Lecture 1 Introduction to Machine Learning

Li Xing

June 13, 2023

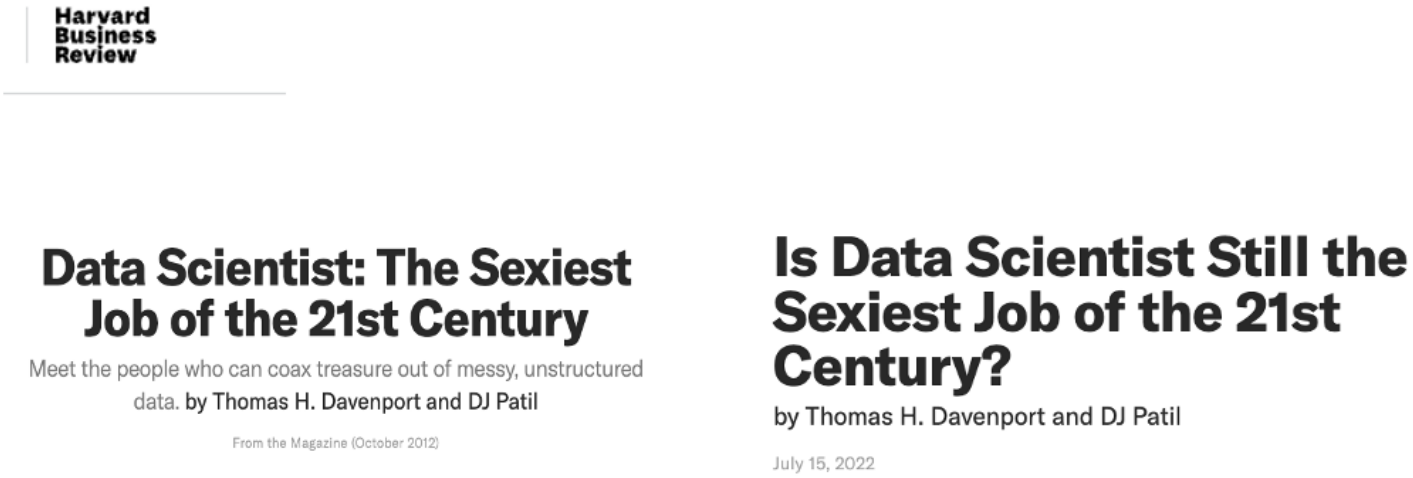
Welcome to the Short Class

Scopes of the short course.

- Introduction to Statistical Machine Learning
- Linear and Logistic Regressions
- Penalized Regressions

And all course materials are here.

Importance of Topics



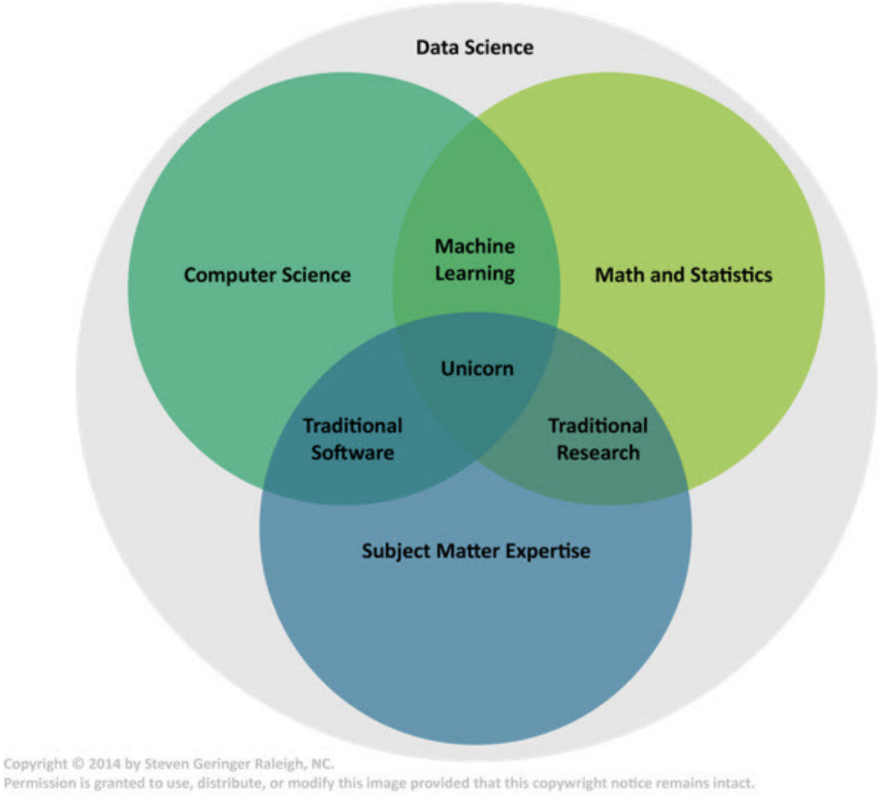
By 2019, postings for data scientists on Indeed had risen by 256%, and the U.S. Bureau of Labor Statistics, predicts data science will see more growth than almost any other field between now and 2029.

What is Statistical Machine Learning?

“The field of study interested in the development of computer algorithms for transforming data into intelligent action is known as machine learning.”
— Brett Lantz

“Statistical learning theory is a framework for machine learning drawing from the fields of statistics and functional analysis.”
— Wikipedia of Statistical Learning Theory

What is Data Science?



1

1. The figure is downloaded from https://www.linkedin.com/pulse/data-science-scientist-simply-put-marie-l-clark/

What is R?

- R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. (From Wikipedia)
- It is open-source free software.
- R and Python

Python is better for...	R is better for...
Handling massive amounts of data	Creating graphics and data visualizations
Building deep learning models	Building statistical models
Performing non-statistical tasks, like web scraping, saving to databases, and running workflows	Its robust ecosystem of statistical packages

2

2. Information is from the coursera website.

Data Scientists' Tasks

- Data Management and Wrangling
- Visualization
- Software Development
- Model Application and Interpretation
- Generalization (Turn into actionable knowledge)
- And more such as Experimental Design, Data Acquisition, Data Storage, Building Interface for End Users, Project Management, etc.

Data Science Process

- From data repositories
- From studies and experiments (Stat 345/834)

- 60%-70% of the time in analysis
- Crucial step

- Explanatory Data Analysis (EDA)
- Explore data in the shallow surface

- Model fitting
- Model/method comparisons

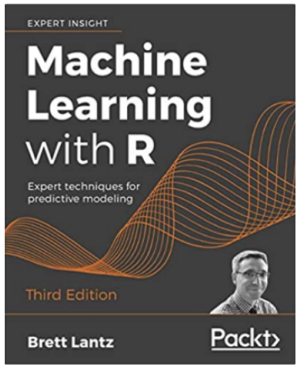
- Presentation, report, and articles.

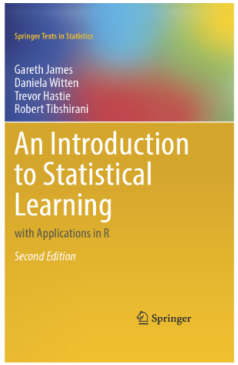
Suggested Learning

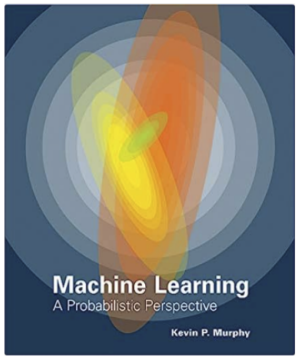
- Knowledge on statistical machine Learning
- Learning R (basic functions, visualization, & reproducibility)
- Learning tools (RStudio, Rmarkdown, ggplot, github, statistical models, & some basic machine learning methods)

Books for Knowledge Learning

- Lantz, B. (2019). Machine Learning with R. Packt Publishing Ltd.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. New York: Springer.
- Murphy, K. P. (2012) Machine Learning: A Probabilistic Perspective. MIT Press.







Books for R Learning

- Wickham, H., & Grolemund, G. (2016). R for data science: import, tidy, transform, visualize, and model data. O'Reilly Media, Inc.
- Grolemund, G. (2015) Hands-On Programming with R: Write Your Own Functions and Simulations. O'Reilly Media, Inc.
- Many other materials online.

Tidyverse for Data Science

A collection of R packages designed for data science

The Leading Developer

Hadley Wickham

- Chief Scientist at Posit (previously called RStudio)
- Adjunct Professor of Statistics at the University of Auckland, Stanford University, and Rice University
- Recipient of the 2019 COPSS Presidents' Award.

Sessions of the Short Course

- **Session 1:** 9:30am - 11:30am
    - 9:30 am-11:00 am: Knowledge Lecture on Introduction to Statistical Machine Learning, Linear Regression and Logistic Regression.
    - 11:10 am-11:30 am: R Lecture on R/Rstudio Installation, R Introduction and Data Management by Ms. Lina Li.

Sessions of the Short Course

- **Session 2:** 1:00 pm - 2:30 pm
- 1:00 pm-1:30 pm: R Lecture on Linear and Logistic Regression by Mr. Kyle Gardiner.
- 1:30 pm-2:00 pm: Hand-on Practice Session on R/Rstudio installation + Introduction + Data Management by Lina Li, Kyle Gardiner and Jing Wang.
- 2:00 pm-2:30 pm: Hand-on Practice Session on Linear and Logistic Regressions with R by Kyle Gardiner, Lina Li and Jing Wang.

Sessions of the Short Course

- **Session 3:** 3:00 pm - 4:30 pm
- 3:00 pm-3:50 pm: Knowlege Lecture on Penalized Regressions by Dr. Xing.
- 3:50 pm-4:10 pm: R Lecture on Penalized Regressions by Kyle Gardiner.
- 4:10 pm-4:30 pm: Hand-on Practice Session on Penalized Regressions by Kyle Gardiner, Lina Li and Jing Wang.

Our Teaching Assistants

Lina Li

Kyle Gardiner

Jing Wang

For Hands-on Sessions

Please note that a laptop is required for all the hands-on sessions.