Learning From Data
Lecture 1: Introduction

Yang Li yangli@sz.tsinghua.edu.cn

TBSI

March 1, 2024

Today's Lecture

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

About this Class

http://yangli-feasibility.com/home/classes/lfd2024spring/

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://cs229.stanford.edu/main_notes.pdf`
- Pattern Recognition and Machine Learning, 2nd Edition, by Christopher Bishop

Staffs

Yang Li
Instructor

Yanru Wu
Head TA

Boshi Tang
TA

Jiahao Lai
TA

Office hours

Name	Time	Location
Yang	Friday 2:00-4:00pm	Info Building 1108a
Yanru	Tuesday 4:00pm-5:00pm	Info Building, 11th floor common area
Boshi	Wednesday 2:00pm-3:00pm	Info Building 1701
Jiahao	Thursday 4:00pm-5:00pm	Info Building, 11th floor common area

You can also make appointments outside office hours.

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~~! slack
- 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

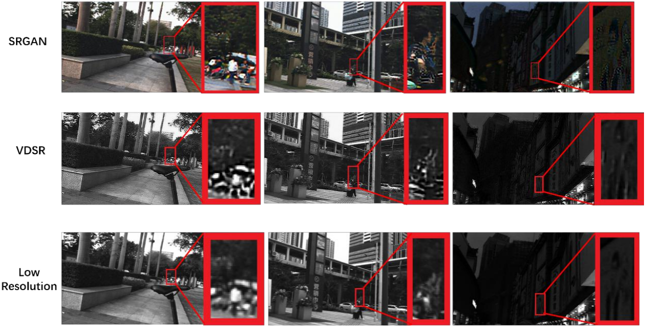
Homework plagiarism (copying) is not tolerated!
Ask for help early and often!

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)



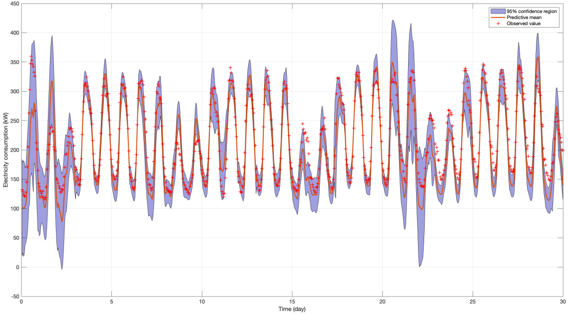
Comparison between two super-resolution models: SRGAN and VDSR

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

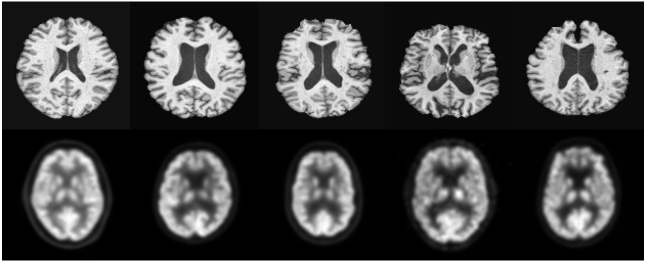
Apply recent machine learning techniques on real-world problems, or explore theoretical problems related to learning from data.

Previous class projects

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

MRI

CT

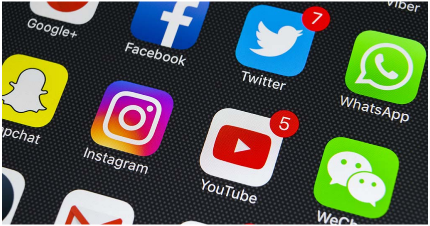


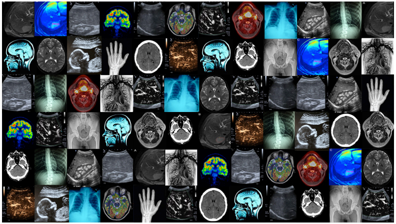
MRI (top) and PET (bottom) scans of normal and Alzheimer patient brains

Section I: What is Machine Learning?

a machine that finds relations/pattern from data.

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

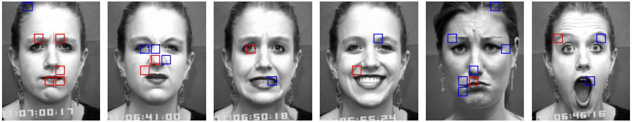
- learn rules from data for some **task**
- adapt to changes
- improve **performance** with **experience**.

relationship
pattern
data

(from "Machine Learning Theory" by Avrim Blum)

Machine Learning Tasks

- Classification

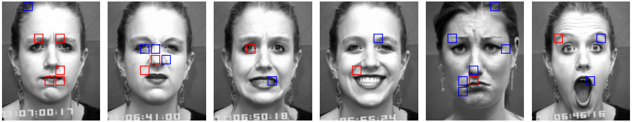


(a) Ang (b) Dis (c) Fea (d) Hap (e) Sad (f) Sur

Facial expression recognization (Liu et al. CVPR 2014)

Machine Learning Tasks

- Classification



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

Facial expression recognization (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

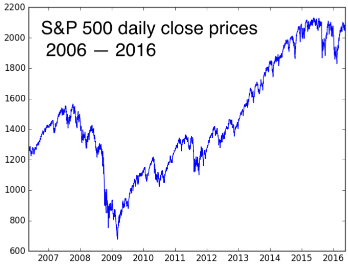
Machine Learning Tasks

- Regression

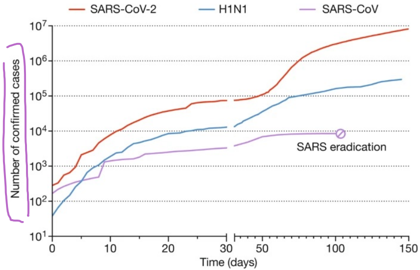
Algorithmic trading: forecast close price, highs and lows

Machine Learning Tasks

- Regression



Algorithmic trading: forecast close price, highs and lows



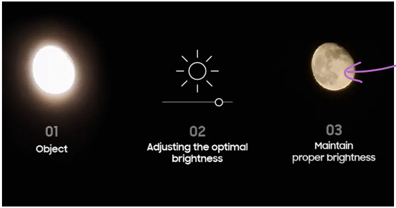
Early-day pandemic case prediction

Machine Learning Tasks

- Recognition (e.g. speech recognition)

Machine Learning Tasks

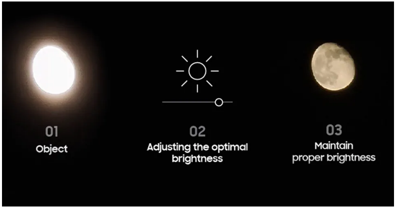
- Recognition (e.g. speech recognition)
- Image denoising/super-resolution



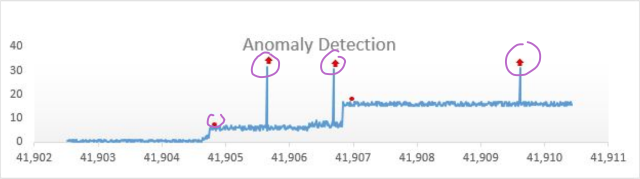
Machine Learning Tasks

- Recognition (e.g. speech recognition)

- Image denoising/super-resolution



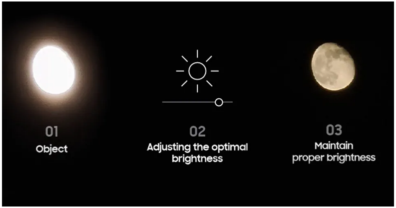
- Anomaly detection: finding abnormal operational activity for network security.



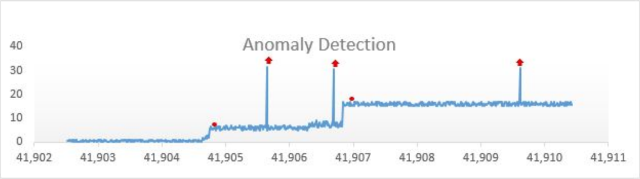
Machine Learning Tasks

- Recognition (e.g. speech recognition)

- Image denoising/super-resolution



- Anomaly detection: finding abnormal operational activity for network security.



Can you name some other tasks?

Machine Learning Experience

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

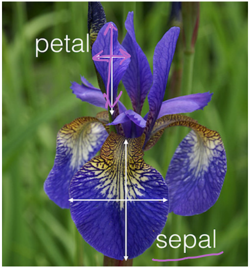
Machine Learning Experience

$x^{(1)} = 5.$

- **Dataset**: a collection of input, $X = \{x^{(1)}, \ldots, x^{(m)}\}$ and optionally, the corresponding output (**labels**) $Y = \{y^{(1)}, \ldots, y^{(m)}\}$
- 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	spieces
$x^{(1)}$	5.1	3.5	1.4	0.2	Setosa
$x^{(2)}$	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
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$



$x^{(1)}$ R G B

1 215 26 . . .

(0, 255)

(3 × w × h)

Machine Learning Performance

machine / prediction function

- Quantitatively evaluate the ability of a machine learning algorithm for a given task, e.g.

predicted value

  - Mean square error (MSE): $\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \boxed{f(x^{(i)})})^2$
  - Mean absolute error (MAE): $\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{y^{(i)} \neq f(x^{(i)})\}$

$\begin{cases} 1 & y^i \neq f(x^i) \\ 0. & y^i = f(x^i). \end{cases}$

Machine Learning Performance

- Quantitatively evaluate the ability of a machine learning algorithm for a given task, e.g.
  - Mean square error (MSE): $\frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - \boxed{f}(x^{(i)}))^2$
  - 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

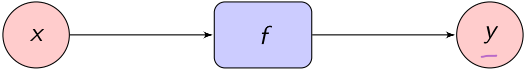
Given some input and output (label) training data, learn the **machine** f from training data

$$x \longrightarrow \boxed{f} \longrightarrow y$$

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)

X : image

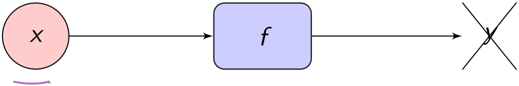
Y : sentence/text

Different Types of Learning

- denoising
- speech recognition
- anomally detection

Unsupervised learning

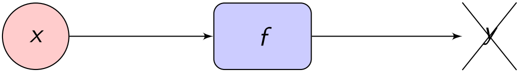
No labels are given in prior, find hidden structure or pattern from the data



Different Types of Learning

Unsupervised learning

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



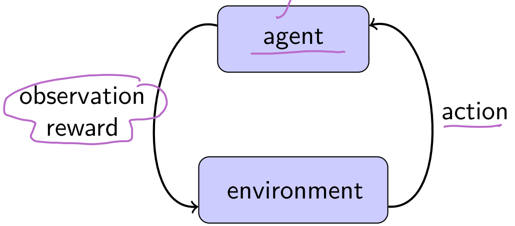
Unsupervised learning tasks:

- Data clustering
- Anomaly detection

Different Types of Learning

Reinforcement learning

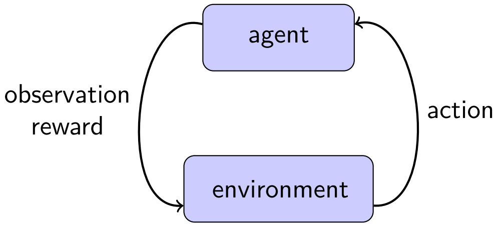
The learning machine is presented in an interactive manner to a dynamic environment, and need to make **sequential decisions**



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**



- Robotic agent (self-driving car, AlphaGo)
- AI Chatbot (Reinforcement learning from Human Feedback)
- Intelligent control system

Inference vs Prediction

$y = \underline{\underline{A}}x + \varepsilon$

$\varepsilon \sim N(0, I)$

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

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

Prediction

given **future** data samples of x, predict the corresponding output data y using f.

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 ← *focus of statistics*

do not confuse with "generation"

Prediction → Focus on generalization to test data

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

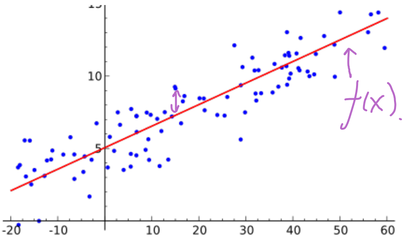
generalize

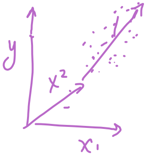
A Brief History of Machine Learning

Development of Statistical Methods (<1950)

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

$$f(x) = b + w_1 x_1 + w_2 x_2 = w^T x + b$$



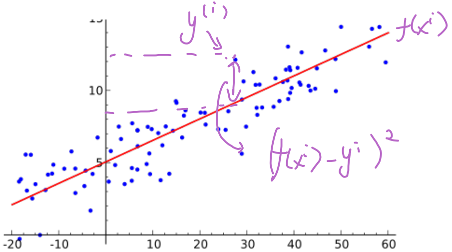


$x \in \mathbb{R}$

Development of Statistical Methods (<1950)

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

$$f(x) = b + w_1 x_1 + w_2 x_2 = w^T x + b$$



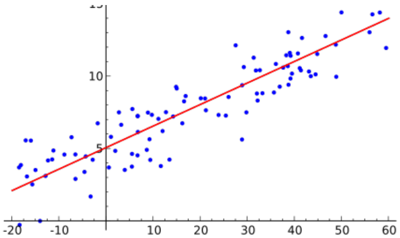
Learn model f by minimizing the **loss function** (MSE):

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

Development of Statistical Methods (<1950)

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

$$f(x) = b + w_1 x_1 + w_2 x_2 = w^T x + b$$



$w_1 x_1^2 + w_2 x_2^2 + w_3 x_1 x_2$
$+ b$ ..

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.

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

Development of Statistical Methods (<1950)

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

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

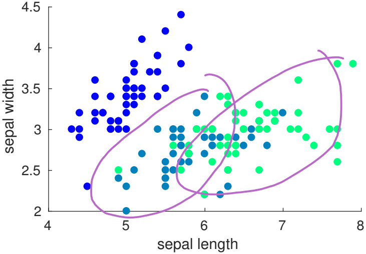
The foundation of **Bayesian estimation**, a core approach in estimating model parameters from data.

Development of Statistical Methods (<1950)

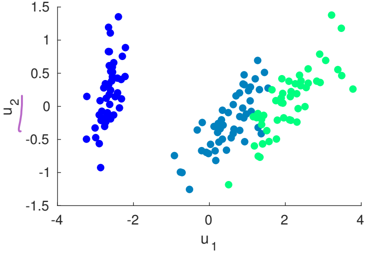
- (1901): Karl Pearson invented **principal component analysis** (PCA), a classic tool in exploratory data analysis and dimension reduction.

PCA

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



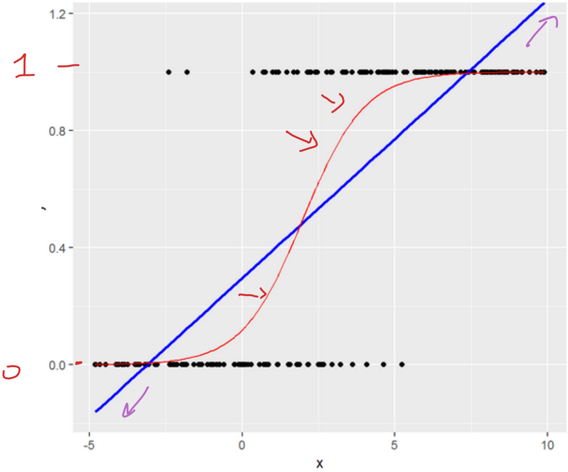
original



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

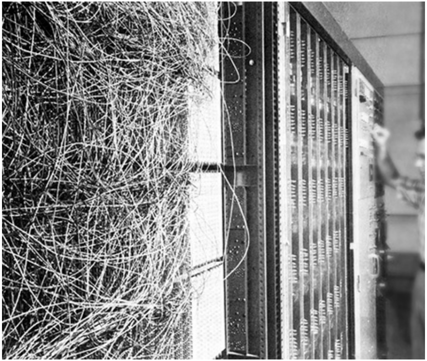
$$f(x) = w^T x + b$$

logistic

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

Simple Learning "Algorithms" (1950) 1945 ENIAC

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

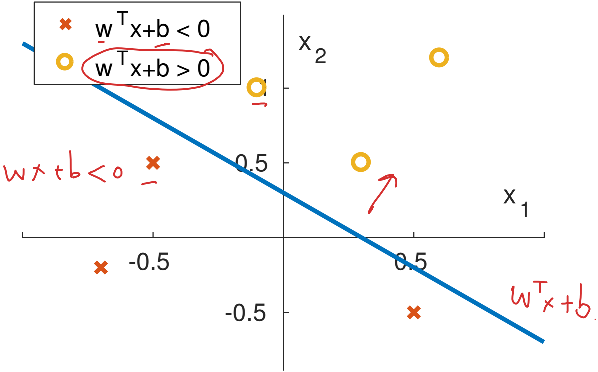


Hardware implementation: Mark I Perceptron

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: $y = "0"$ → small positive coefficient
  - predicted "1": $w_{t+1} := w_t - \alpha x$
  - predicted "0": $w_{t+1} := w_t + \alpha x$

$y = "1"$

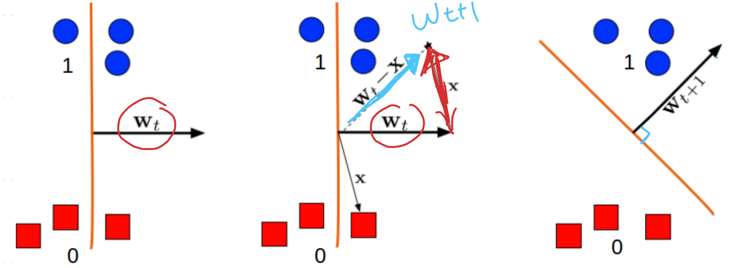
The perceptron learning algorithm

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 rules.

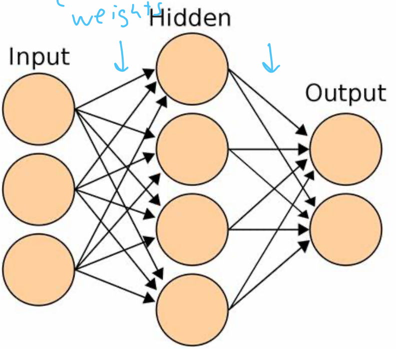


Simple Learning Algorithms (1960s)

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

Learning always involves modifying the connection weights

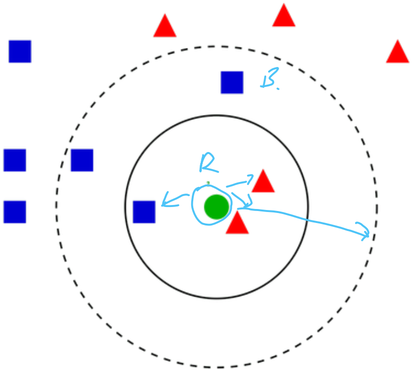
connection weights



ANN with a hidden layer

Simple Learning Algorithms (1960s)

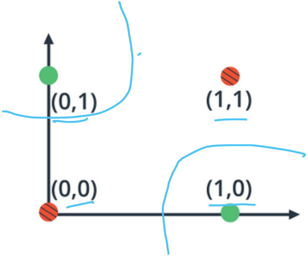
- (1967): Cover and Hart invented **Nearest Neighbor Classification** and the start of Pattern Recognition *One of the first non-parametric learning algorithms*



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



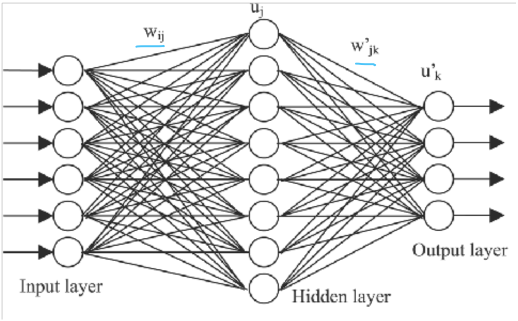
The XOR problem

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**.



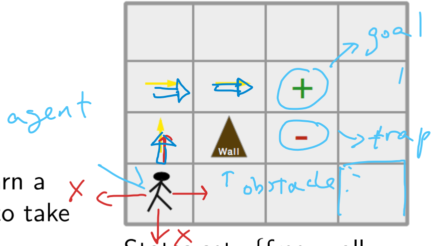
Many hidden units increase expressiveness of ANNs

Rediscovery of Backpropagation (1980s)

- (1989) Christopher Watkins proposed **Q-learning**, fundation 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).

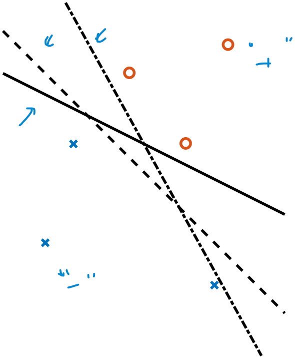


States set: {free, wall, goal, }
Action set: {Left, Right, Top, Down}

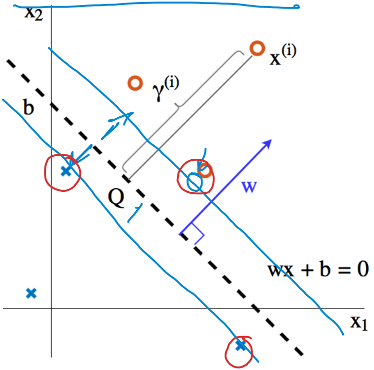
Rise of Data Driven Methods (1990s)

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

Single-layer perceptron may have infinite solutions



Support Vector Classifier

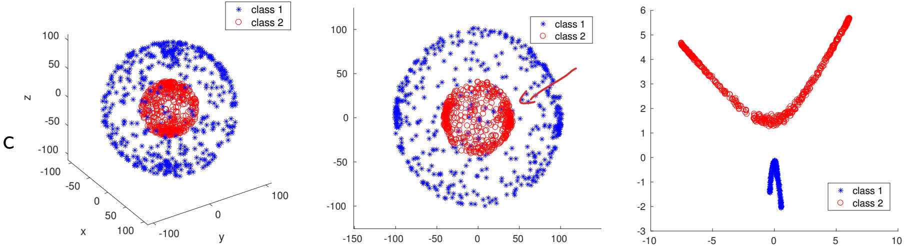


Give accuracy comparable to neural networks with elaborated features in a handwriting task

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 | Gassian-kernel PCA

RBF

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)
- (2017) Transformer (multi-head attention) was proposed by Google Brain
- ...

Deep Reinforcement Learning (2013-Present)

Deep reinforcement learning demonstrates human-level game play

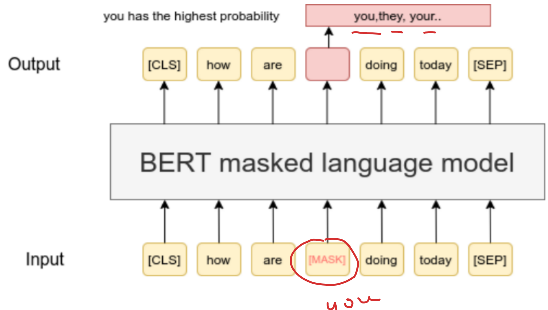
state space is huge

Screenshots of Atari 2600 Challenge

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

Self-supervised Learning & Foundation Model

self-supervised learning: a type of supervised learning where labels comes from the data itself



BERT masked language model

- Pretrain a model on self-supervised tasks (e.g. Masked Language Model (2018), Masked auto-encoder (2021)
- Use the model representation for downstream tasks

Self-supervised Learning & Foundation Model

foundation models are large deep learning models trained on *massive data* such that it can be applied across a wide range of tasks. e.g. GPT-n (2020-Present), DALL-E (2021)

pre-trained

Often used as a starting point to develop new ML models, rather than train from scratch.

GPT-4.0

You

create a simple comic about chat-gpt

ChatGPT

Here's the simple three-panel comic about ChatGPT.

Challenges in Deep Learning

- Overfitting
- Lack of interpretability → interpretable / explainable
- Vulnerable to adversarial attack
- Dependency on data quantity & quality robustness
- Training large models are costly (GPT-4 has about 1.8 trillion parameters)

Machine Learning Research

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.

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.

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.

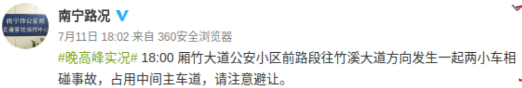
- **Semi-supervised learning**: learn from a small set of labeled data and a rich set of unlabeled data.
- **Active learning**: a type of semi-supervised learning that interactively queries the user to obtain labels at new datapoints.
- **Self-supervised learning** : leverage inherent structures or relationships within the input data to create meaningful features

Heterogeneous Learning

Real world applications encounter a lot of **heterogeneities** in data modalities, 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
- ...



Transfer learning, multi-modal learning and foundational models are motivated by this challenge.

Machine learning theories

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

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

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
- Understand deep neural networks through information theory ...

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
- Understand deep neural networks through information theory ...
- Understanding the generalizing capability of transformer-based models

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
- Understand deep neural networks through information theory ...
- Understanding the generalizing capability of transformer-based models
- How well pre-trained model adapt to future task

Summary

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

data.

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

→ intro foundation model.

written

HW0 → WA0 : basic mathematical exercise (not graded)

→ programming

PA0 : notebook } githubclassroom tool.

Next Lecture: Linear Space Methods

- Linear Regression
- Logistic Regression
- Optimization methods