Course Overview and Introduction

CE-717 : Machine Learning

Sharif University of Technology

M. Soleymani

Fall 2016

Course Info

- Instructor: Mahdieh Soleymani
  - Email: soleymani@sharif.edu

- Lectures: Sun-Tue (13:30-15:00)

- Website: http://ce.sharif.edu/cources/95-96/1/ce717-2

Text Books

- Pattern Recognition and Machine Learning, C. Bishop, Springer, 2006.
- Machine Learning, T. Mitchell, MIT Press, 1998.
- Additional readings: will be made available when appropriate.

- Other books:
  - The elements of statistical learning, T. Hastie, R. Tibshirani, J. Friedman, Second Edition, 2008.
  - Machine Learning: A Probabilistic Perspective, K. Murphy, MIT Press, 2012.

Marking Scheme

Midterm Exam:	25%
Final Exam:	30%
Project:	5-10%
Homeworks (written & programming) :	20-25%
Mini-exams:	15%

Machine Learning (ML) and Artificial Intelligence (AI)

- ML appears first as a branch of AI

- ML is now also a preferred approach to other subareas of AI
  - Computer Vision, Speech Recognition, …
  - Robotics
  - Natural Language Processing

- ML is a strong driver in Computer Vision and NLP

A Definition of ML

- Tom Mitchell (1998): Well-posed learning problem
  - "A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E".
- Using the observed data to make better decisions
  - Generalizing from the observed data

ML Definition: Example

- Consider an email program that learns how to filter spam according to emails you do or do not mark as spam.
  - T: Classifying emails as spam or not spam.
  - E: Watching you label emails as spam or not spam.
  - P: The number (or fraction) of emails correctly classified as spam/not spam.

The essence of machine learning

- A pattern exist
- We do not know it mathematically
- We have data on it

Example: Home Price

- Housing price prediction

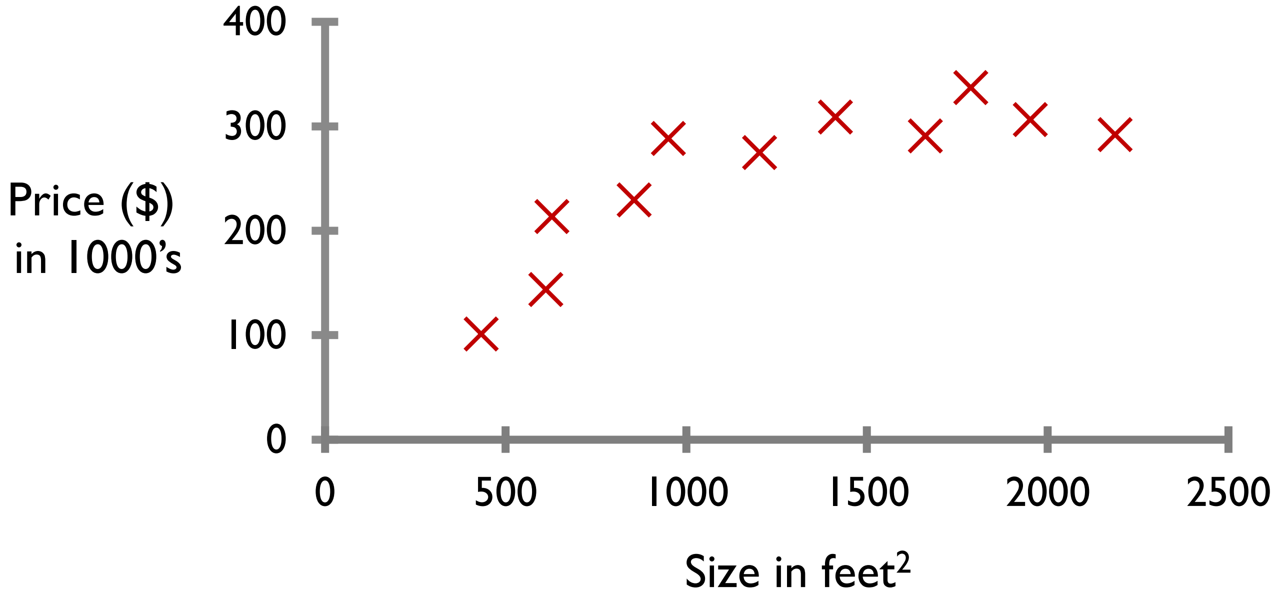


Figure adopted from slides of Andrew Ng,
Machine Learning course, Stanford.

Example: Bank loan

- Applicant form as the input:

age	23 years
gender	male
annual salary	\$30,000
years in residence	1 year
years in job	1 year
current debt	\$15,000
...	...

- Output: approving or denying the request

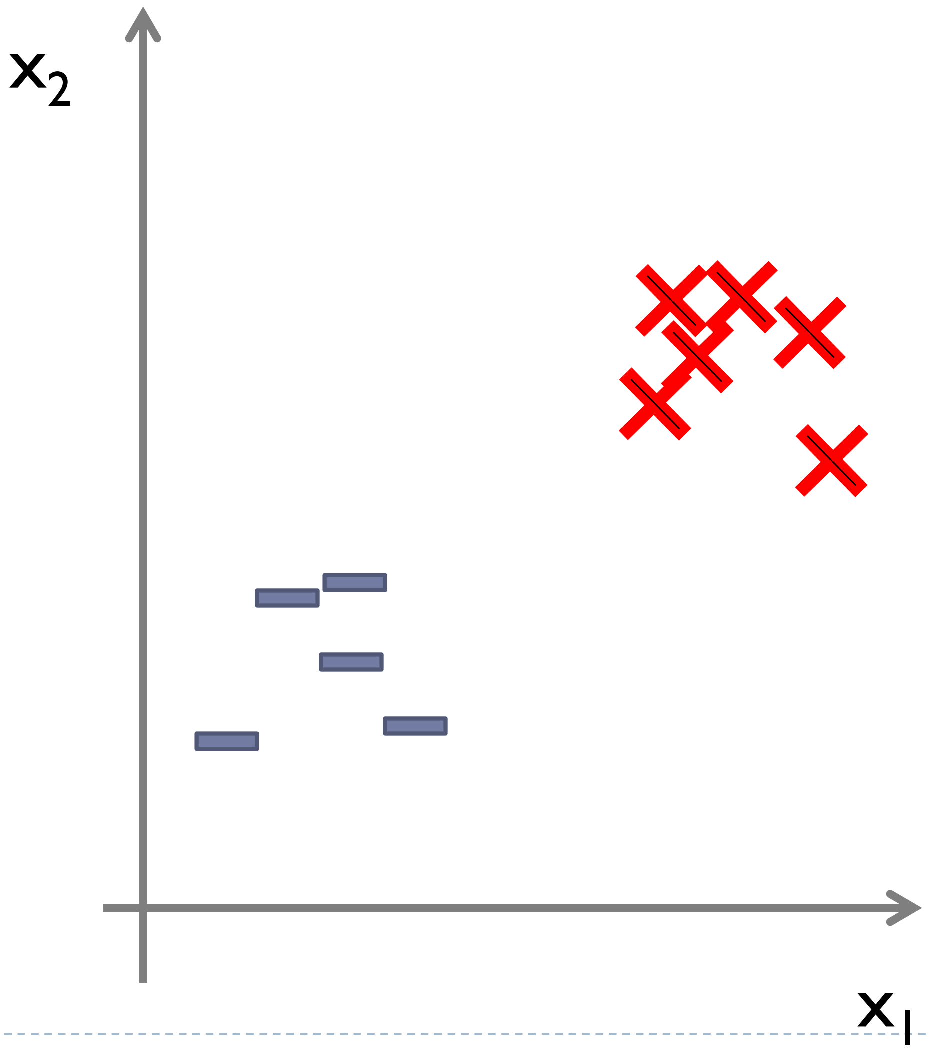
Components of (Supervised) Learning

- Unknown target function: $f: \mathcal{X} \rightarrow \mathcal{Y}$
  - Input space: $\mathcal{X}$
  - Output space: $\mathcal{Y}$

- Training data: $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)$

- Pick a formula $g: \mathcal{X} \rightarrow \mathcal{Y}$ that approximates the target function f
  - selected from a set of hypotheses $\mathcal{H}$

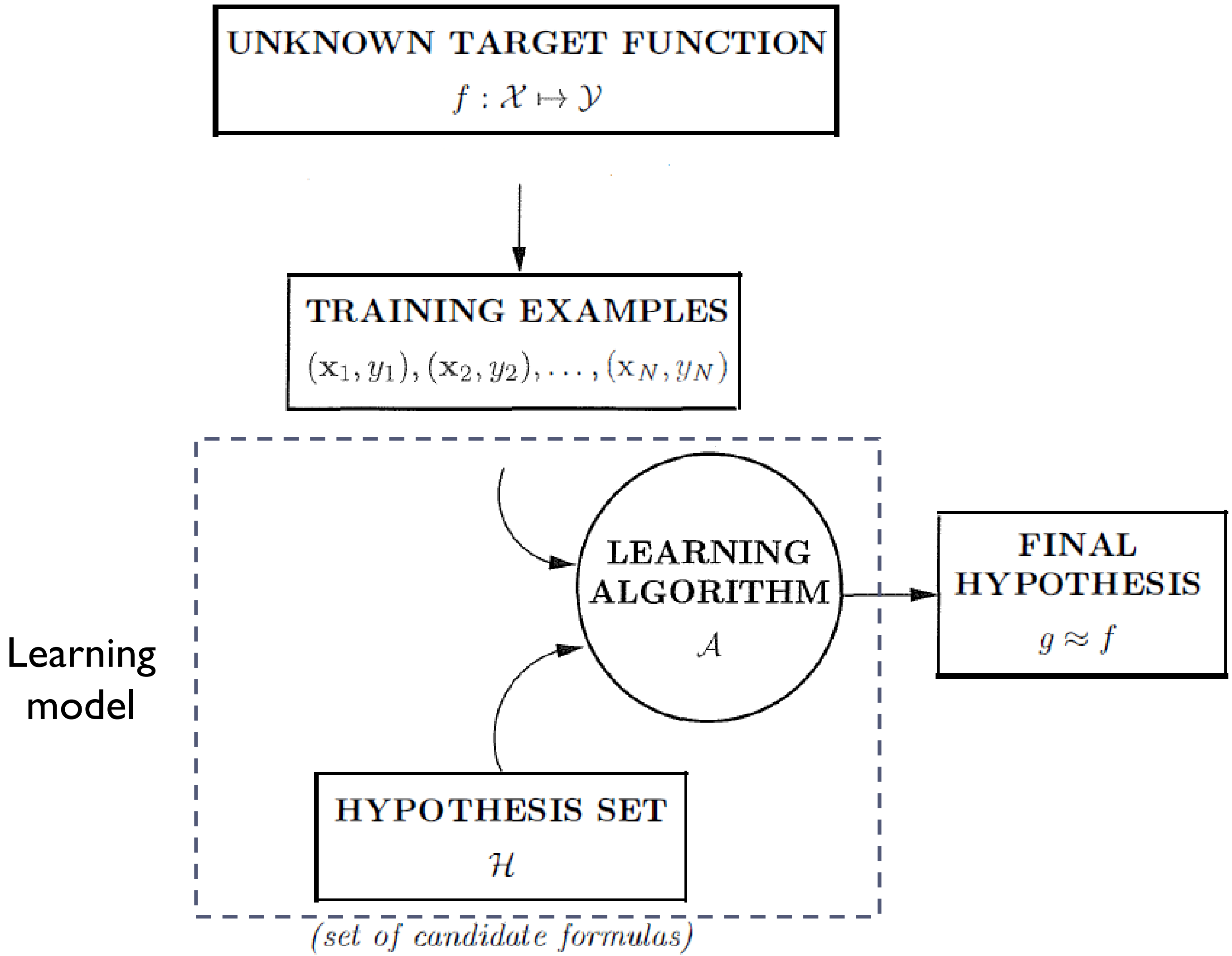
Training data: Example



Training data

x_1	x_2	y
0.9	2.3	1
3.5	2.6	1
2.6	3.3	1
2.7	4.1	1
1.8	3.9	1
6.5	6.8	-1
7.2	7.5	-1
7.9	8.3	-1
6.9	8.3	-1
8.8	7.9	-1
9.1	6.2	-1

Components of (Supervised) Learning



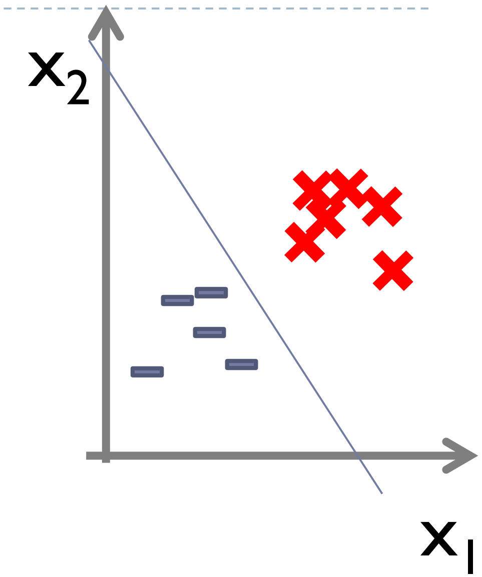
Solution Components

- Learning model composed of:
  - Learning algorithm
  - Hypothesis set

- Perceptron example

Perceptron classifier

- Input $\boldsymbol{x} = [x_1, \ldots, x_d]$
- Classifier:
  - If $\sum_{i=1}^{d} w_i x_i > \text{threshold}$ then output 1
  - else output -1



- The linear formula $g \in \mathcal{H}$ can be written:

$$g(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{d} w_i x_i \quad + \quad w_0 \qquad\right)$$

If we add a coordinate $x_0 = 1$ to the input:

$$g(\boldsymbol{x}) = \text{sign}\left(\sum_{i=0}^{d} w_i x_i\right)$$

Vector form

$$g(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x})$$

Perceptron learning algorithm: linearly separable data

- Give the training data $\left(\boldsymbol{x}^{(1)}, y^{(1)}\right), \ldots, \left(\boldsymbol{x}^{(N)}, y^{(N)}\right)$

- Misclassified data $\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)$:

$$\operatorname{sign}(\boldsymbol{w}^T \boldsymbol{x}^{(n)}) \neq y^{(n)}$$

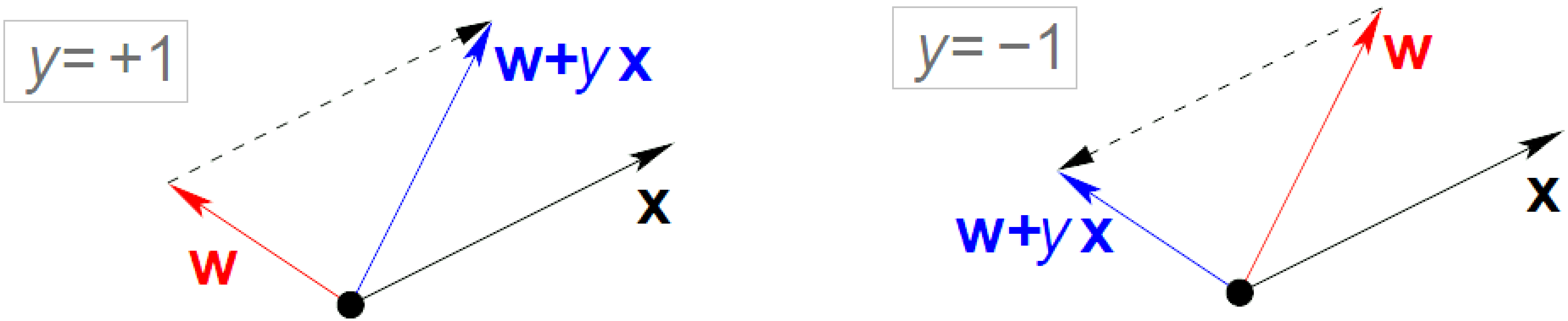
Repeat

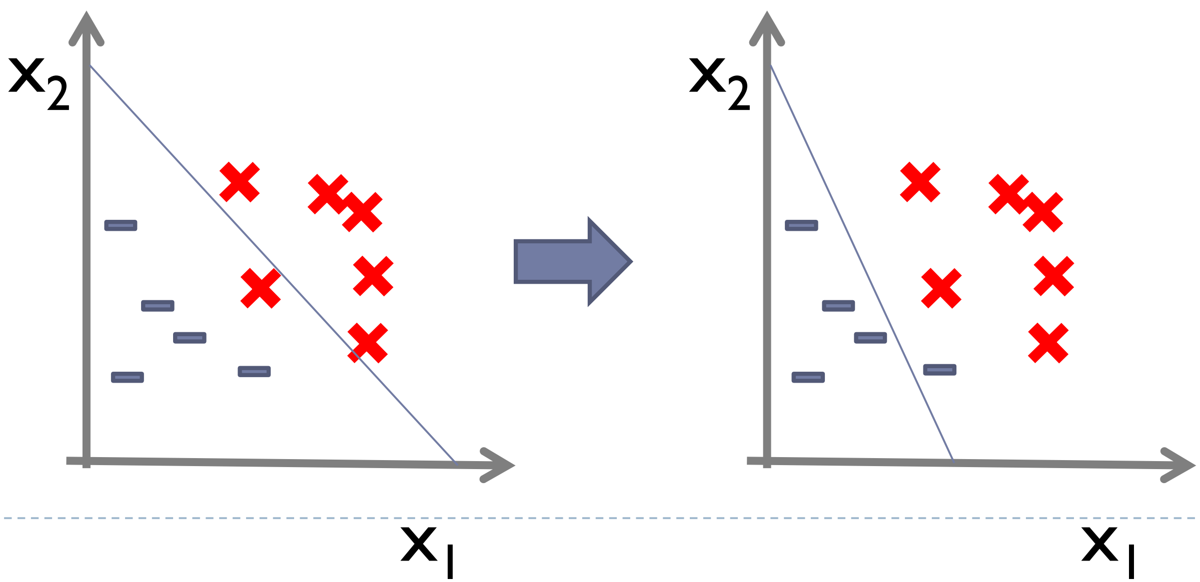
Pick a misclassified data $\left(\boldsymbol{x}^{(n)}, y^{(n)}\right)$ from training data and update $\boldsymbol{w}$:

$$\boldsymbol{w} = \boldsymbol{w} + y^{(n)} \boldsymbol{x}^{(n)}$$

Until all training data points are correctly classified by g

Perceptron learning algorithm: Example of weight update





Experience (E) in ML

- Basic premise of learning:
  - "Using a set of observations to uncover an underlying process"

- We have different types of (getting) observations in different types or paradigms of ML methods

Paradigms of ML

- Supervised learning (regression, classification)
  - predicting a target variable for which we get to see examples.
- Unsupervised learning
  - revealing structure in the observed data
- Reinforcement learning
  - partial (indirect) feedback, no explicit guidance
  - Given rewards for a sequence of moves to learn a policy and utility functions
- Other paradigms: semi-supervised learning, active learning, online learning, etc.

Supervised Learning: Regression vs. Classification

- Supervised Learning
  - **Regression**: predict a <u>continuous</u> target variable
    - E.g., $y \in [0,1]$
  - **Classification**: predict a <u>discrete</u> target variable
    - E.g.,$y \in \{1,2, \dots, C\}$

Data in Supervised Learning

- Data are usually considered as vectors in a d dimensional space
  - Now, we make this assumption for illustrative purpose
    - We will see it is not necessary

Columns:
Features/attributes/dimensions

Rows:
Data/points/instances/examples/samples

Y column:
Target/outcome/response/label

	x_1	x_2	…	x_d	y (Target)
Sample1					
Sample 2					
…					
Sample n-1					
Sample n					

Regression: Example

- Housing price prediction

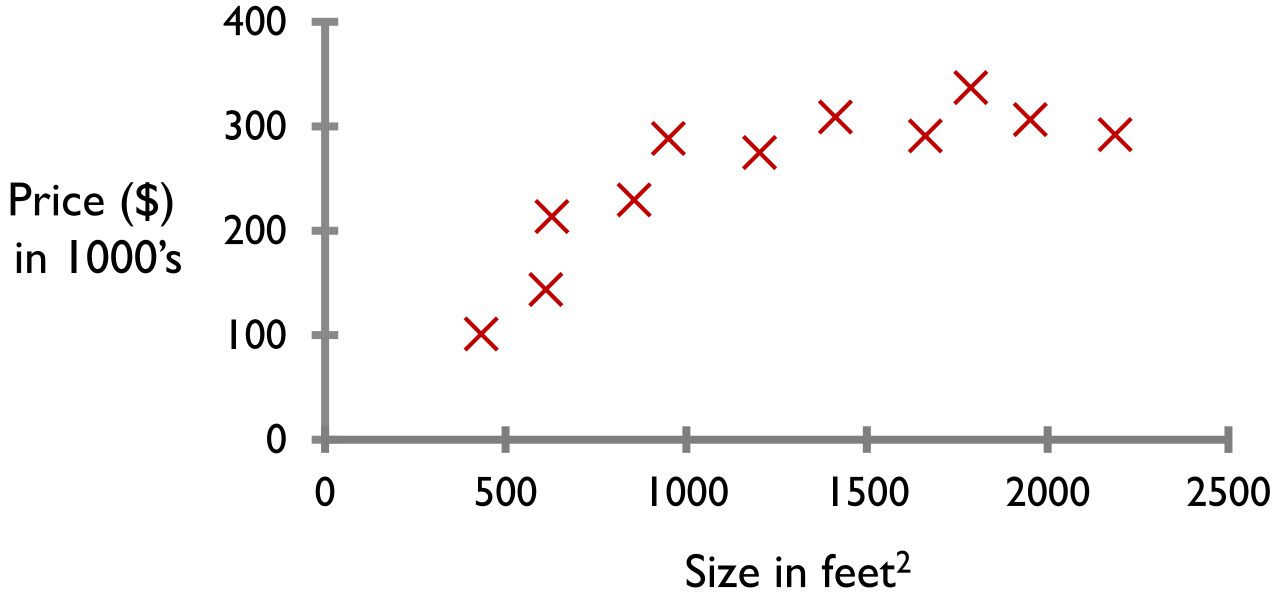
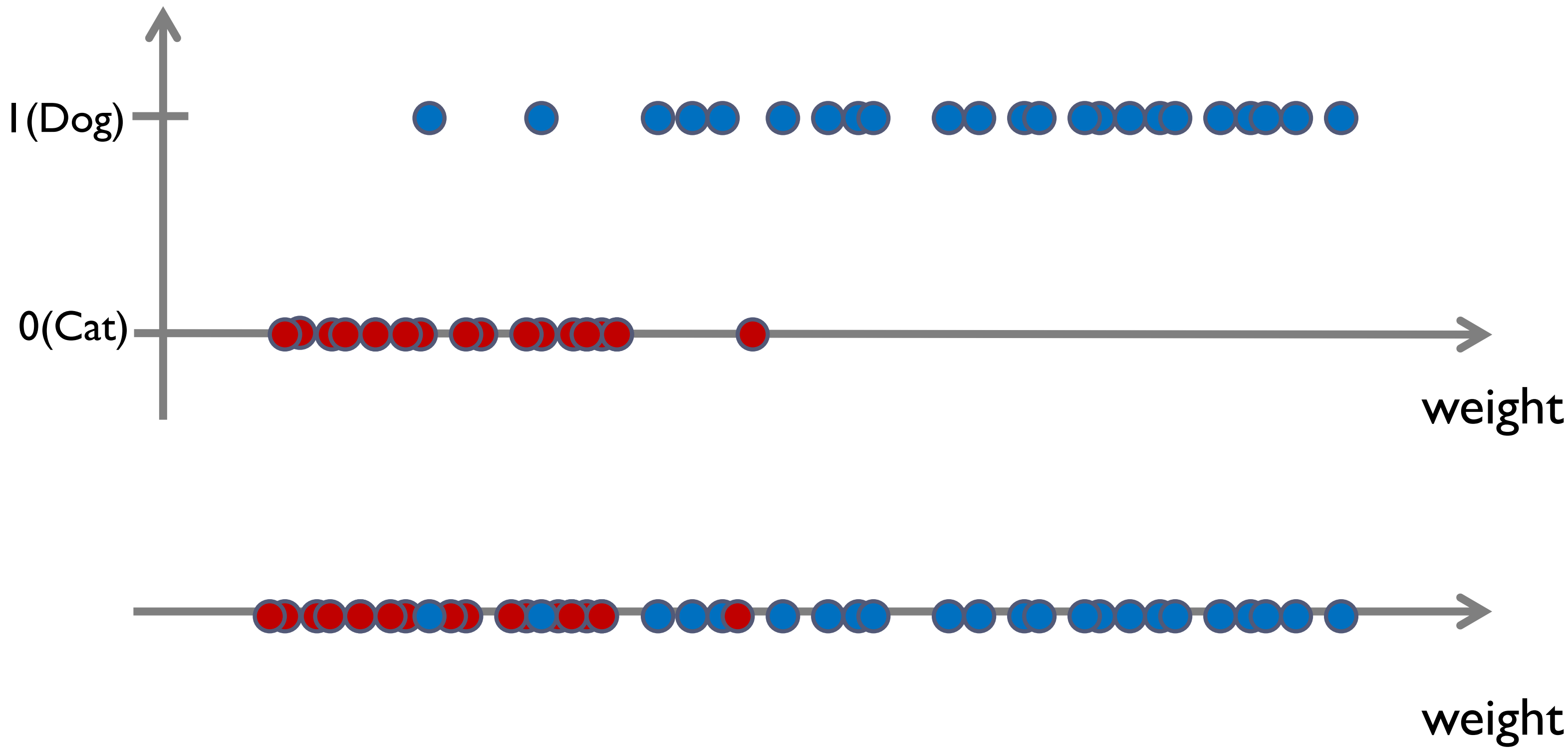


Figure adopted from slides of Andrew Ng

Classification: Example

- Weight (Cat, Dog)

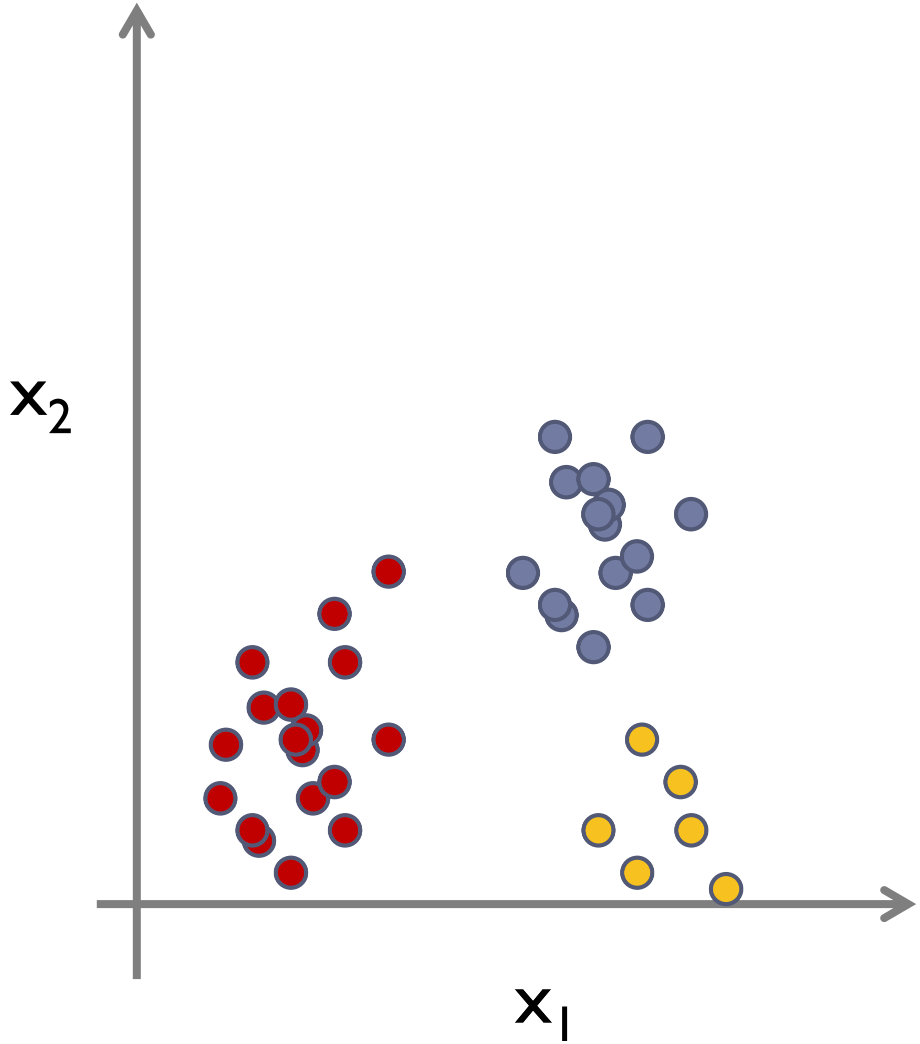


Supervised Learning vs. Unsupervised Learning

- Supervised learning
  - Given: Training set
    - labeled set of N input-output pairs $D = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$
  - Goal: learning a mapping from $\boldsymbol{x}$ to y

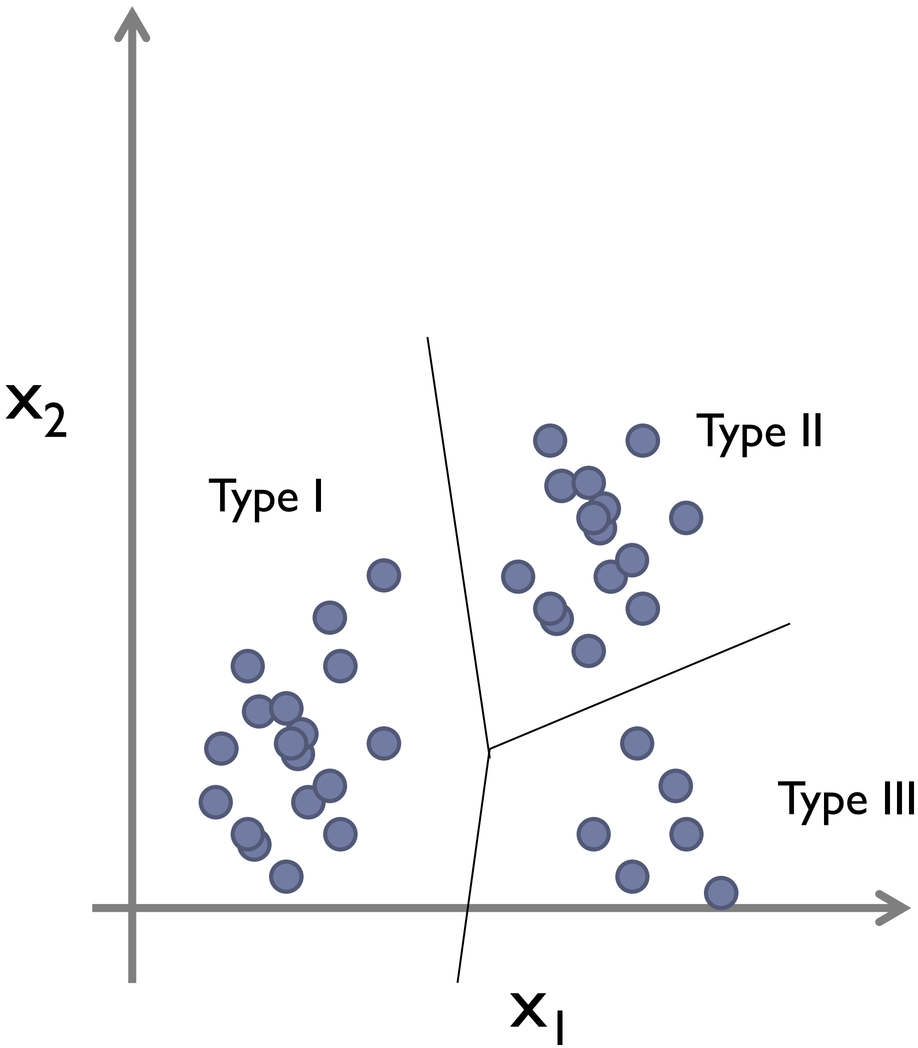
- Unsupervised learning
  - Given: Training set
    - $\{\boldsymbol{x}^{(i)}\}_{i=1}^{N}$
  - Goal: find groups or structures in the data
    - Discover the intrinsic structure in the data

Supervised Learning: Samples



Classification

Unsupervised Learning: Samples



Clustering

Sample Data in Unsupervised Learning

- Unsupervised Learning:

Columns:
Features/attributes/dimensions

Rows:
Data/points/instances/examples/samples

	x_1	x_2	…	x_d
Sample1				
Sample 2				
…				
Sample n-1				
Sample n				

Unsupervised Learning: Example Applications

- Clustering docs based on their similarities
  - Grouping new stories in the Google news site
- Market segmentation: group customers into different market segments given a database of customer data.
- Social network analysis

Reinforcement

- Provides only an indication as to whether an action is correct or not

Data in supervised learning:

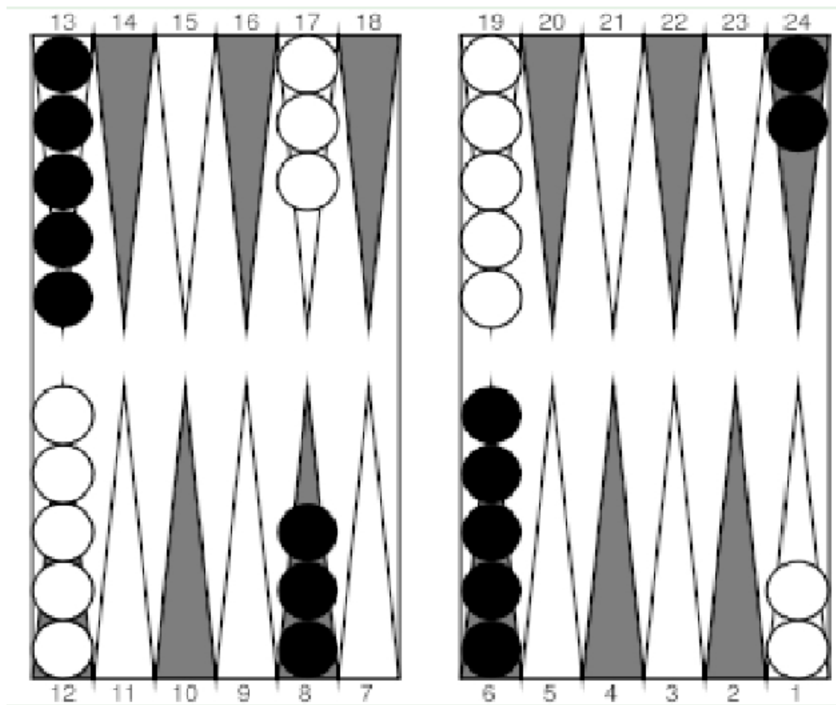
(input, correct output)

Data in Reinforcement Learning:

(input, some output, a grade of reward for this output)

Reinforcement Learning

- Typically, we need to get a sequence of decisions
  - it is usually assumed that reward signals refer to the entire sequence

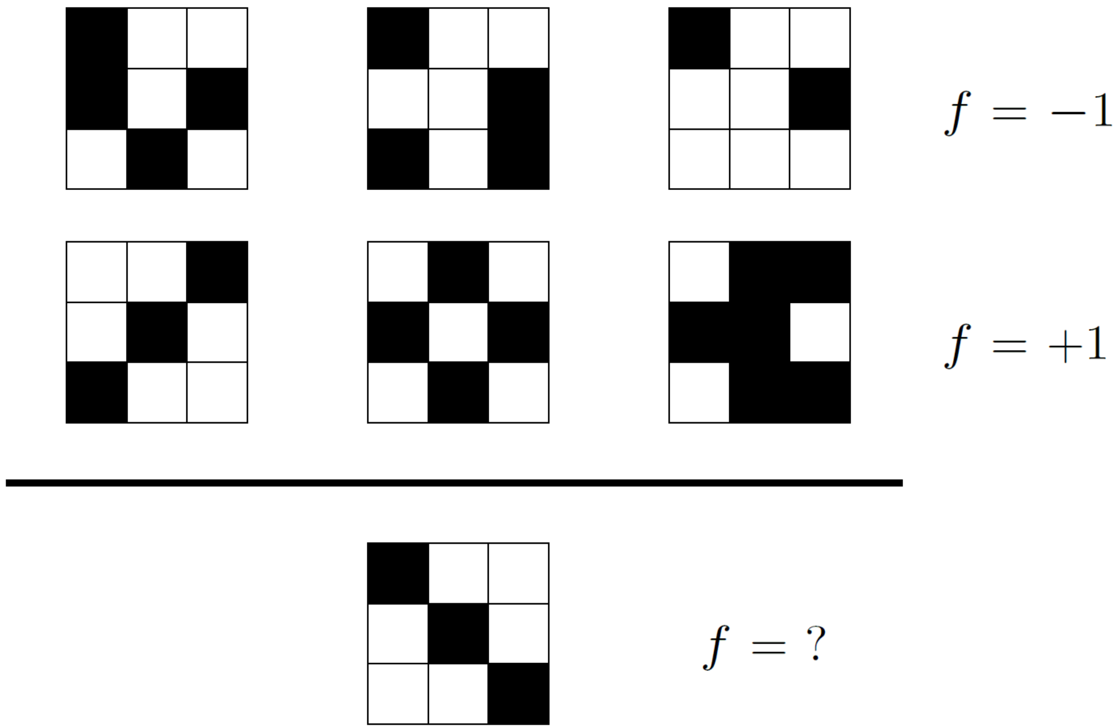


Is learning feasible?

- Learning an unknown function is impossible.
  - The function can assume any value outside the data we have.

- However, it is feasible in a probabilistic sense.

Example

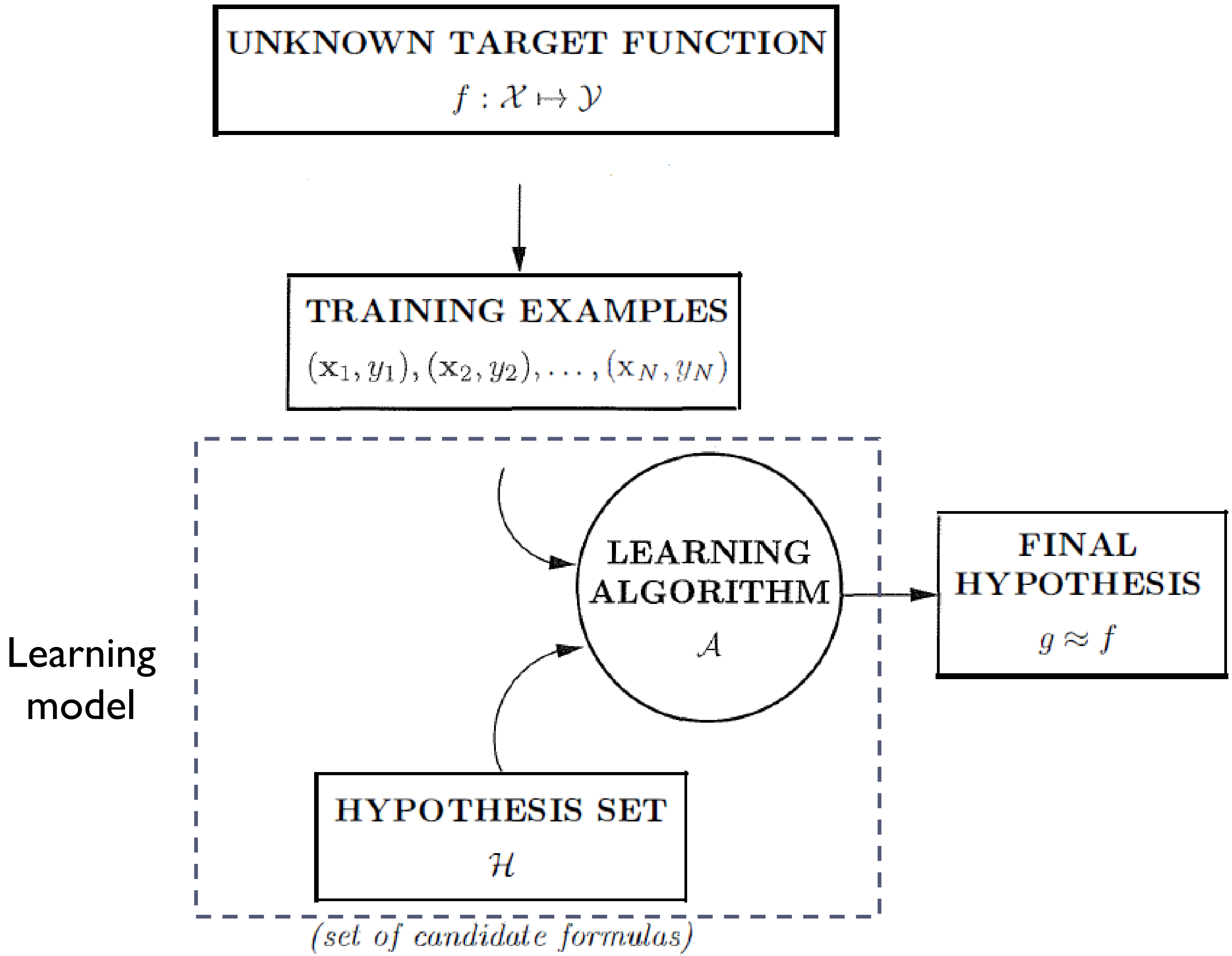


Generalization

- We don't intend to memorize data but need to figure out the pattern.

- A core objective of learning is to generalize from the experience.
  - Generalization: ability of a learning algorithm to perform accurately on new, unseen examples after having experienced.

Components of (Supervised) Learning



Main Steps of Learning Tasks

- Selection of hypothesis set (or model specification)
  - Which class of models (mappings) should we use for our data?

- Learning: find mapping $\hat{f}$ (from hypothesis set) based on the training data
  - Which notion of error should we use? (loss functions)
  - Optimization of loss function to find mapping $\hat{f}$

- Evaluation: how well $\hat{f}$ generalizes to yet unseen examples
  - How do we ensure that the error on future data is minimized? (generalization)

Some Learning Applications

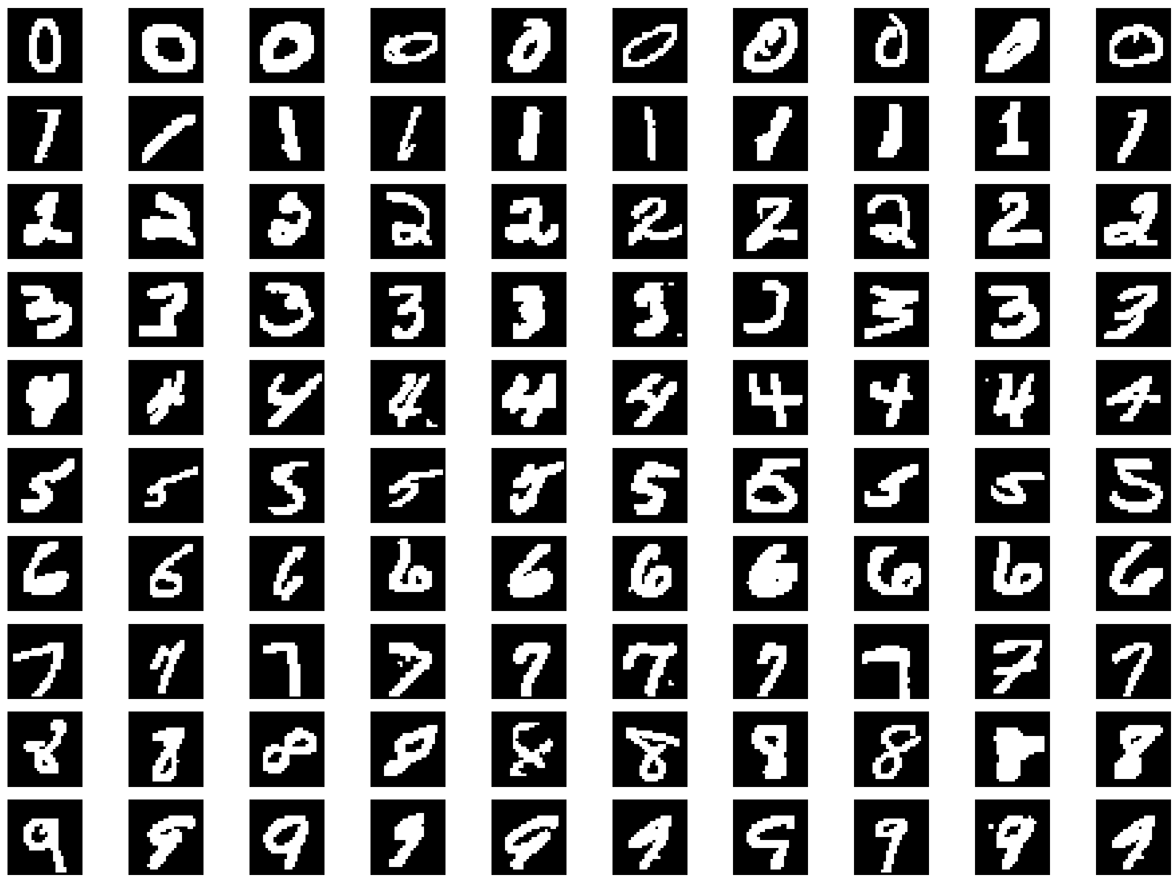
- Face, speech, handwritten character recognition
- Document classification and ranking in web search engines
- Photo tagging
- Self-customizing programs (recommender systems)
- Database mining (e.g., medical records)
- Market prediction (e.g., stock/house prices)
- Computational biology (e.g., annotation of biological sequences)
- Autonomous vehicles

ML in Computer Science

- Why ML applications are growing?
  - Improved machine learning algorithms
  - Availability of data (Increased data capture, networking, etc)
  - Demand for self-customization to user or environment
  - Software too complex to write by hand

Handwritten Digit Recognition Example

- Data: labeled samples

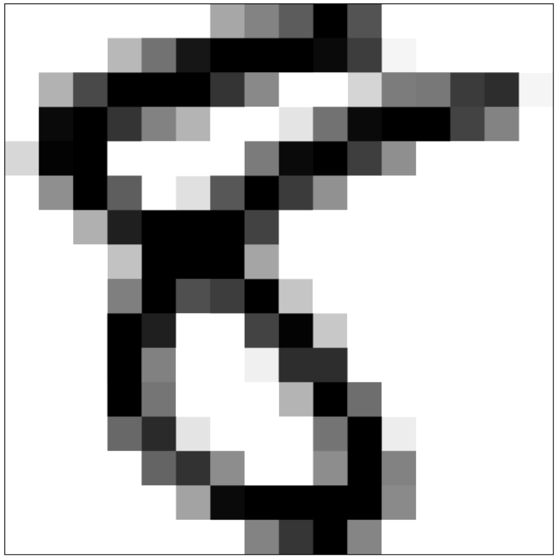


0
1
2
3
4
5
6
7
8
9

Example: Input representation

'raw' input $\mathbf{x} = (x_0, x_1, x_2, \cdots, x_{256})$

linear model: $(w_0, w_1, w_2, \cdots, w_{256})$



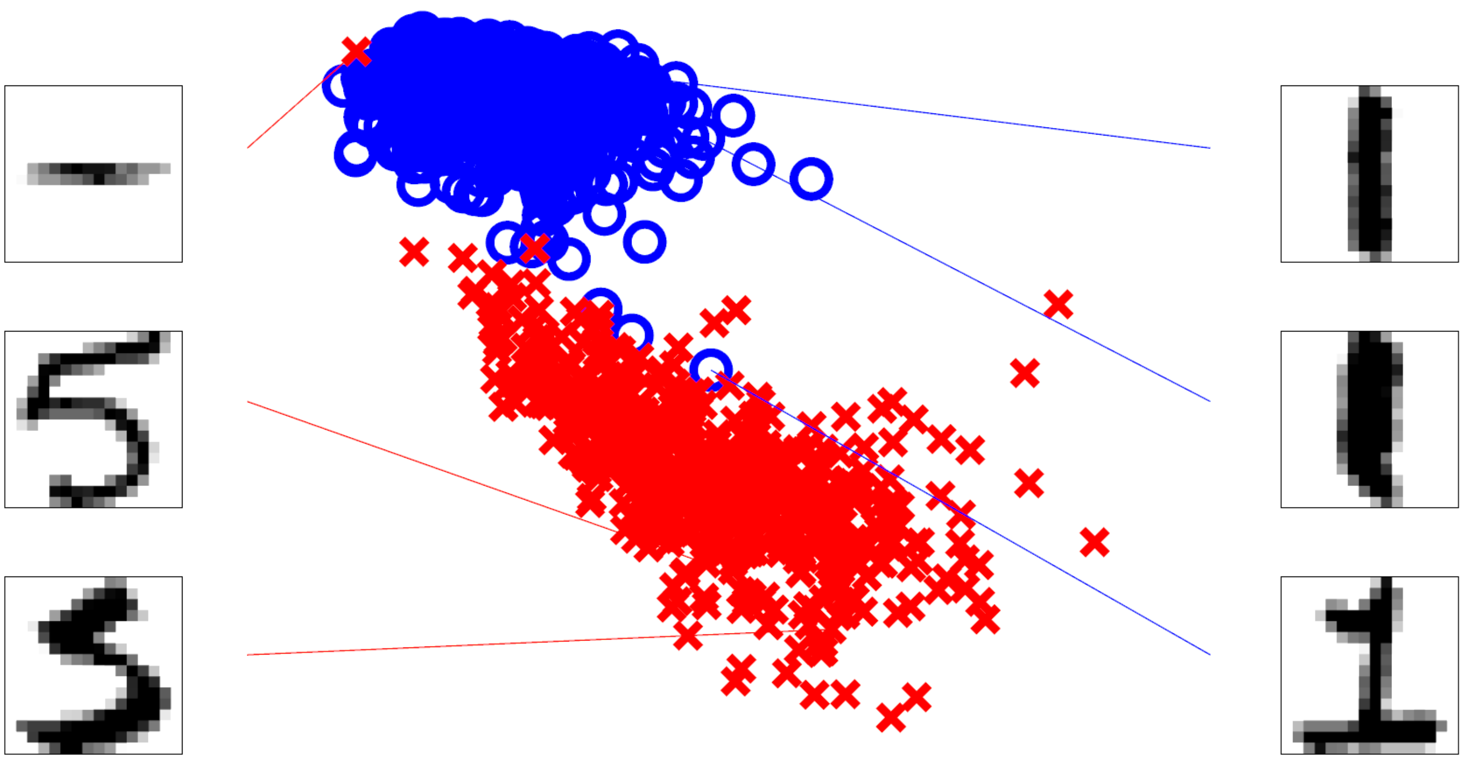
Features: Extract useful information, e.g.,

intensity and symmetry $\mathbf{x} = (x_0, x_1, x_2)$

linear model: (w_0, w_1, w_2)

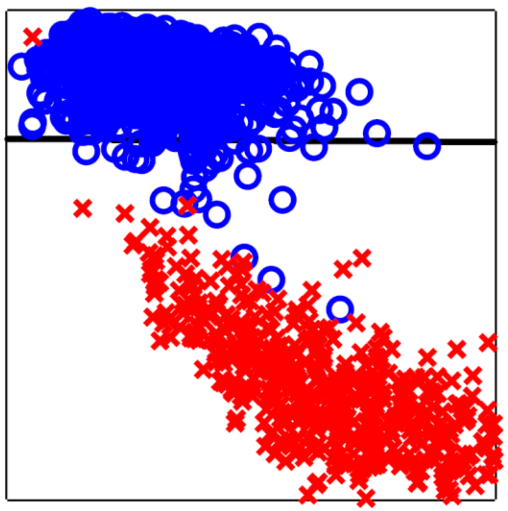
Example: Illustration of features

$\mathbf{x} = (x_0, x_1, x_2)$ $\quad x_1$: intensity $\quad x_2$: symmetry

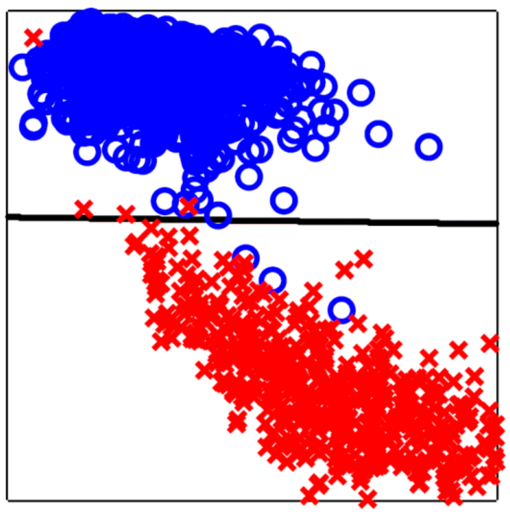


Example: Classification boundary

PLA:



Pocket:



Main Topics of the Course

- Supervised learning
  - Regression
  - Classification (our main focus)
- Learning theory
- Unsupervised learning
- Reinforcement learning
- Some advanced topics & applications

Most of the lectures are on this topic

Resource

- Yaser S. Abu-Mostafa, Malik Maghdon-Ismail, and Hsuan Tien Lin, "**Learning from Data**", 2012.