Introduction

Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

Machine Learning for Economics and Finance

Bachelor in Economics

Marcel Weschke

27.05.2025

Machine Learning for Economics and Finance

This course is an introduction to machine learning in Python with applications in economics and finance.
You will learn how to practically analyze economic data using machine learning techniques.

The best way to learn about data science methods:

- sit down at a computer
- and actually analyze data

The class is structured to emphasize that.

Learning Goals

(1) Use the free software Python to solve key tasks in machine learning including loading and cleaning big datasets, using libraries/Python-packages, summarize and visualize data

(2) Apply supervised learning methods to solve prediction problems in economics and finance

(3) Analyze the main benefits and limitations of the supervised learning methods we cover in class

(4) Evaluate and compare the performance of different methods

Course format

- 3-hour lectures with interactive coding exercises (first half theory; second half applications in Python)
- Bring your laptop to the lectures! We will discuss your results
- Lectures take place on campus
- Several problem sets where you learn to apply the machine learning techniques

Introduction to Programming in Python

Programming Language: RStudio/JupyterLab (IDE for R and Python)

- There will be an introductory Python programming session on April 16
- I will provide practice exercises to get you started in Python

Important:

- Please install Python and RStudio before the introduction to Python session next week

Schedule and Topics

- April 2: No lecture
- April 9: Introduction to machine learning
- April 16: Intro to Python with Marlene Renkel + practice Python exercises
- April 23: Introduction to supervised learning and linear regressions
- April 30: Classification problems
- May 7: Resampling methods (cross-validation)
- May 14: Problem Set 1
- May 28: Shrinkage methods (lasso and ridge regressions)
- June 4: Tree-based methods
- June 11: Tree-based methods
- June 18: Problem Set 2
- June 25: Deep learning (neural networks)
- July 2: Deep learning (neural networks)
- July 9: Problem Set 3

Background

There are no prerequisites but it is helpful if you have already some knowledge about the following:

- **Statistics**: linear regressions, logistic regressions, maximum likelihood estimation
- **Programming**: data handling in Python

Textbooks and Learning Material

Main Reference (ISL):

- James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). An Introduction to Statistical Learning with Applications in Python
- The book as well as Python tutorials, datasets and practice exercises are available at https://statlearning.com/

More advanced (ESL):

- T. Hastie, R. Tibshirani and J. Friedman. The Elements of Statistical Learning, Springer. Free at: https://web.stanford.edu/~hastie/ElemStatLearn/.

Organization

Any questions regarding the organization before we get started with the course material?

What is Big Data and Machine Learning?

Big data:

New storage capacities and the technological progress in computing power allow for the storage and processing of huge datasets

Machine learning:

Machine learning algorithms build a model based on sample data, in order to make predictions or decisions without being explicitly programmed to do so.

Big Data

- Big data is a marketing term rather than a well-defined concept.
- Data storage has become very cheap and efficient.
  - A standard computer can store several terabytes of data.
  - Easy access to large clusters by cloud computing (for example Amazon (EC2) and Microsoft (Azure)).
- Businesses, especially internet businesses, now generate large amounts of data.
  - Google's search index contains hundreds of billions of pages, and takes up over 100,000 terabytes.
  - Twitter sees over 500 million tweets per day.
  - Amazon has 17 million sales per hour on average.
- Big data is not only about storing large datasets, but also about software for managing it (SQL databases, Apache Spark, Hadoop,...)

Machine Learning

Being able to store and manage big data due to the advances in computing power allows us to analyze large datasets and learn from it. For this we can use machine learning techniques.

Some applications of big data and machine learning:

- **Market-basket analysis**: Amazon uses consumer data to find out what their customers like
- **Medicine**: Use image recognition to improve cancer detection
- **Online Fraud Detection**: Paypal uses transactions data to detect money laundering activities
- **Social Media Services**: network analysis of people you may know; face recognition to link pictures,...

Machine Learning in Economics and Finance

This course will focus on applications in economics and finance (which are many).

Wide-spread examples include

- Automated trading and stock return predictions
- Credit scoring: determining which loans are likely to go bad
- Monetary economics: analyse cental banker's statements via natural language processing
- GDP measures in developing countries via satellite data

Machine Learning vs Traditional Statistics

Machine learning aimed at (out of sample) prediction and less at hypothesis testing/inference.

Models can have lots and lots of parameters.

No effort is made (and it is usually hard) to interpret individual parameters.

Some terminology:

Statistics	Machine Learning
dependent variable	outcome/response
independent variables	features/predictors
fitting	learning
coefficients	weights

Machine Learning Methods

In machine learning terminology, methods can be broadly classified as:

- **Supervised learning**
  - Regression
  - Classification
- Unsupervised learning
- Reinforcement learning (not covered in this course)

Supervised Learning

Suppose you have a quantitative response Y and p different predictors $X = (X_1, X_2, \ldots, X_p)$.

In supervised learning, we try to establish a relation

$$Y = f(X) + \epsilon$$

f: unknown function that represents the systematic information that X provides about Y.

ϵ: random, independent error term with mean zero

Key task: find $\hat{f} \approx f$ that 'fits the data well'

Supervised Learning: Example

Predict the income of an individual by its years of education:

- Y: income
- $X = X_1$: years of education
- $p = 1$

We could approximate f using for example a linear regression:

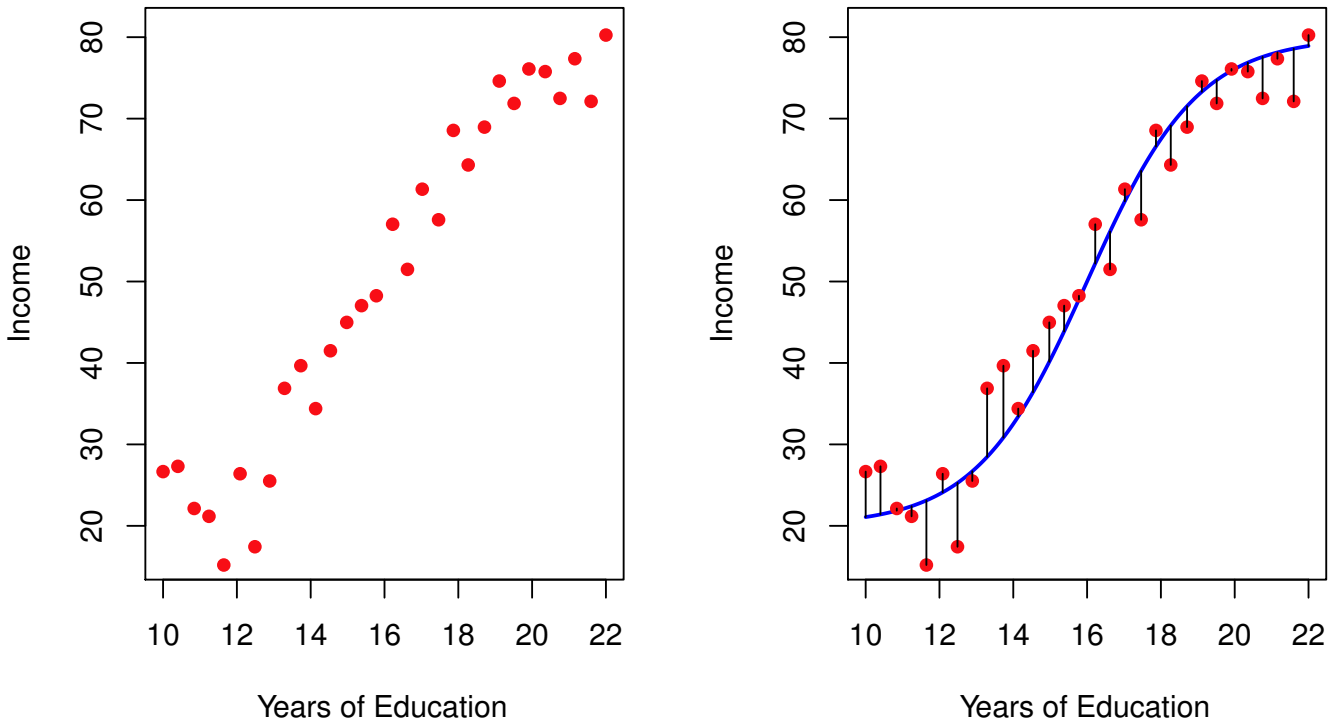
$$\hat{f}(X_1) = \beta_0 + \beta_1 X_1$$

where β_0 and β_1 are the regression coefficients (weights)

Disclaimer

Some of the figures in this course are taken from "An Introduction to Statistical Learning, with applications in Python" (Springer, 2021) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Example: Income data with one predictor



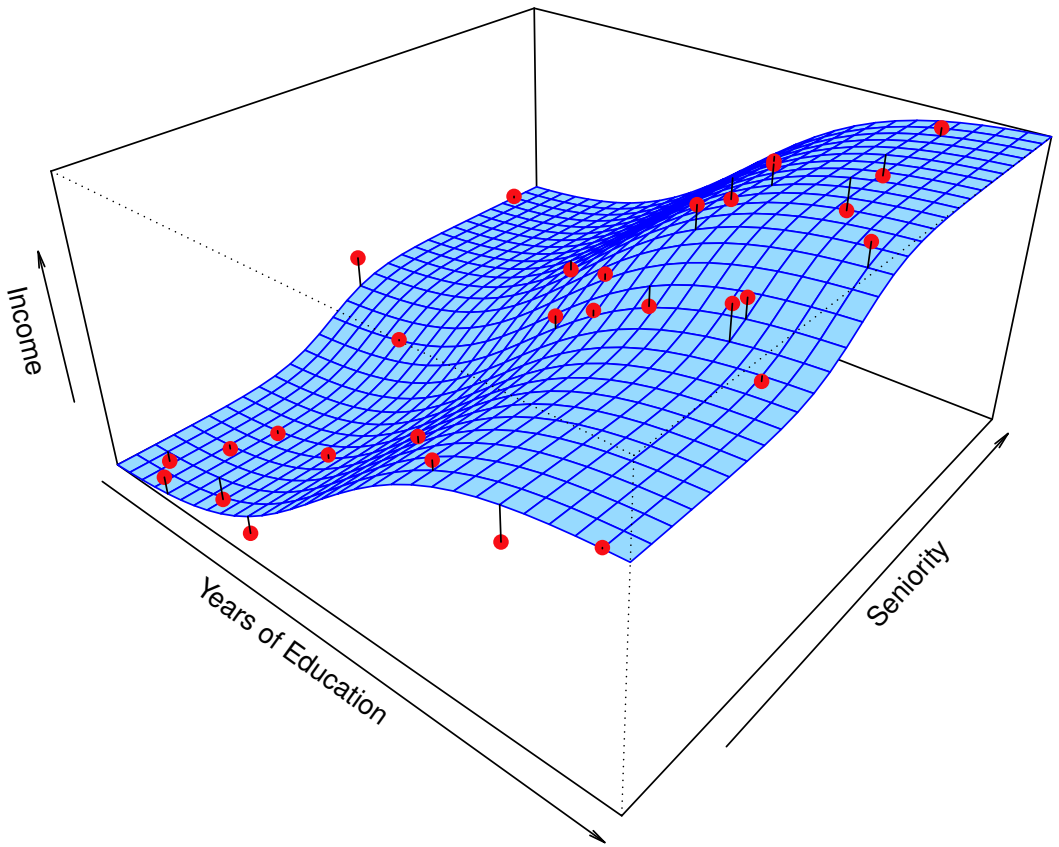
Y: Data on income

X: Years of education

Blue line: f (known as this is simlated data)

Black lines: ϵ

Example: Income data with two predictors



Y: Data on income

X_1: Years of education

X_2: Seniority

Why estimate f? Prediction vs Inference

Prediction:

Often features X are readily available, but outcome Y isn't easily obtained.

Since the error term averages to zero, we can predict Y using

$$\hat{Y} = \hat{f}(X).$$

$\hat{Y}$: prediction for Y

$\hat{f}$: estimate for f

Example: Predict stock returns tomorrow using today's data (returns today, size of a company, market value of a company...)

For prediction, $\hat{f}$ is treated as a *black box* as we do not care about the form of $\hat{f}$ as long as it yields good predictions

Prediction Errors

Accuracy of $\hat{Y}$ depends on two quantities:

1. Reducible error: $\hat{f}$ is not a perfect estimate for f
2. Irreducible error: variability of ϵ affects accuracy of $\hat{Y}$

We will learn about different techniques to estimate f that aim to minimize the reducible error.

Inference

Goal: Estimate f to understand how X affects Y
Example of stock market dataset:

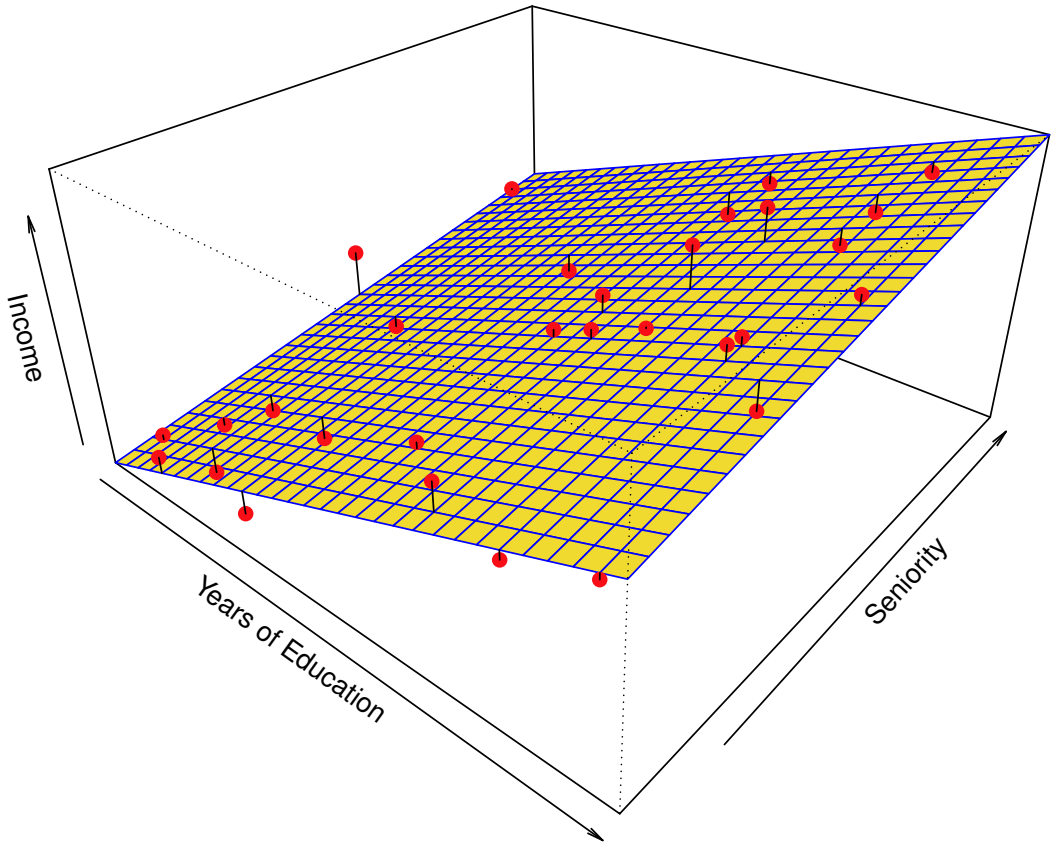
- Which predictors are associated with the response?
  - Learn which of the features (returns today, size of a company, market value of a company...) are important determinants for stock returns?
- What is the relationship between the response and the predictor:
  - Does the size of a company have a negative or positive impact on its stock returns?
  - Is the relation between size and stock returns linear or does it take a more complex form?
  - Are there interactions between size and other predictors?

For inference, $\hat{f}$ cannot be treated as a black box but we are interested in its particular form.

Estimating f

- Different machine learning methods use different functional forms for $\hat{f}$
- This leads to different predictions and also different prediction errors
- There is no general rule which method is best, but it depends on the data problem at hand
- **Key learning goal**: determine which machine learning method works best for a specific problem

Income Data: Linear Regression



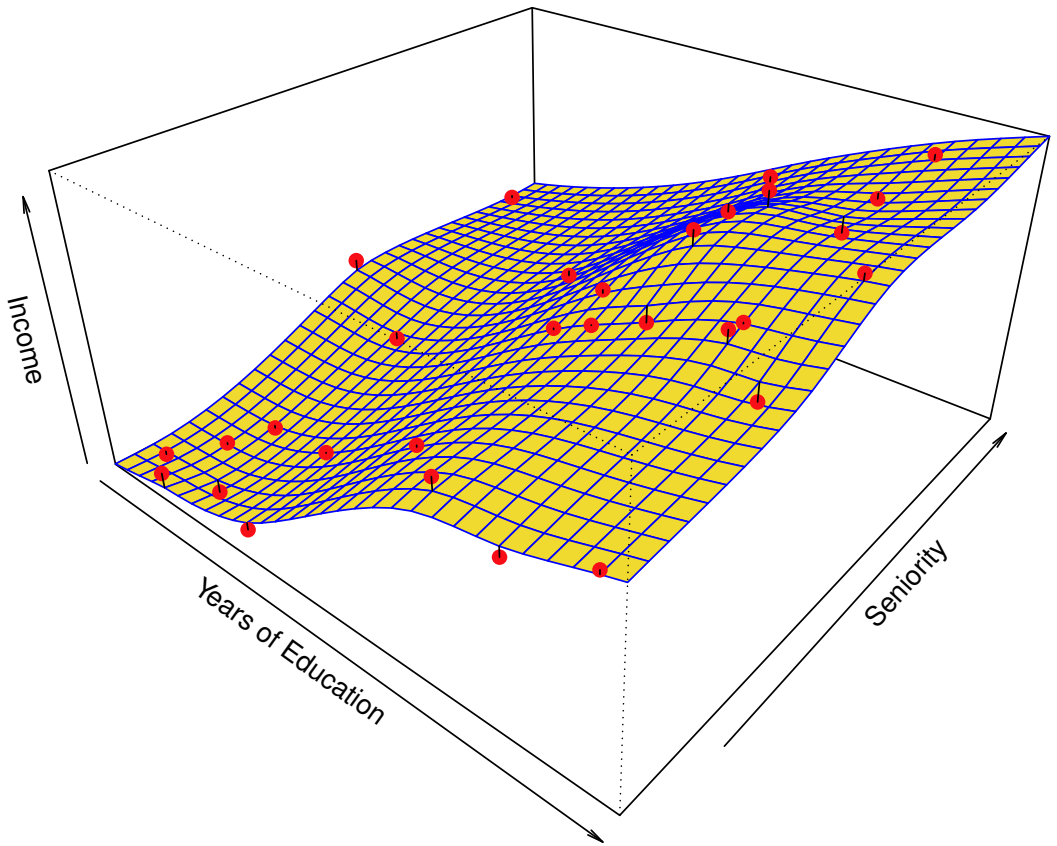
Linear regression: $y_i = \beta_0 + \beta_1 * x_{1,i} + \beta_2 * x_{2,i} + \epsilon_i$

y = income

x_1 = years of education

x_2 = seniority

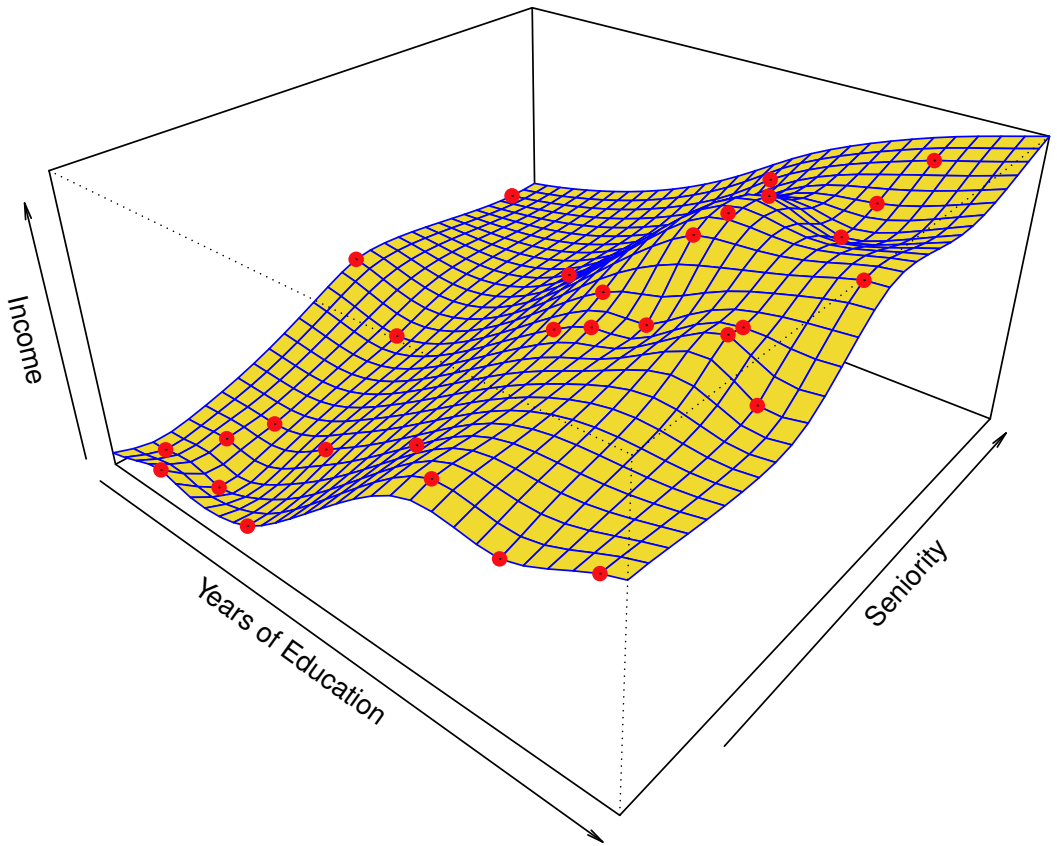
Income Data: Smooth Spline



Better fit and hence smaller errors than linear regression

But: also more parameters

Income Data: Rough Spline



Perfect fit and hence no errors

But: many parameters and less smoothness

Prediction Accuracy and Interpretability

Why would we choose a more restrictive method with higher errors over a more flexible method?

Inference:

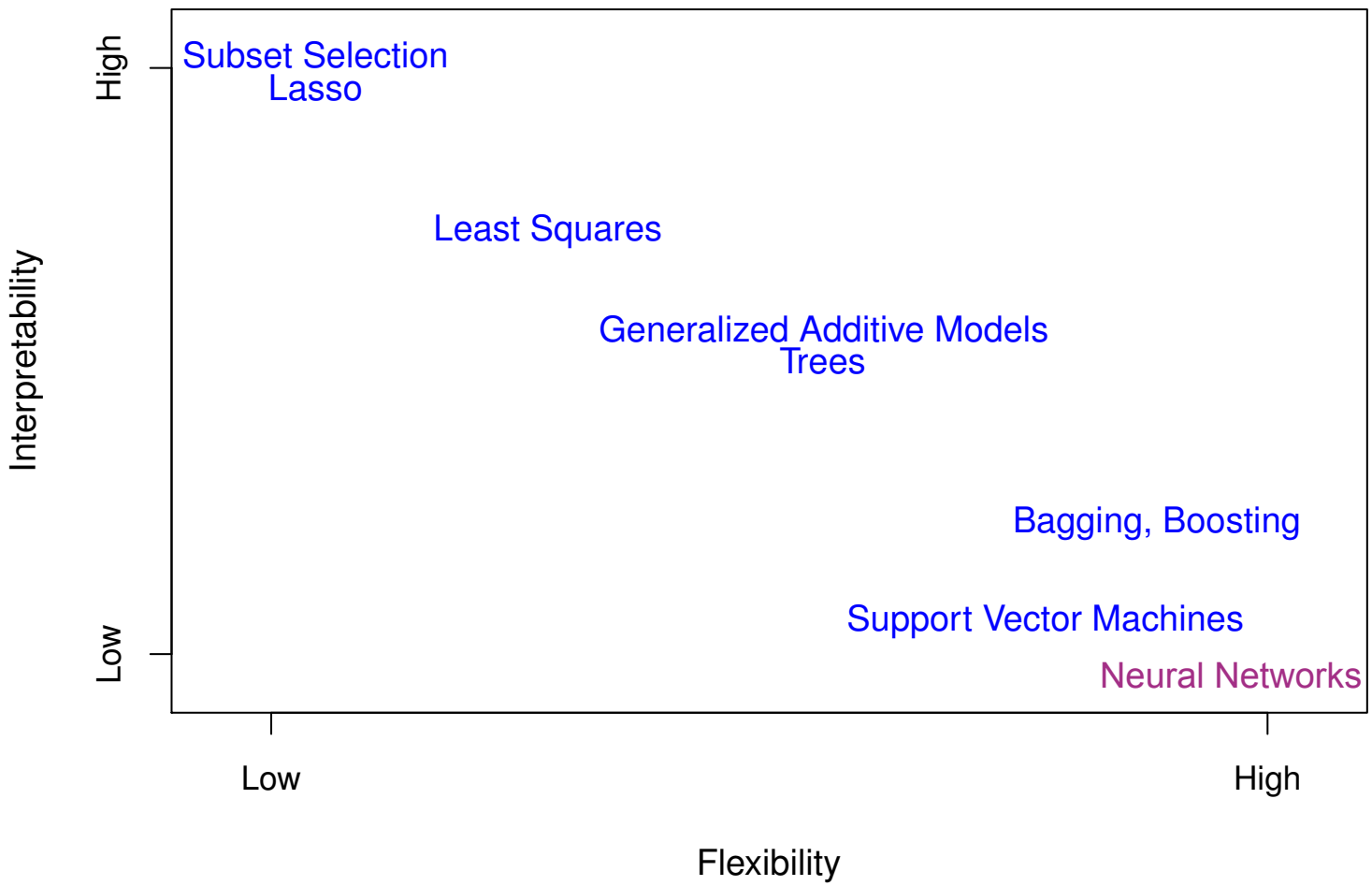
- Flexible models lead to complicated estimates of f
- Hence, influence of individual predictor on the response is difficult to understand

Prediction:

- In-sample fit can be increased by more flexible methods
- But: this doesn't imply better out-of-sample predictions
- Risk of overfitting
- Imposing some structure often increases predictive power

More on this later...

Flexibility vs Interpretability



Supervised Learning: Regression vs Classification

Recall again the general supervised learning setup:

$$Y = f(X) + \epsilon$$

- **Regression**: Y is continuous (as in the linear regression example)
- **Classification**: Y is a category (for example 0 or 1 as in logistic regressions (conflict with statistics terminology))

Example of regression: predict returns tomorrow
Example of classification: predict whether returns tomorrow are

positive or negative

Unsupervised Learning

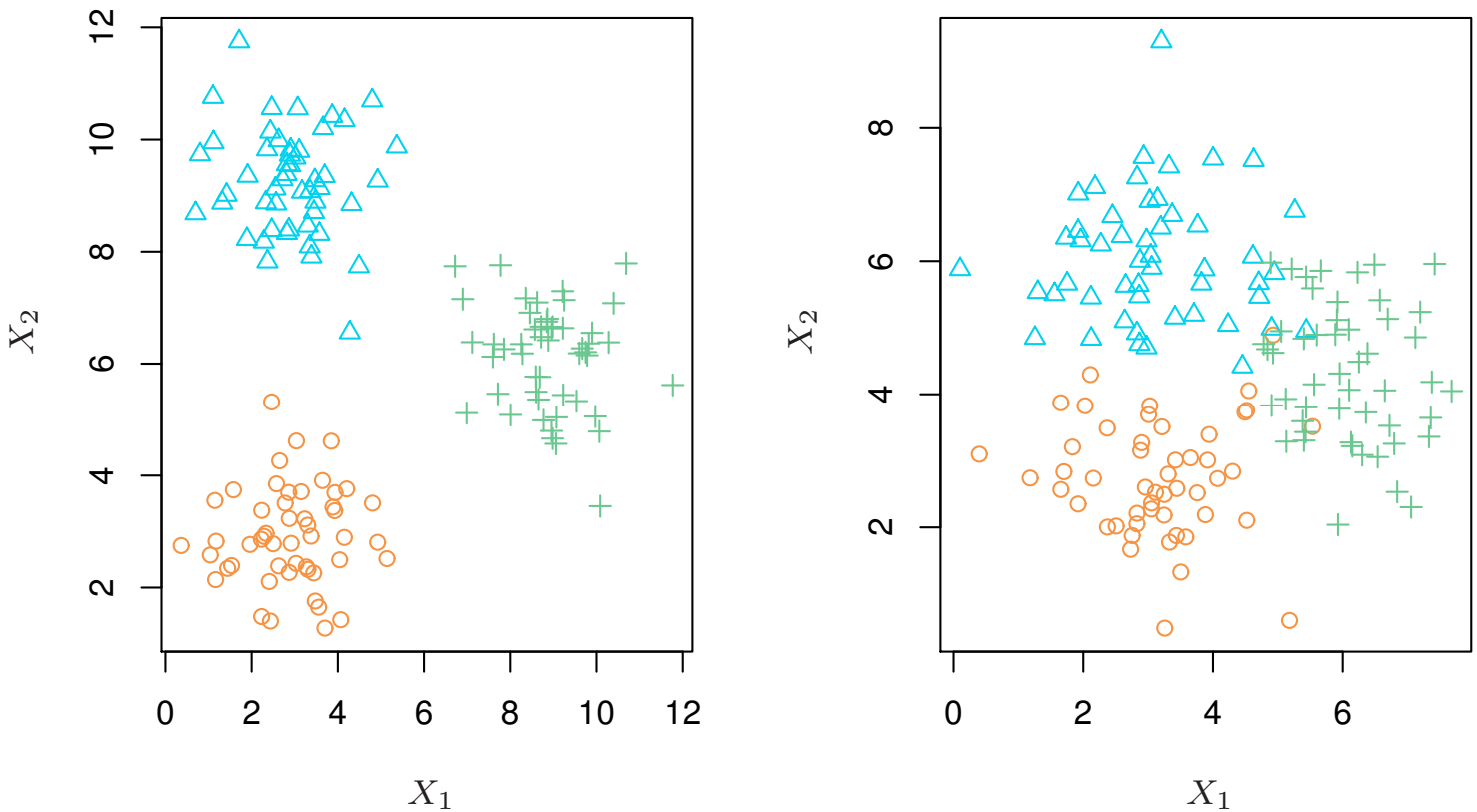
Unsupervised learning denotes the case where we have data x_i, $i = 1, 2, \ldots, n$ but **no** corresponding responses y_i
Hence, the problem is more challenging than the supervised setup

Goals of **unsupervised learning**:

- Try to understand the relationship between the variables
- Identify groups that share common characteristics
- Dimension reduction

Methods: Clustering, Principal component analysis
Examples: Finding customer segments, reduce dimensionality of firm characteristics

Unsupervised Learning: Clustering



Clustering two-dimensional data into three groups

Evaluating Machine Learning Models

Throughout the course we will learn about a wide range of machine learning techniques

Each of these methods has its advantages and disadvantages and

there is no single method that performs 'best' in general

In particular, the performance depends on the problem itself as well

as the amount and type of data available

A key learning goal of the course is to differentiate between machine

learning methods and select the 'best' approach for a specific problem at hand

Getting started for the course

Please install Python and JupyterLab (R and RStudio)

Next week I will provide the introduction to Python starting from the very basics

However, installation is at your own response and you should join the session with Python installed already

If you experience problems, use google and ask your classmates for help

If you are already an expert in Python, feel free to skip next weeks lecture and enjoy your afternoon