Lecture 1: Introduction

COMP 343, Spring 2022

Victoria Manfredi

WESLEYAN
UNIVERSITY

Acknowledgements: These slides are based primarily on content from the book "Machine Learning" by Tom Mitchell, slides created by Vivek Srikumar (U of Utah), Dan Roth (UPenn), and lectures by Balaraman Ravindran (IIT Madras), as well as from courses at Stanford (http://cs229.stanford.edu/) and UC Berkeley (http://ai.berkeley.edu/)

Today's Topics

Course logistics and information

What is machine learning?

- Overview
- A game …
- Paradigms
- Looking forward

vumanfredi@wesleyan.edu

Course Logistics and Information

Course webpage (***not*** moodle)

Course schedule and homework posted on webpage

- http://vumanfredi.wescreates.wesleyan.edu/teaching/comp343-s22/

We'll use Google classroom for announcements, discussion, maybe grades

- I will add you via email

We'll use Google drive for homework submissions

- Each of you will have directory for this course, with homework subdirectories

Grade breakdown

- 60%: approximately 9 homework assignments, no scores dropped
  - Mix of written and (possibly multi-assignment) programming projects
  - homework is due 5p on Wednesdays! Help sessions are Sun, Mon, Tues nights
- 20%: Midterm exam
- 20%: Final project

Textbooks

Not strictly required but helpful

Machine Learning
by Tom Mitchell (1997) ... Yes, it is "old" but it gives a lot of intuition for the key ideas in machine learning that are important today, and is somewhat less heavy on the math.

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems
2nd Edition (2019) by Aurelien Geron

Deep Learning with Python:
2nd Edition (2021) by Francois Chollet

First 2 books are the most important ones to get. All of the books should be available on reserve in the library. If you are having trouble getting hold of a book, please email me and we will figure something out. A number of free online books are also posted on the webpage

Getting started

Python3

- we'll review as needed, see class resources webpage
  - please check you have python3 installed!
    - type python3 at terminal prompt
    - alternatively you can use Google Colab
  - CAs can help you get setup
  - tutorials and other resources posted on course website

Python help available

- at SCIC on 1st floor of Exley

Homework

1st homework out Wednesday

- play with python libraries and datasets (see shared directory for data)
- get familiar with basic machine learning concepts

Submissions

- Google drive: COMP343-s22 is shared directory
- Submit homework by copying to COMP343-s22/hw1/USERNAME
- Substitute your Wesleyan username for USERNAME

Late policy

- 4 free late days you can use without asking, otherwise 15% per day if you don't ask me

Important!

- Put your name inside every file including code!
- File formats: only .py, .ipynb, .pdf! Use .txt or .csv for data you upload

Honor code

Do

- Form study groups (with arbitrary number of people), discuss and work on homework problems in groups
- Lectures are based on material from a variety of places. Feel free to read other notes, watch other lectures, as it can be helpful for understanding to see the same material presented in different ways
- Write down the solutions independently (except for group projects)
- Write down the names of people with whom you've discussed the homework
- Understand the solution well enough to reconstruct it yourself
- Read the longer description on the course website

Don't

- copy, refer to, or look at any official or unofficial previous years' solutions in preparing the answers
- search or submit solution found online for any homework

This course

Focuses on underlying concepts and algorithmic ideas in machine learning

This course is not about

- Using a specific machine learning tool, or any single paradigm
- Why? will be able to better apply algorithms if we really understand them

1st few weeks

- High-level overview of machine learning, probability review
- Covers a lot of material!

Rest of course

- Digging into details of what we talked about in 1st few classes
- Having had high-level should help give context for details

If you have questions or concerns please come talk to me

What will you learn?

1. A broad theoretical and practical understanding of machine learning paradigms and algorithms

2. Ability to implement learning algorithms

3. Identify where machine learning can be applied and make the most appropriate decisions (about algorithms, models, supervision, etc)

What is machine learning?

What do you think machine learning is?

Machine learning is everywhere

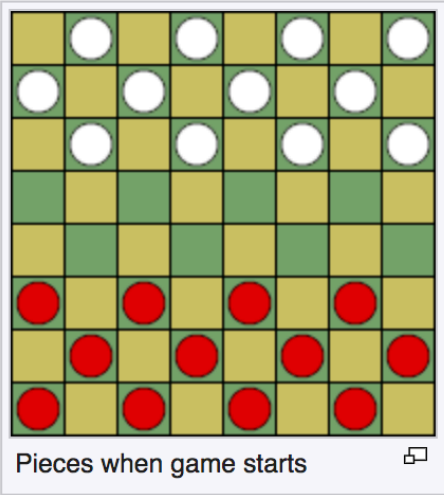
And you are probably already impacted by it!

- Is an email spam?
- Find all the people in this photo
- If I like these 3 movies, what should I like next?
- Based on your purchase history, you might be interested in …
- Will a stock price go up or down tomorrow? By how much?
- Handwriting recognition
- What are the best ads to place on this website?
- I would like to read that Dutch website in English
- Ok, Google, drive this car for me. And while you're at it, fly this helicopter
- Does this genetic marker correspond to Alzheimer's disease?

Writing a program that exhibits intelligence

Arthur Samuel (1959):

- "Machine Learning is the field of study that gives the computer the ability to learn without being explicitly programmed."



Pieces when game starts

Wikipedia

A. L. Samuel

Some Studies in Machine Learning Using the Game of Checkers

Abstract: Two machine-learning procedures have been investigated in some detail using the game of checkers. Enough work has been done to verify the fact that a computer can be programmed so that it will learn to play a better game of checkers than can be played by the person who wrote the program. Furthermore, it can learn to do this in a remarkably short period of time (8 or 10 hours of machine-playing time) when given only the rules of the game, a sense of direction, and a redundant and incomplete list of parameters which are thought to have something to do with the game, but whose correct signs and relative weights are unknown and unspecified. The principles of machine learning verified by these experiments are, of course, applicable to many other situations.

Learning as generalization

Herbert Simon (1993):

- "Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the task (or tasks, drawn from the same population) more effectively the next time."
- Father of Artificial Intelligence. Economist, psychologist, political scientist, sociologist, Nobel Prize (1978), Turing Award (1975), …

Unedited Interview about History of AI at CMU from 1955-1985
https://www.youtube.com/watch?v=r-naBUbhUEs

It's not enough just to repeat someone else's instructions!

Learning as generalization

Tom Mitchell, Machine Learning book (1997)

- "A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E."

Tasks

- Define learning with respect to a specific class of tasks
- E.g.,
  - Detecting spam
  - Playing the game of chess
  - Diagnosing patients with a particular disease

Learning as generalization

Tom Mitchell, Machine Learning book (1997)

- "A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E."

Experience E

- Performance should improve with experience (data)
- E.g.,
  - More emails seen improves spam detection
  - Games played by the program (with itself)
  - More patients you examine, better you get at diagnosing illness

Learning as generalization

Tom Mitchell, Machine Learning book (1997)

- "A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E."

Performance measure P

- To determine whether learning is happening
- E.g.,
  - Spam emails correctly identified
  - Winning rate: whether you won or lost game
  - # of patients that were accurately diagnosed

Learning as generalization

What should our program do with these photos?

From slides of Vivek Srikumar (U of Utah)

Learning as generalization

From slides of Vivek Srikumar (U of Utah)

What should our program do with these photos?

Program should see these pictures of dogs, learn how to identify dogs, and then be able to identify new pictures as dogs or not

Learning as generalization

Given unseen photo, classify it as dog or not!

Learning has to go beyond just memorizing what has been seen so far: this is generalization

From slides of Vivek Srikumar (U of Utah)

Learning as generalization

Gives a system the ability to perform a task in a situation which has never been encountered before!

- New way to think about programming
- Programs that gain new capabilities as they get more experience!

Learning allows programs to interact more robustly with messy data

- Is there noise in our data or is it just a dog wearing sunglasses?

Traditional programming

Machine learning

A new way of programming: write a program to write a program to do the task

Let’s play a game

Let's play a game

Name	Label
Norman Danner	+
Karen Collins	-
Dan Licata	?
Danny Krizanc	?
Saray Shai	?
Wai Kiu Chan	?

Based on a game at a machine learning conference in 1994: attendees received conference name tags labeled + or -. Only conference organizers knew function that generated the labels. Depended only on the attendee's name. The task was to determine the unknown function by looking at examples

Let's play a game

Name	Label
Norman Danner	+
Karen Collins	-
Dan Licata	?
Danny Krizanc	?
Saray Shai	?
Wai Kiu Chan	?

What is the label for Simone Biles?

Can you guess the label for my name? Yours?

How were the labels generated? Many different functions, which is right?

Once you know the labeling function, do you still need to know the data?

Let's play a game

Name	Label
Norman Danner	+
Karen Collins	-
Dan Licata	?
Danny Krizanc	?
Saray Shai	?
Wai Kiu Chan	?

What is the label for Simone Biles?

Can you guess the label for my name? Yours?

How were the labels generated?

Once you know the labeling function, do you still need to know the data?

Think about for badges.txt in shared datasets directory for next class

Machine Learning

PARADIGMS

The main question throughout the semester

What is learning?

Different formal answers to this problem will give us:

- Various families of learning algorithms

There are different kinds of

- learning paradigms
- learning algorithms/models
- data representations

Different learning paradigms

Supervised learning

- Learn with a teacher

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- Learn with and without a teacher

Active learning

- Learner and teacher interact with each other

Reinforcement learning

- Learn by interacting in environment

Different learning paradigms

Supervised learning

- Learn with a teacher

Our focus in this class will be supervised learning!

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- Learn with and without a teacher

Active learning

- Learner and teacher interact with each other

Reinforcement learning

- Learn by interacting in environment

Supervised learning

Learn with a teacher

- **Labeled examples**: learn map from input to an output

Classification

- **categorical output**: e.g., predict type of residence from (lot size, house size)

Regression

- **continuous output**: e.g., predict price of residence from (lot size, house size)

Classification

Learn map from input to categorical output

- **Examples**
  - Medical
    - Input: description of the patient who comes to clinic
    - Output: whether patient has a certain disease or not
  - Real estate
    - Input: housing square feet and lot size
    - Output: type of house: e.g., house or townhouse
- **Experience**: known input and output pairs

Typical performance measure is classification error

- How many of the patients were diagnosed incorrectly?
- How many of the exam answers were incorrect?

⇒ Not possible to learn directly w.r.t classification error so use other forms

Regression

Learn map from input to continuous output

- **Examples**
  - Product life
    - Input: product description
    - Output: how long will product last before it fails
  - Real estate
    - Input: housing square feet and lot size
    - Output: house price
- **Experience**: known input and output pairs

Performance measure is prediction error

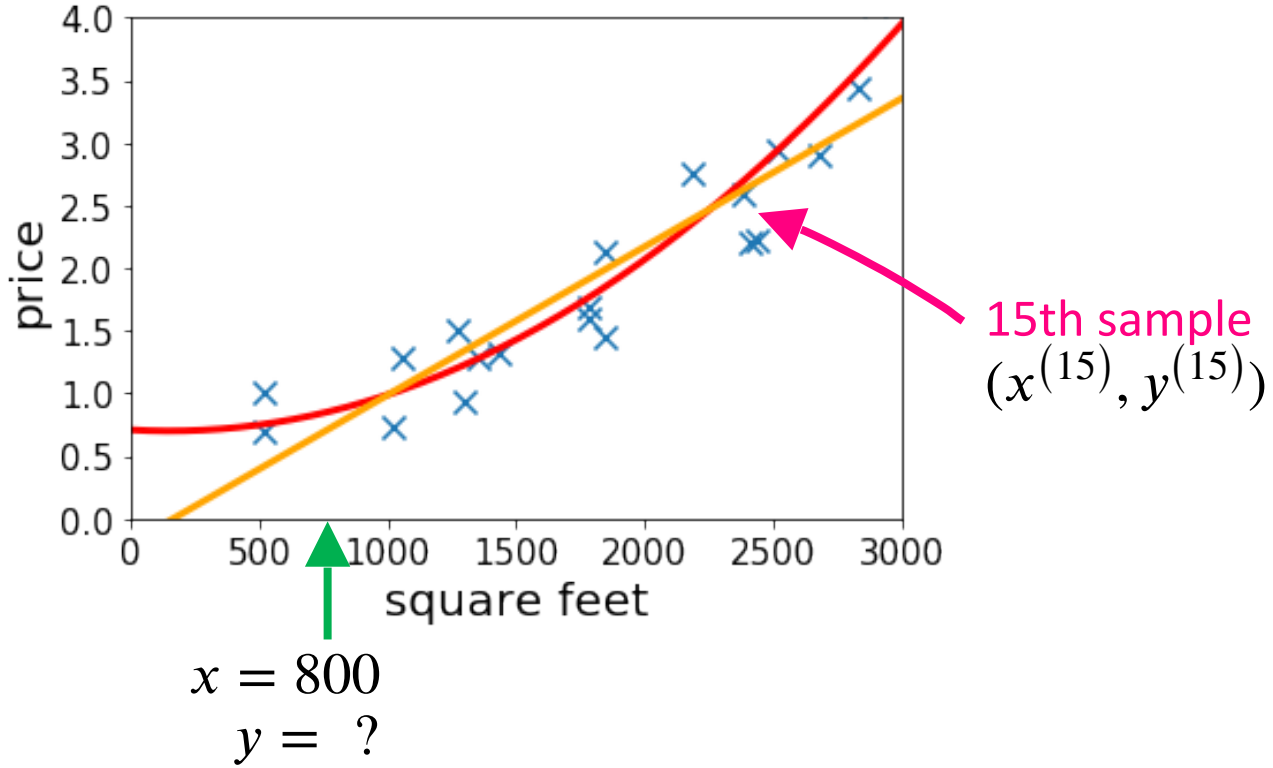
- I say it is going to rain 23 mm, and it rains 49 cm, huge prediction error

Housing Price Prediction

Given: dataset that contains n samples

$$\left(x^{(1)}, y^{(1)}\right), \ \ldots, \ (x^{(n)}, y^{(n)})$$

Task: if a residence has x square feet, predict its price?



More Features … suppose we know lot size

Task: find a function that maps

(house size, lot size) $\rightarrow$ price

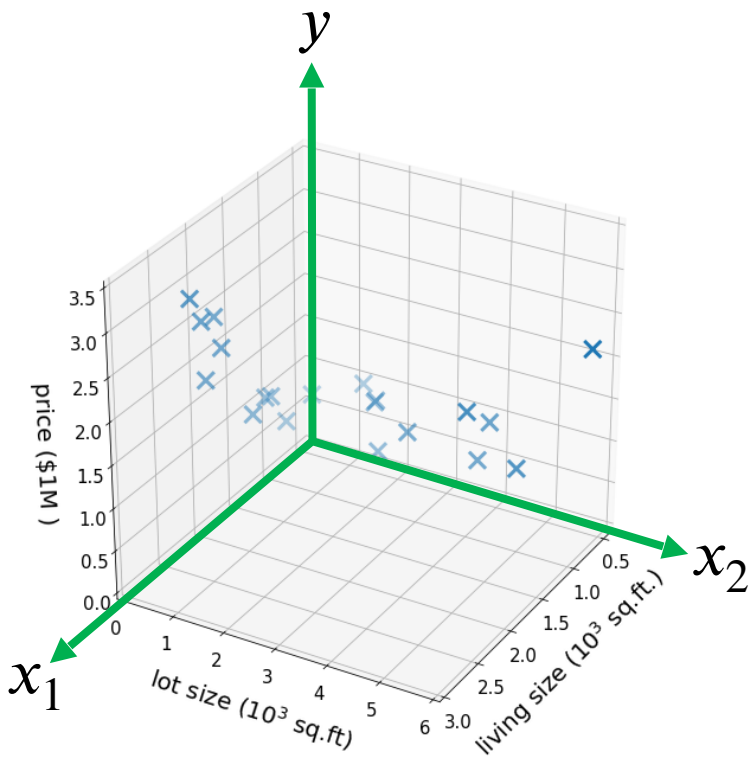
features/input $x \in \mathbb{R}^2$

label/output $y \in \mathbb{R}$

Dataset: $\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$

where $x^{(i)} = \left(x_1^{(i)}, x_2^{(i)}\right)$

Supervision refers to $y^{(1)}, \ldots, y^{(n)}$



Regression vs. Classification

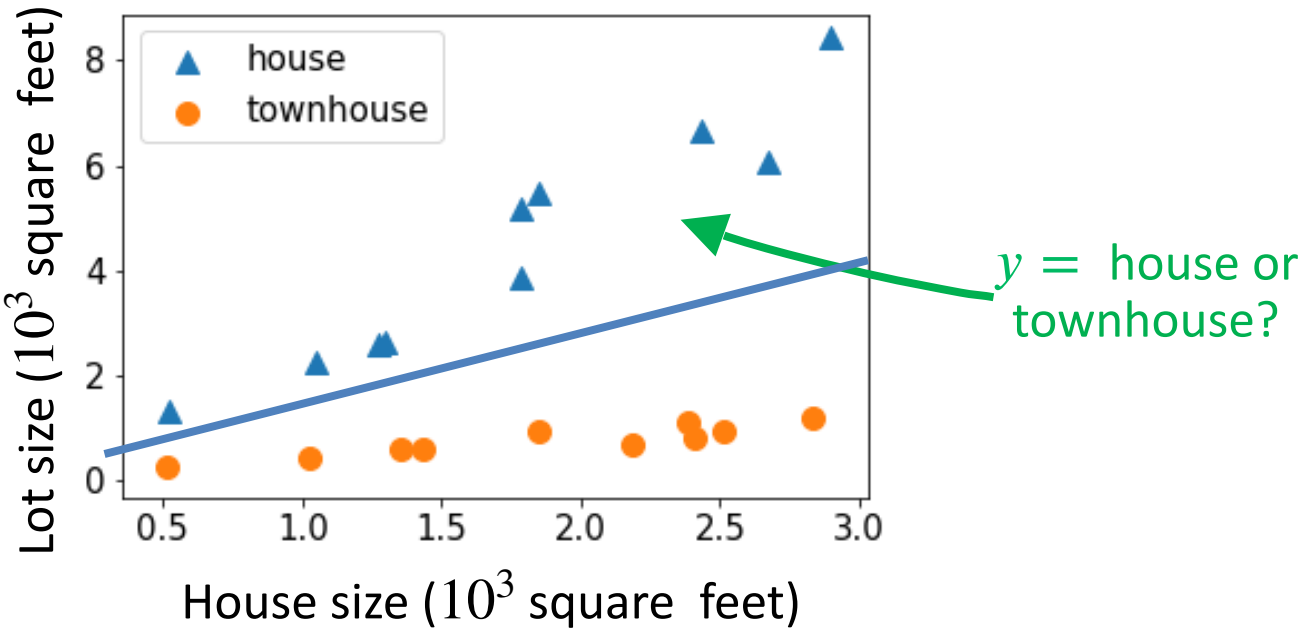
Regression: if $y \in \mathbb{R}$ is a continuous variable

- e.g., price prediction

Classification: the label is a discrete variable

- e.g., the task of predicting the types of residence

  (house size, lot size) $\rightarrow$ house or townhouse?



Different learning paradigms

Supervised learning

- Learn with a teacher

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- Learn with and without a teacher

Active learning

- Learner and teacher interact with each other

Reinforcement learning

- Learn by interacting in environment

Unsupervised learning

Given a set of n data, discover patterns in the data

- No real desired output that we are looking for
- More interested in finding patterns in data

Clustering

- Find cohesive groups among input data
  - E.g., look at customers that come to tech store, figure out if there are categories of customers: college students, IT professionals, …
- Performance measures: scatter/spread of cluster, purity

Association rule mining or frequent pattern mining

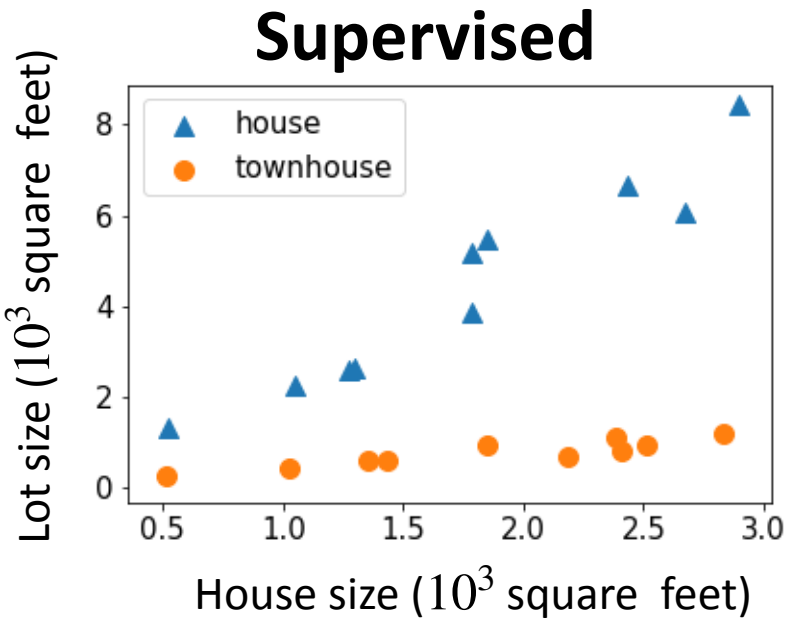
- Find frequent co-occurrence of items in the data
  - Whenever a comes to shop, b also comes to shop
- Performance measures: support and confidence

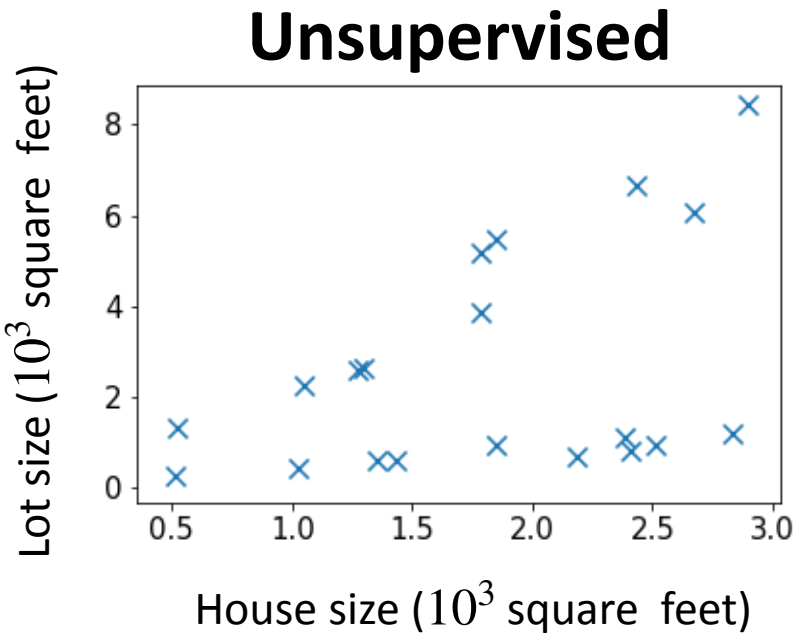
Supervised vs. Unsupervised Learning

Unsupervised dataset contains no labels: $x^{(1)}, \ldots, x^{(n)}$

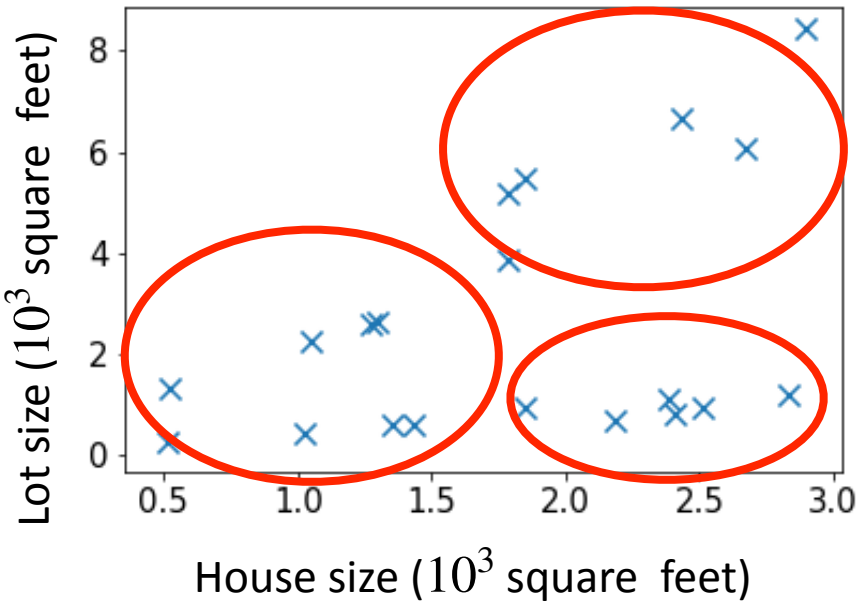
Goal (vaguely-posed):

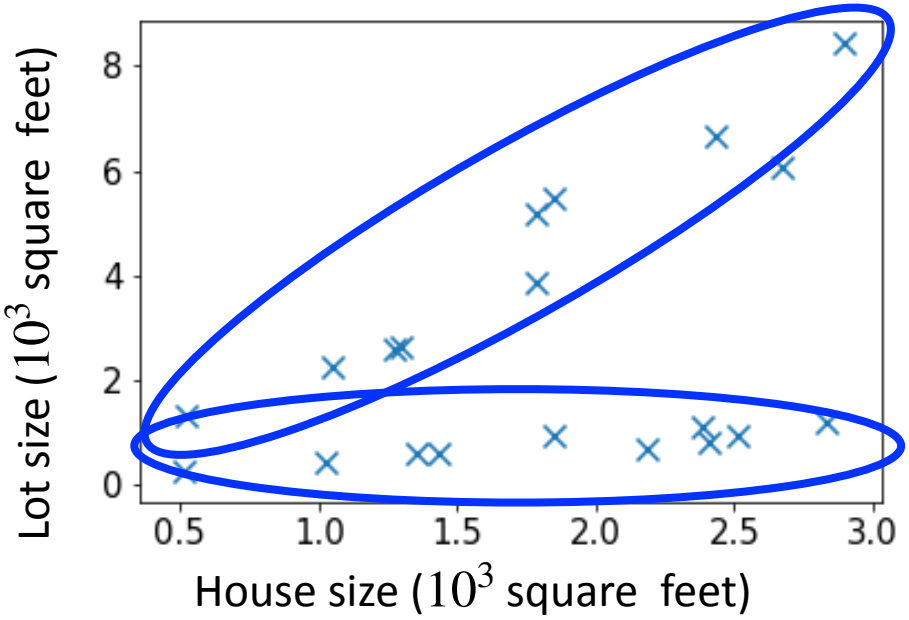
- to find interesting structures in the data



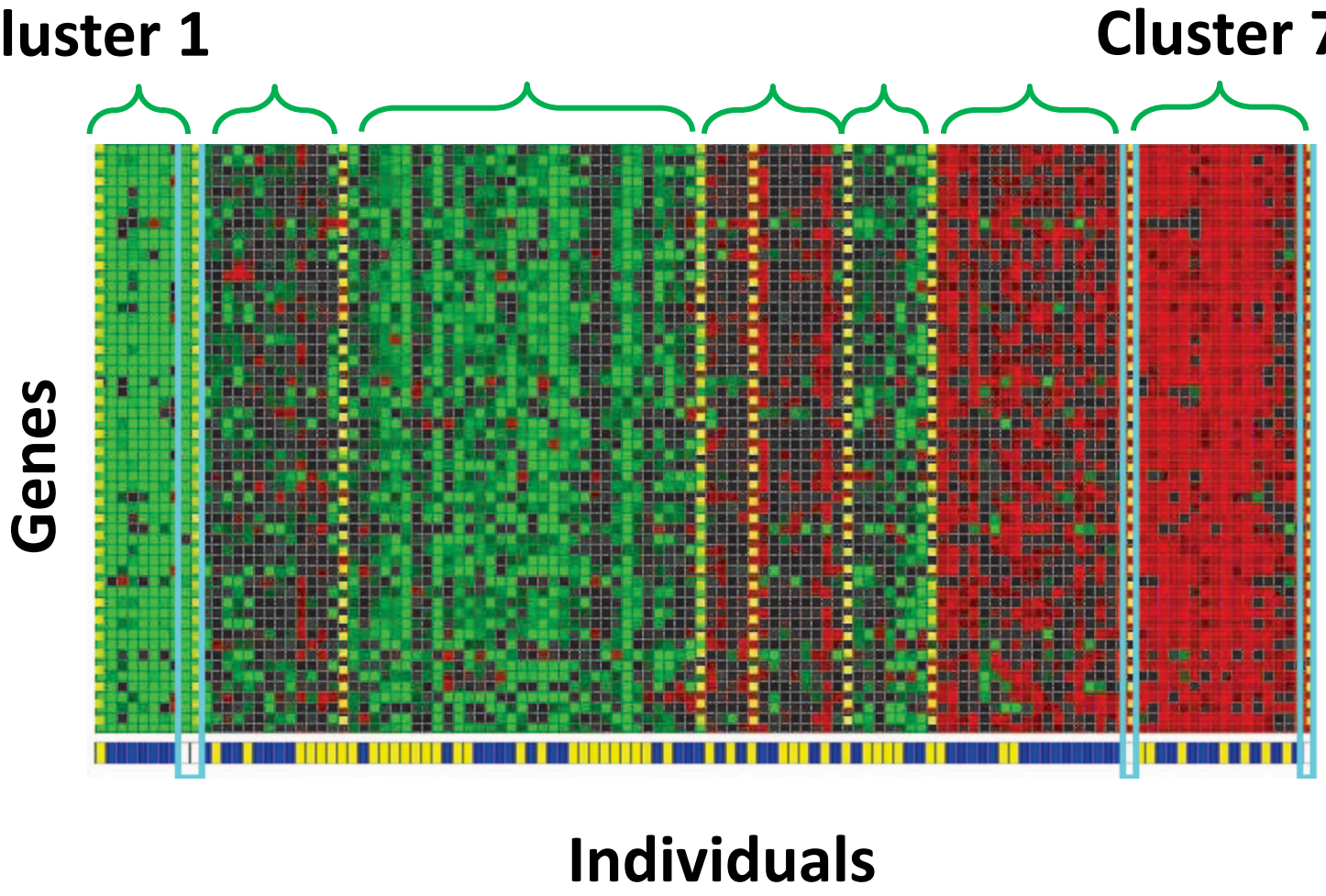


Clustering


Lot size (10^3 square feet)
House size (10^3 square feet)
0
2
4
6
8
0.5
1.0
1.5
2.0
2.5
3.0

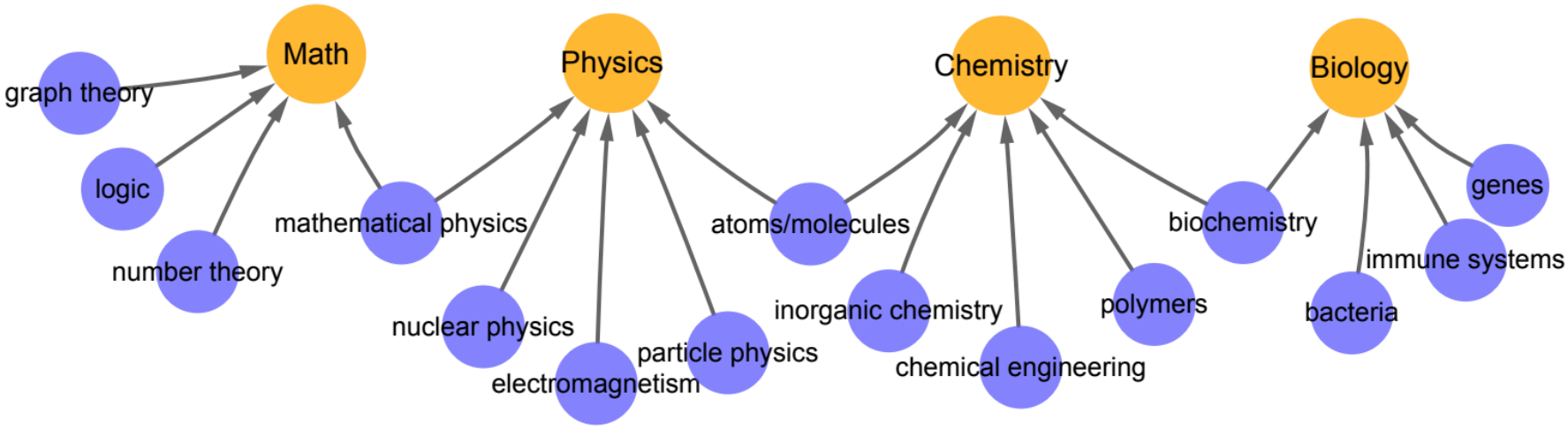

Lot size (10^3 square feet)
House size (10^3 square feet)
0
2
4
6
8
0.5
1.0
1.5
2.0
2.5
3.0

Clustering Genes



Identifying Regulatory Mechanisms using Individual Variation Reveals Key Role for Chromatin Modification. [Su-In Lee, Dana Pe'er, Aimee M. Dudley, George M. Church and Daphne Koller. ’06]

Clustering Words with Similar Meanings (Hierarchically)



	logic deductive propositional semantics	graph subgraph bipartite vertex	boson massless particle higgs	polyester polypropylene resins epoxy	acids amino biosynthesis peptide
tag	*logic*	*graph theory*	*particle physics*	*polymer*	*biochemistry*

[Arora-Ge-Liang-M.-Risteski, TACL'17,18]

Applications

Customer data

- Discover classes of customers

Image pixels

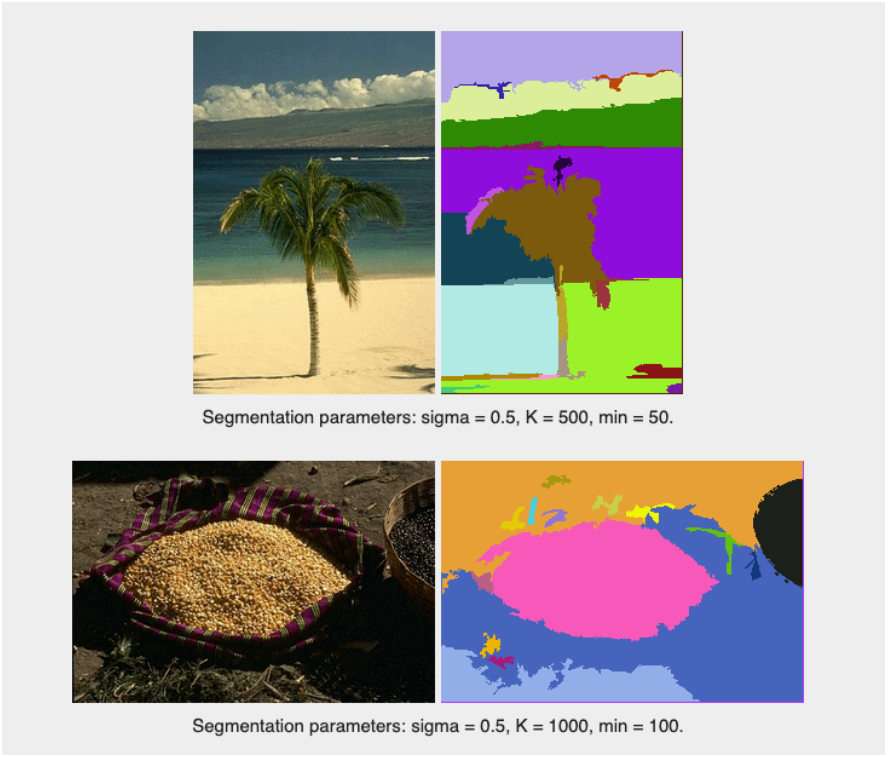
- Discover regions

Words

- Synonyms

Documents

- Topics



http://cs.brown.edu/people/pfelzens/segment/

Different learning paradigms

Supervised learning

- Learn with a teacher

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- **Learn with and without a teacher**: learner gets examples from teacher plus other data

Active learning

- Learner and teacher interact with each other

Reinforcement learning

- Learn by interacting in environment

Different learning paradigms

Supervised learning

- Learn with a teacher

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- Learn with and without a teacher

Active learning

- **Learner and teacher interact with each other**: learner asks teacher questions, teacher answers. Learner has to decide what questions to ask and what to do with the answers.

Reinforcement learning

- Learn by interacting in environment

Different learning paradigms

Supervised learning

- Learn with a teacher

Unsupervised learning

- Learn without a teacher

Semi-supervised learning

- Learn with and without a teacher

Active learning

- Learner and teacher interact with each other

Reinforcement learning

- Learn by interacting in environment

Learning to control

Popular models of machine learning

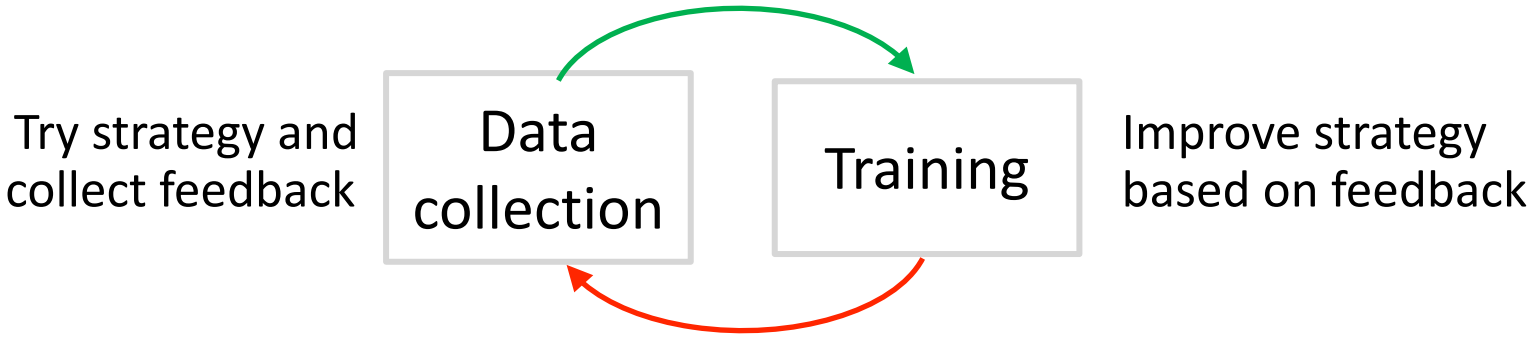
- Supervised: classification, regression, etc
- Unsupervised: clustering, frequent patterns, etc

How did you learn to bicycle?

- Neither of the above
- Trial and error!
- Falling down hurts!

Reinforcement learning

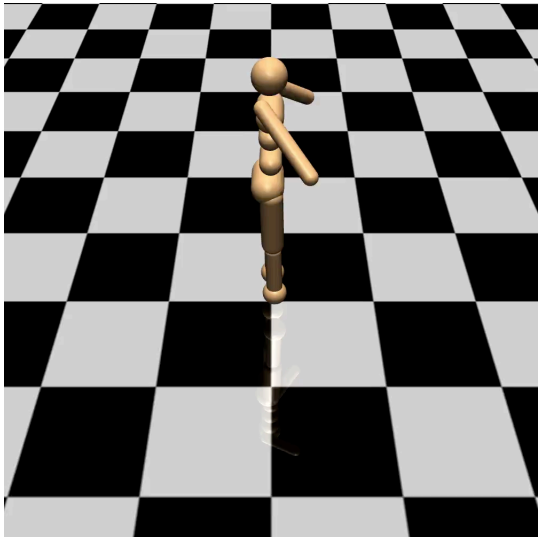
Learn to control the behavior of a system



Performance measure

- Cost for controlling system
- Goal is to minimize the cost that is accrued

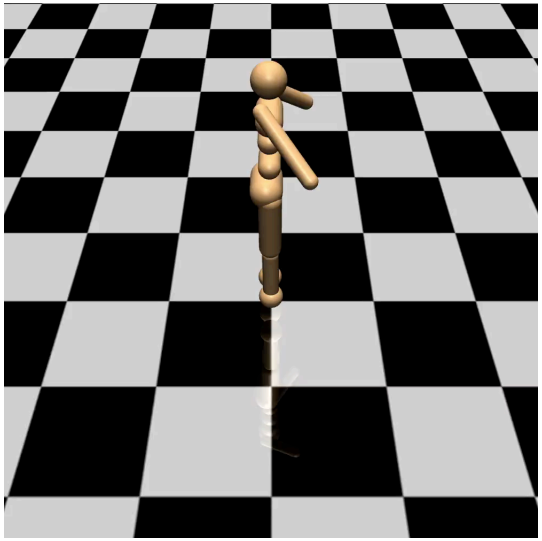
learning to walk to the right



Iteration 10

[Luo-Xu-Li-Tian-Darrell-M.'18]

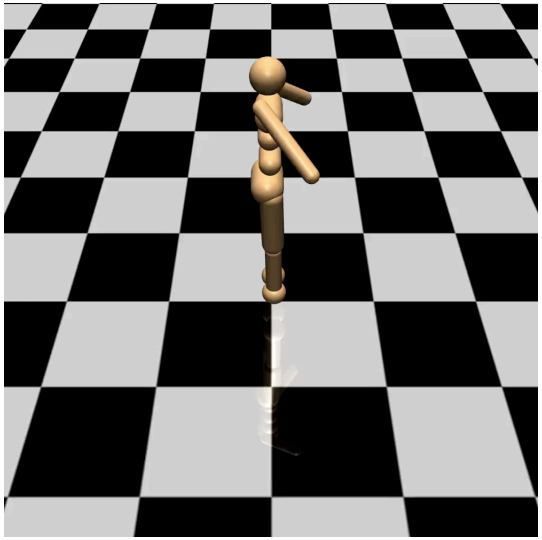
learning to walk to the right



Iteration 20

[Luo-Xu-Li-Tian-Darrell-M.'18]

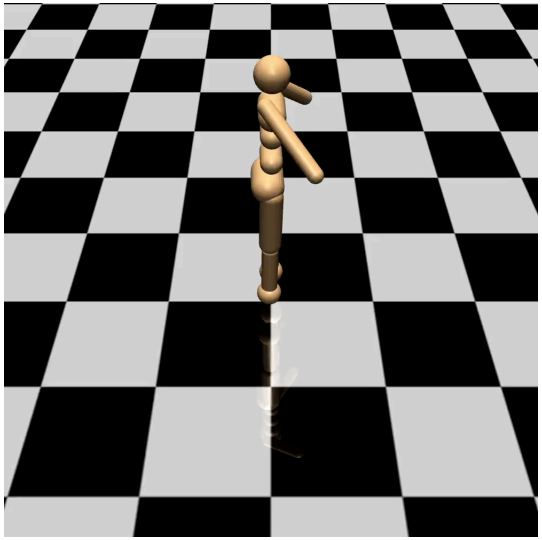
learning to walk to the right



Iteration 80

[Luo-Xu-Li-Tian-Darrell-M.'18]

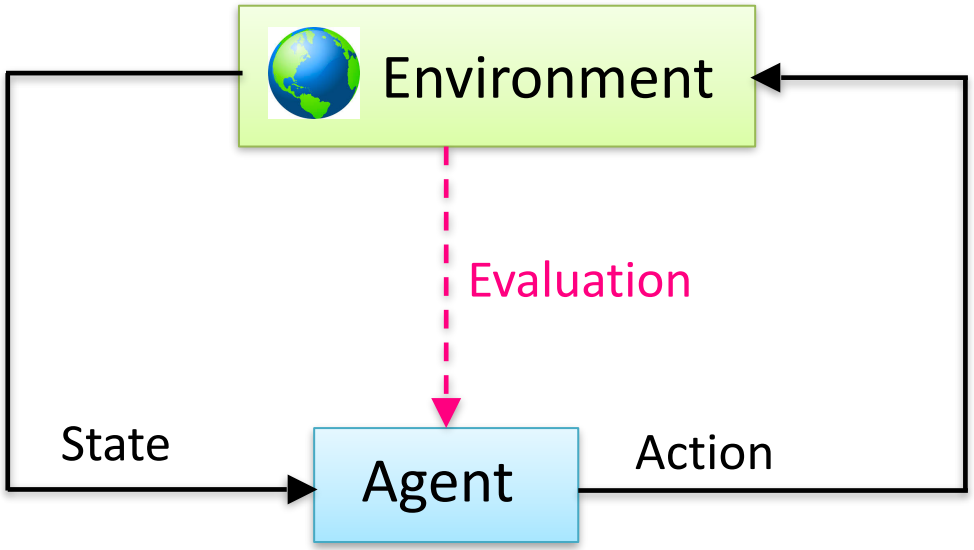
learning to walk to the right



Iteration 210

[Luo-Xu-Li-Tian-Darrell-M.'18]

Reinforcement learning framework



Learn from close interaction

Stochastic environment

Noisy, delayed scalar evaluation

Learn a policy: maximize a measure of long term performance

Applications of reinforcement learning

Game playing

- World's best player: backgammon, chess, go, Atari games from scratch

Autonomous agents

- Robot navigation

Adaptive control

- Helicopter pilot

Combinatorial optimization

- VLSI placement

Intelligent tutoring systems

Machine Learning

LOOKING FORWARD

The main question throughout the semester

What is learning?

Different formal answers to this problem will give us:

- Various families of learning algorithms

There are different kinds of

- learning paradigms
- learning algorithms/models
- data representations

Key issues in machine learning

Modeling

- How to formulate your problem as a machine learning problem?
- How to represent data? Do you have enough data?
- Is the data of sufficient quality (e.g., errors in data, missing values)
- Which algorithms to use?

Representation

- What functions should we learn (hypothesis spaces) ?
- How to map raw input to an instance space?
- Any rigorous way to find these? Any general approach?

Algorithms

- What is a good learning algorithm?
- What is success? How confident can I be of results?
- Generalization vs. overfitting

Coming up …

Different hypothesis spaces and learning algorithms

Decision trees and ID3 algorithm

- (Non-parametric) regression and classification

Provides a "gentle" introduction to machine learning concepts

Rule-based decision algorithm

Linear regression

- Least mean squares regression

Linear classifiers

- Perceptron

Non-linear classifiers

- Multi-layer perceptron (neural networks)

Modeling, evaluation, real world challenges

Cross-topic concepts: bias, variance, feature selection, ML advice