MUNI
ECON

Artificial Intelligence in Finance

Introduction - part A

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

September 26, 2024

Course outline

1. September 27:
   - Lecture 1 - Introduction part A
   - Seminar 1 - Intro to R
2. October 4:
   - Lecture 2 - Introduction part B
   - Seminar 2 - Intro to R
3. October 11:
   - Lecture 3 - Supervised learning cont' outcome part A
   - Seminar 3 - Linear models
4. October 18:
   - Lecture 4 - Supervised learning cont' outcome part B
   - Seminar 4 - Penalized models
5. October 25:
   - Lecture 5 - Supervised learning cont' outcome part C
   - Seminar 5 - Penalized models
6. November 1:
   - Lecture 6 - Supervised learning discrete outcome part A
   - Seminar 6 - Tree-based models
7. November 8 - **reading week**
8. November 15:
   - Lecture 7 - Supervised learning discrete outcome part A
   - **Mid-Term**

Course outline

5. November 22:
   - Lecture 8 - eXplainable AI prof. **B Hadji Misheva**
   - Seminar 7 - logistic regression, penalized models
6. November 29:
   - Lecture 9 - Supervised learning discrete outcome part B
   - Seminar 8 - tree-based models
7. December 6:
   - Lecture 10 - Unsupervised learning part A
   - Seminar 9 - PCA and K-Means (Feature engineering)
8. December 13:
   - Lecture 11 - Unsupervised learning part B
   - Seminar 10 - PCA and K-Means (Feature engineering)
9. December 20:
   - Lecture 12 - Forecast combinations
   - Seminar 11 - Forecast combinations

Outline for Section 1

Terminology

- **A**rtificial **I**ntelligence refers to the use of non-biological tools (machines) to solve complex (non-trivial) problems [3], where solving the problem imitates human behavior [6].
  - What are artificial tools? Mix of hardware and algorithms.
  - What are complex problems? Depends on the field.
- **M**achine **L**earning uses statistical techniques to learn from data and solve problems.
- **Deep learning** refers to a specific class of neural networks (statistical method).

The **distinction** between AI and ML (in this course) **not** that **relevant**.

Terminology

- **Algorithm** is a set (series) of tasks, rules, instructions to follow [6].
- Data types:
  - **Features** are inputs to the algorithm.
    - independent variable(s),
    - characteristics,
    - explanatory variable(s),
    - covariate(s).
  - **Labels** are outputs of interest.
    - dependent variable(s),
    - learning variable(s).
    - response variable(s),
    - target variable(s).
- Data structures:
  - Cross-sectional.
  - Time-series.

Terminology

- Learning algorithms [3]:
  - **U**nsupervised **L**earning - algorithm that learns about the structure of features (inputs).
    - Cluster analysis.
    - K-Means and K-Medoid analysis.
    - Network based clustering algorithms.
  - **S**upervised **L**earning - algorithm that learns about the relationship between features (inputs) and label(s) (outputs).
    - Ordinary Least Squares.
    - Logistic regression.
    - Tree-Based methods: decision trees, random forest, boosted trees, ...
    - Neural networks,...
  - **R**einforcement **L**earning - algorithm based on trial and error.

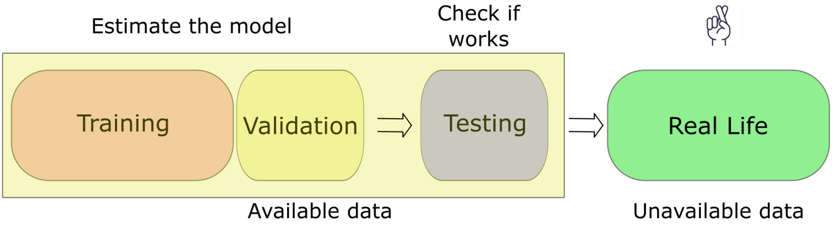
We will cover selected methods in unsupervised and supervised learning.

Terminology

- In supervised learning, we create a statistical model in order to predict future (unseen) [4, 5], unlabelled data.
  - How do we create meaningful statistical (or any other) models? $\rightarrow$ topic of this course.
  - We use inputs from a **training** and **validation** (calibration) database to estimate/train/learn the model.
  - We use inputs/features from **testing** database to assess the suitability of the model.
  - We use unlabelled (before unseen) data with known features to predict future outcomes $\rightarrow$ real life 'test'.

Statistical Learning

Data types during the model building process:



Outline for Section 2

Unsupervised learning

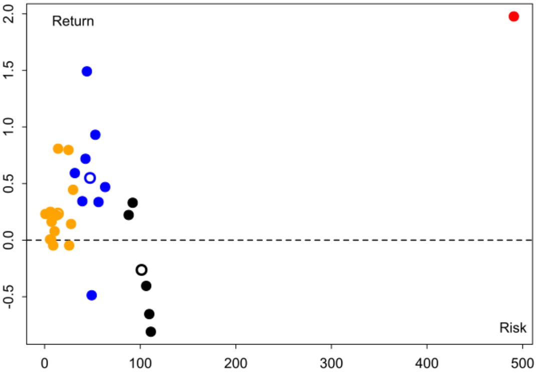
Example

Asset classification - what is a **Bitcoin** anyway?

- We have a sample of n assets with monthly returns from April 2017 to August 2024.
- We find two features of each asset:
  - Average monthly return.
  - Average monthly volatility.
- Using these data, we estimate a K-means classification algorithm with three (*hyper-parameter*) clusters.
- A new asset is introduced to the data, BTC/USD rate and we want to predict to which asset class it belongs.

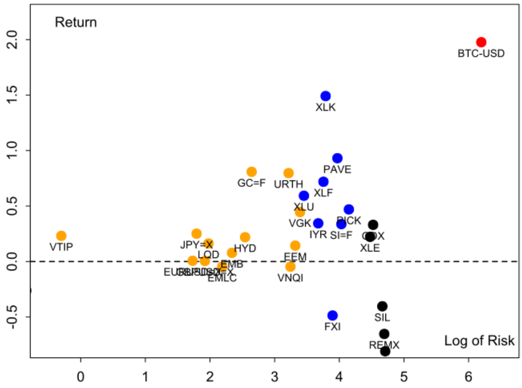
Unsupervised learning

Example



Unsupervised learning

Example



Outline for Section 3

Example

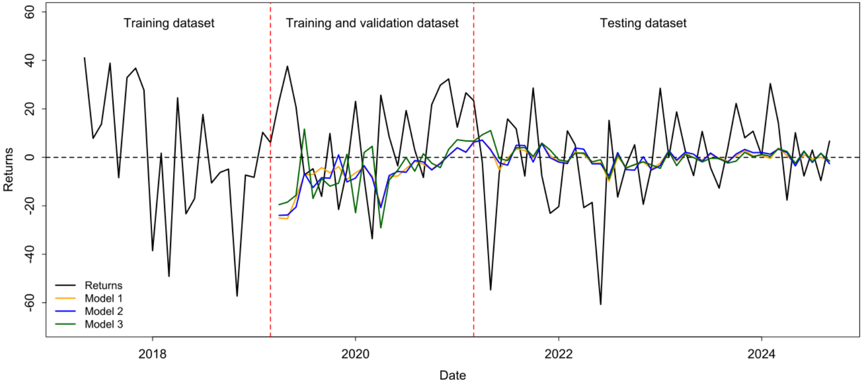
Trading algorithm:

- We have a sample of n monthly returns from April 2017 to August 2024 for BTC/USD rate.
- We consider **three** simple **linear models** to predict next month's returns:
- We use (expanding) **training** and **validation** sample to learn which model should produce most accurate monthly return forecasts in the next month.
- We use data from the **testing** sample to estimate the accuracy of our algorithm.

Statistical Learning

Example

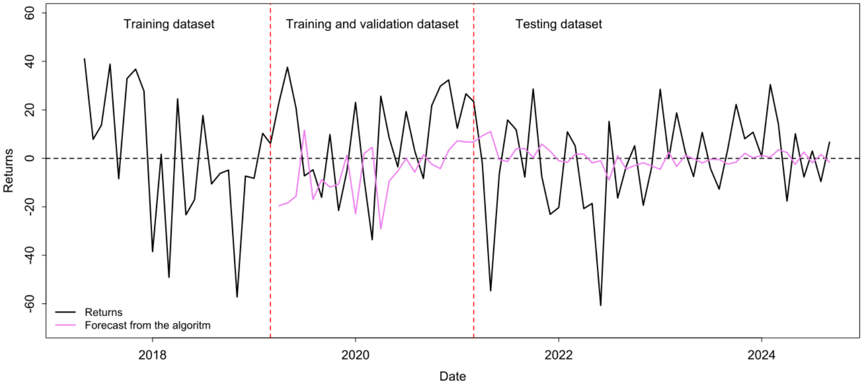
Competing models:



Statistical Learning

Example

Supervised algorithm allows us to learn which model performed the best:



Statistical Learning

Terminology

Common parameter tuning (estimation) techniques in AI/ML [3]:

- **Cross-sectional** data:
  - **Classification**, where outcomes of the variable(s) of interest - labels - are characters representing different classes.
  - **Regression**, where outcomes of variable(s) of interest are real-valued (usually continuous) numbers.
- **Time-series** data:
  - **Classification**, where outcomes of the variable(s) of interest - labels - are characters representing different classes.
  - **Regression**, where outcomes of variable(s) of interest are real-valued (usually continuous) numbers.

Outline for Section 4

Applications

- Trading algorithms.
  - Profit maximization.
  - Risk management.
- Credit risk modelling.
  - loans - mortgage, consumer, car, business,... .
  - corporations.
- Forecasting of macro-financial variables:
  - consumer/business confidence indicators.
  - house prices.
  - credit market growth.
  - interest rates... .
- Customer classification/segmentation.
- Hedonic pricing models.
- ... what else?

Outline for Section 5

Variance bias trade-off

- When we estimate a model, we estimate parameters. Each parameter is estimated with an **error**. Increasing the number of model parameters **might** improve fit in the **training dataset**, but this comes at the expense of the parameter **uncertainty**.
- This **might** lead to a **variance bias trade off**. Model fits well in the **training sample**, but performs poorly in the **testing sample**.
- Predictions from an over-fitted model show high levels of inaccuracy.
- Solution?
  - We often **sacrifice bias** (in-sample accuracy) to **hopefully gain accuracy** on the testing sample.

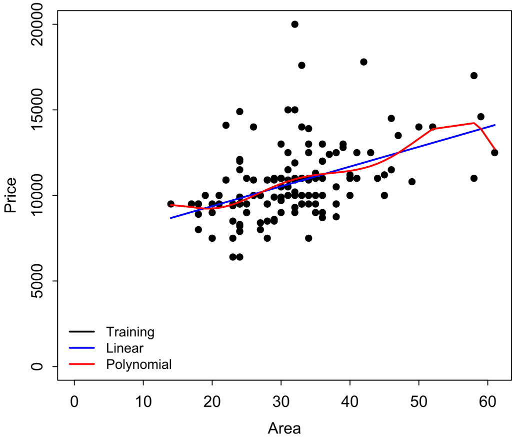
Variance and bias trade-off

Estimate two models that links price of a rent for a one-bed room apartment to the size (in terms of m^2) of the apartment.

- One is a **simple linear model**, the other a more **complex** polynomial **model**.
- We only use training sample to estimate the models.
- Within the training sample, more complex model performs better.

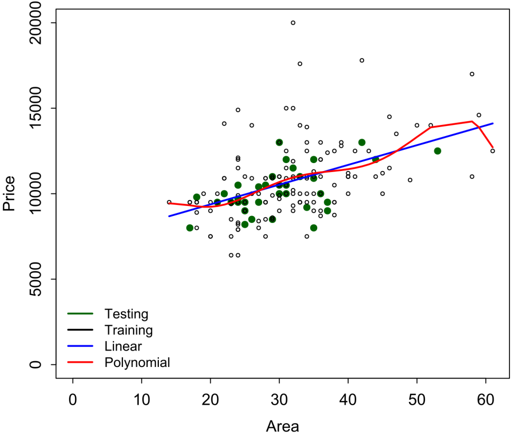
Variance and bias trade-off

More complex model performs better on the **training sample**.



Variance and bias trade-off

- We introduce the **testing sample** (green dots).
- The simple model now performs better.
- How is that possible? Polynomial model was **over-fitting**!



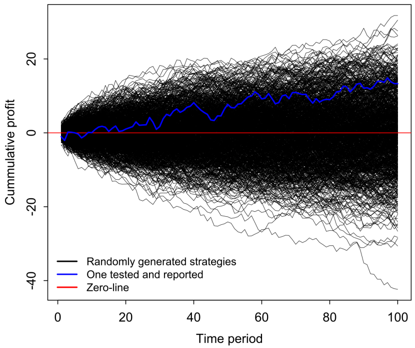
Data-snooping bias

'The first principle is that you must not fool yourself and you are the easiest person to fool.' R.P. Feynman

- Torturing data until a desired (or any) result is found [1] is often referred to as **data-snooping bias** or **data-dreading** or **p-hacking**.
- **Positive outcome bias** - not only in academic literature.
- In trading applications - how many strategies are being tested?
- Solution?
  - Trading strategy based on theory (not a solution - good practice).
  - Thorough statistical testing (e.g. model confidence set [2]).

Data-snooping bias

- We generate 1000 **random** trading strategies.
- Cumulative profit should be around 0.
- Because of the randomness, there are going to be strategies that look profitable (blue line).
- In fact, the signals were generated randomly.



Interpretability

Interpretability refers to the possibility of the analyst to tell, **which variables are relevant** for predicting a certain outcome and in what way (effect size and direction).

- From a policy perspective, forecasting accuracy is relevant, but what should a policy maker do?
- Which parameters are relevant is possible to estimate - yet problematic.
  - **Historical context** from Econometrics and linear models.
  - **Regulatory requirement** for anti-discrimination purposes.

[1] Guillaume Coqueret and Tony Guida. *Machine Learning for Factor Investing: R Version*. Chapman and Hall/CRC, 2020.

[2] Peter R Hansen, Asger Lunde, and James M Nason. “The model confidence set”. In: *Econometrica* 79.2 (2011), pp. 453–497.

[3] Yves Hilpisch. *Artificial Intelligence in Finance*. O’Reilly Media, 2020.

[4] Gareth James et al. *An introduction to statistical learning*. Springer, 2013.

[5] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2021.

[6] Fred Nwanganga and Mike Chapple. *Practical machine learning in R*. John Wiley & Sons, 2020.

MUNI
ECON

Artificial Intelligence in Finance

Introduction - part A

Štefan Lyócsa

Department of Finance, Faculty of Economics and Administration

September 26, 2024

MASARYK
UNIVERSITY