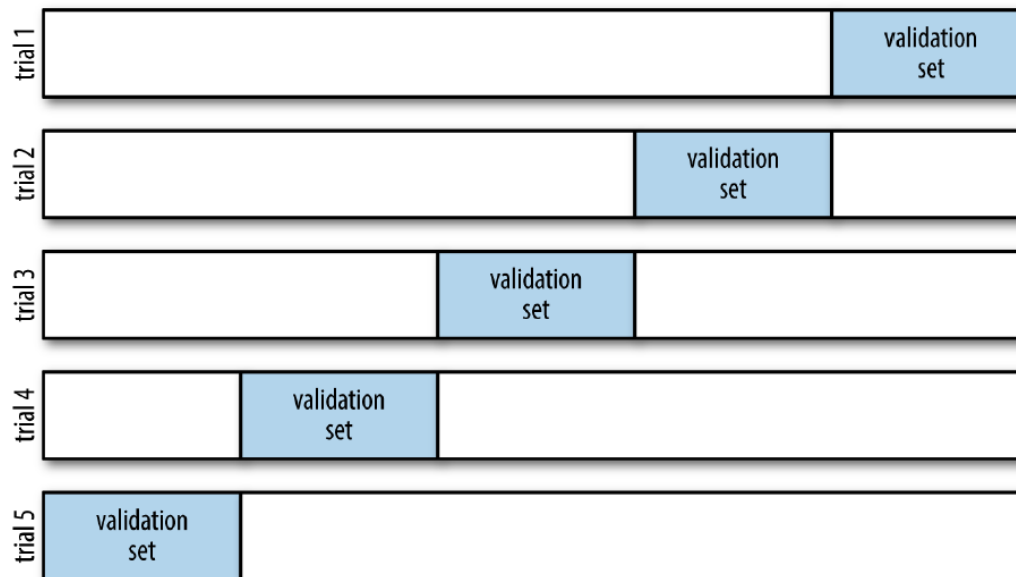# Unsupervised Learning

- Clustering

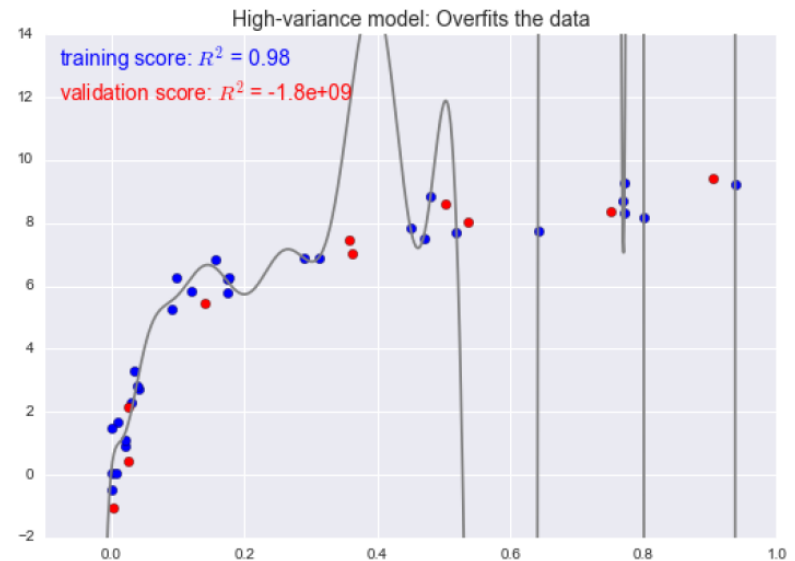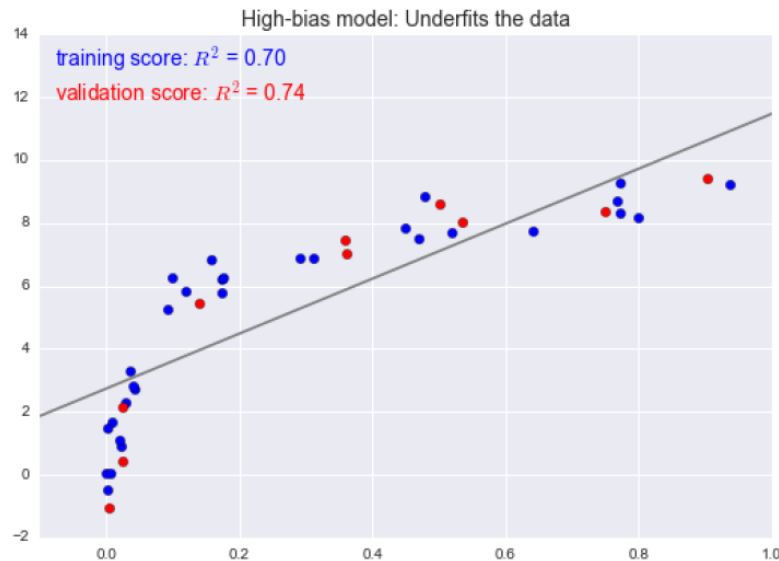# Unsupervised Learning

- Dimension reduction

Model Validation: Cross-Validation

- Wrong way of model validation
  - Entire dataset
- Right way of model validation
  - Holdout sets
- Make use of holdout sets: cross-validation
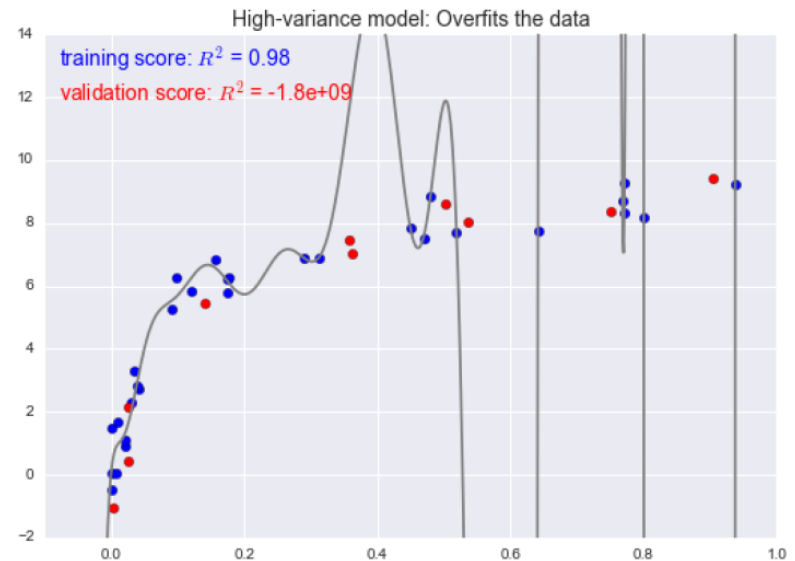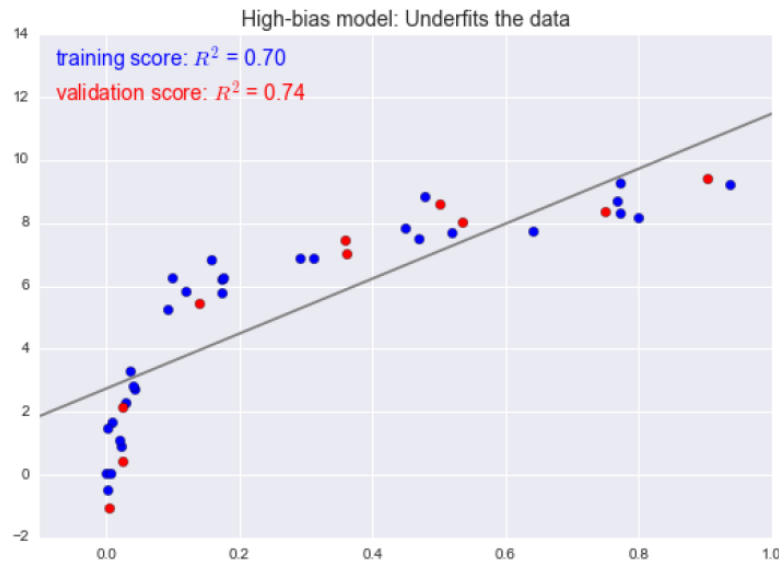
Model Selection: Bias-Variance Trade-Off

High-bias model: Underfits the data

training score: $R^2 = 0.70$
validation score: $R^2 = 0.74$

High-variance model: Overfits the data

training score: $R^2 = 0.98$
validation score: $R^2 = -1.8e+09$

- r^2: the proportion of the variance in the dependent variable that can be explained by the independent variable(s)

Model Selection: Bias-Variance Trade-Off

Validation Curve Schematic
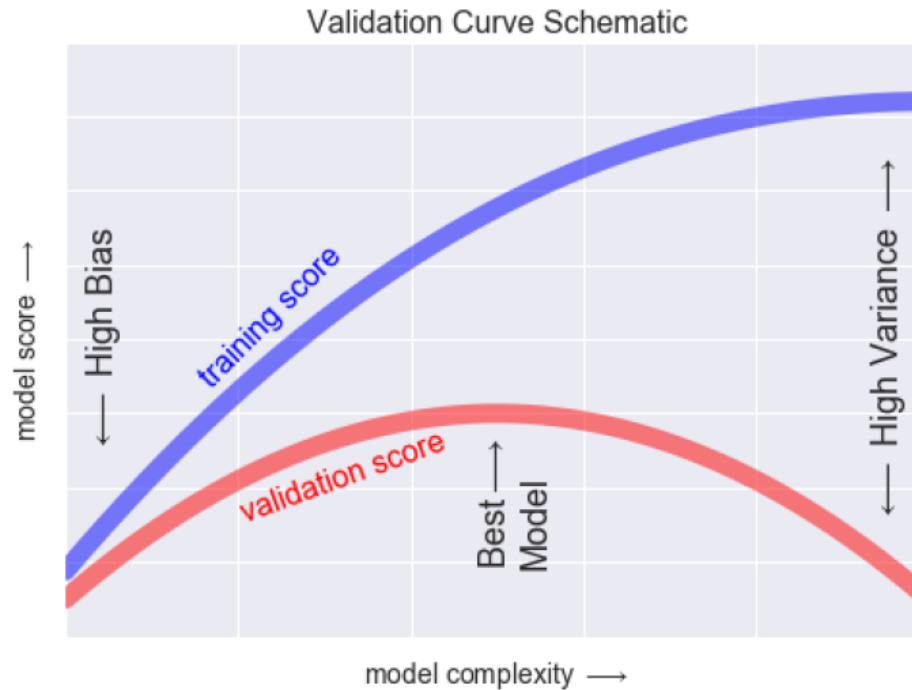
- Bias and variance trade-off

Model Selection: Bias-Variance Trade-Off

- r^2: the proportion of the variance in the dependent variable that can be explained by the independent variable(s)