Machine Learning

Chapter 1: Introduction

Fall 2021

Instructor: Xiaodong Gu, Yue Ding

Course Information

Instructor

顾小东

Rm:1208

丁玥

Rm:5303

Teaching Assistants

柴迤天

Rm:1417

施宇翔

Rm:5303

Website: https://oc.sjtu.edu.cn/courses/34447

E-mail: xiaodong.gu@sjtu.edu.cn

dingyue@sjtu.edu.cn

Course Information

Reference Materials:

Course Information

Useful Online Materials

https://d2l.ai/index.html

Course Information

Prerequisite:

《Linear Algebra》

《Probability and Statistics》

《Mathematical Foundations for Computer Science》

《Python Programming》

Course Information

Grading:

1) 平时作业：10%
2) 编程实验：10%,
3) 大作业：20%,
4) 期末考试：60%。

Today

- History
- What is machine learning?
  - Key Components
- How machine learning works?
  - A generic pipeline
  - A working example
- Overview of machine learning algorithms
  - Categories
  - Typical algorithms
- Important ML concepts that we'll use throughout the semester
  - Generalization、overfitting、regularization

What is Artificial Intelligence (AI) ?

What is intelligence ?

- Capacity for logic, understanding, self-awareness, learning, reasoning, planning , creativity, critical thinking, and problem-solving.

John McCarthy
one of the founders of the discipline of AI

What is artificial intelligence ?

- It is the science and engineering of making intelligent machines, especially intelligent computer programs

A Brief History of AI

1940—1950: Early days

- 1943: McCulloch & Pitts: **Boolean circuit** model of brain
- 1950: Turing's “Computing Machinery and Intelligence”

1950—70: Excitement: Look, Ma, no hands!

- 1950s: Early AI programs, including Samuel's checkers program, Newell & Simon's **Logic** Theorist, Gelernter's Geometry Engine
- 1956: Dartmouth meeting: “Artificial Intelligence” adopted

1970—90: Knowledge-based approaches

- 1969—79: Early development of **knowledge**-based systems
- 1980—88: Expert systems industry booms
- 1988—93: Expert systems industry busts: “AI Winter”

1990—: Statistical approaches

- **Data**-driven machine learning algorithms
- Resurgence of **probability**, focus on uncertainty
- General increase in technical depth
- Agents and learning systems… “AI Spring”?

2012—: Deep learning

- 2012: ImageNet & AlexNet
- **Deep** Neural Networks

Artificial Intelligence v.s. Machine Learning

Artificial Intelligence

Computer systems that perform tasks that would usually require human intelligence

Machine Learning

Statistical techniques that learn from a series of inputs and outputs

Deep Learning

Machine Learning algorithms that enable self learning of data representations

History of Machine Learning

- 1957 — Perceptron algorithm (implemented as a circuit!)
- 1959 — Arthur Samuel wrote a learning-based checkers program that could defeat him
- 1969 — Minsky and Papert's book *Perceptrons* (limitations of linear models)
- 1980s — Some foundational ideas
  - Connectionist psychologists explored neural models of cognition
  - 1984 — Leslie Valiant formalized the problem of learning as PAC learning
  - 1988 — Backpropagation (re-)discovered by Geoffrey Hinton and colleagues
  - 1988 — Judea Pearl's book *Probabilistic Reasoning in Intelligent Systems* introduced Bayesian networks

History of Machine Learning

- 1990s — the "AI Winter", a time of pessimism and low funding
- But looking back, the '90s were also sort of a golden age for ML research
  - Markov chain Monte Carlo
  - variational inference
  - kernels and support vector machines
  - boosting
  - convolutional networks
- 2000s — applied AI fields (vision, NLP, etc.) adopted ML
- 2010s — deep learning
  - 2010–2012 — neural nets smashed previous records in speech-to-text and object recognition
  - increasing adoption by the tech industry
  - 2016 — AlphaGo defeated the human Go champion

Why machine learning?

Traditional Programming

Machine Learning

"The ability to learn without being programmed"

Write a computer program with **explicit rules** to follow

Write a computer program to **learn from examples**

```
if email contains V!agrå
  then mark is-spam;
if email contains …
if email contains …
```

So many decision rules…

Magic!

Why machine learning?

- For many problems, it's difficult to program the correct behavior by hand
  - recognizing people and objects
  - understanding human speech
- Machine learning approach: program an algorithm to automatically learn from data, or from experience
- Why might you want to use a learning algorithm?
  - hard to code up a solution by hand (e.g. vision, speech)
  - system needs to adapt to a changing environment (e.g. spam detection)
  - want the system to perform *better* than the human programmers

Traditional Programming vs. Machine Learning

Machine Learning Applications

Machine learning is very widely used in:

Machine Learning Applications

Computer vision: Object detection, semantic segmentation, pose estimation, and almost every other task is done with ML.

Figure 4. More results of **Mask R-CNN** on COCO test images, using ResNet-101-FPN and running at 5 fps, with 35.7 mask AP (Table 1).

Instance segmentation - Link

DAQUAR 1553
What is there in front of the sofa?
Ground truth: table
IMG+BOW: table (0.74)
2-VIS+BLSTM: table (0.88)
LSTM: chair (0.47)

COCOQA 5078
How many leftover donuts is the red bicycle holding?
Ground truth: three
IMG+BOW: two (0.51)
2-VIS+BLSTM: three (0.27)
BOW: one (0.29)

Machine Learning Applications

Speech Recognition & Conversational Assistants

Machine Learning Applications

User Profiles

Machine Learning Applications

Medical Image Analysis

Breast Cancer Diagnoses

(AI + Pathologist) > Pathologist

Machine Learning Applications

Anomaly Detection

Detect malicious calls

Spam Detection

Machine Learning Applications

NLP & Translations

(Sutskever et al., 2014)

Machine Learning Applications

NLP & Text Generation

Chinese poem example. Can you distinguish?

Human	Machine
南陌春风早，东邻去日斜。	山夜有雪寒，桂里逢客时。
紫陌追随日，青门相见时。	此时人且饮，酒愁一节梦。
胡风不开花，四气多作雪。	四面客归路，桂花开青竹。

Jiaxian Guo, Sidi Lu, Weinan Zhang et al. Long Text Generation via Adversarial Training with Leaked Information. AAAI 2018.
Lantao Yu, Weinan Zhang, et al. Seqgan: sequence generative adversarial nets with policy gradient. AAAI 2017.

Machine Learning Applications

E-commerce & Recommender Systems : Amazon, netflix, ...

Inspired by your shopping trends

Related to items you've viewed See more

Machine Learning Applications

Social Networks

Social based Recommendation

Machine Learning Applications

Self-Driving Cars

Machine Learning Applications

AI as a Movie Maker

What is machine learning?

What is Machine Learning?

Human Learning v.s. Machine Learning

Knowledge

Practice

Exam

What is Machine Learning?

Human Learning v.s. Machine Learning

What is Machine Learning?

Human Learning v.s. Machine Learning

So, what is machine learning after all

A subset of artificial intelligence which uses statistical methods to enable machine to improve towards an objective using data without requiring explicit programming by human.

Key Elements of Machine Learning

Elements:

#1 Data (Experience)

$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(N)}, y^{(N)})\}$

Key Elements of Machine Learning

We **cannot** learn from data

We **can** learn from data + *hypothesis*

What hypothesis do we make about this data?

Elements:

#1 Data (Experience)

#2 Model(Hypothesis)

$$\hat{y} = g(x|\theta) + \epsilon$$

e.g. $\hat{y} = \boldsymbol{w}x + \boldsymbol{b}$

where θ (e.g., W and b) denotes the parameters

Key Elements of Machine Learning

Elements:

#1 Data (Experience)
#2 Model (Hypothesis)

#3 Loss Function (Objective)

$$\mathcal{L}(\theta|\mathcal{D}) = \frac{1}{2}\sum_{n=1}^{N}[\boldsymbol{w}x_n + b - y_n]^2$$

How to evaluate a model? (performance criterion)

Key Elements of Machine Learning

Elements:

#1 Data (Experience)
#2 Model (Hypothesis)
#3 Loss Function (Objective)

#4 Optimization Algorithm
(Improvement)

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\theta|\mathcal{D})$$

How to find the optimal model?

So, what is machine learning after all

A subset of artificial intelligence which uses statistical methods (model) to enable machine to improve (optimization) towards an objective (loss function) using (data) without requiring explicit programming by human.

Generalized to other data

What about other data representations?

Vectorization/
Feature Extraction

Data Points in High-Dimensional Vector Space

Generalized to other data

Vectorization/Feature Extraction

How machine learning works?

"Hello-World" of machine learning

Recall: Key Elements of Machine Learning

- Data (Experience)
- Model (Hypothesis)
- Loss Function (Objective)
- Optimization Algorithm (Improve)

How to Approach a Machine Learning Problem?

① consider the nature of available data **D**

- how much amount of data can you obtain? how would it cost (in time, computation, human efforts)?

② select a representation for the input X

- data preprocessing, feature extraction, etc.

③ choose a set of possible models **H** (hypothesis space)

- set of functions h: X → Y

④ choose the performance measure **P** (loss function)

⑤ choose or design a learning algorithm

- for using examples (**E**) to converge on a member of **H** that optimizes **P**

A Working Example

The problem: Automatic Fruit Classification

Data Collection

Create training samples

- Tell me which type a specific fruit belongs to

Sample	Label
	Apple
	Orange
	Apple
...	...

Pre-processing

Different types of fruit differ in size, lightness, position, etc.

Needs pre-processing!

Example:

remove the background

adjust for light level

Feature Extraction

Consider each fruit as a point in some feature space

Feature:

Color: red
Shape: round
Size: 3
...

$$\begin{bmatrix} 1 \\ 0 \\ 3 \end{bmatrix}$$

Feature:

Color: orange
Shape: round
Size: 2
...

$$\begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix}$$

Learning (Training)

Partition the feature space into 2 regions, one for each type of fruit

- decision boundary

Test and Evaluation

How to handle new images?

Classifying a new image using the trained classifier .

How to measure the classifier performance?

- classification error rate
  - % patterns that are assigned to the wrong category
- other aspects may be important too
  - e.g. computational complexity
  - e.g. user-friendliness

Pipeline of an ML System

Data Preparation

Data Collection

Feature Extraction

Training

Train model

Validation

Evaluation

Prediction

Evaluation

Overview of Machine Learning Algorithms

Categories of Machine Learning Algorithms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

The learner is provided with a set of inputs together with the corresponding desired outputs.

- has a "teacher"

Example

- teaching kids to recognize different animals.
- graded examinations with correct answers provided.

Supervised Learning

Classification

- Face detection and recognition
- Character recognition
- Speech recognition
- Object recognition and image classification
- Biometric authentication
- Multi-touch gesture classification
- Document classification
- Spam detection and filtering
- Intrusion detection
- Terrorism detection
- Medical diagnosis
- Weather forecasting
- ...

Supervised Learning

Example: Face Detection

Supervised Learning

Regression

Applications

- Navigation of autonomous vehicle: angle of steering wheel
- Kinematics of robot arm: joint angles
- Recommender system: movie ratings
- Age estimation from facial image: age
- Chemical manufacturing process: yield
- Risk prediction from financial reports: risk
- ...

Supervised Learning

Example: Recommendation Systems

Supervised Learning

Example: Control

Unsupervised Learning

Training examples as input patterns, with no associated output.

- no teacher
- similarity measure exists to detect groupings / clusterings

Unsupervised Learning

Clustering

- in the early stages of an investigation, it may be helpful to perform exploratory data analysis to gain some insight into the nature or structure of the data.

Unsupervised Learning

Probability Estimation

Applications

- Density estimation
- Data generation
- Abnormal Detection

Unsupervised Learning

Data Generation

Distribution of images generated by the model

The goal of the generative model is to find a $p_{model}(x)$ that approximates $p_{data}(x)$ well.

Distribution of actual images

Unsupervised Learning

Example: Face Generation

Unsupervised Learning

Dimension Reduction and Feature Selection

- find features or preprocess existing features for the subsequent pattern classification problem (supervised learning)

principal component analysis (PCA)

Example (eigenface)

Unsupervised Learning

Example: Auto-encoder

Useful for dimension reduction, information compression, etc

Unsupervised Learning

Topic Modeling – Learning latent topics from documents

For documents in a large collection of text, model p(Word|Topic), p(Topic)

Unsupervised Learning

Find the least likely observations from a dataset (outlier detection)

Example (network intrusion detection)

Detect whether someone is trying to hack the network or doing anything else unusual on the network

File Logs Settings Help

Application	Protocol	Local Address	Remote Address	State	Creation Time	Rx [Bytes]	Type	Tx [Bytes]
SVCHOST.EXE	TCP	all:135	----------	Listening	27/Feb/2007 09:41:27	0	0	0
SVCHOST.EXE	TCP	all:3389	----------	Listening	27/Feb/2007 09:41:33	0	0	0
SVCHOST.EXE	TCP	all:1025	----------	Listening	27/Feb/2007 09:41:29	0	0	0
SYSTEM	TCP	192.168.123.128:139	----------	Listening	27/Feb/2007 09:41:30	0	0	0
PERSFW.EXE	TCP	all:44334	----------	Listening	27/Feb/2007 09:41:43	0	0	0
PERSFW.EXE	TCP	all:44334	localhost:1986	Connected In	27/Feb/2007 16:45:42	2931	<attack>	152516
SVCHOST.EXE	TCP	all:5000	----------	Listening	27/Feb/2007 09:42:15	0	0	0
LSASS.EXE	TCP	all:27155	----------	Listening	27/Feb/2007 09:42:17	0	0	0
PUTTY.EXE	TCP	all:1898	uststu1.ust.hk:22	Connected Out	27/Feb/2007 15:54:54	445204	<normal>	141398
CCAPP.EXE	TCP	localhost:1035	----------	Listening	27/Feb/2007 09:42:23	0	0	0
SSHCLIENT.EXE	TCP	all:1639	lcpu2.cse.ust.hk:22	Connected Out	27/Feb/2007 11:43:48	135198	<normal>	3764
ICQ.EXE	TCP	all:21470	----------	Listening	27/Feb/2007 14:41:02	0	0	0
ICQ.EXE	TCP	all:1759	64.12.24.205:5190	Connected Out	27/Feb/2007 14:41:04	59285	<normal>	12191
MSNMSGR.EXE	TCP	all:1722	by2msg1104003.ph...	Connected Out	27/Feb/2007 14:38:33	176880	<normal>	129012
PFWADMIN.EXE	TCP	all:1986	localhost:44334	Connected Out	27/Feb/2007 16:45:42	278215	<normal>	2931
LSASS.EXE	UDP	all:500	----------	Listening	27/Feb/2007 09:41:40	0	0	0
SYSTEM	UDP	192.168.123.128:138	----------	Listening	27/Feb/2007 09:41:30	205	0	40995
LSASS.EXE	UDP	localhost:1033	----------	Listening	27/Feb/2007 09:42:23	0	0	2
SYSTEM	UDP	192.168.123.128:137	----------	Listening	27/Feb/2007 09:41:30	1088	0	17982
PERSFW.EXE	UDP	all:44334	----------	Listening	27/Feb/2007 09:41:43	0	0	0
IEXPLORE.EXE	UDP	localhost:1600	----------	Listening	27/Feb/2007 11:17:52	1173	0	1173
IEXPLORE.EXE	UDP	localhost:1697	----------	Listening	27/Feb/2007 14:09:48	15	0	15

Supervised v.s. Unsupervised Learning

Supervised Learning

Data: (x, y)
x is data, y is label

Goal: Learn function to map
$x \rightarrow y$

Examples: Classification, regression, object detection, semantic segmentation, etc.

Unsupervised Learning

Data: x
x is data, no label

Goal: Learn the hidden or underlying structure of the data.

Examples: Clustering, feature or dimensionality reduction, etc.

Reinforcement Learning

Learning from interacting with an environment to achieve a goal

Learning a mapping from states to actions to maximize long-term reward

- a set of environment states S;
- a set of actions A;
- rules of transitioning between states;
- rules that determine the scalar immediate reward of a transition; and
- rules that describe what the agent observes.

Goal: Maximize the expected long-term payoff

Example: graded examinations with only overall scores but no correct answers

Example

internal state
reward
environment
action
learning rate α
inverse temperature β
discount rate γ
observation

Monte-Carlo Reinforcement Learning

$$v(s_t) \leftarrow v(s_t) + \alpha[R_t - v(s_t)]$$

R_t is the actual long-term return following state s_t in a sampled trajectory

Reinforcement Learning

Example: AlphaGo

Reinforcement Learning

Example: Robots

- input: sensory information and reinforcement signal.
- output: avoid negative reward and try to have positive reward (obstacle avoidance, wall following, etc.)

Important ML concepts that we'll use throughout the semester

Generalization, Overfitting, Cross Validation, VC dimension, etc

Generalization

Will the classifier work for unseen fruits?

- The issue of generalization

Which model is better?

Bias and Variance

Bias – How close does the assumed model fit for the observed data?

Variance – How complex (e.g., freedom) the assumed model is?

Low Variance
High Variance
Low Bias
High Bias
$E[f^*] = \bar{f}$
f^*
Variance
Bias
$\hat{f}$

The Bias-Variance Dilemma

Bias – How close does the assumed model fit for the observed data?

Variance – How complex (e.g., freedom) the assumed model is?

Overfitting

Underfitting – the model isn't complex enough to capture the real knowledge, the assumption may not be true.

Overfitting – the model is too complex and thus describes the details of data (e.g., random noise) instead of underlying knowledge

Underfitted Good Fit/Robust Overfitted

Example: https://ml.berkeley.edu/blog/posts/crash-course/part-4/

Overfitting

A real-world example (driving test)

1. 机动车遇到行人时应 A. 快速通过 B. **减速通过** C. 正常行驶
2. 在如图所示的场景中，驾驶员应 A. 停车等待 B **减速让行** C.
3. 遇到下图所示的情况，下面哪个操作正确？ A.
4. …

30

三短一长选最长

90

非正常行驶的场合，应选择谨慎的选项

100

- 题中有”**应**”字样的选B,
- 有”**操作**”字样的选C,
- 最长的题目选D,
- 看图题有女孩的选A, 没人的选C

Prevent Overfitting

Selecting appropriate model complexity

In general, design a good loss function to indicate the performance of the predictive model over the whole-data distribution instead of the training data can help achieve compromise between simplicity and complexity of the model structure

Widely used approaches (to prevent overfitting)

- Increase training data
- Hold-out & cross validation (unseen data to ensure generalization)
- Regularization (penalizing model complexity)
- Early stop
- Prior knowledge (e.g., Bayesian prior)
- …

Prevent Overfitting

Fit the overall distribution instead of the training set.

- To estimate the generalization error, we need data unseen during model training.
- Data Splitting (Hold-Out):
  - Training set (e.g., 50%) 平时练习
  - Validation set (e.g., 25%) 模拟考试
  - Test set (e.g., 25%) 高考!
- Resampling is needed when data is limited.

Cross Validation

What if we want to make the most use of the data?

(e.g., the dataset is small)

- K-folder CV

- Allows for more train-test splits
- A better indication of how well your model performs on unseen data
- Takes more computational power and time

Regularization

Add a penalty term of the parameters to prevent the model from overfitting the training data.

$$L(\boldsymbol{W}, \lambda_1, \lambda_2) = \|y - \boldsymbol{W}X\|^2 + \lambda_2\|w\|_2^2 + \lambda_1\|w\|_1$$

Penalizing model complexity in the loss function

No free lunch theorem:

There is no universally best model.
Different types of models have to be developed to suit the nature of the data in real applications.

The Next Lecture: Linear Regression

A simple and well-known machine learning algorithm.
Similar to the least squares method (最小二乘法)