Lecture 1: Introduction

COMP 411, Fall 2021

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

Administrivia

What is machine learning?

- What is a well-defined learning problem?
- An example: learning to play checkers
- What questions should we ask about Machine Learning?

Types of machine learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

vumanfredi@wesleyan.edu

Administrivia

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

Course schedule and homework posted on webpage

- http://vumanfredi.wescreates.wesleyan.edu/teaching/comp412-f21/

We'll use Google classroom for announcements and discussion

- 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

- 80%: 10 homework assignments, no scores dropped
  - Mix of written and (possibly multi-assignment) programming projects
  - Some flexibility in what is done for homework
- 10%: Feedback on slides and homework
- 10%: Writing up questions and answers for one homework (in latex)

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, 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 of these homework

Homework

1st homework out Thursday

- play with python libraries and datasets
- get familiar with basic machine learning concepts

Submissions

- Google drive: COMP412-f21 is shared directory
- Submit homework by copying to COMP412-f21/hw1/USERNAME
- Substitute your Wesleyan username for USERNAME

Important!

- Put your name inside every file including code!
- File formats: only .py, pdf, .txt so my printing script works
  - If I can't print it, I can't grade it :-)
- Filename should match what is specified

Getting started

Python3

- we'll review as needed, see class resources webpage
  - please check you have python3 installed!
    - type python3 at terminal prompt
  - tutorials and other resources posted on course website

Python help available

- at SCIC on 1st floor of Exley

vim and python

- create a .vimrc file in your home directory
- put lines in block in .vimrc and save it
- open new terminal and use vim
  - should see color, line numbers, etc.

```
syntax on
filetype indent plugin on
set modeline
set number
autocmd BufWritePre * %s/\s\+$//ei
au BufNewFile,BufRead *.py
\ set tabstop=4
\ set softtabstop=4
\ set shiftwidth=4
\ set textwidth=79
\ set expandtab
\ set autoindent
\ set fileformat=unix
```

Looking forward

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 is machine learning?

INFORMALLY

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?

Machine learning

… is at the core of

- understanding high level cognition
- performing knowledge intensive inferences
- building adaptive, intelligent systems
- dealing with messy, real world data
- analytics

… has multiple purposes

- knowledge acquisition
- integration of various knowledge sources to ensure robust behavior
- adaptation (human, systems)
- decision making (predictions)

Why use machine learning?

Recent progress in algorithms and theory

- Growing flood of online data
- Computational power is available
- Budding industry

ML is used when:

- Human expertise does not exist (navigating on Mars)
- Humans cannot explain their expertise (speech recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)

Learning is not always useful:

- There is no need to “learn” to calculate payroll

Why use machine learning?

Some examples of tasks that are best solved by using a learning algorithm

Learning patterns

- Predicting passenger survivability on the *Titanic*
- Recognizing tweets as positive or negative
- Clustering faces by identity

State of the art applications

- Autonomous cars
- Automatic speech recognition
- Anomalies in credit card transactions
- Stock prediction

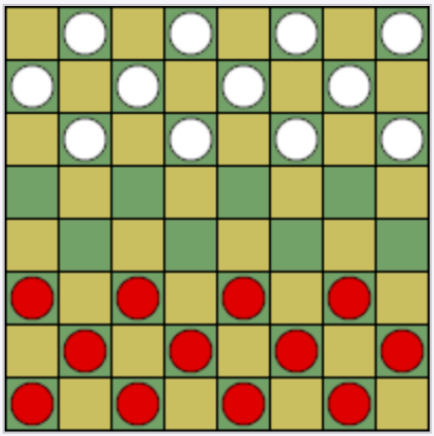
What is machine learning?

SLIGHTLY MORE FORMALLY

What is machine learning?

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

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.,
  - 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."

Performance measure P

- To determine whether learning is happening
- E.g.,
  - Winning rate: whether you won or lost game
  - # of patients that were accurately diagnosed
  - # of marks you get on the exam

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.,
  - Games played by the program (with itself)
  - More patients you examine, better you get at diagnosing illness
  - More exams you take, the better you get at taking them

Learning as generalization

From slides of Vivek Srikumar (U of Utah)

Machine learning is the future

Traditional programming

Machine learning

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

Machine learning paradigms

Supervised learning

- **Learn with a teacher** (labeled examples)
- Construct map from input to an output
  - Classification: categorical output
  - Regression: continuous output

Unsupervised learning

- **Learn without a teacher** (unlabeled examples)
- Data mining
  - Discover patterns and structure in data

Reinforcement learning

- **Learn by interacting in environment**
- Learn control policy

Inductive learning: learning to improve performance based on experience (data)

Machine Learning Paradigms

SUPERVISED LEARNING

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

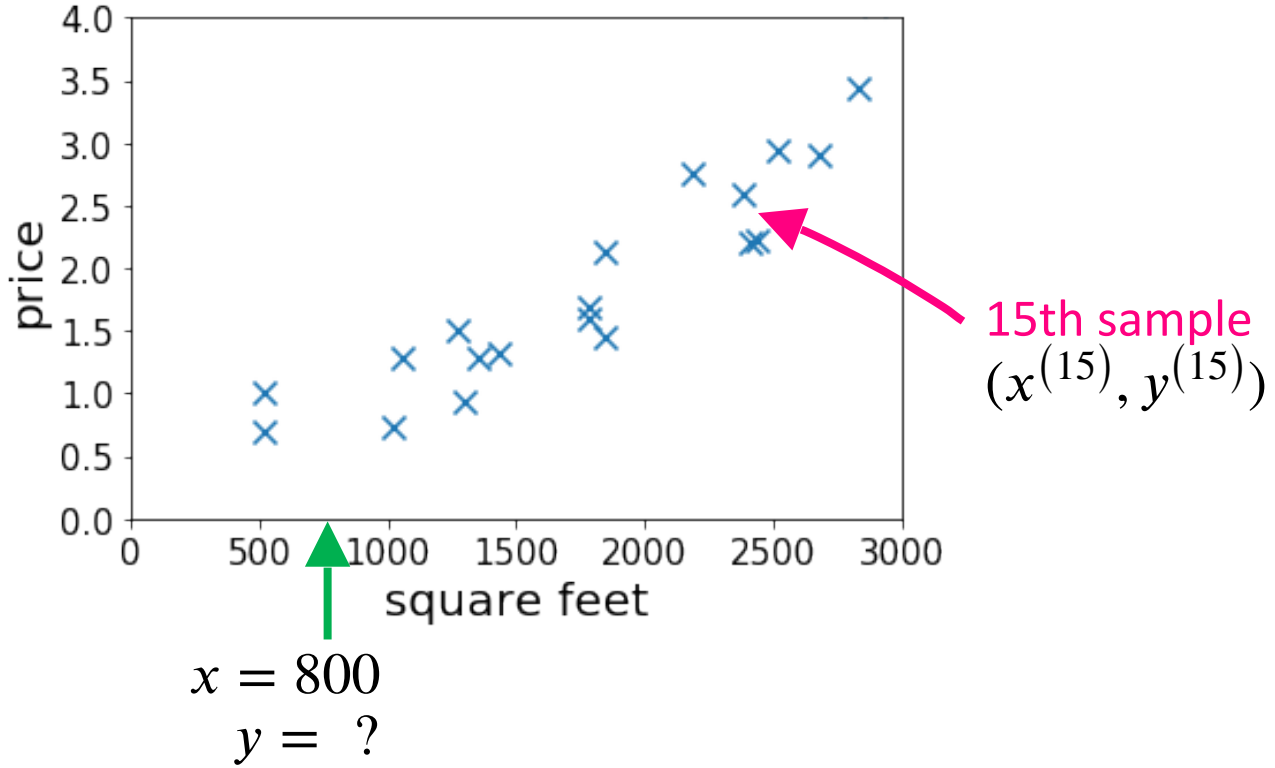
- I say it is going to rain 23 mm, and it rains 49 cm, huge prediction error

Housing Price Prediction

Given: dataset that contains n samples

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

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

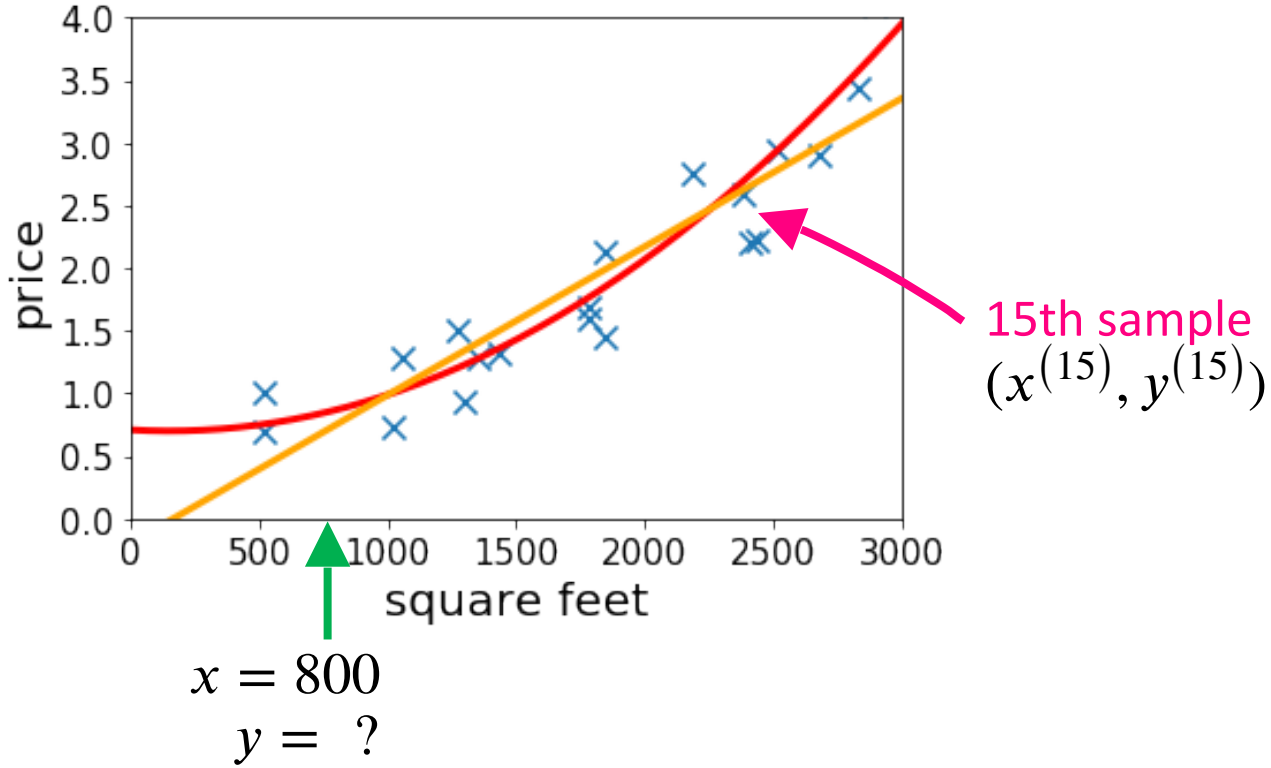


Housing Price Prediction

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

(size, lot size) $\rightarrow$ price

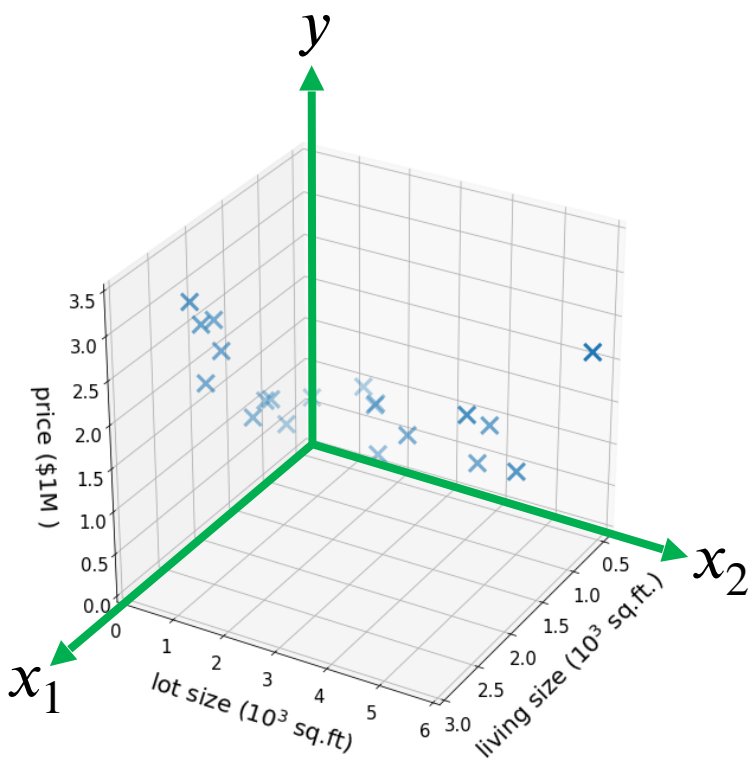
features/input $x \in \mathbb{R}^2$

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



High-dimensional Features

$x \in \mathbb{R}^d$ for large d

For example:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ \vdots \\ \vdots \\ x_d \end{bmatrix}$$

--- living size
--- lot size
--- # floors
--- condition
--- zip code
$\vdots$

$y \longrightarrow$ --- price

Regression vs. Classification

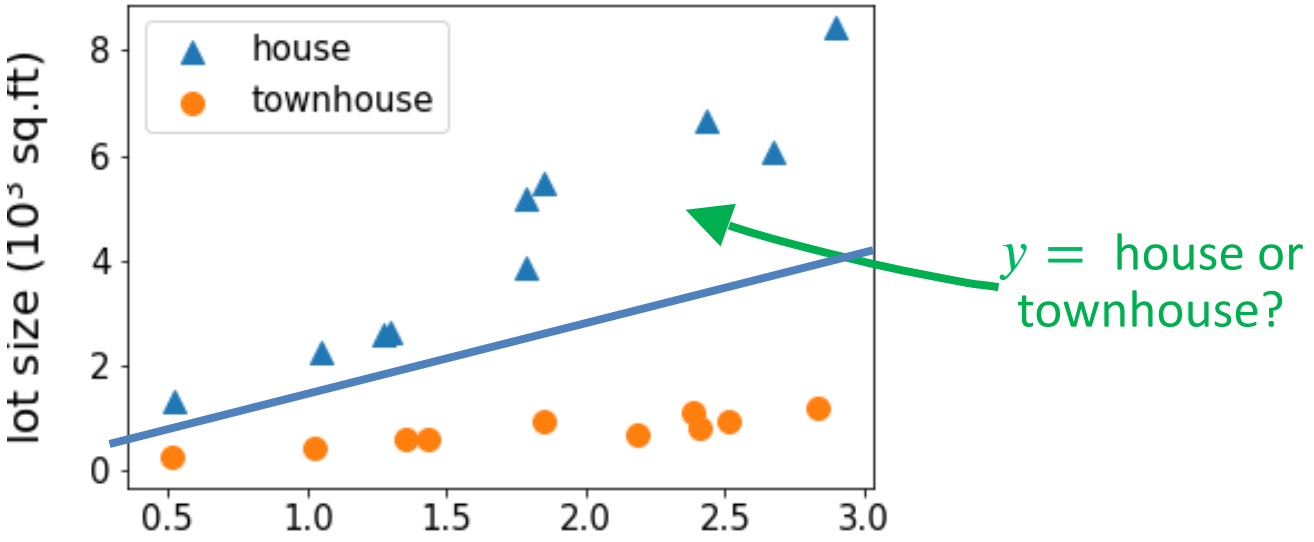
Regression: if $y \in \mathbb{R}$ is a continuous variable

- e.g., price prediction

Classification: the label is a discrete variable

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

  (size, lot size) → house or townhouse?



Machine Learning Paradigms

UNSUPERVISED LEARNING

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

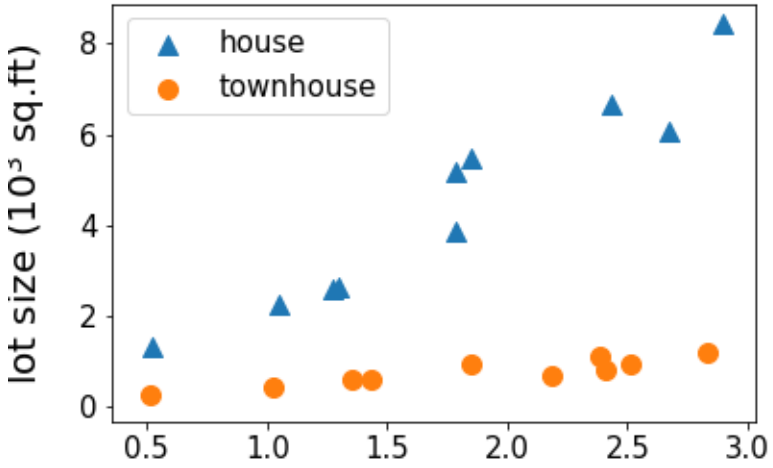
Supervised vs. Unsupervised Learning

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

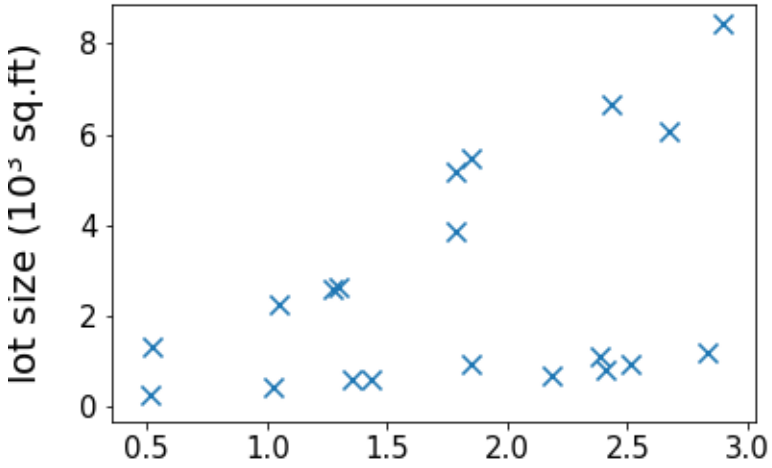
Goal (vaguely-posed):

- to find interesting structures in the data

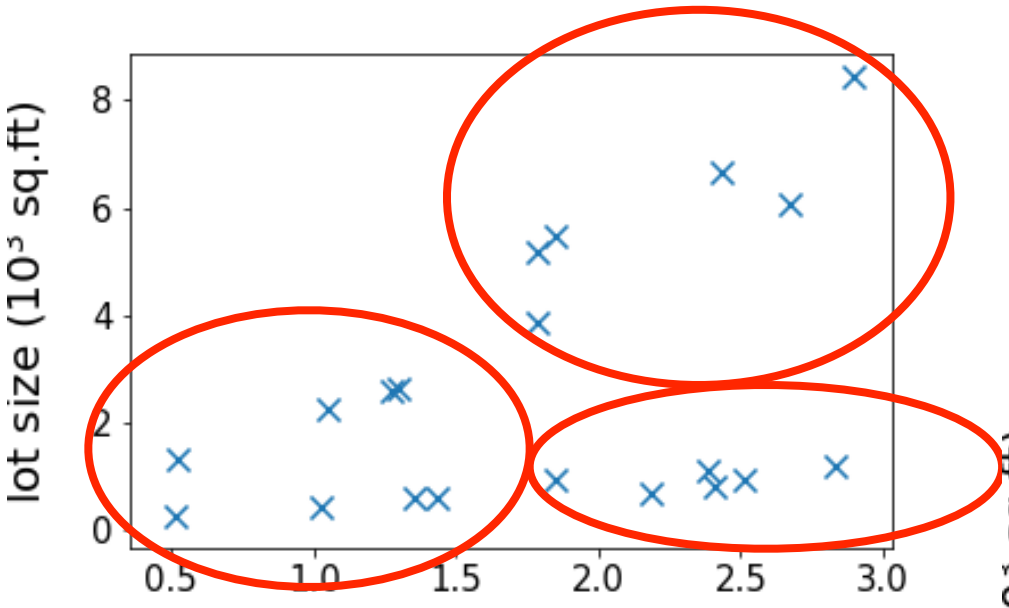
Supervised

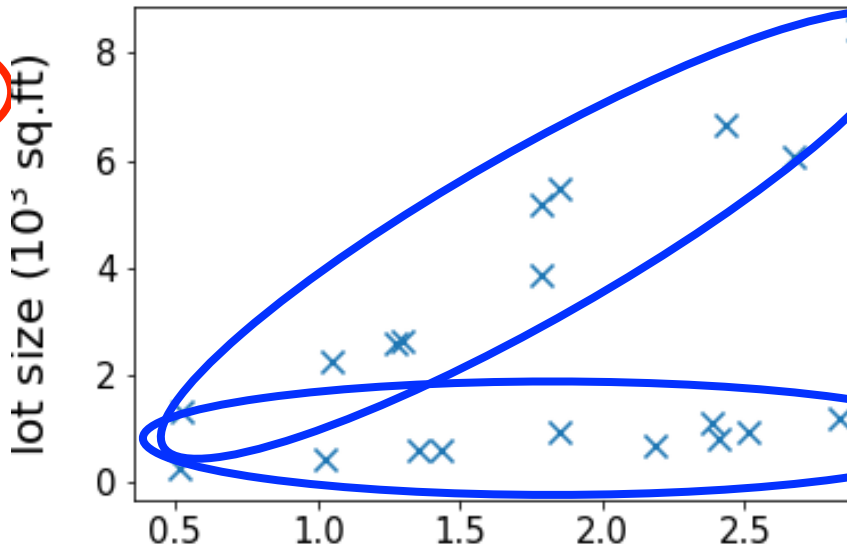


Unsupervised

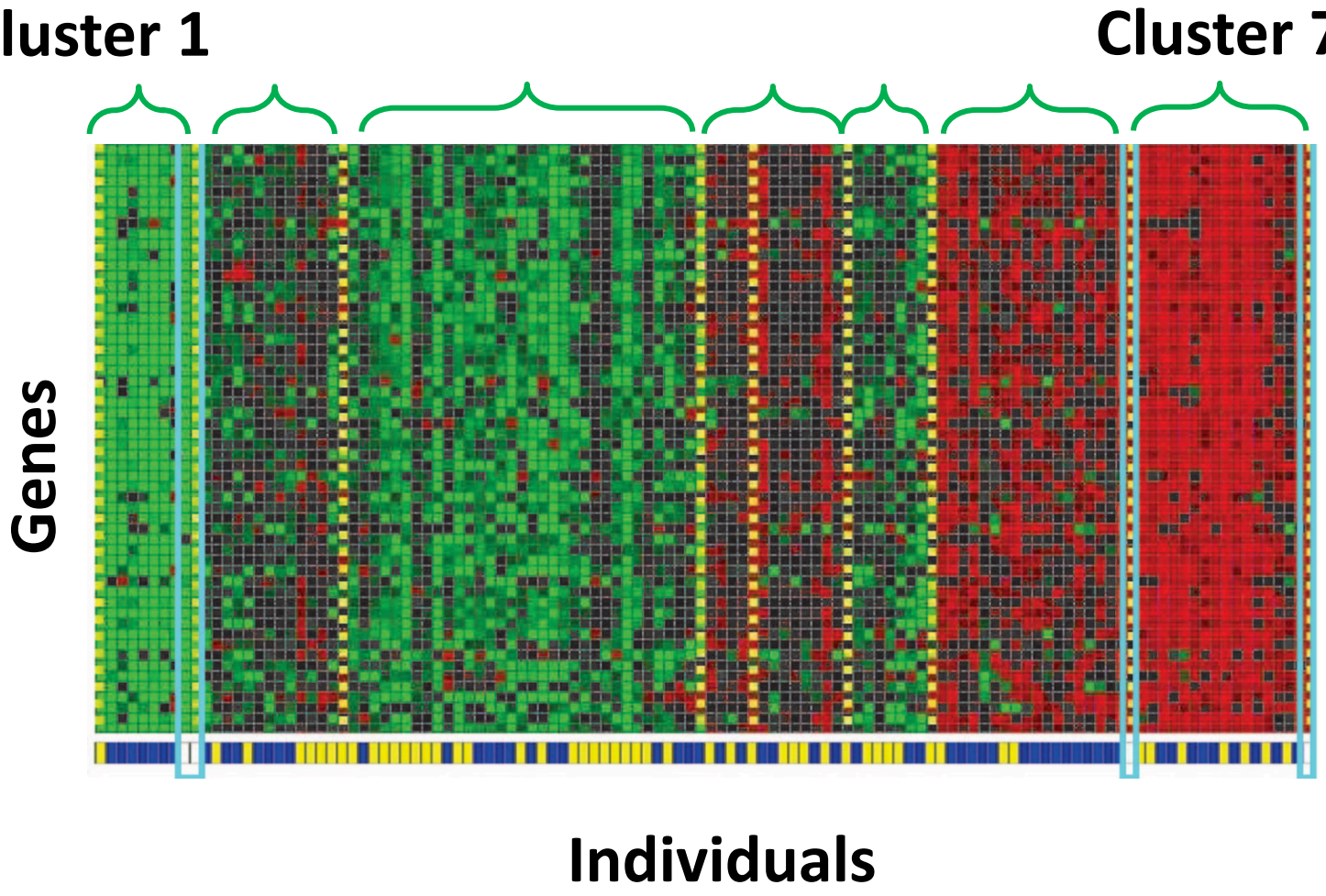


Clustering


lot size (10^3 sq.ft)
0
2
4
6
8
0.5
1.0
1.5
2.0
2.5
3.0

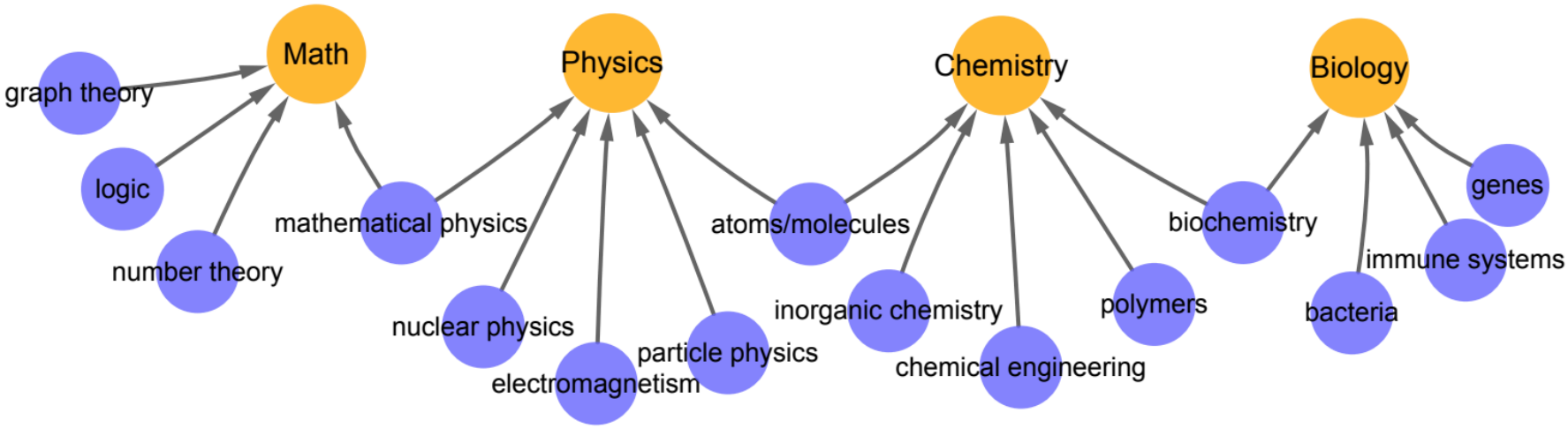

lot size (10^3 sq.ft)
0
2
4
6
8
0.5
1.0
1.5
2.0
2.5

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

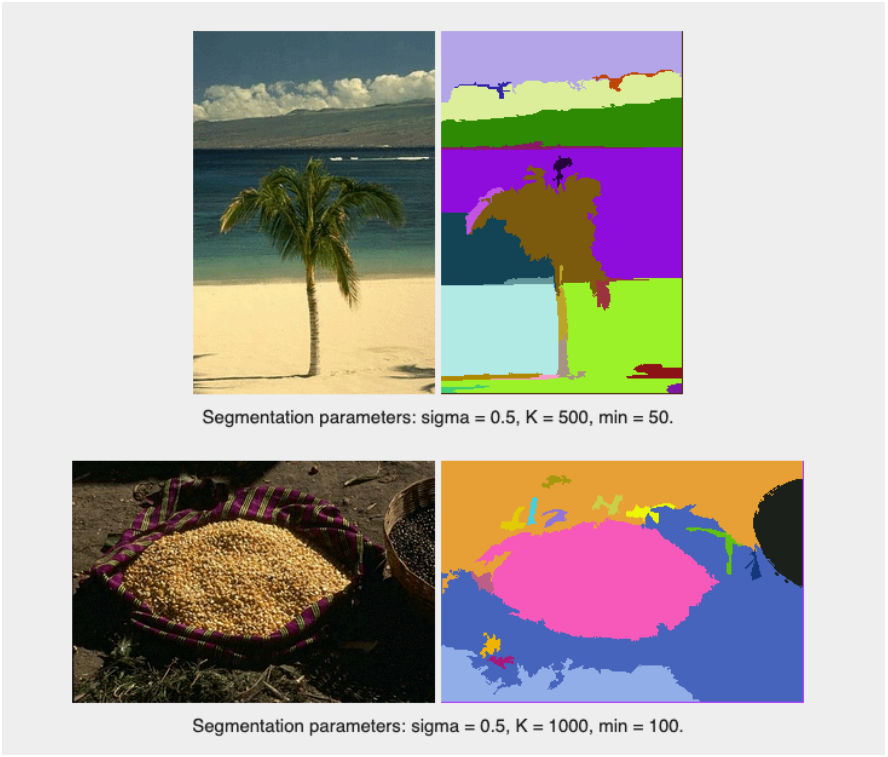
- Discover regions

Words

- Synonyms

Documents

- Topics



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

Machine Learning Paradigms

REINFORCEMENT LEARNING

Learning to control

Popular models of machine learning

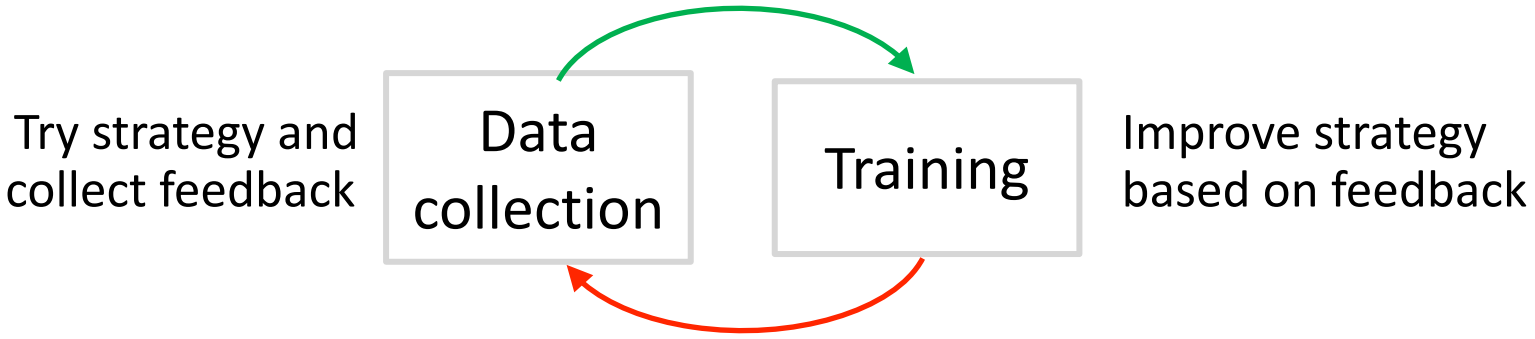
- Supervised: classification, regression, etc
- Unsupervised: clustering, frequent patterns, etc

How did you learn to bicycle?

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

Reinforcement learning

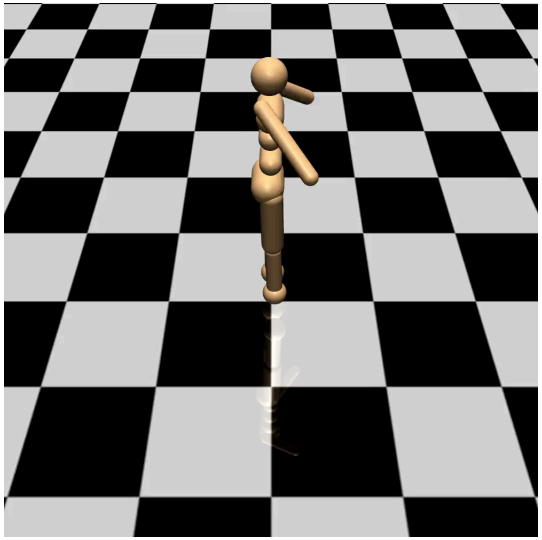
Learn to control the behavior of a system



Performance measure

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

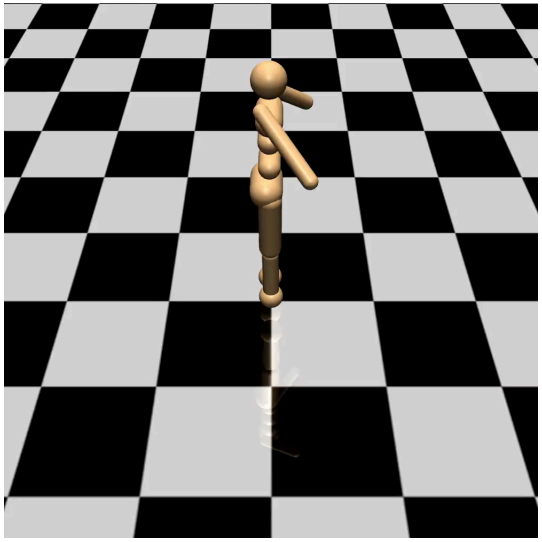
learning to walk to the right



Iteration 10

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

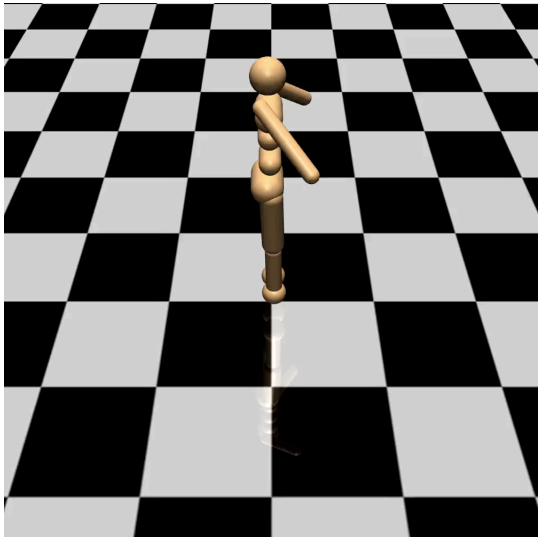
learning to walk to the right



Iteration 20

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

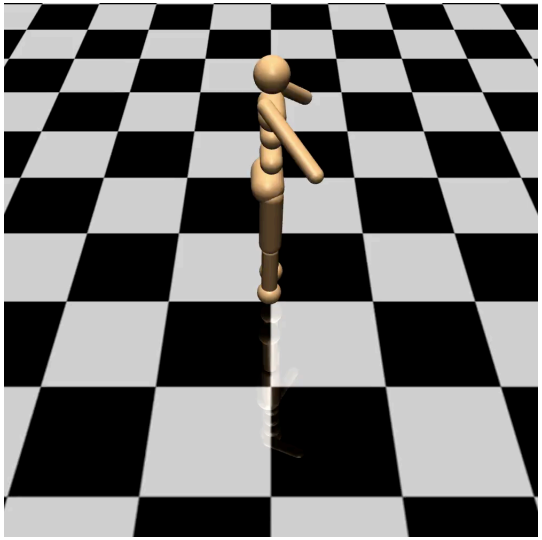
learning to walk to the right



Iteration 80

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

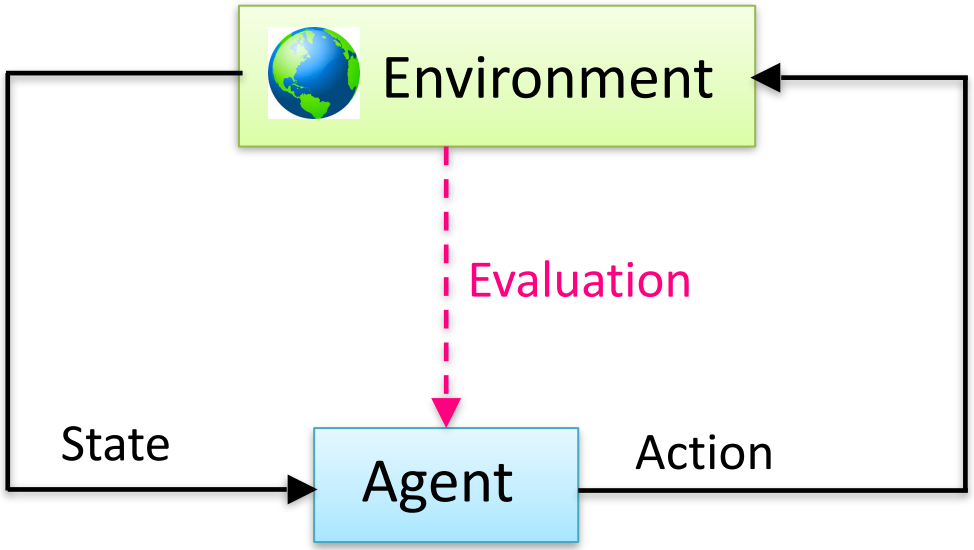
learning to walk to the right



Iteration 210

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

RL framework



Learn from close interaction

Stochastic environment

Noisy, delayed scalar evaluation

Learn a policy: maximize a measure of long term performance

Applications of RL

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 Paradigms

LOOKING FORWARD

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

Linear regression

- Least mean squares regression

Linear classifiers

- Perceptron

Non-linear classifiers

- Decision trees and ID3 algorithm
- Multi-layer perceptron (neural networks)

Reinforcement learning

Modeling, evaluation, real world challenges

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