Learning From Data
Lecture 1: Overview

Yang Li yangli@sz.tsinghua.edu.cn

TBSI

September 19, 2020

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

Grading

Your overall grade will be determined roughly as follows:

ACTIVITIES	PERCENTAGES
Midterm	20 %
Final Project	30 %
Problem sets (written & programming)	50 %

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!
- Ask a TA in person or in Wechat

Staffs

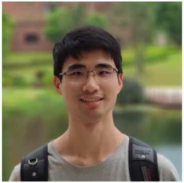
See Open Office Hour on the website.

Yang Li (Instructor)

Office: Info Building 1108A

Weida Wang (TA)

Office: Info Building 1111A



Feng Zhao (TA)

Office: Info Building 1111A

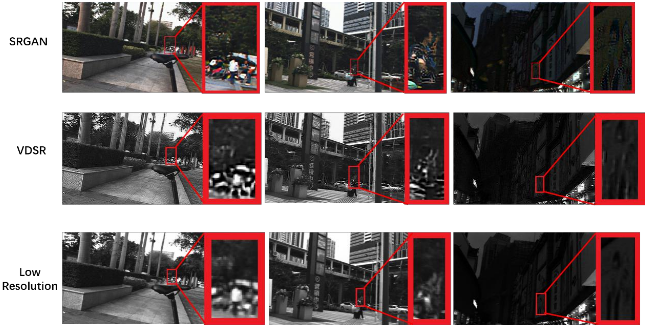
TAs will stop responding to HW questions at 7pm the night before deadline!

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)



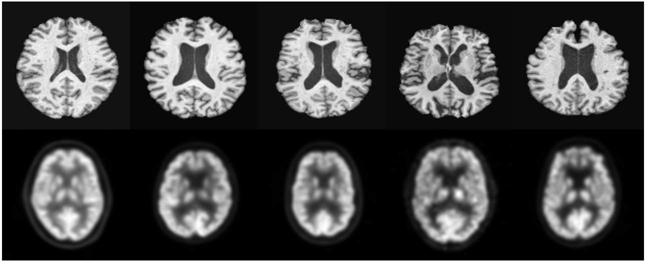
Comparison between two super-resolution models: SRGAN and VDSR

Final Group Project

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

Previous class projects

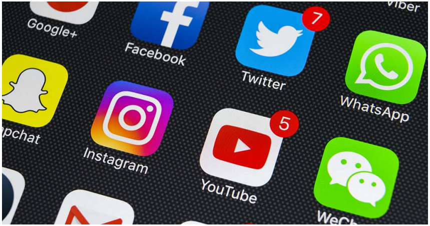
- Missing Data Imputation for Multi-Modal Brain Images (Wangbin Sun)

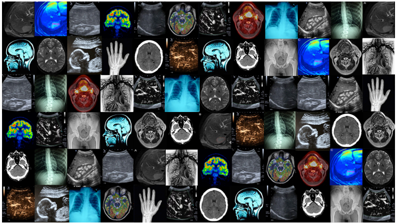


MRI (top) and PET (bottom) scans of normal and Alzheimer patient brains

Section I: What is Machine Learning?

The age of big data





brings challenge to data storage, data analysis, search and information privacy

What is Machine Learning?

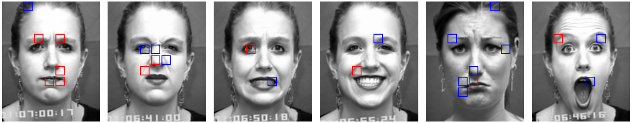
Learning Algorithm

A computer program, a.k.a. **machine** f, learns from experience E with respect to some task T , if its performance P while performing task T improves over E.

— Tom Mitchell (1989)

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)

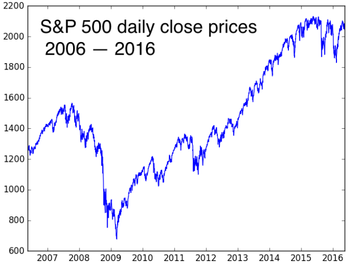
"The earphone broke in two days." (Negative)

Product review sentiment classification

Machine Learning Tasks

- Regression

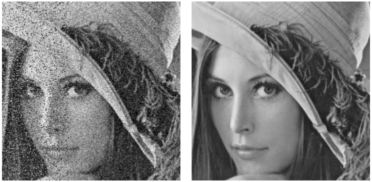
Highway travel time prediction



Algorithmic trading: forecast close price, highs and lows

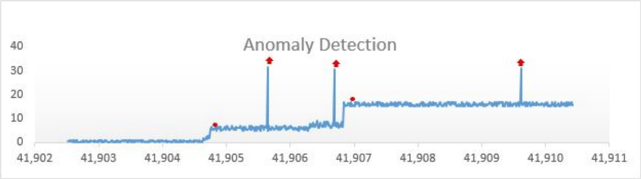
Machine Learning Tasks

- Data denoising



- Pattern recognition (e.g. spam filter)

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

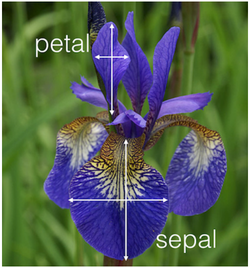


Machine Learning Experience

- **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	spieces
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
⋮	⋮	⋮	⋮	⋮



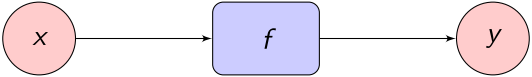
Machine Learning Performance

- 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



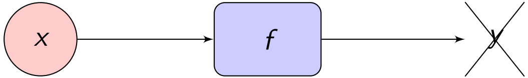
Supervised learning tasks:

- 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



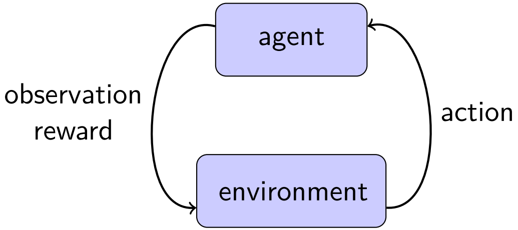
Unsupervised learning tasks:

- 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 sequntial 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 ← *focus of statistics*

Prediction

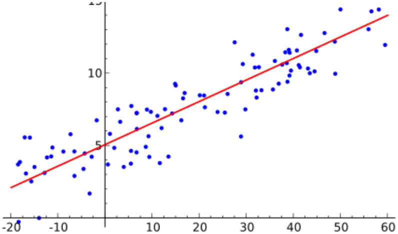
given **future** data samples of x, predict the corresponding output data y using f. ← *focus of machine learning*

A Brief History of Machine Learning

Development of Statistical Methods (<1950)

- (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):

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

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.

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

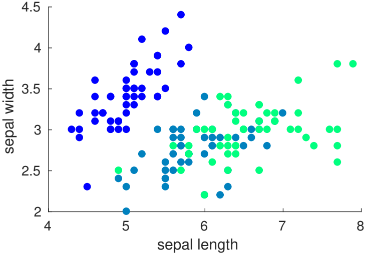
the foundation of **Bayesian estimation**, a core approach in estimating model parameters from data.

Development of Statistical Methods (<1950)

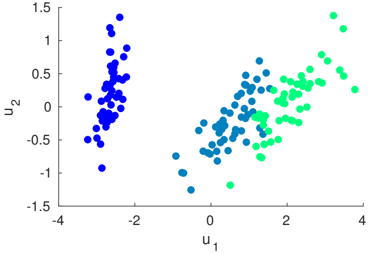
- (1901): Karl Pearson invented **principal component analysis** (PCA), a classic tool in exploratory data analysis and dimension reduction.

PCA

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



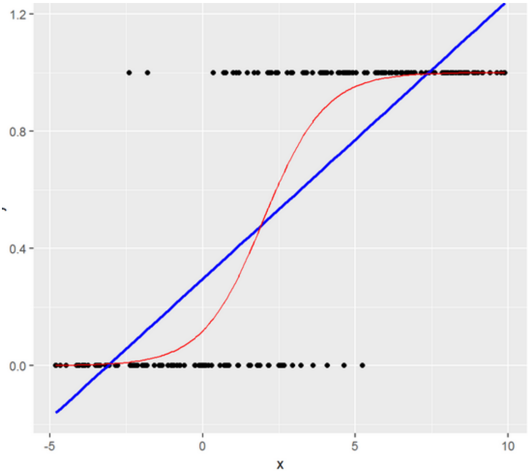
original



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

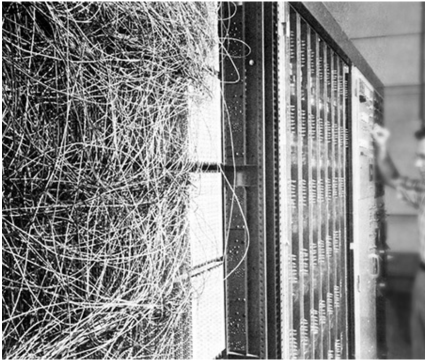
$$f(x) = w^T x + b$$

▬ 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 nueral network *Brings the popularity of Connectionism*

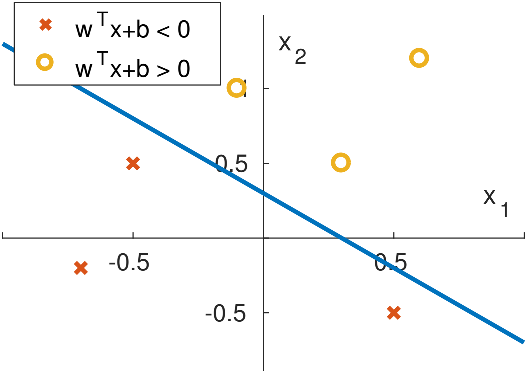


Hardware implementation: Mark I Perceptron

The perceptron learning algorithm

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

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

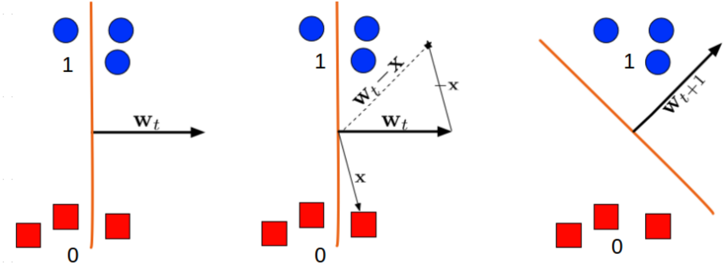


The perceptron learning algorithm

Training a perceptron

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

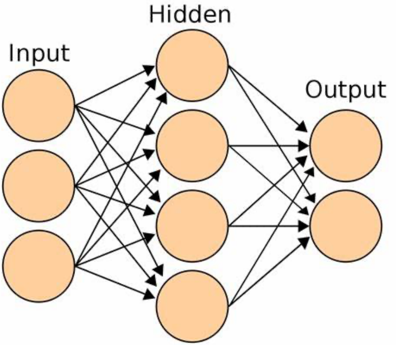
- When prediction is correct: $w_{t+1} = w_t$
- When prediction is incorrect:
  - predicted "1": $w_{t+1} := w_t - \alpha x$
  - predicted "0": $w_{t+1} := w_t + \alpha x$



Simple Learning Algorithms (1960s)

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

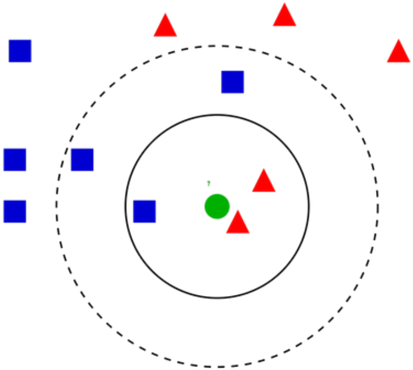
Learning always involves modifying the connection weights



ANN with a hidden layer

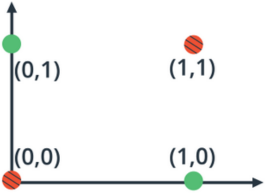
Simple Learning Algorithms (1960s)

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



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



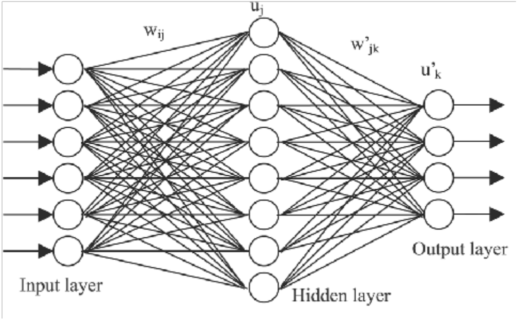
The XOR problem

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**.



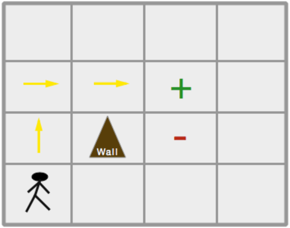
Many hidden units increase expressiveness of ANNs

Rediscovery of Backpropagation (1980s)

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

Q-learning

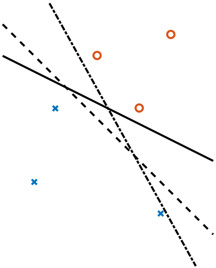
Given any **Markov decision process**, learn a policy, which tells an agent what action to take under what circumstances (states).



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

Rise of Data Driven Methods (1990s)

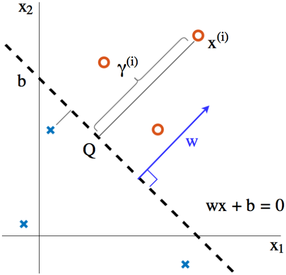
- (1992): Corinna Cortes and Vladimir Vapnik discovered **Support Vector Machine** (SVM)



SVM

- Single-layer perceptron has infinite solutions if data are separable
- Geometric "margin" $\gamma^{(i)}$ of hyperplane $w^T x + b = 0$ to sample $(x^{(i)}, y^{(i)})$:

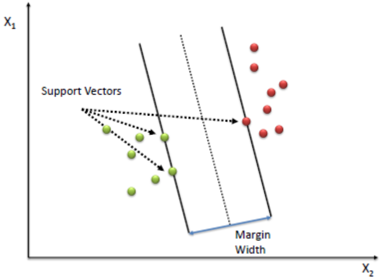
$$\gamma^{(i)} = y^{(i)} \left(\frac{w}{||w||}^T x^{(i)} + \frac{b}{||w||} \right)$$



SVM: Optimal Margin Classifier

SVM finds the hyperplane (w, b) that maximizes $\hat{\gamma}$, *margins of closest points to the hyperplane*. Such points are called **support vectors**

c



$$\max_{\gamma,w,b} \frac{\hat{\gamma}}{||w||}$$

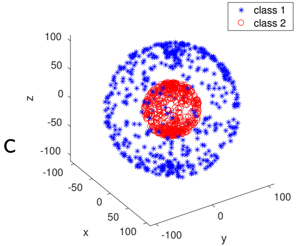
$$\text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

- Kernel SVM is proposed to support non-linear SVM

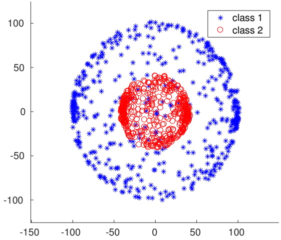
Kernel Methods (2000s)

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

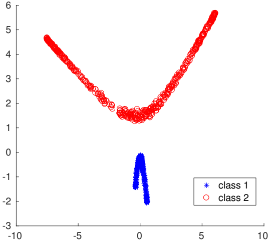
- (1998) **Kernel principal component analysis** (kernel PCA) was proposed by Schölkopf
- (2010) Radio Basis Function (RBF) kernel for SVM



original data



linear PCA



Gassian-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
- (2015) Further development in CNN: e.g. ResNet (image classification) and UNet (semantic segmentation)
- (2020) language model GTP-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
- Vulnerbility to adversarial attack
- Fairness
- Highly dependent on data (GTP-3 is the current largest deep neural network with 175,000,000,000 parameters)