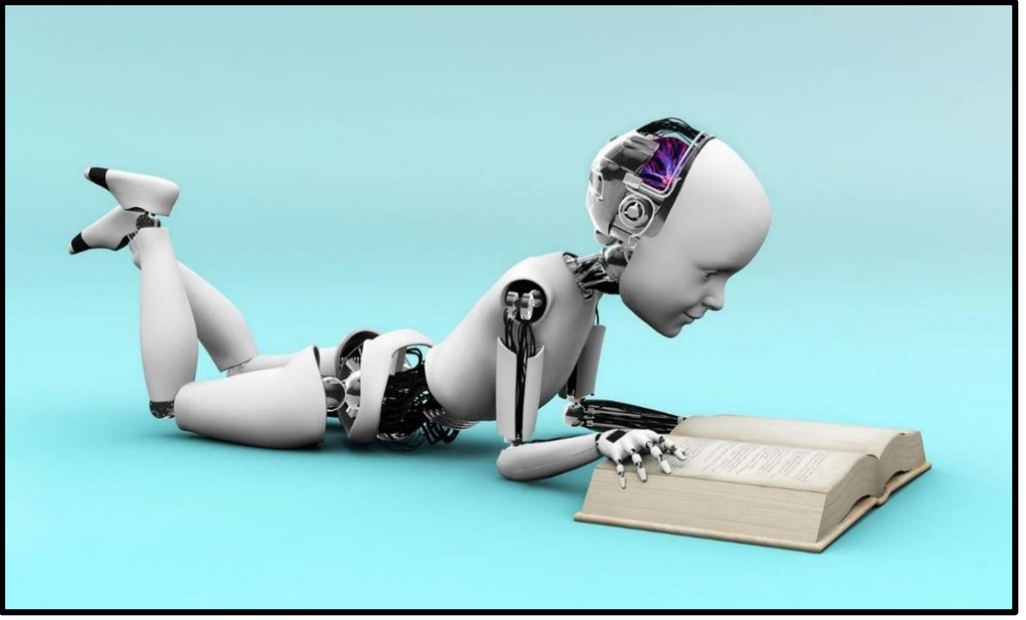
Machine Learning from Data

Zohar Yakhini

IDC Herzliya

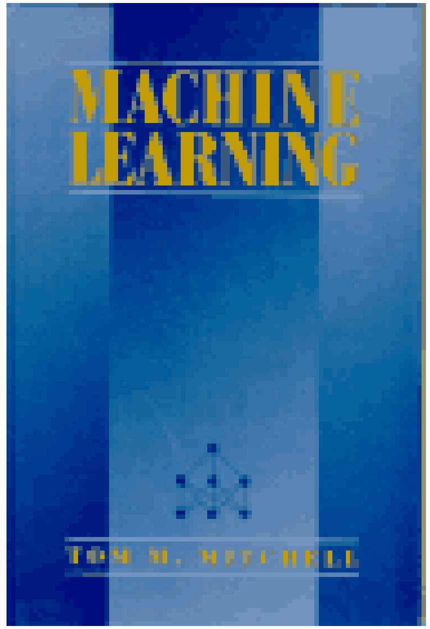
2022/TASHPAB



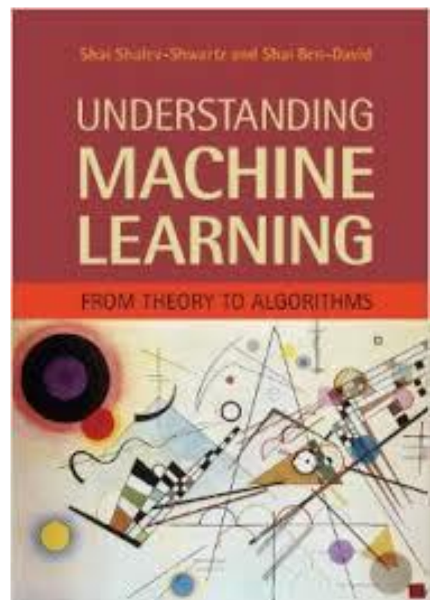
General

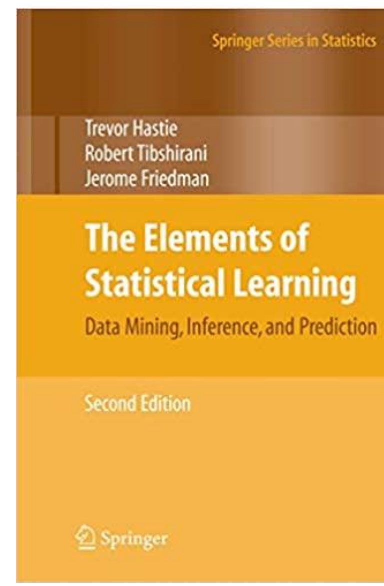
- Instructors: Zohar Yakhini, Ariel (Arik) Shamir
- Slides (all in English) all prepared by Zohar Yakhini, Arik Shamir, Leon Anavy and Ben Galili + others as noted
- Principal TA: Ben Galili
- TAs and recitation instructors:
  Saar Buchnik (MSc student IDC),
  Shuli Finley (MSc student IDC),
  Yarden Rachamim (MSc student IDC),
  Yinnon Meshi (PAN, MSc EE Technion)
- Grade:
  - + 0.5*HW + 0.5*exam
  - + Must pass BOTH exam AND HW
  - + Returning students w high HW grades (>80) can be exempt from some HW.

IDC HERZLIYA



1. Tom Mitchel: Machine Learning
2. Duda, Hart, Stork:
   Pattern Classification
3. C Bishop:
   Pattern recognition and machine learning
4. T Hastie, R Tibshirani, J Friedman:
   Elements of Statistical Learning
5. Ben-David and Shalev-Schwarz:
   Understanding ML
6. Partial lecture notes (will be distributed as alpha version)
7. Internet …





HW Assignments

- 6 homework assignments, only 5 will count (n-1)
- But
  - **HWA5 (theory assignment) is mandatory**
    (at least 4 more out of the other 5 are required)
  - **There will be NO changes to the due dates of the HWAs**
- First one assigned this week, after the Thu English class
- Programming assignments will be done in Python using Jupyter notebooks.
  Help on Python will be available in the first two weeks of the semester
  (by appointment; please write if you are interested)
- Assignments in PAIRS
  (Singles only in exceptional cases and with specific explicit email approval from Ben)

Office hours

- TAs: will publish on Moodle and Piazza and in the recitation slides
- Zohar Yakhini Thu 10-11, by email appointment
- Special Python office hours will be held by volunteer students. By email appointment.

IDC HERZLIYA

Background and pre-requisites

- Linear algebra and calculus
- Probability and statistics
- Algorithms, data structures
- Programming skills
(Python will be further developed, at the necessary level, during the course)

Today …

- Why machine learning?
- What is machine learning?
- Types of Learning Problems/Tasks
- First example of a learning algorithm: how to learn a function?
  - ➢ <u>Linear Regression</u>

Why Machine Learning?

- Complex tasks that require processing that can not be done manually
- Data Mining:
  Convert data into information and knowledge.
  Find patterns in the data that lead to insights.
  Sometimes actionable insights.
- ML:
  Construct mechanisms for supporting data driven predictions and decisions.
  Replace human generated or curated rules.
- Rapidly developing industry – almost everything today needs or uses learning and data science techniques

IDC HERZLIYA

ML Applications - examples

- Spam filtering
- Credit card fraud detection
- Face recognition/detection/identification in images
- Recommendation systems (books, movies, travel destinations)
- Bioinformatics and analysis of scientific data
- Healthcare and medical science applications
- Self driving cars
- Predictive maintenance
- NLP: translation, text completion
- …

IDC HERZLIYA

Example: hospital ER

- Consider patients arriving in the ER with, say, chest pain.
- When considering their discharge, can we predict, based on features we may be able to measure, if they will need to be re-hospitalized within the coming week (day, month, etc …)?
- When admitted in the ER can we predict, based on features we can measure, whether they will develop an acute event within the next 4 hours?

ER acute events – Rules based approach

- In the rules based approach we will collect input from doctors and other sources and write an algorithm/protocol:

    - Rule1:
      If <BP > 150 OR HR at rest >150 >
      then RE-Hospitalized within one week
    - Rule 2: ….
    - Etc …

IDC HERZLIYA

Acute event

- Collect data with features including, for example:
  - Glucose level
  - Type of surgery performed, if any
  - Age
  - Type of insurance
  - Gender
  - Medications taken and adherence
  - Blood neutrophils, other serological parameters
  - Weight, BMI
  - ECG characteristics
  - Address, last name
- Also collect labels – acute event YES or NO
- Some measured features are more important than others in affecting the prediction values.
- Some may turn out to be not useful at all …
- How do we know?

IDC HERZLIYA

LEARNING

Use the collected data and apply a (learning) algorithm to output a set of rules (protocol) as above

Example: Spam Filtering

In 2016 spam accounted for 14.5 billion messages globally per day, which constitutes ~50% of all emails.

We would like to build effective and efficient filters.

IDC HERZLIYA

How To Filter Spam?

- Every email containing some suspicious words. “Viagra”, “cheap”, ”free tickets”?
- Flashy text in the message
- Other attributes/features:
  - + Where did the message come from?
  - + Software that had sent the message
  - + Date? Time? Length?
- Rule based approach:
  develop a protocol based on the above.
- Learning approach:
  - + Collect (a lot of) data
  - + Determine feature values and a label (spam or not)
  - + LEARN a protocol – which words, features, attributes are associated with spam and what function to apply

IDC HERZLIYA

Learning Essentials

1. Collect data, examples: both good and bad (positive and negative).
   These are called "**samples**" or "**instances**".
2. Gather information regarding the instances.
   Values for selected measurable "**features**" (note: some features are relevant and some are irrelevant; we don't necessarily know apriori which is which). Assign **"labels"** to the instances.
3. Then what? - <u>the inference stage</u>.
   This stage uses a
   LEARNING "**technique**" or "**algorithm**"

IDC HERZLIYA

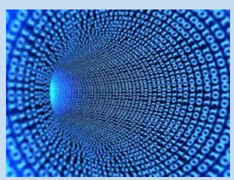
Supervised Learning

Output of the learning process:
The Execution Algorithm
AKA the Model, the Hypothesis, the Predictor, the Prediction Rule/Function

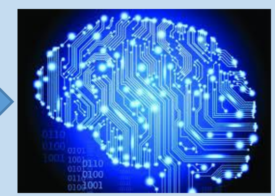
Data Gathering

Learning Algorithm

Complicated Task



Define, measure, collect features and labels



Execution Algorithm

Complicated Task (new instance)

Execution Algorithm

Results

IDC HERZLIYA

Learning

- We do not provide the rules of the execution algorithm ourselves
- We expect the “computer” to learn them by itself, from the data provided
- How?
- We apply a “**<u>learning algorithm</u>**” or a “**<u>learning technique</u>**”
- This is the primary subject of this course!

IDC HERZLIYA

Performance Evaluation

- How do we measure if we learned a good execution algorithm?
- Can we prove that we have learned something?
- What is the error of a proposed model?
- How much error do we allow? Do we need 100% performance? Can we expect that?
- How can we assess if a body of data and a technique is expected to allow for sufficiently accurate learning?
  (cast in a probabilistic framework)

IDC HERZLIYA

Types of learning tasks that will be addressed in the course

- Regression
  + Given $\{x_i, y_i\}$ find f such that $y = f(x)$
- Classification
  + Given $\{x_i, y_i\}$ where $y_i \in \{0,1\}$ for training, determine, for any new x, if $x \in C_0$ or $x \in C_1$
- Density Estimation
  + Given $\{x_i\}$ find a probability distribution that best explains the data (possibly from within a family of distributions)
- Clustering
  + Given $\{x_i\}$ find a “good” partition into k subsets

IDC HERZLIYA

First Example: Regression

- We want to represent a quantity as a function of other observed/measurable quantities.
- For example: determine the price of a used Subaro based on Kms driven, age of the car, age of driver(s), color integrity ….
- Assume that our <u>explaining features</u> are numerical – each instance is represented by a vector of features $\boldsymbol{x} \in \mathbb{R}^n$
- We want to learn a <u>function</u> that predicts a real valued output from these features –

$$f\colon \mathbb{R}^n \rightarrow \mathbb{R}$$

IDC HERZLIYA

Regression Slide-set …