

# Learning From Data

## Lecture 1: Introduction

Yang Li   yangli@sz.tsinghua.edu.cn

TBSI

September 17, 2022

# Today's Lecture

- ▶ About This Class
- ▶ What is Machine Learning?
- ▶ Course Preview: a Brief History of Machine Learning

# About this Class

## Course Goal

- ▶ In-depth understanding of key concepts, algorithms for machine learning.
- ▶ Practical applications of learning from data.

# Course Material

The primary course materials are the lecture slides.

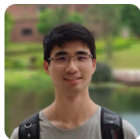
Reference Text :

- ▶ (Recommended) Machine Learning Lecture Notes by Andrew Ng: <https://github.com/mxc19912008/Andrew-Ng-Machine-Learning-Notes>
- ▶ Pattern Recognition and Machine Learning, 2nd Edition, by Christopher Bishop

# Staffs



**Yang Li**  
Instructor



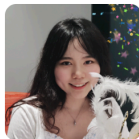
**Weida Wang**  
TA



**Dexu Kong**  
TA



**Zhiyuan Peng**  
TA



**Wanda Li**  
TA

## Office hours

Name	Time	Location
Yang	Friday 2:00-4:00pm	Info Building 1108a
Zhiyuan	Tuesday 3:00-5:00pm	Info Building, 11th floor common area)
Weida	Wednesday 4:00-6:00pm	(same as above)
Dexu	Thursday 7:00-9:00pm	(same as above)

# Grading

Your overall grade will be determined roughly as follows:

ACTIVITIES	PERCENTAGES
Midterm	15 %
Final Project	25 %
Problem sets (written & programming)	60 %

## Homework advice

- ▶ Form study groups (2-3 people) to discuss homework problems. Do homework **independently**, indicate your study group members on your submitted file.
- ▶ Use "Online Learning" Q&A discussion board!
- ▶ Come to office hours
- ▶ Attend recitations

# Class Policy

## Late homeworks

- ▶ **2 free chances** to turn in a late homework assignment (except for the final project).
- ▶ Late homework must be handed in within 3 days of the deadline.

# Class Policy

## How to give credits

- ▶ Write your collaborators' names in the homework (*this includes receiving/giving explicit help from/to others on any part of the homework*)
- ▶ Note any online resource (e.g. wiki, github, stackoverflow) you've used for the assignment

**Homework plagiarism (copying) is not tolerated!**

Ask for help early and often!



# Final Group Project

Apply recent machine learning techniques on real-world problems, or explore theoretical problems related to learning from data.

## Previous class projects

- ▶ Camera lens super-resolution (Dinjian Jin & Xiangyu Chen)



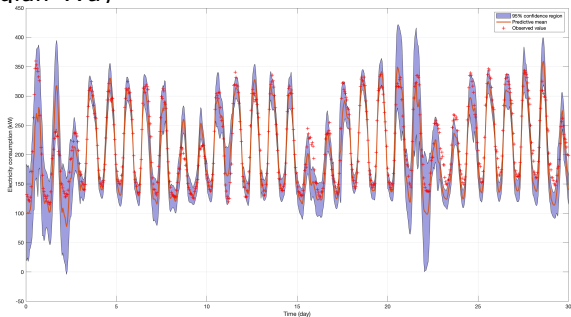
Comparison between two super-resolution models: SRGAN and VDSR

# Final Group Project

Apply recent machine learning techniques on real-world problems, or explore theoretical problems related to learning from data.

## Previous class projects

- ▶ A Gaussian Process Regression Based Approach for Predicting Building Cooling and Heating Consumption (Xiaoting Wang & Yiqian Wu)



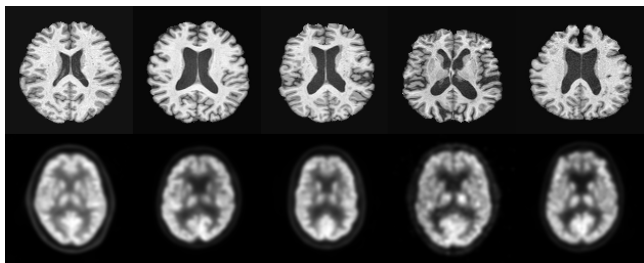
1-month prediction of electricity consumption

# Final Group Project

Apply recent machine learning techniques on real-world problems, or explore theoretical problems related to learning from data.

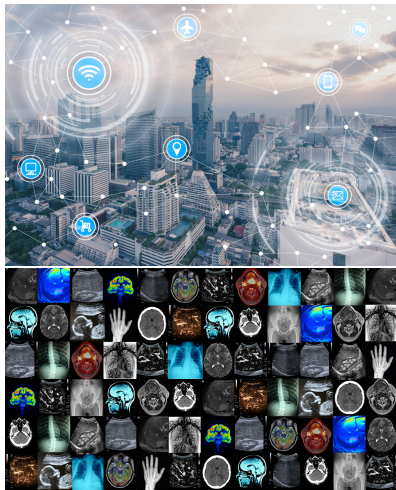
## Previous class projects

- ▶ Missing Data Imputation for Multi-Modal Brain Images (Wangbin Sun)



MRI (top) and PET (bottom) scans of normal and Alzheimer patient brains

# The age of big data



How does a computer program learn “knowledge” from data ? *i.e.*  
*machine learning*



## What is Machine Learning?

Design programs that can ...



## What is Machine Learning?

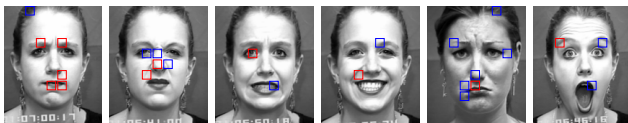
Design programs that can

- ▶ learn rules from data for some **task**
- ▶ adapt to changes
- ▶ improve **performance** with **experience**.

(from "Machine Learning Theory" by Avrim Blum )

# Machine Learning Tasks

## ► Classification



(a) Ang (b) Dis (c) Fea (d) Hap (e) Sad (f) Sur

Facial expression recognition (Liu et al. CVPR 2014)

“The voice quality of this phone is amazing.” (Positive)

“The earphone broke in two days.” (Negative)

Product review sentiment classification

# Machine Learning Tasks

## ► Regression



Highway travel time prediction

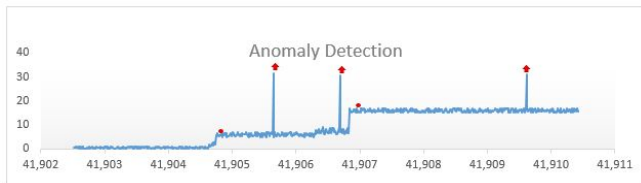


Algorithmic trading: forecast close price, highs and lows



# Machine Learning Tasks

- ▶ Data denoising
- ▶ Pattern recognition (e.g. spam filter)
- ▶ Anomaly detection: finding abnormal operational activity for network security.



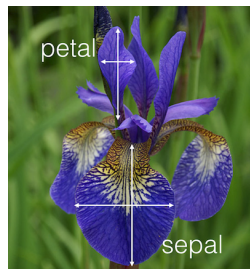
*Can you name some other tasks?*

# Machine Learning Experience

- ▶ **Dataset:** a collection of input,  $X = \{x^{(1)}, \dots, x^{(m)}\}$  and optionally, the corresponding output (**labels**)  $Y = \{y^{(1)}, \dots, y^{(m)}\}$
- ▶ Each input (data point)  $x^{(i)}$  is represented by  $n$  **features**

Example: features of an iris flower

sepal length	sepal width	petal length	petal width	species
5.1	3.5	1.4	0.2	Setosa
4.9	3.0	1.4	0.2	Setosa
6.4	3.5	4.5	1.2	Versicolor
5.9	3.0	5.0	1.8	Virginica
⋮	⋮	⋮	⋮	⋮



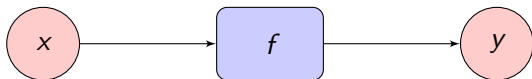
# Machine Learning Performance

- ▶ Quantitatively evaluate the ability of a machine learning algorithm for a given task, e.g.
  - ▶ Mean square error (MSE):  $\frac{1}{m} \sum_{i=1}^m (y^{(i)} - f(x^{(i)}))^2$
  - ▶ Mean absolute error (MAE):  $\frac{1}{m} \sum_{i=1}^m \mathbf{1}\{y^{(i)} \neq f(x^{(i)})\}$
- ▶ Must perform well on new, previously unseen input!
  - ▶ Separate **test dataset** from training data

# Different Types of Learning

## Supervised learning

Given some input and output (label) training data, learn the **machine**  $f$  from training data



Supervised learning tasks:

- ▶ Classification:  $y$  is discrete
- ▶ Regression:  $y$  is continuous (predict stock market closing price, image captioning, automated video transcription)

# Different Types of Learning

## Unsupervised learning

No labels are given in prior, find hidden structure or pattern from the data



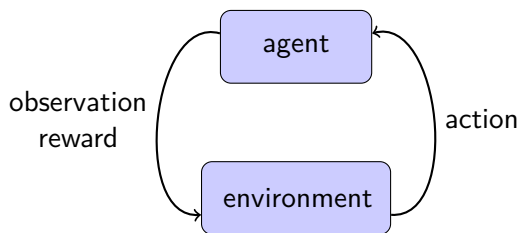
Unsupervised learning tasks:

- ▶ Data clustering
- ▶ Anomaly detection

# Different Types of Learning

## Reinforcement learning

The learning machine is presented in an interactive manner to a dynamic environment, and need to make **sequential decisions**



- ▶ Robotics (self-driving car)
- ▶ AI for sequential decision making (AlphaGo)
- ▶ Intelligent control system

# Inference vs Prediction

Given training data of  $x$  and  $y$ ,

## Inference

knowing the structure of  $f$ , find good models to describe  $f$ . i.e. model the data generation process  $\leftarrow$  *focus of statistics*

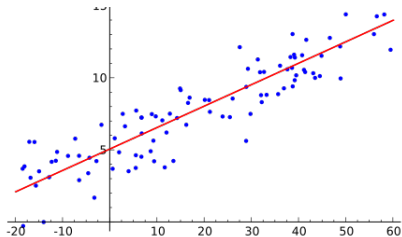
## Prediction

given **future** data samples of  $x$ , predict the corresponding output data  $y$  using  $f$ .  $\leftarrow$  *focus of machine learning*

## Development of Statistical Methods (<1950)

- ▶ (1805): Adrien-Marie Legendre proposed the **least squares** method for data fitting. (**e.g. linear regression**)

$$f(x) = b + w_1x_1 + w_2x_2 = w^T x + b$$



Learn model  $f$  by minimizing the **loss function** (MSE):

$$J(w, b) = \frac{1}{2} \sum_{i=1}^m (f(x^{(i)}) - y^{(i)})^2$$

Can be generalize to nonlinear least squares



## Development of Statistical Methods (<1950)

- ▶ (1812): Pierre-Simon Laplace defined **Bayes Theorem**, based on earlier works of Thomas Bayes.

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

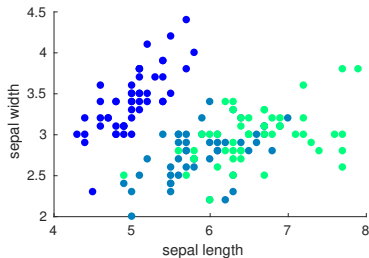
The foundation of **Bayesian estimation**, a core approach in estimating model parameters from data.

# Development of Statistical Methods (<1950)

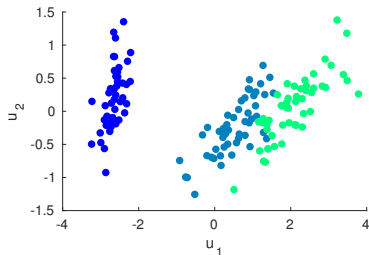
- ▶ (1901): Karl Pearson invented **principal component analysis** (PCA), a classic tool in exploratory data analysis and dimension reduction.

## PCA

Convert observations of possibly correlated variables into a set of *linearly uncorrelated variables* called **principal components**.



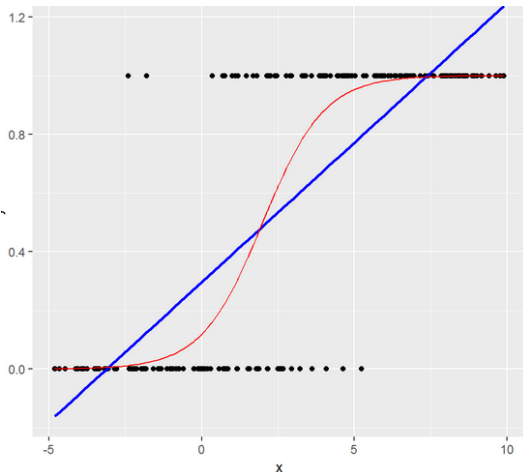
original



PCA transformed

## Development of Statistical Methods (<1950)

- ▶ (1935): Ronald A. Fisher fit the **Probit** model using maximal likelihood estimation for binary classification problem (a.k.a. **Logistic Regression** )



Regression model

— linear

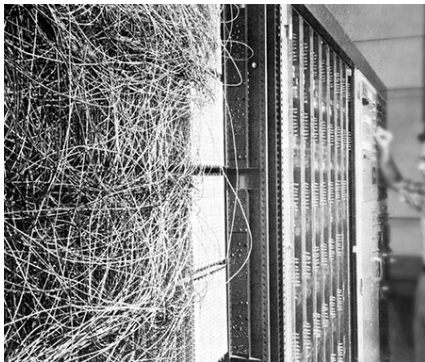
$$f(x) = w^T x + b$$

— logistic

$$f(x) = \frac{1}{1 + e^{-z(w^T x + b)}}$$

# Simple Learning Algorithms (1950)

- ▶ (1957): Frank Rosenblatt invented the **Perceptron** algorithm, the first artificial neural network

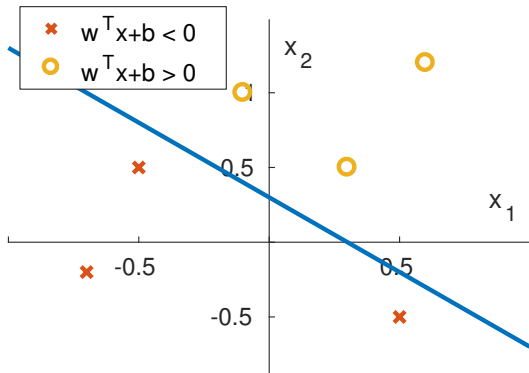


Hardware implementation: Mark I Perceptron

# The perceptron learning algorithm

Given  $x$ , predict  $y \in \{0, 1\}$

$$f(x) = \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

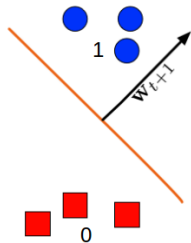
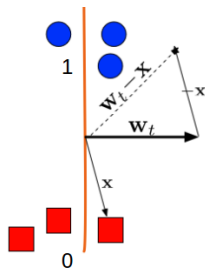
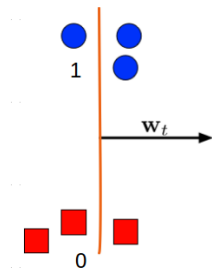


# The perceptron learning algorithm

## Training a perceptron

For each  $x$ , compare  $y$  and the prediction  $f(x)$

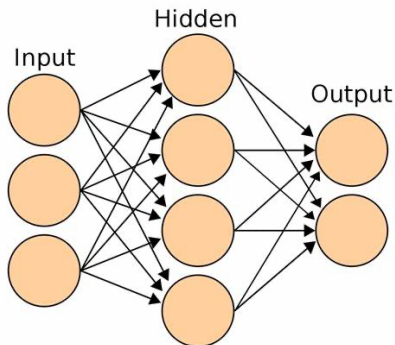
- ▶ When prediction is correct:  $w_{t+1} = w_t$
- ▶ When prediction is incorrect:
  - ▶ predicted "1":  $w_{t+1} := w_t - \alpha x$
  - ▶ predicted "0":  $w_{t+1} := w_t + \alpha x$



## Simple Learning Algorithms (1960s)

- ▶ Rise of **Connectionism**: an approach to explain mental phenomena using artificial neural networks (ANN)

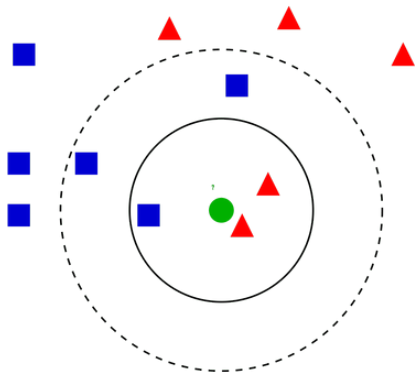
Learning always involves modifying the connection weights



ANN with a hidden layer

## Simple Learning Algorithms (1960s)

- ▶ (1967): Cover and Hart invented **Nearest Neighbor Classification** and the start of Pattern Recognition *One of the first non-parametric learning algorithms*

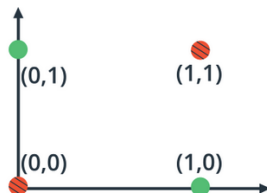


When  $k=3$ , target is classified as 1; When  $k=5$ ,



## The “AI Winter”(1970s)

- ▶ (1969): Minsky and Papert’s 1969 book *Perceptrons* presented limitations to what perceptrons could do
  - ▶ Single-layer network can not solve the XOR problem
  - ▶ Difficult to update weights in neural networks with multiple hidden layers



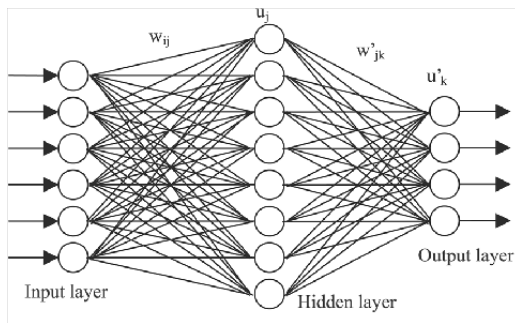
The XOR problem

Virtually no research at all was done in connectionism for 10 years

## Rediscovery of Backpropagation (1980s)

- ▶ (1976) David Rumelhart, Geoff Hinton and Ronald J. Williams rediscovered of **Backpropagation** (first proposed by Linnainmaa in 1970) *an efficient way to calculate the derivative of the loss function with respect to the weights of the network*

Allows efficient training of **multi-layer perceptrons**.



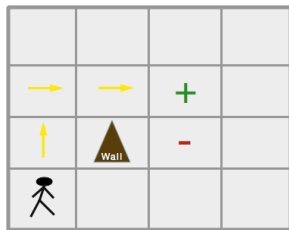
Many hidden units increase expressiveness of ANNs

# Rediscovery of Backpropagation (1980s)

- ▶ (1989) Christopher Watkins proposed **Q-learning**, foundation of modern **Reinforcement Learning**

## Q-learning

Given any **Markov decision process**, learn a policy, which tells an agent what action to take under what circumstances (states).



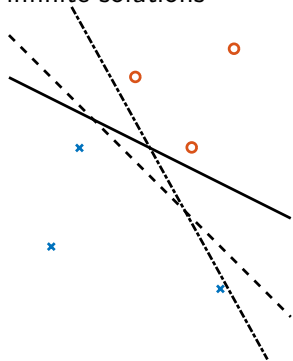
States set: {free, wall, goal, }

Action set: {Left, Right, Top, Down}

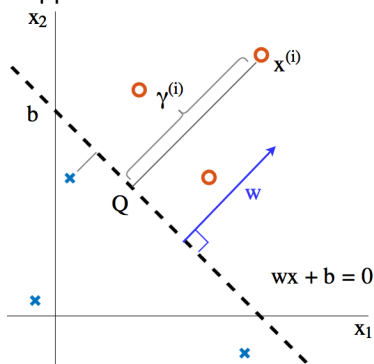
# Rise of Data Driven Methods (1990s)

- ▶ (1992): Corinna Cortes and Vladimir Vapnik discovered **Support Vector Machine**

Single-layer perceptron may have infinite solutions



Support Vector Classifier

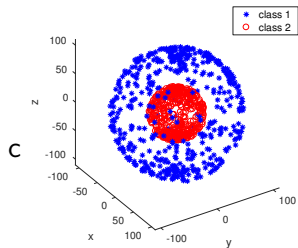


*Give accuracy comparable to neural networks with elaborated features in a handwriting task*

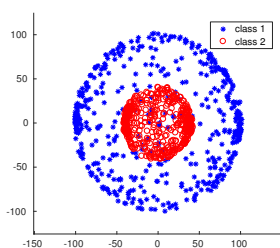
# Kernel Methods (2000s)

**Kernel method:** learn feature representations of data from pairwise similarity, defined by some (family of) kernel functions

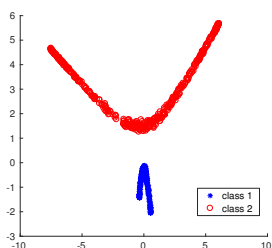
- ▶ (1998) **Kernel principal component analysis** (kernel PCA) was proposed by Schölkopf
- ▶ (2010) Radio Basis Function (RBF) kernel for SVM proposed by Yin-Wen Chang et. al.



original data



linear PCA



Gaussian-kernel PCA

# Deep Neural Networks (2010s-Present)

Notable events and achievements in computer vision and NLP:

- ▶ (2006) First GPU-implementation CNN by K. Chellapilla et al.
- ▶ (2009) Nvidia GPUs were used for deep learning, drastically speedup training
- ▶ (2012) ImageNet dataset by Feifei Li's team, greatly facilitated vision recognition research
- ▶ (2013) Word2Vec word embedding model released by Google
- ▶ (2014) Generative Adversarial Network (GAN) was invented by Ian Goodfellow and his colleagues
- ▶ (2016) Further development in CNN: e.g. ResNet (image classification) and UNet (semantic segmentation)
- ▶ (2020) language model GPT-3 generates human-like text

# Deep Neural Networks (2010s-Present)

Deep reinforcement learning demonstrates human-level game play



Screenshots of Atari 2600 Challenge

- ▶ (2013) AI plays Atari games
- ▶ (2016) AlphaGo beats human at Go
- ▶ (2018) AlphaStar reaches grandmaster level at Starcraft

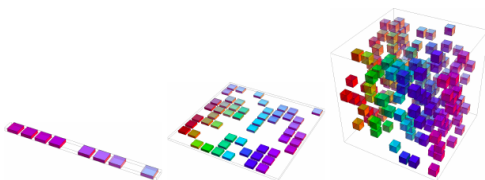
# Challenges in Deep Learning

- ▶ Overfitting
- ▶ Lack of interpretability
- ▶ Vulnerability to adversarial attack
- ▶ Highly dependent on data (GPT-3 is the current largest deep neural network with 175,000,000,000 parameters )



# Important Challenges in Machine Learning Research

## Curse of dimensionality



In high dimensional space, the possible configuration of  $x$  is much larger than the number of training examples.

- ▶ **Semi-supervised learning:** learn from a small set of labeled data and a rich set of unlabeled data.
- ▶ **Active learning:** a type of semi-supervised learning that interactively queries the user to obtain labels at new datapoints.
- ▶ **Deep convolutional neural networks :** learn efficient representations from data with multiple levels of abstraction.

## Heterogeneous learning

Real world applications encounter a lot of **heterogeneities** in data representations and tasks.

e.g. Road traffic status are partially observed by heterogeneous sources:

- ▶ Static sensors
- ▶ Mobile sensors
- ▶ Real-time social media content related to traffic condition
- ▶ Accident report
- ▶ ...



南宁路况 ✓

7月11日 18:02 来自 360安全浏览器

#晚高峰实况# 18:00 厢竹大道公安小区前路段往竹溪大道方向发生一起两小车相碰事故，占用中间主车道，请注意避让。

Transfer learning, domain adaption, and multi-modal learning are motivated by this challenge.

# Machine learning theories

Provides theoretical supports on why machine learning algorithms work, improves learning performances, and discovers potential pitfalls.

## Open theoretical questions

- ▶ How data quality affects learning performance
- ▶ How auxiliary information (unlabeled data, similar tasks) improves the ability to learn from new things
- ▶ Understand deep neural networks through information theory
- ...

# Summary

Machine learning: learn rules from data, adapt to changes and improves performance with experience.

- ▶ Machine learning themes in history
  - ▶ Statistical methods
  - ▶ Perceptrons and ANN
  - ▶ SVM, kernel methods, ensemble methods
  - ▶ Deep neural networks

## Next Lecture: Linear Space Methods

- ▶ Linear Regression
- ▶ Logistic Regression
- ▶ Optimization methods