

Learning From Data

Lecture 1: Introduction

Yang Li yangli@sz.tsinghua.edu.cn

TBSI

September 13, 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/lfd2024fall/>

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



Weida Wang
Head TA



Yuanbo Tang
TA



Tong Wu
TA



Chengfeng Wu
TA

Office hours

Name	Time	Location
Yang	Monday 2:00-4:00pm	Info Building 1108a
Weida	Thursday 17:00-18:00	Info Building, 11th floor common area
Yuanbo	Wednesday 17:00-18:00	Info Building, 11th floor common area
Tong	Friday 17:00-18:00	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)	57 %
Continuous Performance	3 %

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 Course Q&A Document
<https://docs.qq.com/doc/DVFJ1SWd5cmhTQ051>
- ▶ 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, ChatGPT) 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)



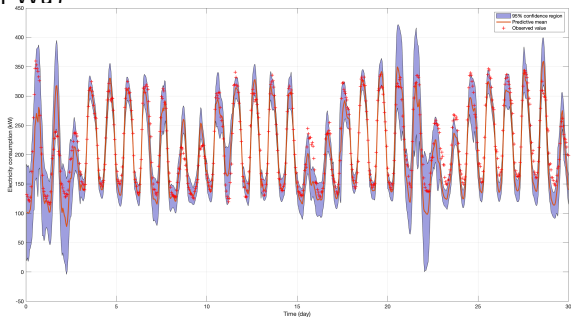
Comparison between two super-resolution models: SRGAN and VDSR

Final Group Project

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

Sample class projects

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



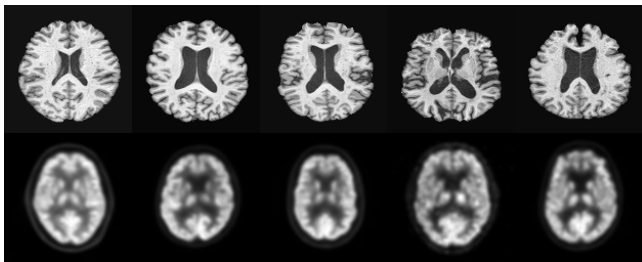
1-month prediction of electricity consumption

Final Group Project

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

Sample class projects

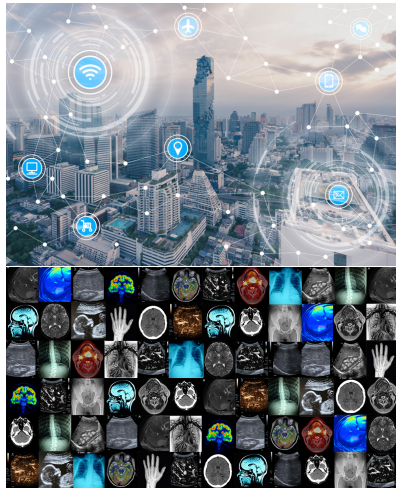
- ▶ Missing Data Imputation for Multi-Modal Brain Images (Wangbin Sun)



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

Section I: What is Machine Learning?

The age of big data



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

Machine Learning Experience

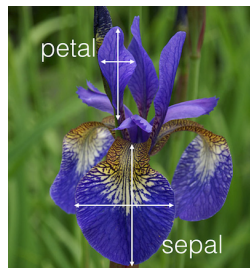
- ▶ **Dataset:** a collection of input, $X = \{x^{(1)}, \dots, x^{(m)}\}$ and optionally, the corresponding output (**labels**) $Y = \{y^{(1)}, \dots, y^{(m)}\}$
- ▶ Each input (data point) $x^{(i)}$ is represented by n **features**

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): $\frac{1}{m} \sum_{i=1}^m \mathbf{1}\{y^{(i)} \neq f(x^{(i)})\}$

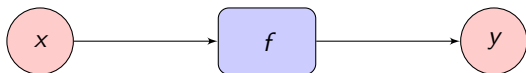
Machine Learning Performance

- ▶ 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



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

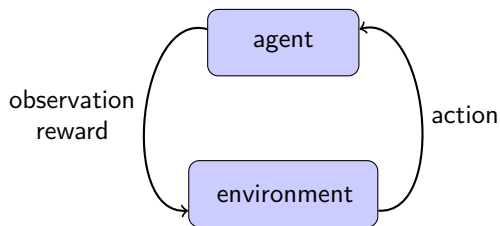
No labels are given in prior, find hidden structure or pattern from the data



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**



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

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

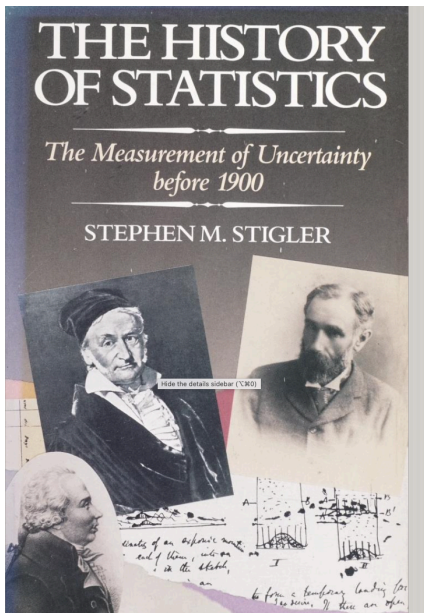
knowing the structure of f , find good models to describe f . i.e. model the data generation process ← *focus of statistics*

Prediction

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

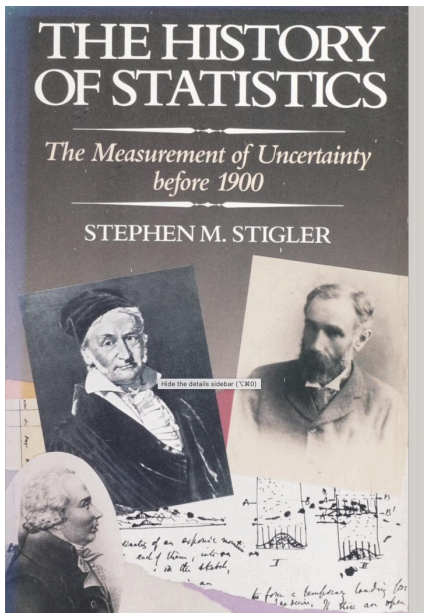
A Brief History of Machine Learning

Development of Statistical Methods (<1950)



What was data fitting first used for?

Development of Statistical Methods (<1950)



What was data fitting first used for?

Astronomy and geodesy for helping ocean navigation during the Age of Discovery.

Development of Statistical Methods (<1950)

Fitting ellipsoid to the earth's surface

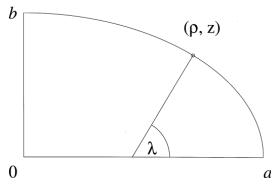
Development of Statistical Methods (<1950)

Fitting ellipsoid to the earth's surface

- ▶ (<1800) Marquis Pierre Simon de Laplace estimates the eccentricity

$e = \sqrt{1 - (b/a)^2}$ of the earth polar cross section using a linear system

- ▶ $c_0 := a(1 - e^2)$, $c_1 := \frac{3}{2}a(1 - e^2)e^2$



Laplace concluded that the earth's surface was not exactly an ellipsoid but the maximum error was within the measurement accuracy, with a squared eccentricity $e^2 < 0 : 01$

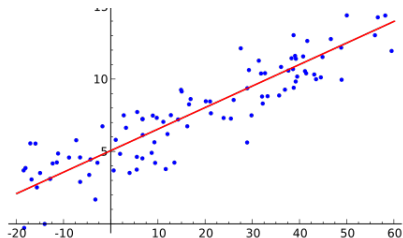
Example 1. With lengths in double toises (1/0.256 537 m) and angles in grads ($2\pi=400^\circ$), Laplace considered the following system [31, Book III, Section 41]:

$\Delta s / \Delta \lambda = c_0 + c_1 [\sin(\lambda)]^2$;	location;	latitude λ ;	arc $\Delta \lambda$;
$25538.85 = c_0 + c_1 * 0.00000$;	Peru;	00.0000° ;	3.4633° ;
$25666.65 = c_0 + c_1 * 0.30156$;	Good Hope;	37.0093° ;	1.3572° ;
$25599.60 = c_0 + c_1 * 0.39946$;	Pennsylvania;	43.5556° ;	1.6435° ;
$25640.55 = c_0 + c_1 * 0.46541$;	Italy;	47.7963° ;	2.4034° ;
$25658.28 = c_0 + c_1 * 0.52093$;	France;	51.3327° ;	10.7487° ;
$25683.30 = c_0 + c_1 * 0.54850$;	Austria;	53.0926° ;	3.2734° ;
$25832.25 = c_0 + c_1 * 0.83887$;	Lapland;	73.7037° ;	1.0644° ;

Development of Statistical Methods (<1950)

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

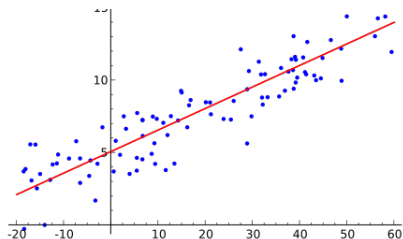
$$f(x) = b + w_1x_1 + w_2x_2 = w^T x + b$$



Development of Statistical Methods (<1950)

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

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



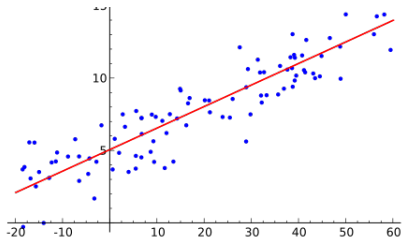
Learn model f by minimizing the **loss function** (MSE):

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

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.

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

Development of Statistical Methods (<1950)

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

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

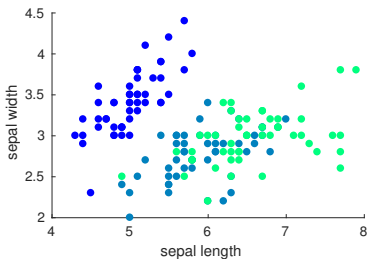
The foundation of **Bayesian estimation**, a core approach in estimating model parameters from data.

Development of Statistical Methods (<1950)

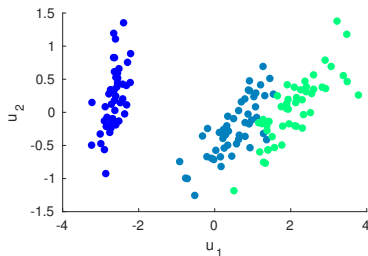
- ▶ (1901): Karl Pearson invented **principal component analysis** (PCA), a classic tool in exploratory data analysis and dimension reduction.

PCA

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



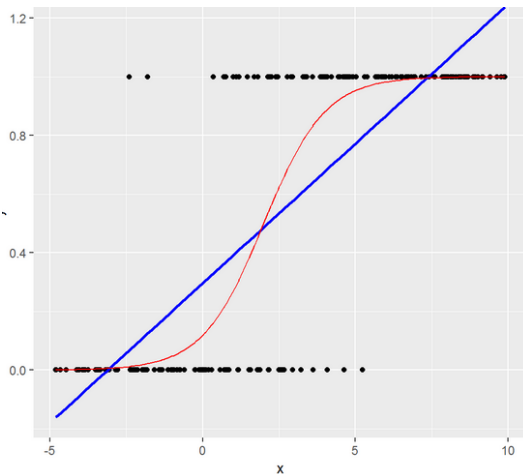
original



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

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

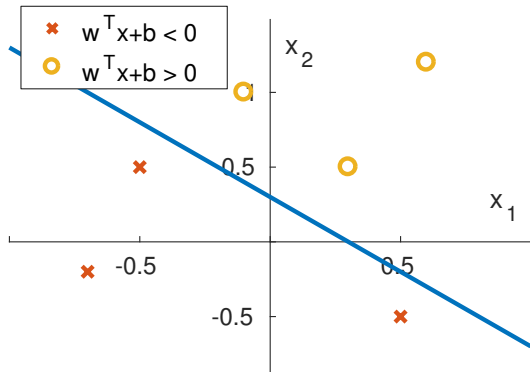
— logistic

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

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 $f(x)$

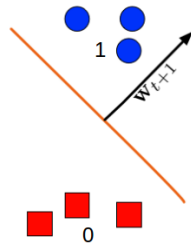
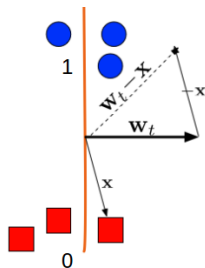
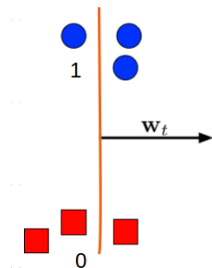
- ▶ When prediction is correct: $w_{t+1} = w_t$
- ▶ When prediction is incorrect:
 - ▶ predicted "1": $w_{t+1} := w_t - \alpha x$
 - ▶ predicted "0": $w_{t+1} := w_t + \alpha x$

The perceptron learning algorithm

Training a perceptron

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

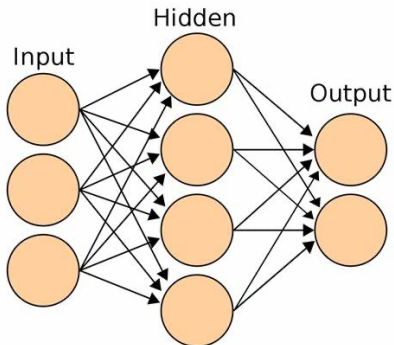
- ▶ When prediction is correct: $w_{t+1} = w_t$
- ▶ When prediction is incorrect:
 - ▶ predicted "1": $w_{t+1} := w_t - \alpha x$
 - ▶ predicted "0": $w_{t+1} := w_t + \alpha x$



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

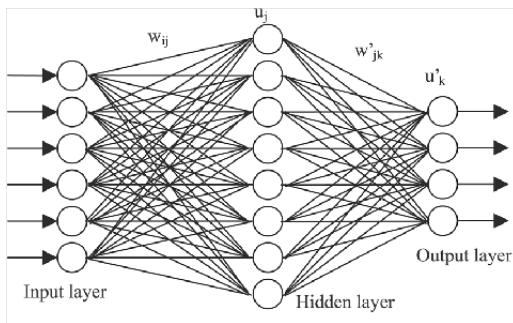


ANN with a hidden layer

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**.

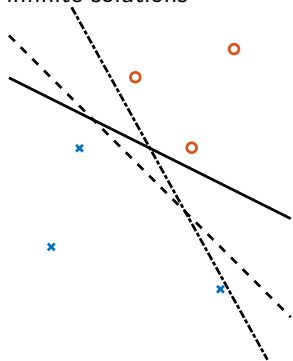


Many hidden units increase expressiveness of ANNs

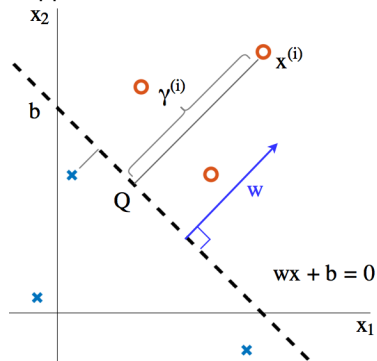
Rise of Data Driven Methods (1990s)

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

Single-layer perceptron may have infinite solutions



Support Vector Classifier



Rise of Data Driven Methods (1990s)

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

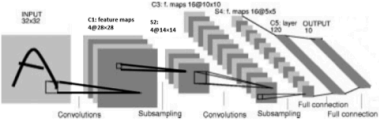
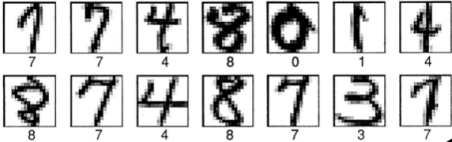
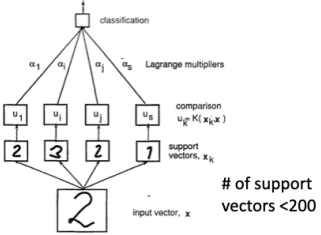
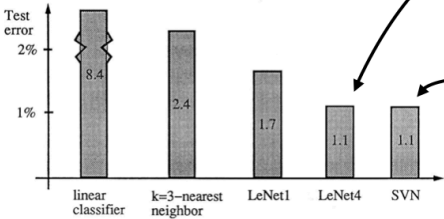


Figure 6. Examples of patterns with labels from the US Postal Service digit database.



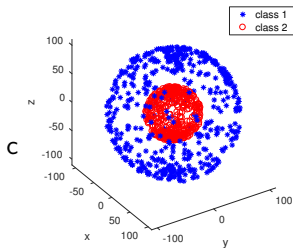
US Postal Service dataset: 7,300 training and 2,000 test images of size 16 x 16

¹Corinna Cortes and Vladimir Vapnik, Support-Vector Networks, Machine Learning, 1995

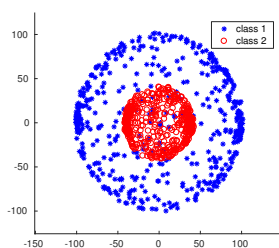
Kernel Methods (2000s)

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

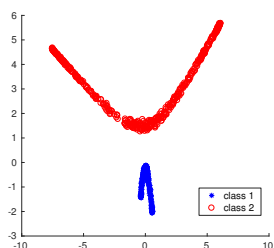
- ▶ (1998) **Kernel principal component analysis** (kernel PCA) was proposed by Schölkopf
- ▶ (2010) Radio Basis Function (RBF) kernel for SVM proposed by Yin-Wen Chang et. al.



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

Deep Reinforcement Learning (2013-Present)

Deep reinforcement learning demonstrates human-level game play

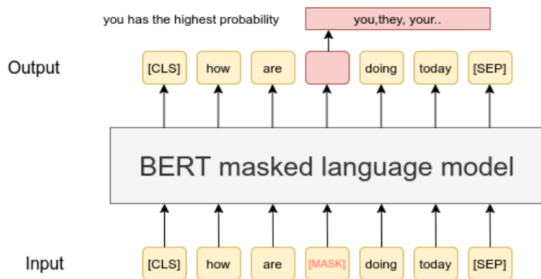


Screenshots of Atari 2600 Challenge

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

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)

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

 You
create a simple comic about chat-gpt

 ChatGPT



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

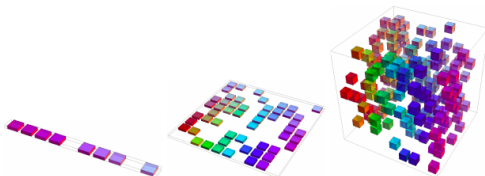
Challenges in Deep Learning

- ▶ Overfitting
- ▶ Lack of interpretability
- ▶ Vulnerable to adversarial attack
- ▶ Dependency on data quantity & quality
- ▶ 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.

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
- ▶ ...



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, 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.

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