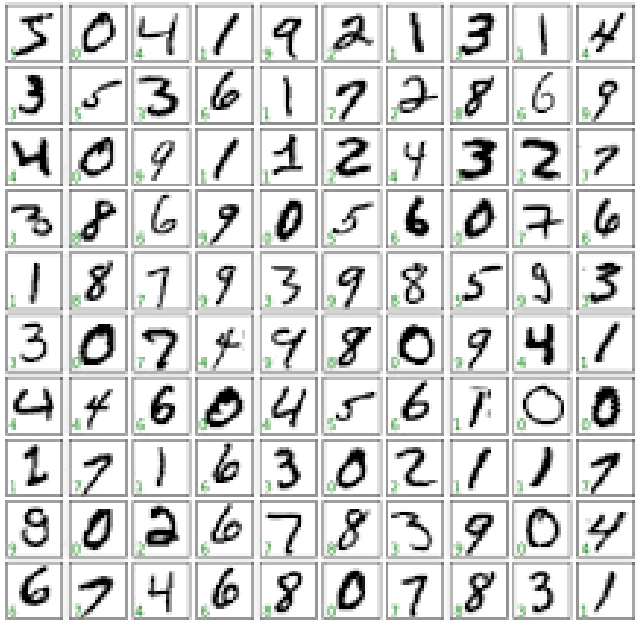
Lecture 1

CS436/536: Introduction to Machine Learning

Zhaohan Xi
Binghamton University

zxi1@binghamton.edu

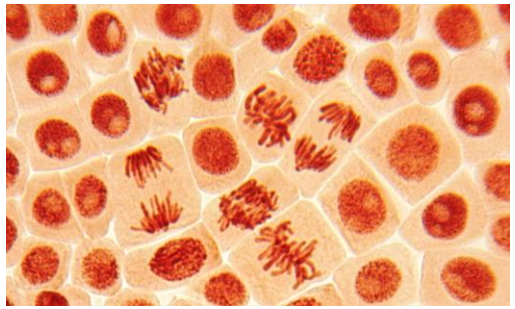
Why? Huh!? What is it good for?

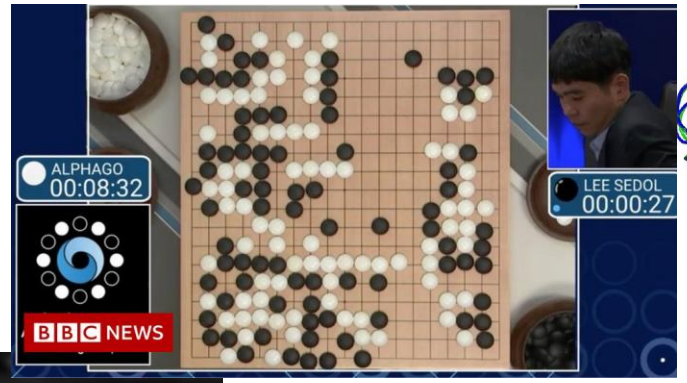


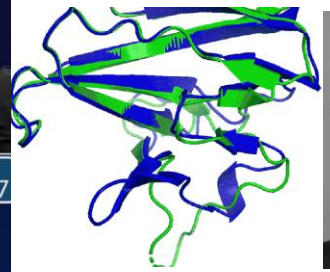
Iris Versicolor

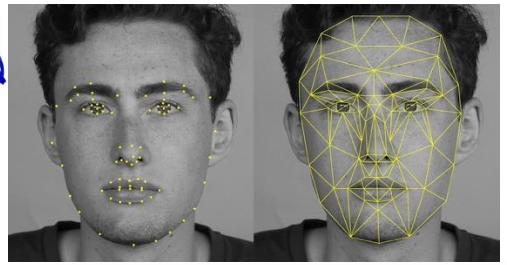
Iris Setosa

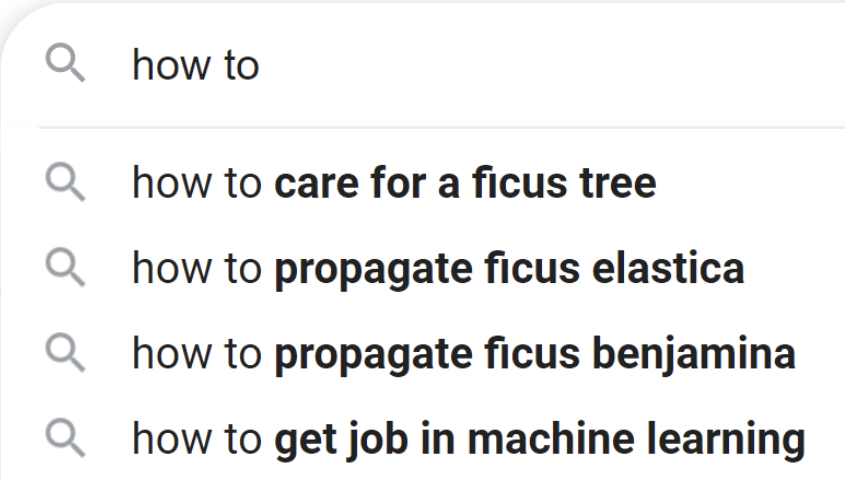
Iris Virginica











Learning to Solve Hard Problems with Data

When is Machine Learning Useful?

- Human expertise does not exist or cannot be conveniently translated into an algorithm (navigating on Mars, protein structure)
- Humans can’t explain their expertise (speech / face recognition)
- Models must be customized (personalized medicine)
- Models are based on huge amounts of data (genomics)
- No “analytical” solution

Some Other Applications

- Recognizing patterns
  - Facial identities or facial expressions
  - Handwritten or spoken words
  - Medical images
- Generating patterns
  - Generating images or motion sequences
- Recognizing anomalies
  - Unusual credit card transactions
  - Unusual patterns of sensor readings in a nuclear power plant
- Prediction
  - Future stock prices or currency exchange rates

Central to

- Natural language processing (NLP)
- Computer vision
- Computational biology
- Robotics

…

Example: Credit Approval

- Using salary, debt, years in residence, etc., approve for credit or not
- Nobody has an optimal credit approval formula
- But banks have data
  - Customer information
  - Credit history

age	33 years
salary	50,000
debt	27,500
years employed	1
years at residence	2
…	…

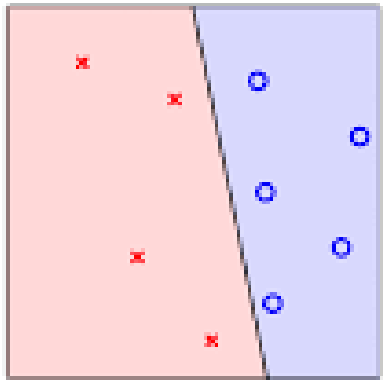
Approve for credit?

Outline

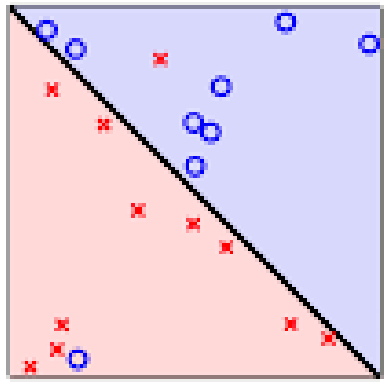
1. What is Learning?

   "Learning is a process by which a system improves performance from experience"

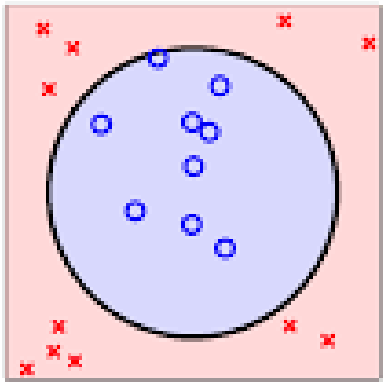
   - Herbert Simon
2. Can we Learn?
3. How to Learn?
4. How to Learn well?
5. Advanced Techniques.
6. Other Learning Paradigms.



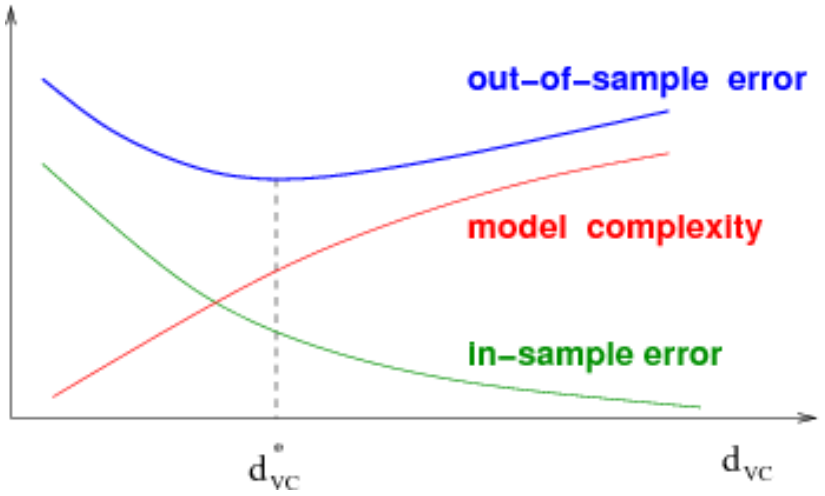
(linear separable)



(not linear separable)

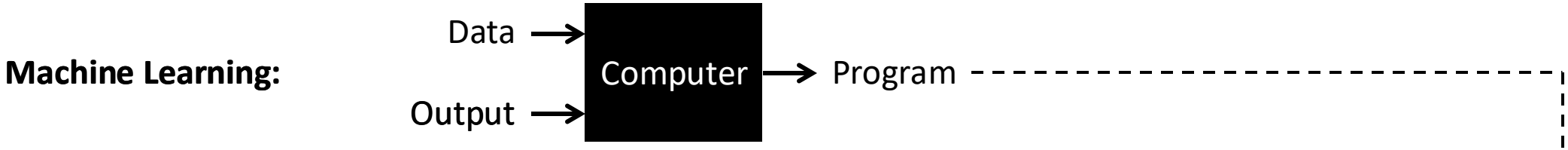


(not linear separable)

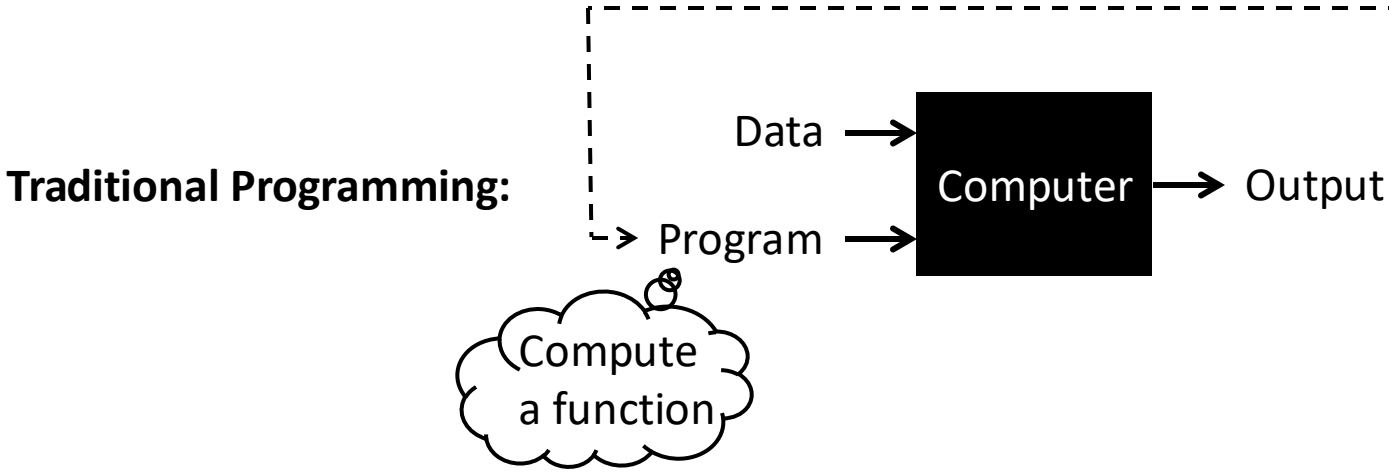


What is Machine Learning?

"*computer programs that automatically improve their performance through experience*" – Herbert Simon



"… the ability to learn without being explicitly programmed" – Arthur Samuel



What it is not:

The *design* approach

How to Learn?

- Memorization
  - Accumulate all the facts in a large lookup table

- Generalization
  - Deduce new facts from old facts
  - Recognize a pattern and use it to *make a prediction*

  - A pattern exists
  - The pattern is **UNKNOWN**
  - But we have data

Credit Approval

- Using salary, debt, years in residence, etc., approve for credit or not
- Nobody has an optimal credit approval formula
- But banks have data
  - Customer information
  - Credit history

age	33 years
salary	50,000
debt	27,500
years employed	1
years at residence	2
…	…

Approve for credit?

Components of the **Learning Problem**

- age, salary, debt, years employed, …
- Approve credit or not
- **True** relationship between $\boldsymbol{x}$ and y
- Data on customers

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

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

target function $f: \mathcal{X} \to \mathcal{Y}$ [unknown]

Optimal credit approval function

data set $\mathcal{D}_N = (x_1, y_1), \ldots (x_N, y_N)$ $[y_n = f(x_n)]$

$\mathcal{X}, \mathcal{Y}$ and $\mathcal{D}$ are *given* by the learning problem

The target function f is fixed but unknown

Learn the target function f from the data $\mathcal{D}$

What is Learning?

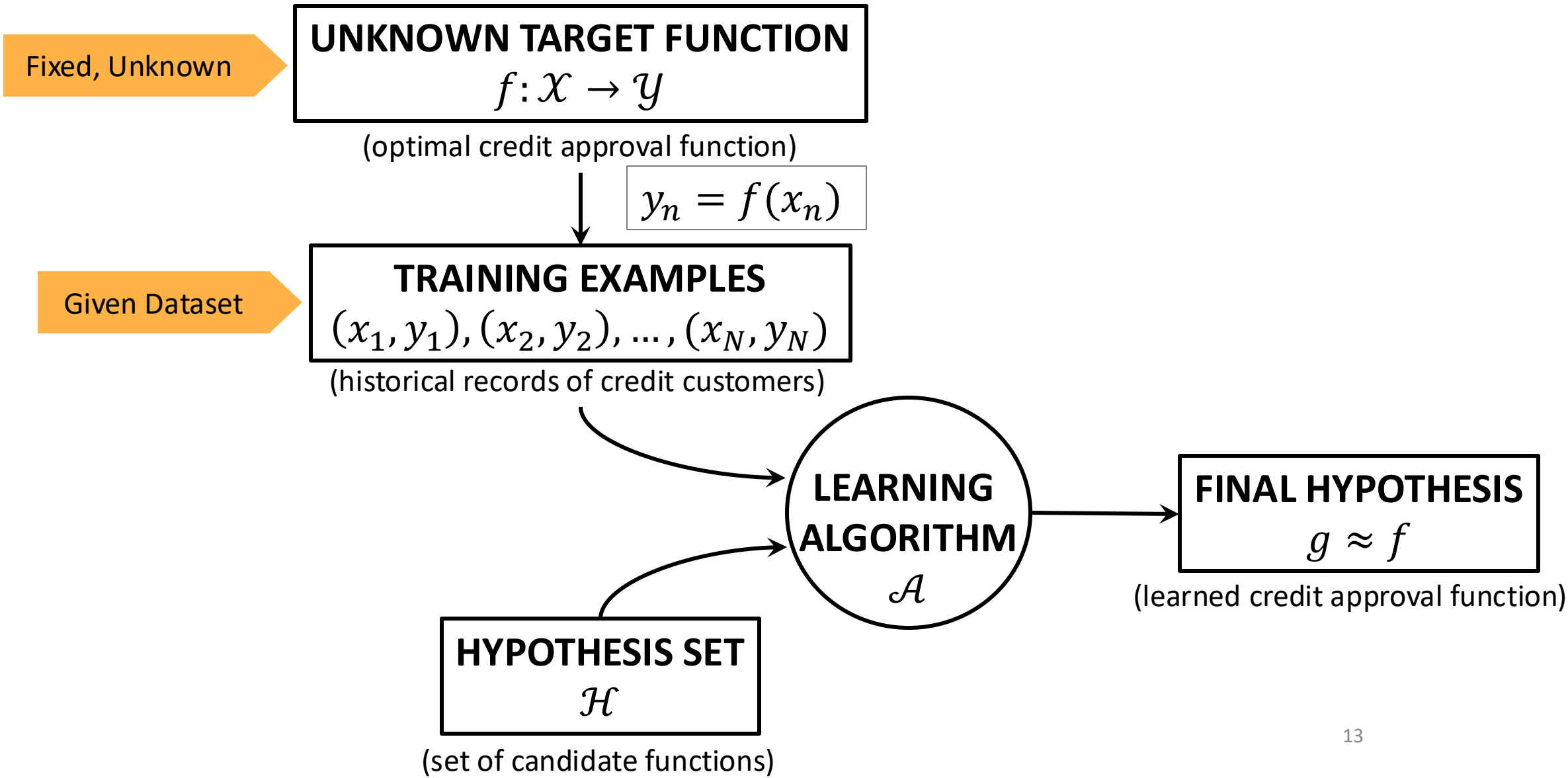
- Obtaining f
- Result of learning is an approximation of f

$$g: \mathcal{X} \to \mathcal{Y}$$

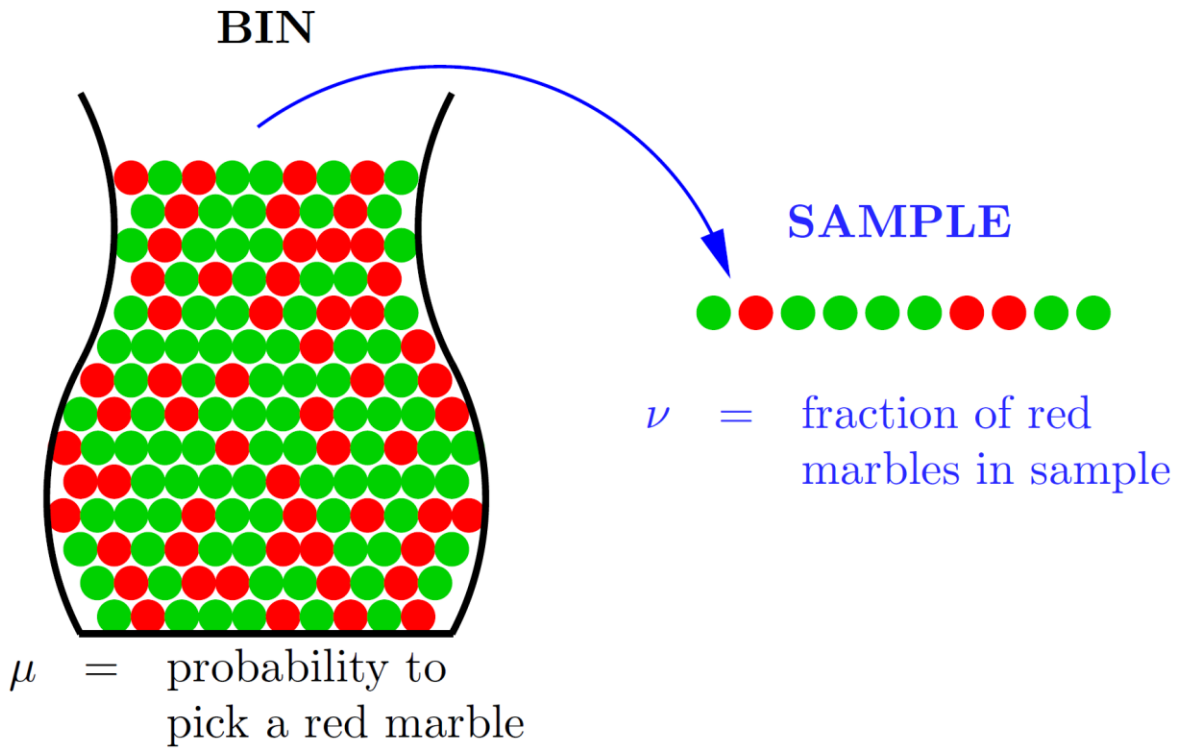
- Want:

$g \approx f$ i.e. $g(\boldsymbol{x}_*) \approx f(\boldsymbol{x}_*)$ where $\boldsymbol{x}_*$ is the next *test data point*

Learning Problem Setup



Estimating Population Mean from Sample Mean



Pick a *random* sample of N marbles with replacement *independently*

Observe the fraction of red marbles ν

Note: the only random quantity here is ν. μ is fixed (albeit unknown)

What does ν tell us about μ?
Nothing for sure.
But…

Estimating Population Mean from Sample Mean

Can we say anything **<u>for certain</u>** about μ (outside the data)
having observed ν (the data)?

- No.

It is *possible* to pick only red marbles while the bin has mostly green marbles

But not probable

- See the binomial distribution

- What is the relationship between ν and μ?

Learning

- Start with a set of candidate hypotheses $\mathcal{H}$ which likely represent f

$\mathcal{H} = \{h_1, h_2, \dots\}$ The hypothesis set or *model*

- Select a hypothesis g from $\mathcal{H}$

Using a *learning algorithm*

- Use g for new customers

Hope that $g \approx f$

$\mathcal{X}, \mathcal{Y}$ and $\mathcal{D}$ are ***given*** by the learning problem

The target function f is **fixed but unknown**

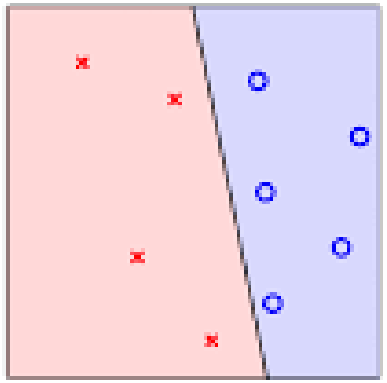
We choose $\mathcal{H}$ and the learning algorithm

Outline

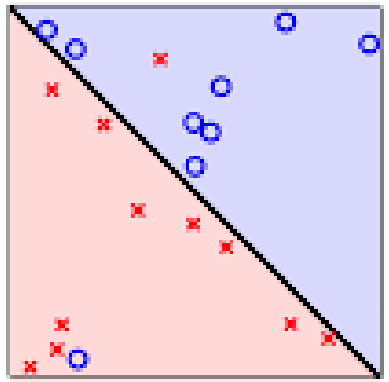
1. What is Learning?

   "Learning is a process by which a system improves performance from experience"

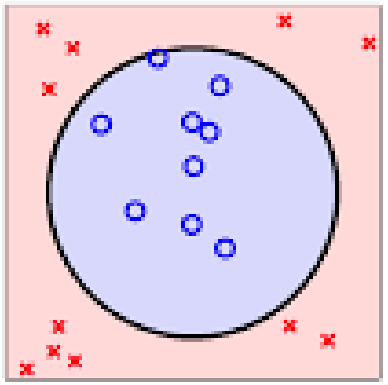
   - Herbert Simon
2. Can we Learn?
3. How to Learn?
4. How to Learn well?
5. Advanced Techniques.
6. Other Learning Paradigms.



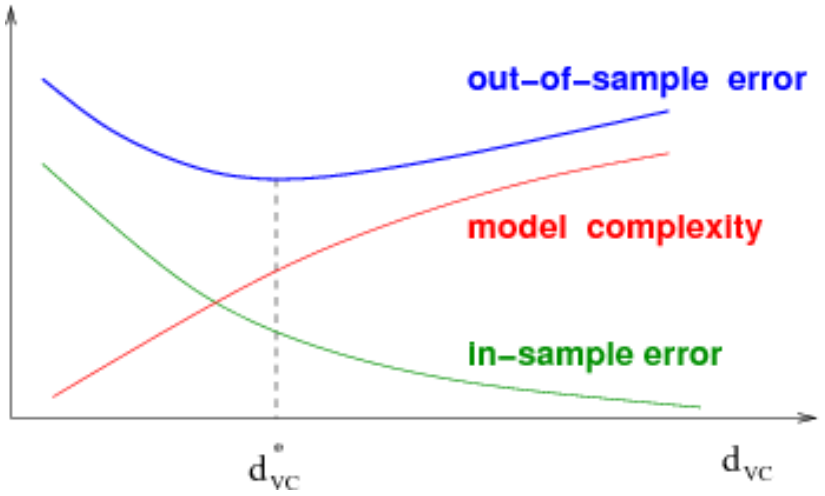
(linear separable)



(not linear separable)



(not linear separable)



Prerequisites

- CS 375 Design and Analysis of Algorithms or equivalent
- MATH 327 Probability with Statistics Methods

- Comfort with Linear Algebra, Calculus

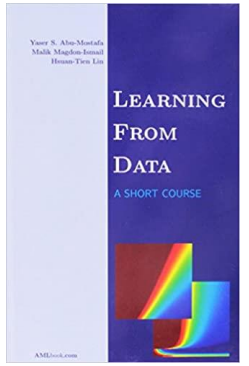
- Proficiency with at least one of the following high level programming language:

Python, Matlab, Java, C++

Textbooks

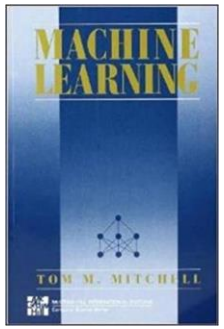
- *Learning from Data*

by Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin



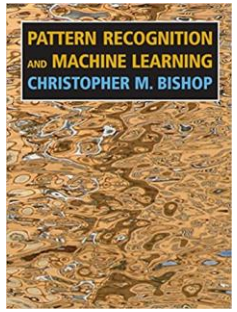
- *Machine Learning*

by Tom Mitchell



- *Pattern Recognition and Machine Learning*

by Christopher M. Bishop



Grading

- (70 points) 7-10 Take-home Homework and Programming Assignments
- (5 points) 4 In-class Quizzes
  - Open book, open notes, NO electronics
- (30 points) Two Exams (open-book)

- Please talk to each other
- Write solutions / implement program yourself

- Check Brightspace for all announcements and course content

Grading

- A/A-: 100-90
- B+/B/B-: 75-90
- C+/C/C-: 65-75
- D: 55-65

Office hours

Zhaohan Xi

Tuesday & Thursday

1:00PM – 2:50PM

EB G06

Taoyue Wang

Friday

2:00PM – 4:00PM

EB N00

Additional appointments arranged by email:

zxi1@binghamton.edu

twang61@binghamton.edu

Syllabus available on Brightspace

- Please review the syllabus
- Please read the CS department Academic Honesty letter to students
- Please review Watson College and University Academic Integrity policy

Academic Integrity/Honesty

- The Watson College Student Academic Honesty Code
- Student Academic Honesty Code

Today’s Tasks

- Say hello
- Please talk to each other
- Find some people to work on homework with
- Form some initial agreements to work in pairs