Machine Learning for Social Science

Christoph Kern & Malte Schierholz

University of Mannheim, MZES
Institute for Employment Research (IAB)

c.kern@uni-mannheim.de
Malte.Schierholz@iab.de

March 21 and 22, 2018

What is Machine Learning?

The term was coined by Arthur Samuel (1959) in a paper titled

> *Some Studies in Machine Learning Using the Game of Checkers*

It starts as follows

> *The studies reported here have been concerned with the programming of a digital computer to behave in a way which, if done by human beings or animals, would be described as involving the process of learning. [...] Programming computers to learn from experience should eventually eliminate the need for much of [the] programming effort.*

What is Machine Learning?

A prominent definition:

> *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

– Tom Mitchell (1997)

A historical perspective

- ML originates from artificial intelligence / computer science
- 1980s goal: develop intelligent systems (problem solving, reasoning)
- Since then, ideas from pattern recognition and statistics were adopted and changed the field ...

Langley (2011)

This course will focus on statistical learning.

What is Statistical Learning?

What is statistical learning?

> *[Use data] to extract important patterns and trends, and understand "what the data says". We call this learning from data.*

– Hastie, Tibshirani, Friedman (2009)

Course Outline

Today

- Introduction (9:30-10)
- Method 1: Variable selection and the Lasso (10-12)
- Method 2: Recursive Partitioning and Decision Trees (13-16)
- General methodology (16-17)

Tomorrow

- General methodology (9:30-10)
- Method 3: Random Forests (10-12)
- Method 4: Boosting (13-15)
- Supervised learning applications in the social sciences (15-16)
- (Method 5: Support Vector Machines)
- Method 6: Deep Learning and Neural Networks (16-17)

Recommended Literature

Hastie, Trevor, Tibshirani, Robert & Friedman, Jerome (2009)
The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer. `https://web.stanford.edu/~hastie/ElemStatLearn/`

James, Gareth, Witten, Daniela, Hastie, Trevor & Tibshirani, Robert (2013)
An Introduction to Statistical Learning. Springer.
`http://www-bcf.usc.edu/~gareth/ISL/`

and many more references given throughout this course...

Resources and Links

Slides and code from this class

- https://github.com/chkern/

Machine Learning with R: Overviews and Metapackages

- Overview: https://cran.r-project.org/web/views/MachineLearning.html
- Caret Documentation: http://topepo.github.io/caret/index.html
- mlr Documentation: https://mlr-org.github.io/mlr-tutorial/devel/html/
- H2O Documentation: http://docs.h2o.ai/ (not easily installable at IAB)

Just for fun

- How Machines Learn https://www.youtube.com/watch?v=R9OHn5ZF4Uo

IAB intro to R

1. Find course material in Maltes Quickablage: `\Iab.baintern.de\dfs\017\Ablagen\D01700-Quickablage\Schierholz\MachineLearning`
2. Copy `.Rprofile` to your personal directory `Z:\EigeneDateien`
3. Connect to a server and open RStudio on your computer
4. Change Tools → Global Options → General → Default Working Directory to `Z:\EigeneDateien`
5. Restart RStudio and install the packages needed for this course (see file install_packages.Rmd)

Text as Data

Text as Data-Basics

A short digression

Text as Data

Text Mining, Text as Data or "Distant Reading"

- Information overload: Text is everywhere, but it is too much to read it all
- How can we still gain insights from it?

(Grimmer et al. 2013; Gentzkow et al. 2017)

Text as Data-Framework

Document		**Outcome**	
		Estimate	Unknown value
	$\longrightarrow$	$\hat{V}_1$	V_1
	$\longrightarrow$	$\hat{V}_2$	V_2
$\vdots$			$\vdots$
	$\longrightarrow$	$\hat{V}_n$	V_n

- Computers can calculate numbers (estimates) from large documents
- Humans must evaluate if the estimates are useful

Text as Data

Some examples

- Authorship: Did Philip Wright or his son Sewall write an appendix in which instrumental variables were invented?
- Stock Prices: Can one forecast changing stock prices from companies' annual reports or from newspaper articles?
- Google Flu: Using billions of search queries, can one estimate the flu prevalence for specific regions?

(Gentzkow et al. 2017)

Preprocessing

Preprocessing is needed for most text mining methods

Document		**Numeric Vector**		**Outcome** Est.	Unknown
	$\longrightarrow$	C_1	$\longrightarrow$	$\hat{V}_1$	V_1
	$\longrightarrow$	C_2	$\longrightarrow$	$\hat{V}_2$	V_2
$\vdots$		$\vdots$			$\vdots$
	$\longrightarrow$	C_n	$\longrightarrow$	$\hat{V}_n$	V_n

Preprocessing

Document:
Time flies like an arrow. Fruit flies like a banana.

Same document after cleaning and processing:

	arrow	banana	fli	fruit	like	time
$C_i =$	1	1	2	1	2	1

Steps taken:

1. Remove punctuation
2. Lowercase letters
3. Remove stopwords (like "a", "the")
4. Stemming ("flies" $\rightarrow$ "fli", based on a linguistic algorithm)
5. Count word frequency

Preprocessing

Document:
Time flies like an arrow. Fruit flies like a banana.

Same document after cleaning and processing:

	arrow	banana	fli	fruit	like	time
$C_i =$	1	1	2	1	2	1

Preprocessing aims to simplify the document without losing important information, but

- Meaning of words is ignored (e.g. "flies")
- Word order is ignored (so-called "bag-of-words" representation)

Many more ways exist for processing (e.g. N-grams, letterwise, tf-idf)

→ Optimal approach depends on the research question

Document-Term Matrix

Preprocessing converts a *corpus* (= a set of documents) into a *Document-Term Matrix*

$$C = \begin{pmatrix} C_1 \\ \vdots \\ C_i \\ \vdots \\ C_n \end{pmatrix} = \begin{array}{c} \begin{array}{ccccccc} \textit{arrow} & \textit{banana} & \textit{fli} & \textit{fruit} & \textit{like} & \textit{time} & \ldots \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 2 & 1 & 2 & 1 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 1 & \ldots \end{pmatrix} \end{array} \quad (1)$$

Matrix is ...

- sparse (= many zeros) → Do fast algorithms exist?
- high-dimensional (= several thousand variables / columns)

→ Statistical learning useful

References

Gentzkow, Matthew, Bryan T. Kelly & Matt Taddy (2017)
Text as data. *NBER Working Paper No. 23276*. 1–53

Grimmer, Justin & Brandon M. Stewart (2013)
Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis 21*(3). 267–297

Hastie, Trevor, Tibshirani, Robert & Friedman, Jerome (2009)
The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Langley, Pat (2011)
The changing science of machine learning. *Machine Learning* **82**. 275–279

Mitchell, Tom M. (1997)
Machine Learning. McGraw-Hill.

Samuel, Arthur L. (1959)
Some studies in machine learning using the game of Checkers. *IBM Journal of Research and Development* **3**(3). 210–229