CS60050: Machine Learning

Autumn 2023

Sudeshna Sarkar

Introduction to ML

3-4 Aug 2023

Course Website

https://sites.google.com/view/cs60050-fall-2023/home

Course Timings

- Wed 11-12 PM, Thu 12-1 PM, Fri 8-9 AM
- Section 1 (Dr. Somak Aditya, NC442)
  - Even Roll Numbers
- **Section 2 (Dr. Sudeshna Sarkar, NC 243)**
  - **Odd Roll Numbers**

Contact
MS Teams
Email: sudeshna@cse.iitkgp.ac.in Subject “CS60050 ML23”

Course Evaluation Plan

- Mid Term - 25%
- Final Exam - 40%
- Assignments (3-4) - 20%
- Class-Tests (Two) - 15%

Attendance Policy

Students must attend all classes

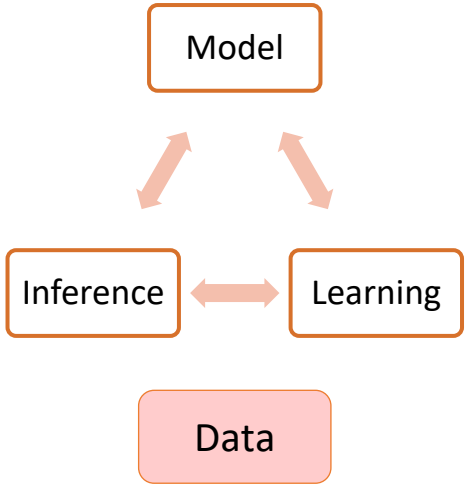
Attendance Policy: Attendance record will be maintained and uploaded on the website.

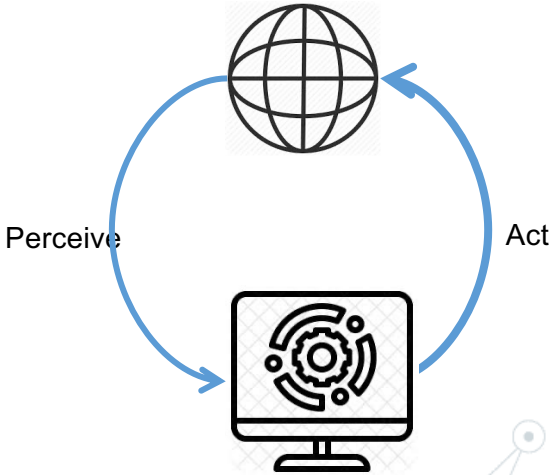
All students must maintain 75% attendance to continue the course.

What is Machine Learning?

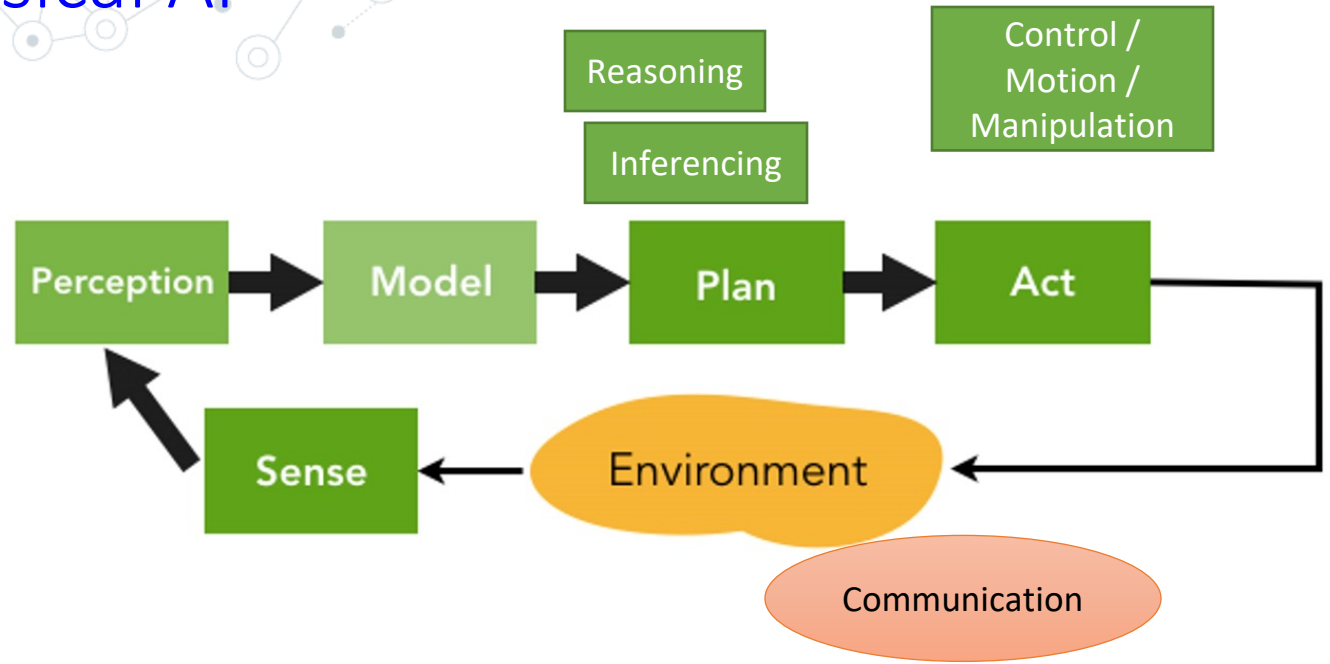
Artificial Intelligence

- The quest for building Intelligent Agents that can think and act rationally (or humanlike)





Classical AI



Learning

Data

Machine Learning Toolbox

- Machine Learning
- Statistics
- Probability
- Computer Science
- Optimization

What is ML?

Speech Recognition

1. Learning to recognize spoken words

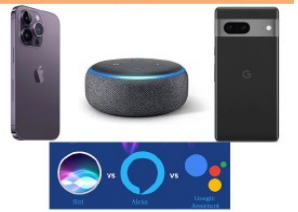
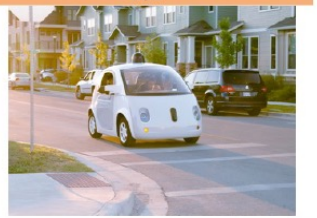
THEN	NOW
"…the SPHINX system (e.g. Lee 1989) learns speaker-specific strategies for recognizing the primitive sounds (phonemes) and words from the observed speech signal…neural network methods…hidden Markov models…" (Mitchell, 1997)	

Figure from https://botpenguin.com/alexa-vs-siri-vs-google-assistant/

Robotics

2. Learning to drive an autonomous vehicle

THEN	NOW
"…the ALVINN system (Pomerleau 1989) has used its learned strategies to drive unassisted at 70 miles per hour for 90 miles on public highways among other cars…" (Mitchell, 1997)	 waymo.com

Games / Reasoning

3. Learning to beat the masters at board games

THEN	NOW
"…the world's top computer program for backgammon, TD-GAMMON (Tesauro, 1992, 1995), learned its strategy by playing over one million practice games against itself…" (Mitchell, 1997)	

Computer Vision

4. Learning to recognize images

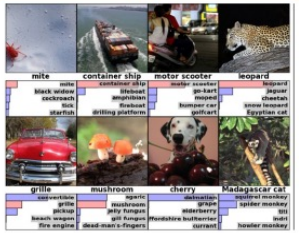
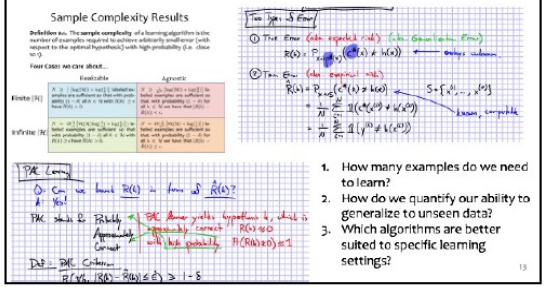
THEN	NOW
"…The recognizer is a convolution network that can be spatially replicated. From the network output, a hidden Markov model produces word scores. The entire system is globally trained to minimize word-level errors…." (LeCun et al., 1995)	

Figure from https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Learning Theory

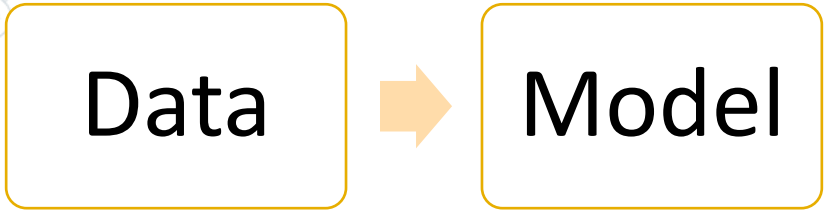
- 5. In what cases and how well can we learn?



1. How many examples do we need to learn?
2. How do we quantify our ability to generalize to unseen data?
3. Which algorithms are better suited to specific learning settings?

[Slides from Dr. Matt Gormley (CMU)]

Learning



- Requires a leap of faith: generalization

ML Big Picture

Learning Paradigms:

What data is available and when? What form of prediction?

- supervised learning
- unsupervised learning
- semi-supervised learning
- reinforcement learning
- active learning
- imitation learning
- domain adaptation
- online learning
- density estimation
- recommender systems
- feature learning
- manifold learning
- dimensionality reduction
- ensemble learning
- distant supervision
- hyperparameter optimization

Problem Formulation:

What is the structure of our output prediction?

boolean	Binary Classification
categorical	Multiclass Classification
ordinal	Ordinal Classification
real	Regression
ordering	Ranking
multiple discrete	Structured Prediction
multiple continuous	(e.g. dynamical systems)
both discrete & cont.	(e.g. mixed graphical models)

Application Areas

Key challenges?

NLP, Speech, Computer Vision, Robotics, Medicine, Search

Theoretical Foundations:

What principles guide learning?

- ❑ probabilistic
- ❑ information theoretic
- ❑ evolutionary search
- ❑ ML as optimization

Facets of Building ML Systems:

How to build systems that are robust, efficient, adaptive, effective?

1. *Data prep*
2. *Model selection*
3. *Training (optimization / search)*
4. *Hyperparameter tuning on validation data*
5. *(Blind) Assessment on test data*

Big Ideas in ML:

Which are the ideas driving development of the field?

- *inductive bias*
- *generalization / overfitting*
- *bias-variance decomposition*
- *generative vs. discriminative*
- *deep nets, graphical models*
- *PAC learning*
- *distant rewards*

Slides from Dr. Matt Gormley (CMU)

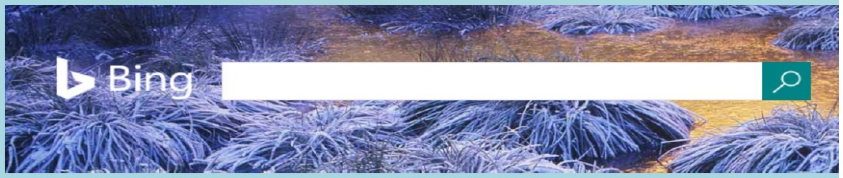
ML and Ethics

What ethical responsibilities do we have as machine learning experts?

Question: What are the possible societal impacts of machine learning for each case below?

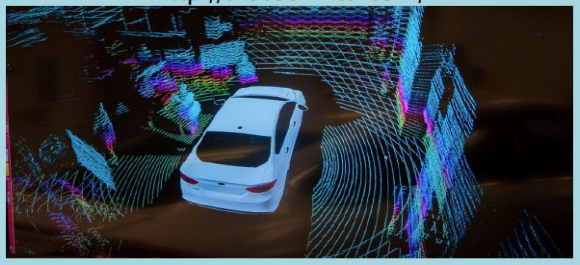
Answer:

1) Search results for news are optimized for ad revenue.



http://bing.com/

http://arstechnica.com/



2) An autonomous vehicle is permitted to drive unassisted on the road.

3) A doctor is prompted by an intelligent system with a plausible diagnosis for her patient.

https://flic.kr/p/HNJHzV

Slides from Dr. Matt Gormley (CMU)

Define Machine Learning

Well-Posed Learning Problem

Three components *<T,P,E>:*

1. *Task, T*
2. *Experience, E*
3. Performance measure, P

Definition of learning*:*

A computer program **learns** if its performance at task T, as measured by P, improves with experience E.

Example Learning Problem(s)

Learning to respond to voice commands (Siri)

- *1. Task, T:*
- *2. Performance measure, P:*
- *3. Experience, E:*

Example Learning Problems

Learning to respond to voice commands (Siri)

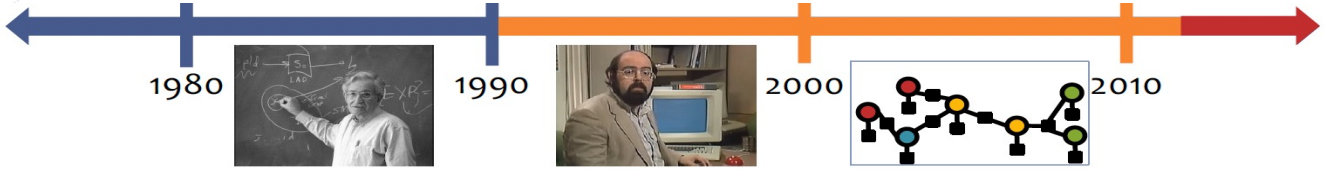
1. Task, T: predicting action from speech

2. Performance measure, P: percent of correct actions taken in user pilot study

3. Experience, E: examples of (speech, action) pairs

Slides from Dr. Matt Gormley (CMU)

Capturing the Knowledge of Experts



Solution #1: Expert Systems

- Over 20 years ago, we had rule-based systems:
  1. Put a bunch of linguists in a room
  2. Have them think about the structure of their native language and write down the rules they devise

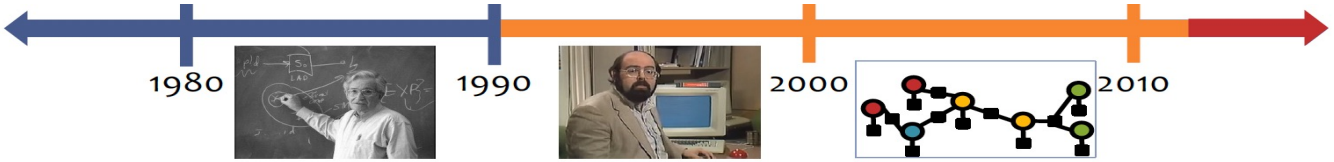
Give me directions to Starbucks
`If: "give me directions to X"` `Then: directions(here, nearest(X))`

How do I get to Starbucks?
`If: "how do i get to X"` `Then: directions(here, nearest(X))`

Where is the nearest Starbucks?
`If: "where is the nearest X"` `Then: directions(here, nearest(X))`

Slides from Dr. Matt Gormley (CMU)

Capturing the Knowledge of Experts

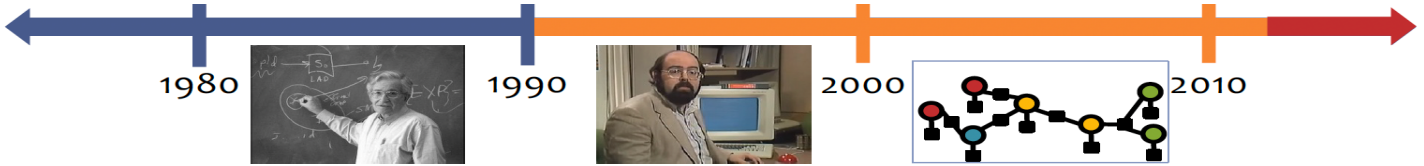


Solution #2: Annotate Data and Learn

- Experts:
  - **Very good at** answering questions about specific cases
  - **Not very good at** telling **HOW** they do it
- 1990s: So why not just have them tell you what they do on **SPECIFIC CASES** and then let **MACHINE LEARNING** tell you how to come to the same decisions that they did

Slides from Dr. Matt Gormley (CMU)

Capturing the Knowledge of Experts



Solution #2: Annotate Data and Learn

1. Collect raw sentences $\{x^{(1)}, \ldots, x^{(n)}\}$
2. Experts annotate their meaning $\{y^{(1)}, \ldots, y^{(n)}\}$

$x^{(1)}$: How do I get to Starbucks?
$y^{(1)}$: `directions(here, nearest(Starbucks))`

$x^{(2)}$: Show me the closest Starbucks
$y^{(2)}$: `map(nearest(Starbucks))`

$x^{(3)}$: Send a text to John that I'll be late
$y^{(3)}$: `txtmsg(John, I'll be late)`

$x^{(4)}$: Set an alarm for seven in the morning
$y^{(4)}$: `setalarm(7:00AM)`

Slides from Dr. Matt Gormley (CMU)

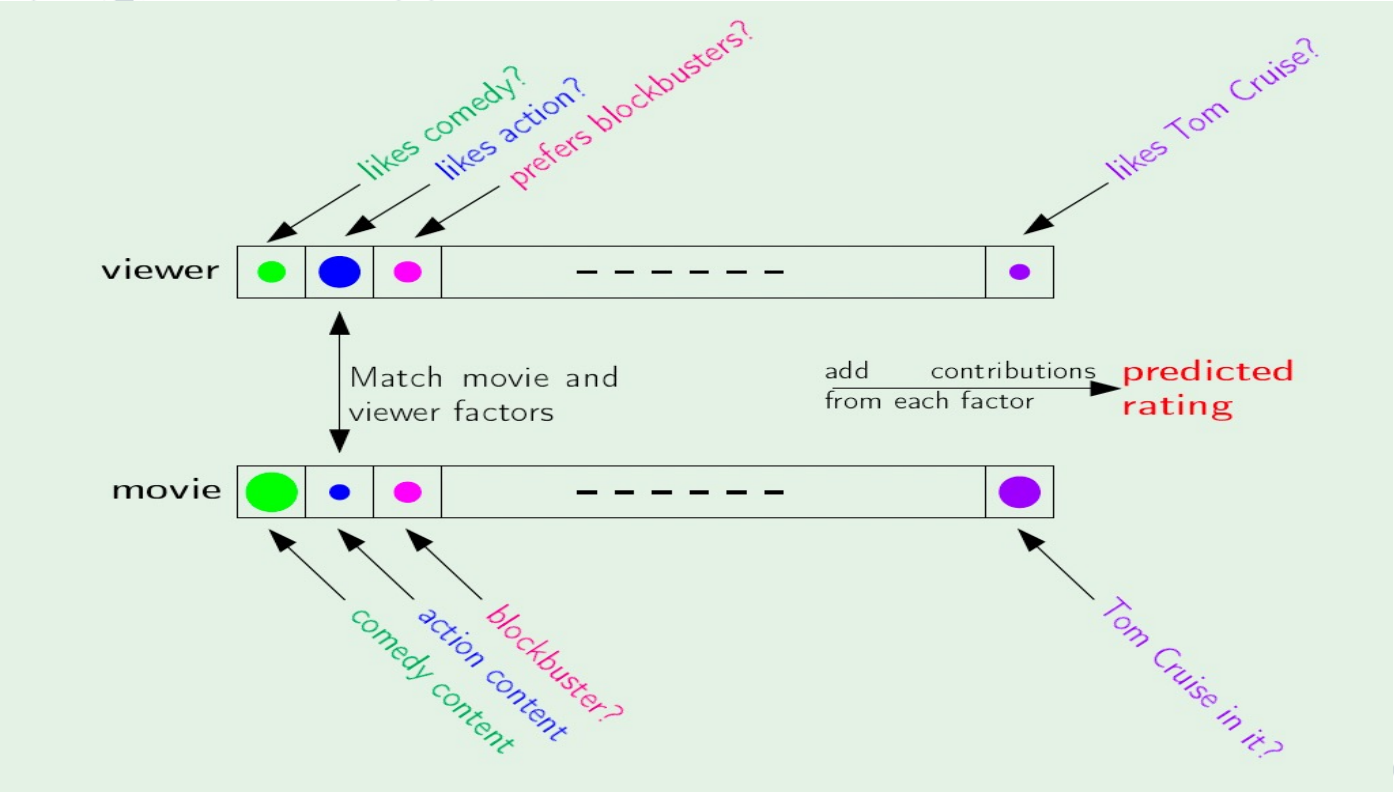
Example: Predicting how a viewer will rate a movie

10% improvement = 1 million dollar prize

- *The essence of machine learning:*
  - *A pattern exists.*
  - *We cannot pin it down mathematically.*
  - *We have data on it.*

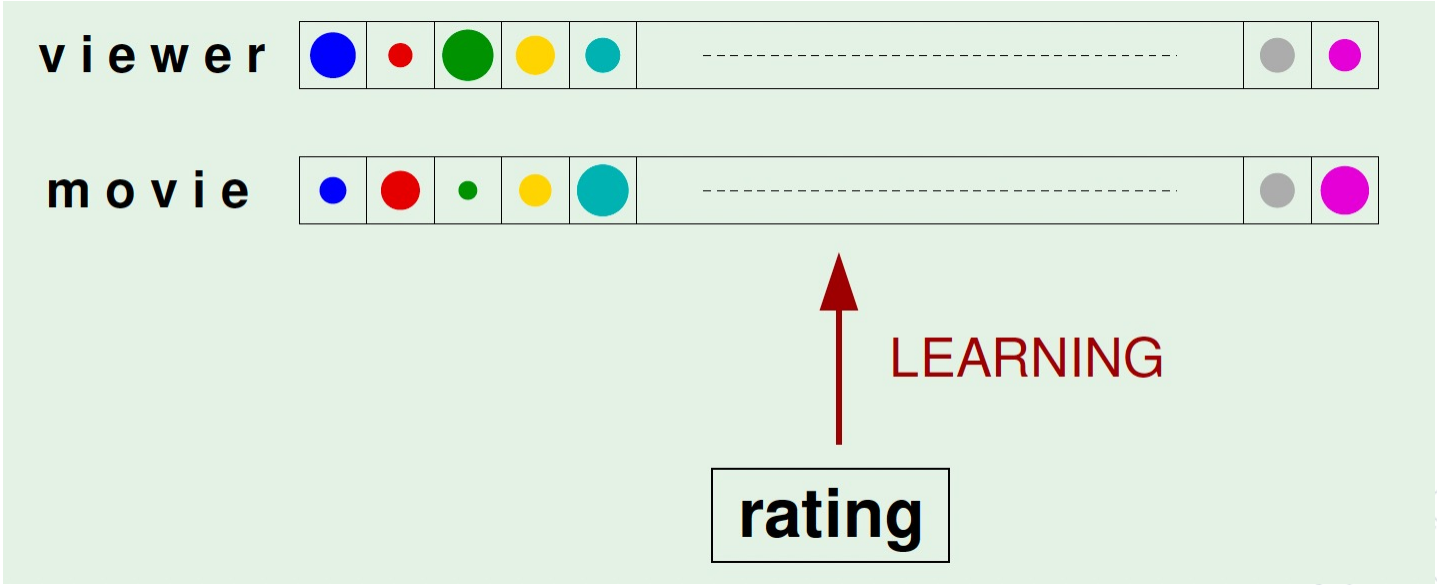
Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Movie Rating – A Solution



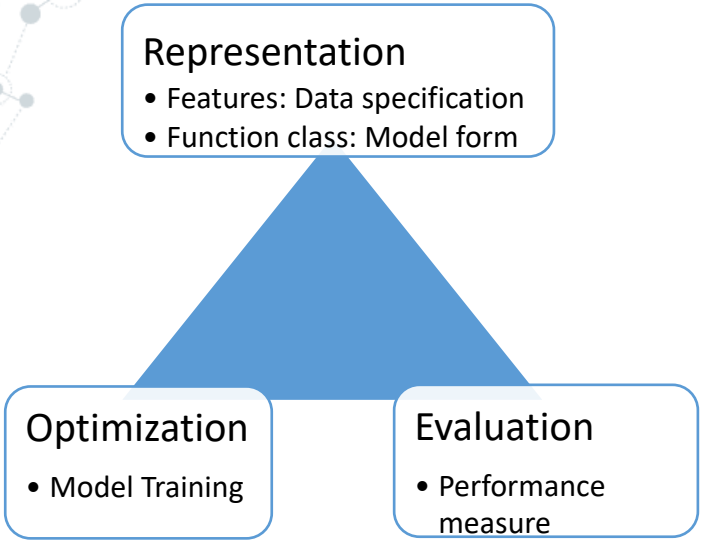
Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

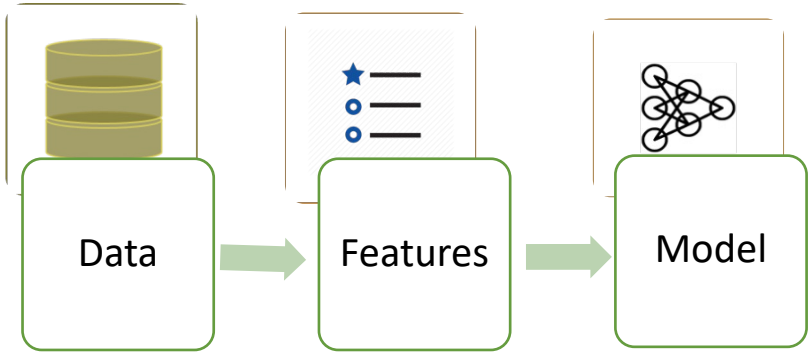
The Learning approach



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Components of a ML application





A. Representation of Data

How is the data specified?

A. Features

- Feature vector of *n* features

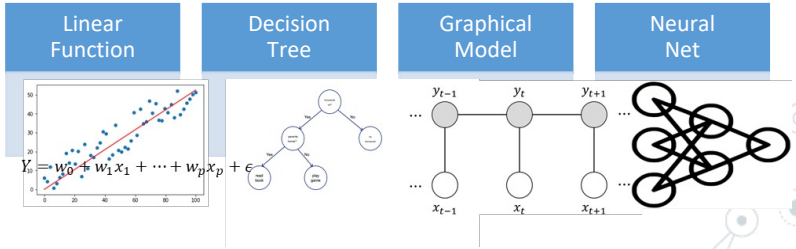
$$\bar{x} = (x_1, x_2, \ldots, x_n)$$

The richer the representation, the more useful it is for subsequent problem solving.

The richer the representation, the more difficult it is to learn.

Hypothesis Set

$\mathcal{H} = \{h\}$



2. Evaluation

1. Accuracy $= \frac{\text{\# correctly classified}}{\text{\# all test examples}}$
2. Logarithmic Loss:

$$L_i = -\log(P(Y = y_i | X = x_i))$$

$$L = \sum_{c=1}^{M} y_{oc} \log(p_{oc})$$

3. Mean Squared error

$$MSE = \frac{1}{m} \sum \left(y_{pred} - y_{true}\right)^2$$

3. Optimization

- Define loss function
- Optimize loss function

- Stochastic Gradient Descent (Convex functions)
- Combinatorial optimization
  - E.g.: Greedy search
- Constrained optimization
  - E.g.: Linear programming

- Task: Credit approval
- Applicant information:

Age	23 years
Gender	male
Annual salary	$30000
Years in residence	1 year
Years in job	1 year
Current debt	$15000
…	…

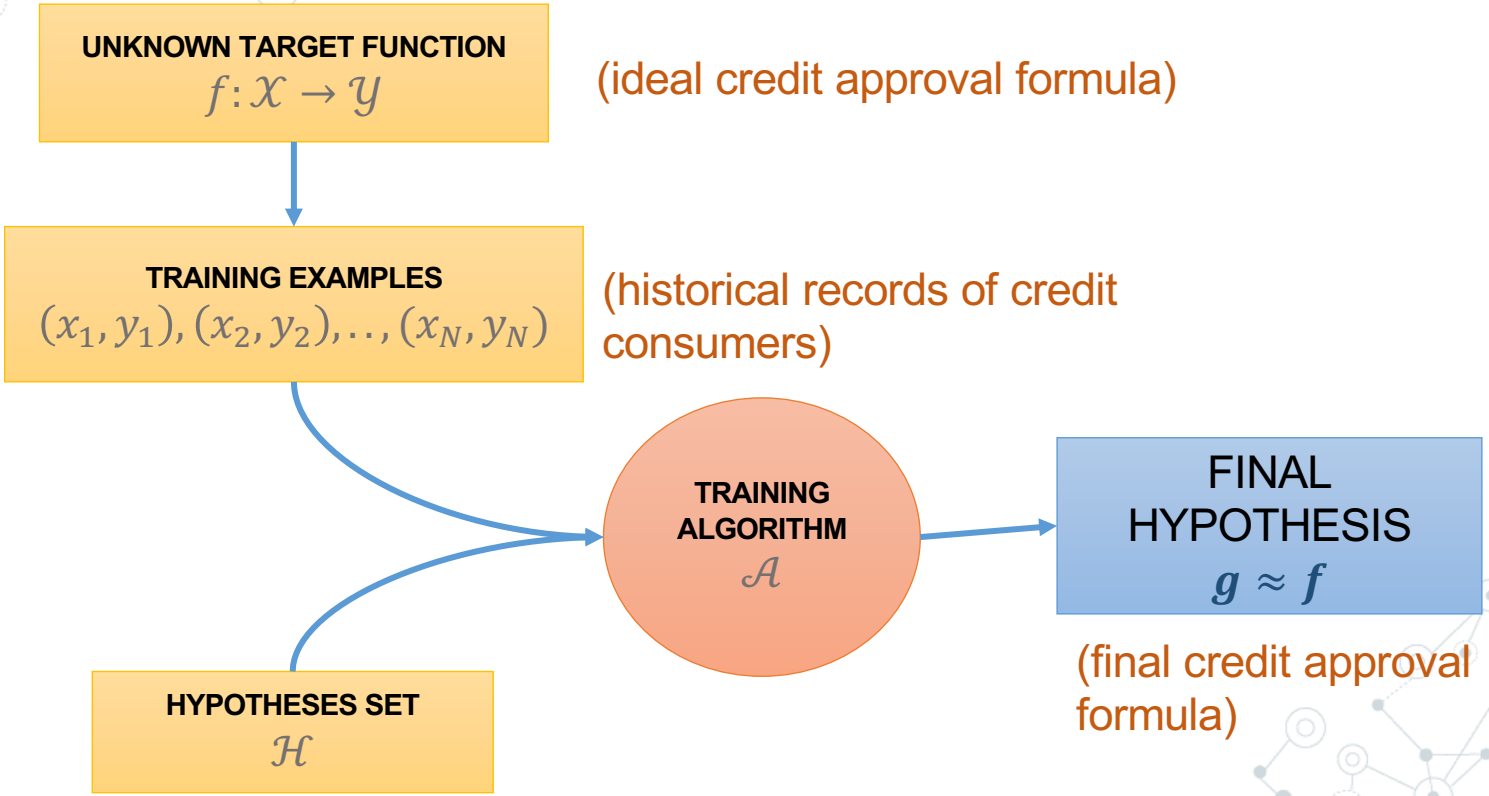
- Approve Credit?

Formalization

- Input x (customer application)
- Output y (good/bad customer?)
- Target function $f: \mathcal{X} \rightarrow \mathcal{Y}$ (ideal credit approval formula)
- Data $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ (historical records)
- Hypothesis g: $\mathcal{X} \rightarrow \mathcal{Y}$ (formula to be used)

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Components of Learning



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Solution Components

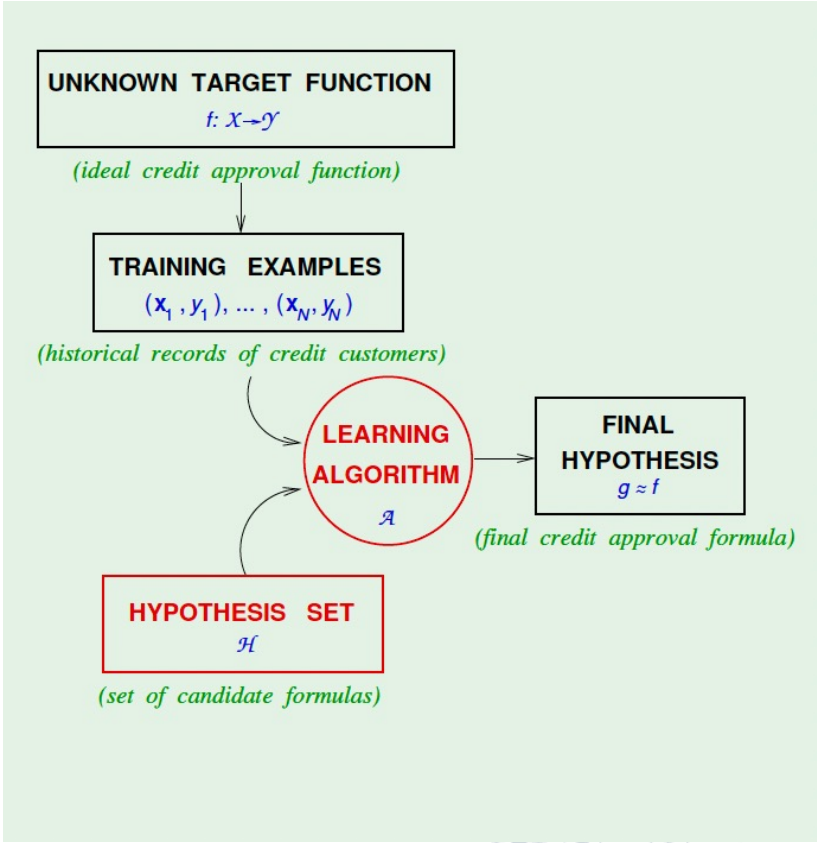
2 solution components of the learning problem

- The Hypothesis Set

  $\mathcal{H} = \{h\}, g \in \mathcal{H}$

- The Learning Algorithm

Together, they are referred to as the *learning model*.



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

A simple hypothesis set – the 'perceptron'

For input $x = (x_1, \ldots, x_d)$ 'attributes of a customer'

Approve credit if $\Sigma_{i=1}^{d} w_i x_i >$ threshold,

Deny credit if $\Sigma_{i=1}^{d} w_i x_i <$ threshold,

Linear formula $h \in \mathcal{H}$ can be written as

$$h(x) = sign(\Sigma_{i=1}^{d} w_i x_i - threshold)$$

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Separability

The perceptron implements

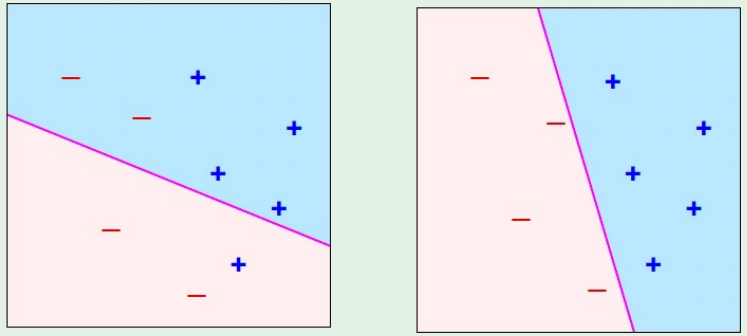
$$h(x) = sign(\Sigma_{i=1}^{d} w_i x_i + \quad w_0)$$

Introducing an artificial coordinate $x_0 = 1$

$$h(x) = sign(\Sigma_{i=0}^{d} w_i x_i)$$

In vector form, the perceptron implements

$$h(x) = sign(w^T x)$$



'linearly separable' data

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

A simple Learning Algo - PLA

The perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\intercal}\mathbf{x})$$

Given the training set:

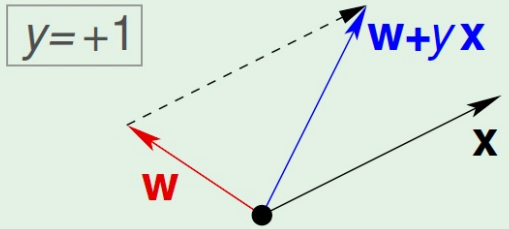
$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

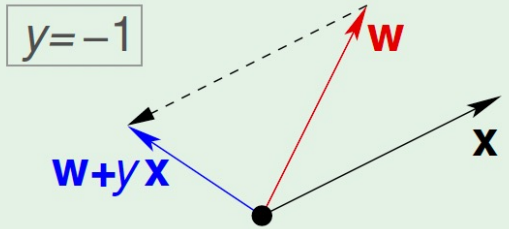
pick a misclassified point:

$$\text{sign}(\mathbf{w}^{\intercal}\mathbf{x}_n) \neq y_n$$

and update the weight vector:

$$\boxed{\mathbf{w} \leftarrow \mathbf{w} + y_n\mathbf{x}_n}$$





Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Iterations of PLA

- One iteration of the PLA:

$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

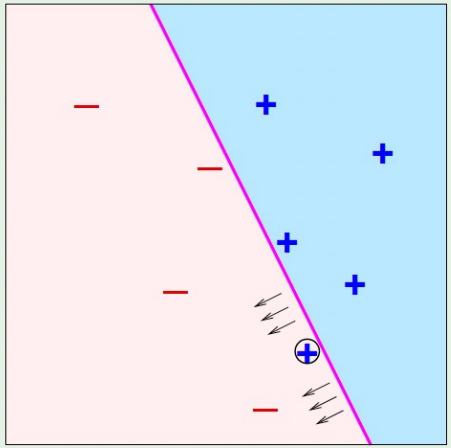
where $(\mathbf{x}, y)$ is a misclassified training point.

- At iteration $t = 1, 2, 3, \cdots$, pick a misclassified point from

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

and run a PLA iteration on it.

- That's it!



Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Broad types of machine learning

- Supervised Learning
  - Training Data with labels: X,y (pre-classified)
  - Given an observation x, what is the best label for y?
- Unsupervised learning
  - Training Data without labels: X
  - Given a set of x's, find hidden structure
- Semi-supervised Learning
  - Training Data + some Labels
- Reinforcement Learning
  - Given: observations and periodic rewards as the agent takes sequential action in an environment
  - Determine optimum policy

Supervised Learning

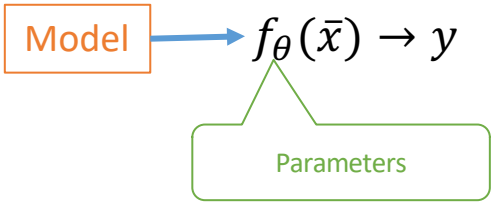
- Given data containing the inputs and outputs:

Training Data:

$\{(\overline{x_1}, y_1), (\overline{x_2}, y_2), \ldots, (\overline{x_m}, y_m)\}$

- Learn a function $f(x)$ to predict y given x

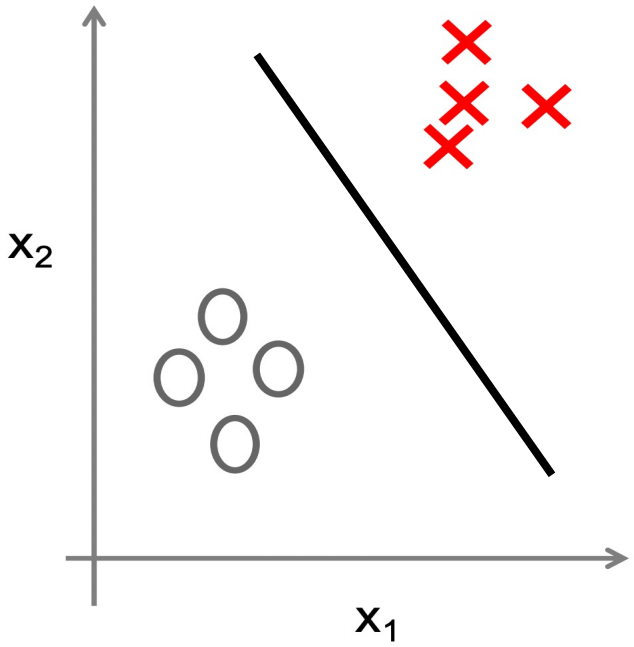
$\bar{X}$	Y
$\overline{x_1}$	y_1
$\overline{x_2}$	y_2
...	..
$\overline{x_m}$	y_m



Training: Learn the model from the Training Data

Given Test instance $\overline{x'}$, predict $y' = f_\theta(\overline{x'})$

Supervised Learning



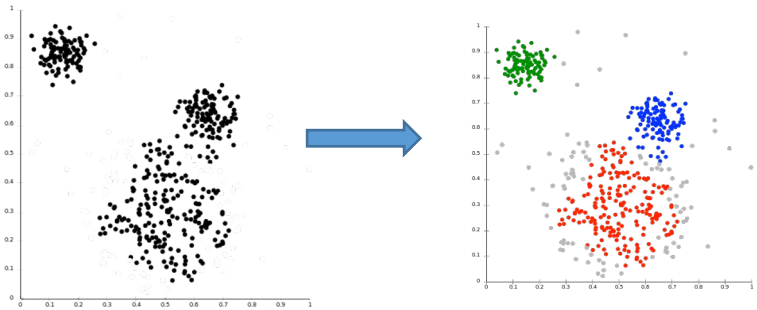
Classification (one example)

- Input: (input-features, correct output)
  - <size, #rooms>, <cheap/costly>
- Output of learning algorithm
  - Function maps features to output
  - F(<size, #rooms>) = cheap/costly

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

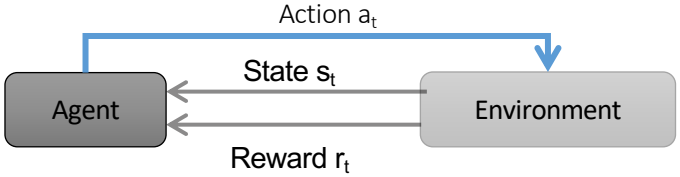
Unsupervised Learning (Clustering)

- Given $\{\overline{x_1}, \overline{x_2}, \ldots \overline{x_m}, \}$ without labels
- Find hidden structure in the data
  - Clustering
  - Dimensionality Reduction
- Clustering: Grouping similar objects



Reinforcement Learning

- Given a sequence of states and actions with (delayed) rewards, output a policy.



  - Receive feedback in the form of rewards
  - Agent's utility is defined by the reward function
  - Must (learn to) act so as to maximize expected rewards

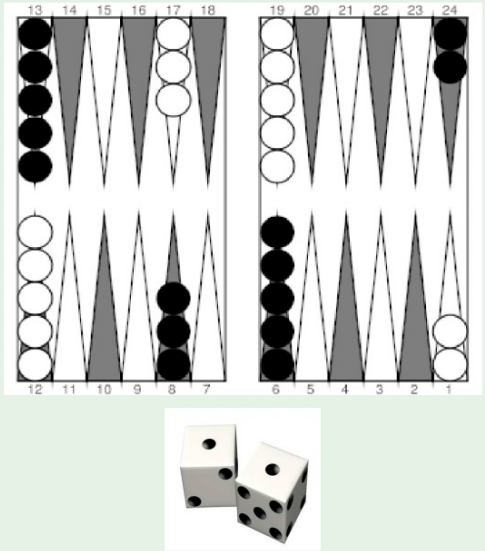
- Examples:
  - Game playing (Go)
  - Robot grasping
  - Controlling aircraft and robotic motion

Goal: Constantly learn to make 'optimal' predictions based on real-time feedback from past predictions

Reinforcement Learning

Instead of (input, correct output),

we get (input, *some* output, grade for this output)



The world champion was a neural network!

Slides from Dr. Yaser S. Abu-Mostafa (CalTech, USA)

Why Machine Learning?

- Human expertise not sacrosanct
  - Pricing, Promotions
- Human expertise cannot be coded
  - face/handwriting/speech recognition
  - driving a car, flying a plane
- Rapidly changing situation
  - Credit scoring, financial modeling
  - Fraud detection
- Need for personalization
  - Product recommendation
- Too much Data