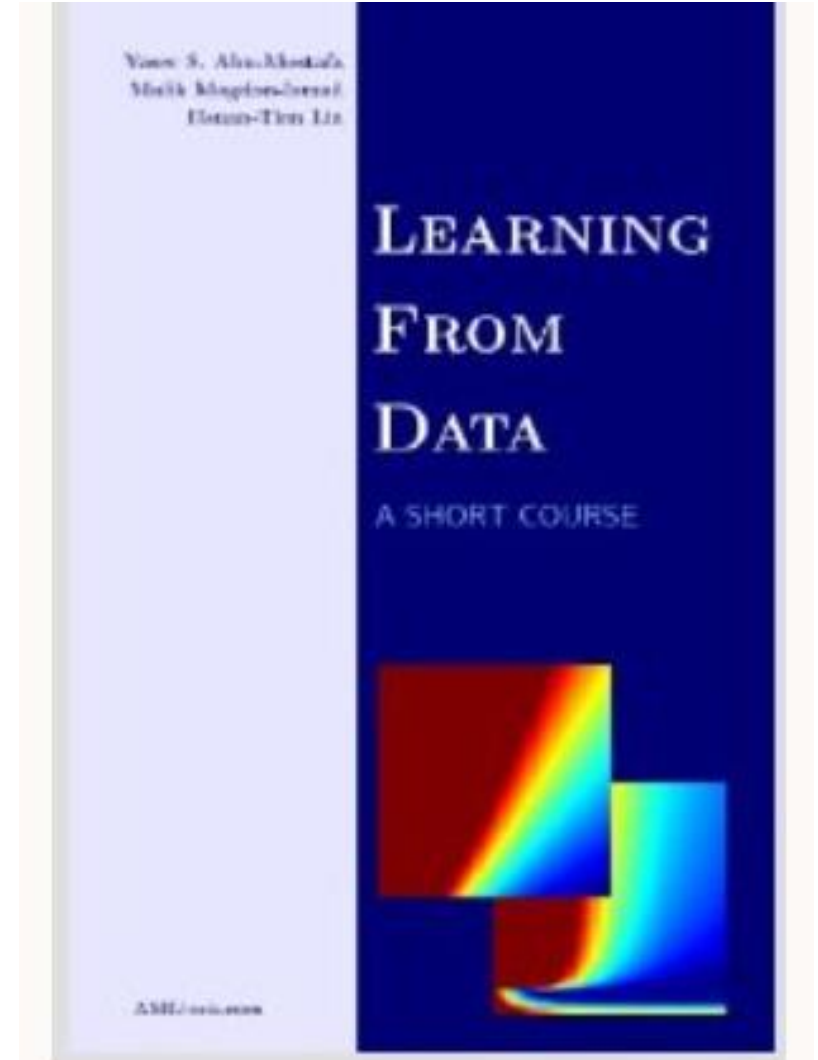


Machine Learning from Data

Lecture 1: Spring 2021

Resources

- Textbook (Yaser S. Abu-Mostafa, Malik Magdon-Ismail)
- Website
- Homework Submission: Submittity



Topics Covered in the Course

- What is Learning?
- Can we Do it?
- How to Do it?
- How to do it well?
- General Principles of Learning
- Advanced techniques
- Other Learning Paradigms

Today's lecture

- Motivation
- Learning Vs. Defining
- Formalize Learning
- Set-up a Machine Learning Problem

Application of Machine Learning



Machine Learning Everywhere

pic source: eduCBA

What is Machine Learning in General

- Ask a 5-year-old, is this a dog?
- Most likely, the answer is Yes





Are these Dogs?

- It is easy for humans to identify.
- Has anyone ever defined dogs for us?
- We have learned from data.



Can we define a dog?


- Let us try.
 - Something that has 4 legs.
 - Runs with a certain speed
 - Something about facial features.



Learning: “Which ones are dogs?”

- Defining is hard.
- Recognizing is Easy.
- It is hard to give a mathematical definition of a Dog.
- A 5-year-old can tell the difference (they learned from Data).
- Learning from Data is used when we do not have an analytic solution.
 - We have data to construct an analytic solution.

The Netflix Problem



The screenshot shows the Netflix Prize website. At the top, the Netflix logo is on the left, and a large yellow banner with the text "Netfix Prize" and a "COMPLETED" stamp is on the right. Below the banner is a navigation bar with links: Home, Rules, Leaderboard, and Update. The main content area is dark and shows a "Movies For You" section with a list of movies. On the right side of the main content area, there is a white box with a blue header "Congratulations!" and text announcing the \$1M Grand Prize awarded to team "BellKor's Pragmatic Chaos" on September 21, 2009. The text also mentions the algorithm and the leaderboard.

NETFLIX

Netfix Prize

COMPLETED

Home Rules Leaderboard Update

NETFLIX

Browse Recommendations Friends Queue Buy DVDs

Home Genres New Releases Previews Netflix Top 100

Movies For You

Randy, the following movies were chosen based on your interest in:

- Waiting for Golyard
- Survivor: Season 1
- ahrenheit 9/11

The Big One

★★★★★

For subversive

from

Original and

Now only for just \$5.98

Shop as low

titles

Original and

Other

Light

Learn More

and more

Disc Series

Congratulations!

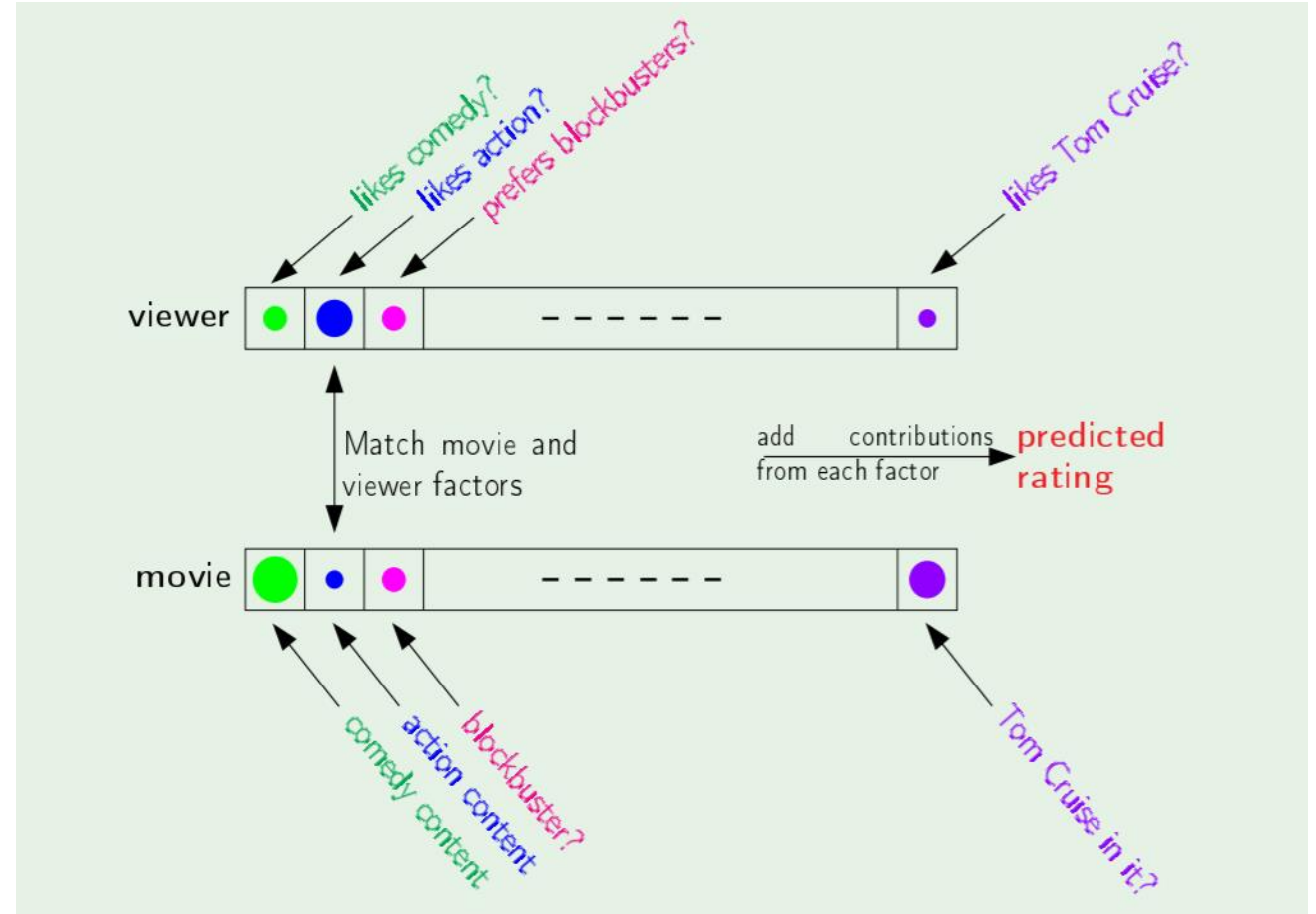
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about [their algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

Problem Setup

- Netflix Problem: Predict recommendations, get more subscriptions.
- Criteria used to rate movies – Unknown/Complex
- Create user and movie profiles.
- Calculate predicted rating.
- The learning algorithm ‘reverse engineers’ the factors based on past ratings (starting with random factors mostly).
- It tunes these factors to make them more aligned with real ratings of viewers.



Components of Learning

- **The Credit Approval Problem:**
- Approve or not?
- No magic formula exists.
- Banks have data: customer information like salary and debt; whether they defaulted on their credit or not.

age	23 years
gender	male
annual salary	\$30,000
years in residence	1 year
years in job	1 year
current debt	\$15,000
...	...

Key Takeaway

- A pattern exists
- We do not know it
- We have data to learn it

Formalize Components of Learning:

input $\mathbf{x} \in \mathbb{R}^d = \mathcal{X}$.

output $y \in \{-1, +1\} = \mathcal{Y}$.

target function $f : \mathcal{X} \mapsto \mathcal{Y}$.

(The target f is *unknown*.)

data set $\mathcal{D} = (\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$.

($y_n = f(\mathbf{x}_n)$.)

- **Input:** Salary, debt, years
Output: Approve or not
Target function: Relationship between X and Y
- Data on customers
- X, Y and D will be given by the learning problem.

The Learning Process

- Start with a set of possible Hypothesis that are most likely to represent the target f .
- $H = \{h_1, h_2, \dots\}$ is the hypothesis set or the ***model***.
- Select a hypothesis g from H . The way we did this selection (process) is the ***learning algorithm***.
- Use this selected hypothesis to predict for new data (new customers). Our goal is to bring g as close to f as possible. The target f is fixed but unknown.
- NOTE: We as ML practitioners will choose H and the learning Algorithm.

Summary of the Learning Set-Up

