Machine Learning (CSE 446): Introduction

Noah Smith

University of Washington
nasmith@cs.washington.edu

September 27, 2017

What VIPs are Saying about Machine Learning

"A breakthrough in machine learning would be worth ten Microsofts"

—Bill Gates

"Machine learning is the next Internet"

—Tony Tether (DARPA director)

"Machine learning is the hot new thing"

—John Hennessy (Stanford president)

"Web rankings today are mostly a matter of machine learning"

—Prabhakar Raghavan (Google VP)

"Machine learning is going to result in a real revolution"

—Greg Papadopoulos (Sun CTO)

"Machine learning is today's discontinuity"

—Jerry Yang (Yahoo founder)

What is Learning?

- Predicting the future, given the past
- Generalizing to new scenarios
- Getting better with practice

What is Learning?

- Predicting the future, given the past
- Generalizing to new scenarios
- Getting better with practice

To measure how well an algorithm has learned, we give it a test (sound familiar?).

Examples

Examples

- Categorizing documents (e.g., "is this email spam?")

Examples

- Categorizing documents (e.g., “is this email spam?”)
- Labeling images (e.g., “who’s in this picture?”)

Examples

- Categorizing documents (e.g., “is this email spam?”)
- Labeling images (e.g., “who’s in this picture?”)
- Predicting the future: weather, finance, medical outcomes

Examples

- Categorizing documents (e.g., "is this email spam?")
- Labeling images (e.g., "who's in this picture?")
- Predicting the future: weather, finance, medical outcomes
- Collect sensor data, predict values everywhere (e.g., energy use in a building)

Examples

- Categorizing documents (e.g., “is this email spam?”)
- Labeling images (e.g., “who’s in this picture?”)
- Predicting the future: weather, finance, medical outcomes
- Collect sensor data, predict values everywhere (e.g., energy use in a building)
- Recommending products (e.g., movies and books)

Examples

- Categorizing documents (e.g., “is this email spam?”)
- Labeling images (e.g., “who’s in this picture?”)
- Predicting the future: weather, finance, medical outcomes
- Collect sensor data, predict values everywhere (e.g., energy use in a building)
- Recommending products (e.g., movies and books)
- Decision-making in the face of uncertainty (e.g., self-driving cars)

Examples

- Categorizing documents (e.g., "is this email spam?")
- Labeling images (e.g., "who's in this picture?")
- Predicting the future: weather, finance, medical outcomes
- Collect sensor data, predict values everywhere (e.g., energy use in a building)
- Recommending products (e.g., movies and books)
- Decision-making in the face of uncertainty (e.g., self-driving cars)
- Given an instance, find similar ones (e.g., images)

Examples

- Categorizing documents (e.g., "is this email spam?")
- Labeling images (e.g., "who's in this picture?")
- Predicting the future: weather, finance, medical outcomes
- Collect sensor data, predict values everywhere (e.g., energy use in a building)
- Recommending products (e.g., movies and books)
- Decision-making in the face of uncertainty (e.g., self-driving cars)
- Given an instance, find similar ones (e.g., images)
- Find structure or patterns in large datasets (e.g., clustering)

Today

ML is required for ...

- Video and image processing
- Speech and language processing
- Search engines
- Robot control
- Sensor networks
- Computational biology
- Medical and health analysis

When people say "AI" they almost always mean "ML."

Trends: more data, faster processing and networks, new sensors and IO devices, demand for customization.

Software is becoming too complex to write by hand.

Is it Magic?

Is it Magic?

More like **gardening**.

Growing successful plants (programs) requires:

- seeds (algorithms)
- nutrients (data)
- a gardener (ML expert)

Is it Magic?

More like **gardening**.

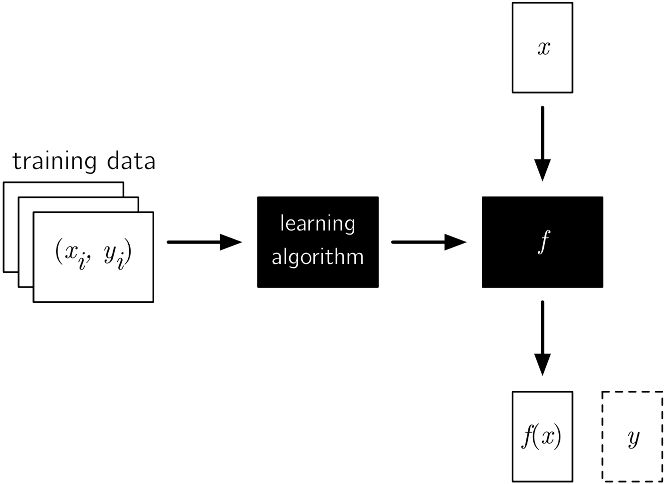
Growing successful plants (programs) requires:

- seeds (algorithms)
- nutrients (data)
- a gardener (ML expert)

Gardens are somewhat predictable, but not entirely, and our scientific understanding is still improving!

Inductive, Supervised Machine Learning

- *Training:* a learning algorithm is given a set of example input-output pairs (x, y) and produces a function f; the goal is for $f(x)$ to recover y, for each example, and on future examples
- *Testing:* we apply f to new test examples (x, y) and measure how well $f(x)$ matches y



Inputs and Output

- x can be pretty much anything we can represent
  - To start, we'll think of x as a bundle of attribute-value pairs, e.g., $\phi(x) = v$.
- y can be
  - a real value (regression)
  - a label (classification)
  - an ordering (ranking)
  - a vector (multivariate regression)
  - a sequence/tree/graph (structured prediction)
  - ...

Examples

Examples

- Predict rainfall in Seattle tomorrow.

Examples

- Predict rainfall in Seattle tomorrow.

- Is this email spam?

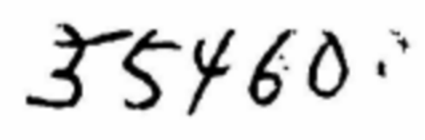
  `From: 6cq0ybi1otqmtyidobfsrd2r8dwkhea@mx7.besthappydayes.com`
  `Subject: We Have Found Your Missing Money`
  `You are Owed Cash That You Dont Know About Find Unclaimed Money`

Examples

- Predict rainfall in Seattle tomorrow.

- Is this email spam?

  `From: 6cq0ybi1otqmtyidobfsrd2r8dwkhea@mx7.besthappydayes.com`
  `Subject: We Have Found Your Missing Money`
  `You are Owed Cash That You Dont Know About Find Unclaimed Money`
- What zip code is in this image?



Administrivia

Bookmark These

Course website: `http://courses.cs.washington.edu/courses/cse446/17au/`

Canvas: `https://canvas.uw.edu/courses/1173938`

Textbook: `http://ciml.info`

Your Instructors

Noah (instructor):

- UW CSE professor since 2015, NIPS & ICML papers since 2008, professor since 2006, using ML since 1998
- Research interests: machine learning for structured problems in NLP, ML & NLP for social science

TAs: Kousuke, John, Deric, Patrick, Andrew, and Jane

Outline of CSE 446

- Problem formulations: classification, regression
- Techniques: decision trees, nearest neighbors, perceptron, linear models, probabilistic models, neural networks, kernel methods, clustering
- "Meta-techniques": ensembles, expectation-maximization
- Understanding ML: limits of learning, practical issues, bias & fairness
- Recurring themes: (stochastic) gradient descent, bullshit detection

Project

- Teams of three
- Parts:
  1. Build and justify a new regression or binary classification **dataset** (due 10/17)
  2. Dataset review (part of A2) & class-wide selection (official datasets announced 11/3)
  3. Implement ML algorithms and compete in a bakeoff on $\sim$5 datasets (due 12/5)
- Don't wait! Part 1 is already available on the course website.

Grading

- Assignments (five, 11% each)
- Project (30%)
- Final exam (15%)

Grading

- Assignments (five, 11% each)
  - Some pencil and paper, mostly programming
  - Graded mostly on attempt, not correctness
  - Five late days; no credit for late work after they are used up.
- Project (30%)
- Final exam (15%)

Grading

- Assignments (five, 11% each)
  - Some pencil and paper, mostly programming
  - Graded mostly on attempt, not correctness
  - Five late days; no credit for late work after they are used up.
- Project (30%)
  - dataset and writeup (10%)
  - final writeup (15%)
  - bakeoff performance (5%)
- Final exam (15%)

Grading

- Assignments (five, 11% each)
  - Some pencil and paper, mostly programming
  - Graded mostly on attempt, not correctness
  - Five late days; no credit for late work after they are used up.
- Project (30%)
  - dataset and writeup (10%)
  - final writeup (15%)
  - bakeoff performance (5%)
- Final exam (15%) tentatively Wed. Dec. 13, 8:30–10:20 am

"Can I Take This Class?"

- Short answer: yes (if you can get past the wait list), but be warned.
- Official prerequisites (and linear algebra) are strongly advised.
  - Be forthcoming with your potential teammates!
- We assume you're a strong programmer and comfortable with math.
- We will move fast; lectures will focus on concepts and mathematics, quizzes are for review and implementation discussions.
- "Sink or swim."

I've been told to give The Link on Friday.

To-Do List

- Quiz section meetings start tomorrow. **Bring your laptop!**
- Read: Daume (2017, ch. 1)
- Academic integrity statement: on the course web page; upload your signed scan through Canvas.
- Form groups and register them on Canvas (People $\rightarrow$ Groups $\rightarrow$ Project Groups)

References I

Hal Daume. *A Course in Machine Learning (v0.9)*. Self-published at `http://ciml.info/`, 2017.