Machine Learning for Social Scientists

Momin M. Malik, PhD <momin_malik@cyber.harvard.edu>
Data Science Postdoctoral Fellow
Berkman Klein Center for Internet & Society at Harvard University

Fairness, Accountability & Transparency/Asia, 11 January 2019

Slides: https://mominmalik.com/ml_socsci.pdf

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Learning goals by background

- No background in social statistics:
  - See what doing machine learning looks like in practice
- Linear regression, in Excel, SPSS, or Stata:
  - Identify use cases for machine learning
  - Use cross-validation
- Logistic regression, and/or Python or R:
  - Build and evaluate a basic machine learning model

About me

History of science →

Social science →

Machine learning →

Social science

Structure

- Preliminaries
- What is machine learning?
- When use machine learning?
- Key concepts
  - "Prediction"
  - Overfitting, Cross-validation
  - Confusion matrix
  - Feature engineering
- Interactive, live demonstration in R

Preliminaries

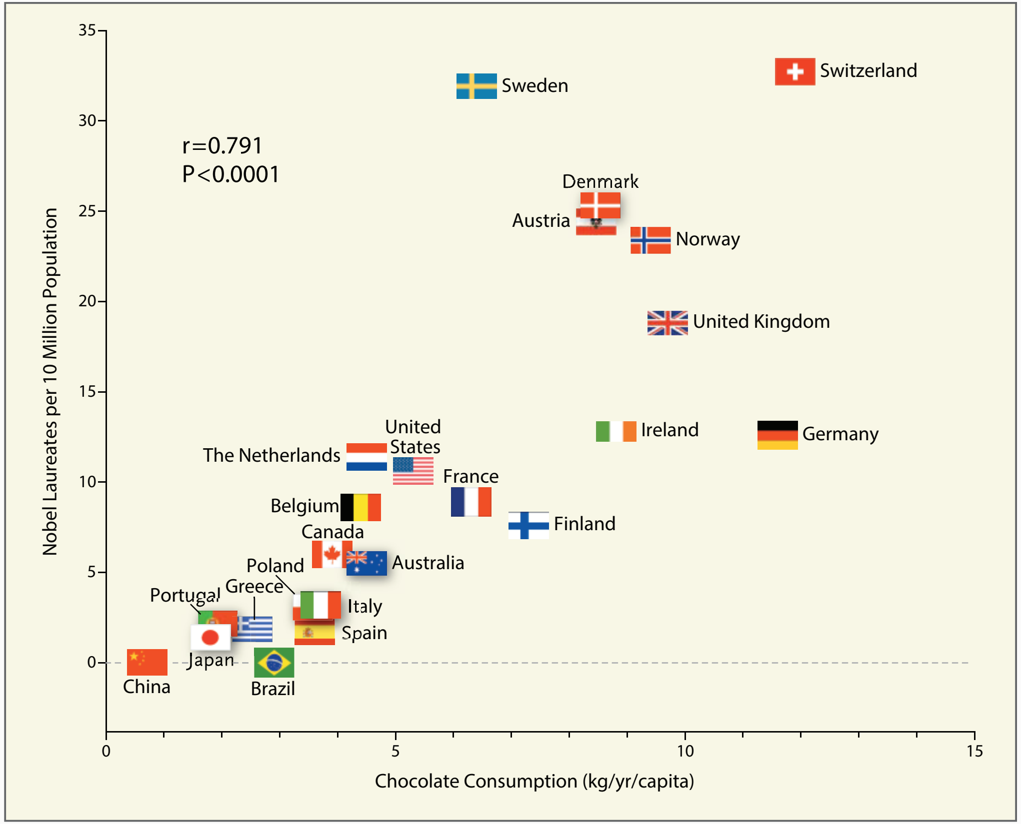
BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Follow along with the demonstration!

- If you don't have it already, download and install R (search: "install R")
- Also install RStudio (search: "install RStudio")
- Installation will take about as long as the introduction

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Basic background: Correlation



Messerli, 2012, *NEJM*

BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Basic background: Idea of model "fit"

- All machine learning and statistics models take in data, process them via some assumptions, and then give out something: relationships, and/or likely future values.
- The processing is called "fitting", and the output is called a "fit." Machine learning uses "learning" or "training," but it's the same.

What is machine learning?

ML = *Using correlations for prediction*

- Textbook definitions are aspirational. In practice, machine learning is about *finding correlations that we can use for prediction*
- Spurious correlations are fine, so long as they are robust
- Machine learning is not well suited for modeling or understanding the world (although people assume it is)

Machine learning is all statistical

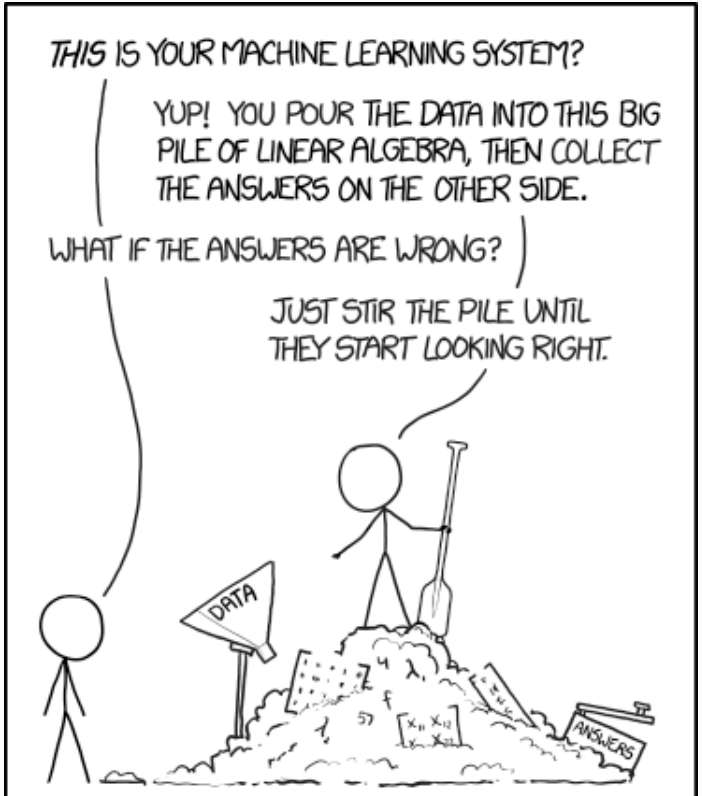
Baron Schwartz @xaprb Follow

When you're fundraising, it's AI
When you're hiring, it's ML
When you're implementing, it's linear regression
When you're debugging, it's printf()

12:52 AM - 15 Nov 2017

5,545 Retweets **12,654** Likes

90 5.5K 13K



xkcd.com/1838

Statistics vs. machine learning

- Same underlying principles, many of the same models, techniques, and tools
- Used for different ends, and used in very different ways (ML: no *p*-values!)
- Folded into machine learning: data mining, pattern recognition, some Bayesian statistics

(Questions so far?)

When use machine learning?

Recover a hard-to-get signal via proxy

- E.g., 500,000 tweets, only two human coders
- Have both coders label 1,000 random tweets
  - Inter-coder reliability (CS: “inter-annotator agreement”)
- Find correlations between *word frequencies* in the tweets and the *human-given labels*
- Use correlations to label other 499,000 tweets

Key components of a good use case

1. We have "ground truth" (e.g., human labels, previous failures), and
2. Ground truth is hard to collect, and
3. We have some readily available proxy measure, and
4. *We don't care how or what in the proxy recovers the ground truth, only that it does*

ML: When *only* accuracy* matters

* Or other relevant metric of success

The surprising part

- *The best-fitting* (most accurate*) *model does not necessarily reflect how the world works*
- This has been shocking in statistics for decades (Stein's paradox, Leo Breiman's "two cultures"), but little known outside
- We can "predict" without "explaining"!

* Or other relevant metric of success

Most useful for building systems

- Narrow people's choices to "relevant" ones (friend connections, search results, products)
- Detection (facial recognition, fraud)
- Anticipation (customer demand, equipment failure)
- ...Seldom happens in social science

For exploratory analysis

- The best fitting model is worthwhile to explore
  - E.g., *variable selection* or *variable importance*
- Unsupervised learning (synonymous with clustering) techniques
  - Topic models

(Questions so far?)

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Background needed

How much math?

- To be a practitioner, same as what you need to do social statistics: algebra and a bit of calculus
- To understand underlying *mechanics*: linear algebra, multivariate calculus
- To understand underling *principles*: learn probability and mathematical statistics

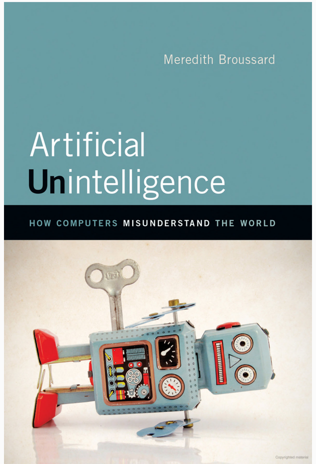
How much programming?

- For personal use: at least be able to write loops and functions, know up to sorting algorithms
- For production: some software development principles
- Alternatives: Weka and Rapid Miner have graphical interfaces, no programming required

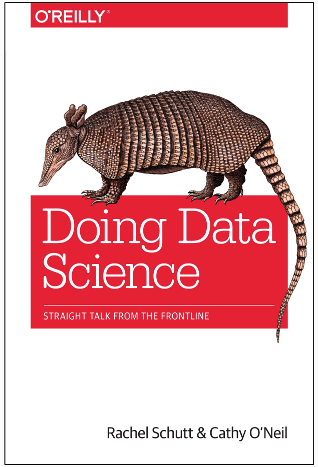
Which language/environment?

- Weka, Rapid Miner
  - Basic use
- Python (numpy, scipy, scikitlearn, pandas)
  - Scale, integrating into production, best visualizations (sometimes), deep learning
- R
  - More flexibility in how to use techniques, a self-contained environment, and better integration with (social) statistics

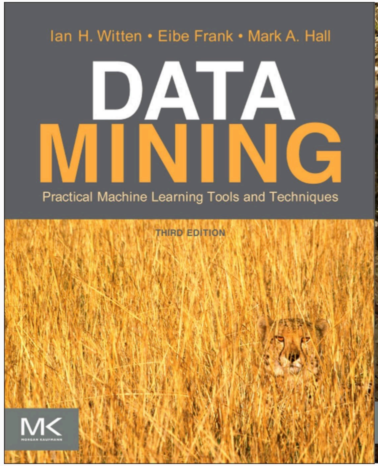
Resources

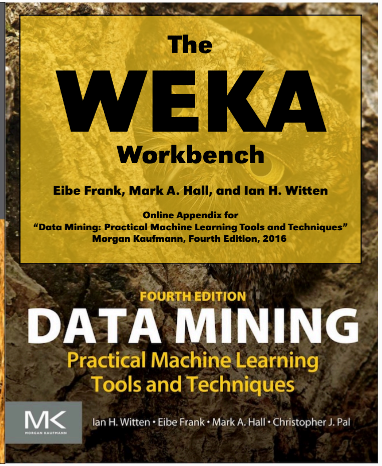


Chapter 7: ML in action

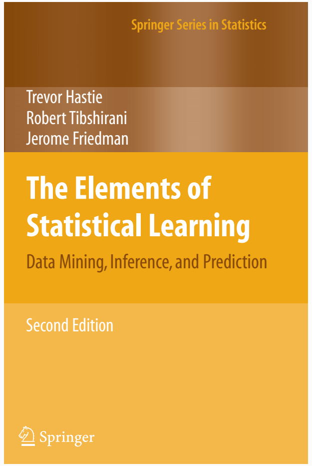


Basics





Machine learning without needing to know any programming



Theory

Unfortunately, I haven't spent time looking through online courses to have one I recommend.

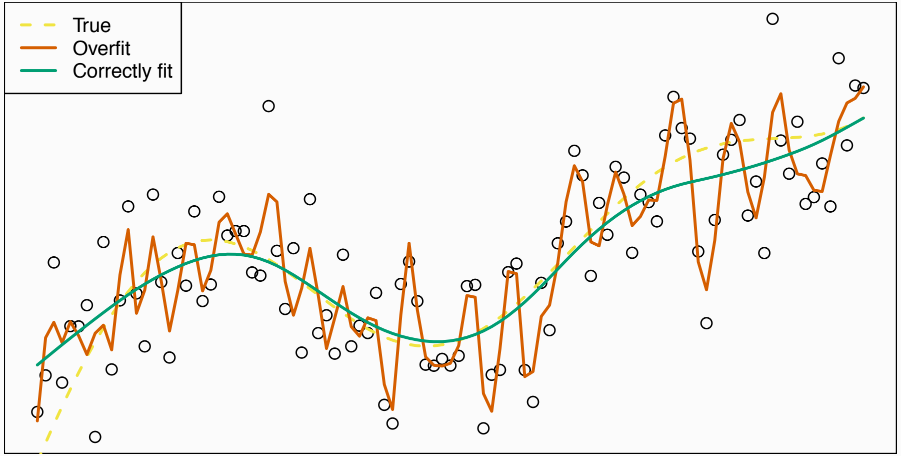
(Questions so far?)

Key concepts

"Prediction" means correlation

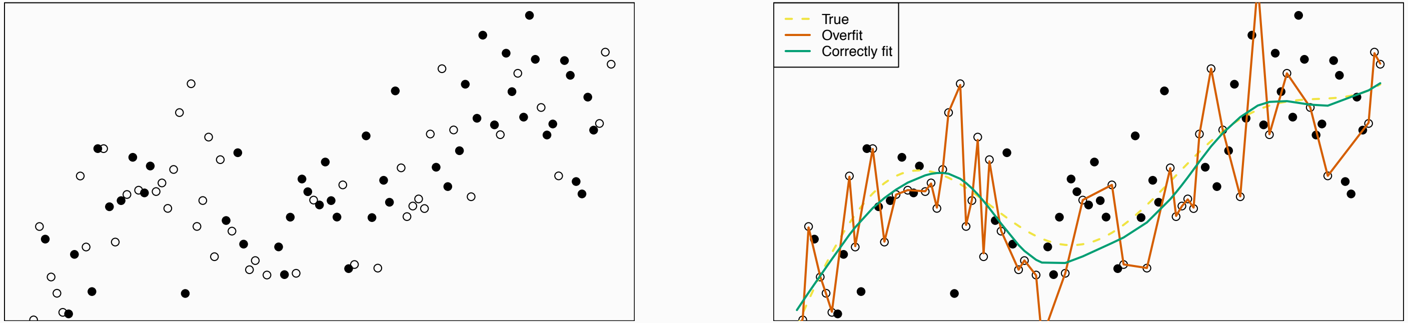
- *Prediction* is a technical term, meaning "fitted values" in both statistics and machine learning
- "X predicts Y" is better read as "In a model, X correlates with Y"
- *A prior correlation does not necessarily predict!* Hopefully it does, but testing is key

Overfitting: fit to noise



- If we are no longer guided by theory, and use automatic methods, we risk overfitting: fitting to the the noise, not the data

Data splitting: Catch overfitting



- Idea: if we split data into two parts, the signal should be the same but the noise would be different
- *Cross validation:* Fitting the model on one part of the data, and "testing" on the other

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

Confusion matrix

		True label	
	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

Confusion matrix

		True label	
	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

$\boxed{\text{Accuracy} = (TP+TN)/N}$

↑ Overall correct

Confusion matrix

		True label	
	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative

$\boxed{\text{Accuracy} = (\text{TP}+\text{TN})/N}$

↑ Overall correct

"accuracy paradox": if 5 out of 1000 are positive, a useless (all negative) classifier is 99.5% accurate

Confusion matrix

True label

	N	Positive	Negative
Predicted label	Predicted positive	True positive	False positive
	Predicted negative	False negative	True negative
		Recall/ sensitivity = TP/(TP+FN)	← How many you detect

Accuracy = (TP+TN)/N

↑ Overall correct

"accuracy paradox": if 5 out of 1000 are positive, a useless (all negative) classifier is 99.5% accurate

Confusion matrix

		True label			
	N	Positive	Negative		Accuracy = (TP+TN)/*N*
Predicted label	Predicted positive	True positive	False positive	Precision = TP/(TP+FP)	↑ Overall correct
	Predicted negative	False negative	True negative	↑ How much is relevant	“accuracy paradox”: if 5 out of 1000 are positive, a useless (all negative) classifier is 99.5% accurate
		Recall/ sensitivity = TP/(TP+FN)	← How many you detect		

Confusion matrix

True label

	N	Positive	Negative		Accuracy = (TP+TN)/N
Predicted label	Predicted positive	True positive	False positive	Precision = TP/(TP+FP)	↑ Overall correct
	Predicted negative	False negative	True negative	↑ How much is relevant	"accuracy paradox": if 5 out of 1000 are positive, a useless (all negative) classifier is 99.5% accurate
		Recall/ sensitivity = TP/(TP+FN)	← How many you detect		
		How many → you correctly reject	Specificity = TN/(TF+TN)		

Confusion matrix

		True label			
	N = 165	Positive: 105	Negative: 60		Accuracy = 0.91
Predicted label	Predicted positive: 110	TP = 100	FP = 10	Precision = 0.91	↑ Overall correct
	Predicted negative: 55	FN = 5	TN = 50	↑ How much is relevant	
		Recall/ sensitivity = 0.95	← How many you detect		
		How many → you correctly reject	Specificity = 0.83		

Feature engineering

- In social science, we have the variables (e.g., the survey responses)
- In machine learning, you might have lots of text data, or lots of sensor data, for a single outcome
- "Feature engineering": heuristics to extract variables to summarize the data. Huge part of ML, no systematic solution for every data type

(Questions so far?)

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

Demo (Background)

Topic: Datacamp "Titanic" example

New York American

EDITION FOR GREATER NEW YORK.

BUSINESS PROPERTY

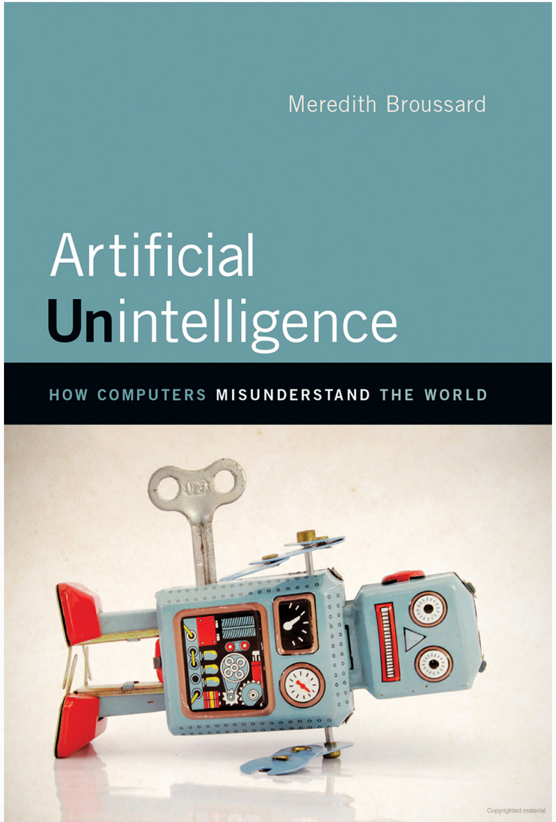
No. 10,499. TUESDAY, APRIL 16, 1912. 16 PAGES PRICE ONE CENT

J. J. ASTOR LOST ON TITANIC
1,500 TO 1,800 DEAD

John Jacob Astor was among the passengers who went down with the ship, according to a wireless dispatch received by Bradstreets last night from the liner Olympic. Mrs. Astor was saved and is being brought to shore by the Carpathia.

The Wireless Operator at Cape Race, Newfoundland, Flashes: "Eighteen Hundred Lives Have Been Lost in the Wreck of the Titanic."

Commentary by Meredith Broussard



- Captain: “Put the women and children in and lower away.”
- First Officer: women and children *first*
- Second Officer: women and children *only*
- “the lifeboat number isn’t in the data. This is a profound and insurmountable problem. Unless a factor is loaded into the model and represented in a manner a computer can calculate, it won’t count… The computer can’t reach out and find out the extra information that might matter. A human can.”

Social science baseline for comparison

- 5 econometrics papers from Frey, Savage, and Torgler (2009-2011) give a comparative "social statistics" approach

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY AT HARVARD UNIVERSITY

Demo time!

Data:
https://github.com/momin-malik/guides/raw/master/titanic.csv

https://github.com/momin-malik/guides/raw/master/titanic.csv