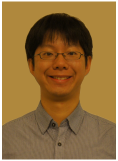
Machine Learning Overview and Applications

Hsuan-Tien Lin (林軒田)
htlin@csie.ntu.edu.tw

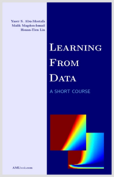
Computational Learning Lab (CLLab)
Department of Computer Science
& Information Engineering
National Taiwan University
(國立台灣大學資訊工程系計算學習實驗室)

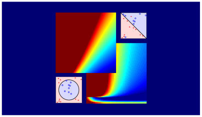
materials mostly taken from my “Learning from Data” book, my “Machine Learning Foundations” free online course, and works from NTU CLLab and NTU KDDCup teams

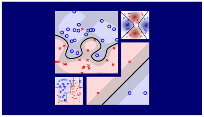
About Me
Hsuan-Tien Lin



- Associate Professor, Dept. of CSIE, National Taiwan University
- Leader of the Computational Learning Laboratory
- Co-author of the textbook "*Learning from Data: A Short Course*" (often **ML best seller on Amazon**)
- Instructor of the NTU-Coursera Mandarin-teaching ML Massive Open Online Courses
  - "*Machine Learning Foundations*": `www.coursera.org/course/ntumlone`
  - "*Machine Learning Techniques*": `www.coursera.org/course/ntumltwo`







What is Machine Learning

From Learning to Machine Learning

learning: acquiring skill
with experience accumulated from observations

observations → learning → skill

machine learning: acquiring skill
with experience accumulated/**computed** from data

data → ML → skill

What is skill?

A More Concrete Definition

skill
$\Leftrightarrow$ improve some performance measure (e.g. prediction accuracy)

machine learning: improving some performance measure
with experience **computed** from data

data $\longrightarrow$ $\boxed{\text{ML}}$ $\longrightarrow$ improved performance measure

An Application in Computational Finance

stock data $\longrightarrow$ $\boxed{\text{ML}}$ $\longrightarrow$ more investment gain

Why use machine learning?

Yet Another Application: Tree Recognition

- 'define' trees and hand-program: **difficult**
- learn from data (observations) and recognize: a **3-year-old can do so**
- 'ML-based tree recognition system' can be **easier to build** than hand-programmed system

ML: an **alternative route** to build complicated systems

The Machine Learning Route

ML: an **alternative route** to build complicated systems

Some Use Scenarios

- when human cannot program the system manually
  —navigating on Mars
- when human cannot 'define the solution' easily
  —speech/visual recognition
- when needing rapid decisions that humans cannot do
  —high-frequency trading
- when needing to be user-oriented in a massive scale
  —consumer-targeted marketing

Give a **computer** a fish, you feed it for a day;
teach it how to fish, you feed it for a lifetime. **:-)**

Key Essence of Machine Learning

machine learning: improving some performance measure
with experience **computed** from data

data $\longrightarrow$ ML $\longrightarrow$ improved performance measure

1. exists some 'underlying pattern' to be learned
—so 'performance measure' can be improved
2. but no programmable (easy) definition
—so 'ML' is needed
3. somehow there is data about the pattern
—so ML has some 'inputs' to learn from

key essence: help decide whether to use ML

Snapshot Applications of Machine Learning

Communication

data → ML → skill

for 4G LTE communication

- data:
  - **channel information** (the channel matrix representing mutual information)
  - **configuration** (precoding, modulation, etc.) that reaches the highest throughput
- skill: predict **best configuration to the base station** in a new environment

ongoing work of my student Yi-An Lin
as intern @ MTK

Daily Needs: Food, Clothing, Housing, Transportation

data ⟶ ML ⟶ skill

1. Food (Sadilek et al., 2013)
   - data: Twitter data (words + location)
   - skill: tell food poisoning likeliness of restaurant properly
2. Clothing (Abu-Mostafa, 2012)
   - data: sales figures + client surveys
   - skill: give good fashion recommendations to clients
3. Housing (Tsanas and Xifara, 2012)
   - data: characteristics of buildings and their energy load
   - skill: predict energy load of other buildings closely
4. Transportation (Stallkamp et al., 2012)
   - data: some traffic sign images and meanings
   - skill: recognize traffic signs accurately

ML is everywhere!

Education

data $\longrightarrow$ $\boxed{\text{ML}}$ $\longrightarrow$ skill

- data: students' records on quizzes on a Math tutoring system
- skill: predict whether a student can give a correct answer to another quiz question

A Possible ML Solution

answer correctly $\approx$ $\llbracket$recent strength of student $>$ difficulty of question$\rrbracket$

- give ML 9 million records from 3000 students
- ML determines (**reverse-engineers**) strength and difficulty automatically

key part of the **world-champion** system from National Taiwan Univ. in KDDCup 2010

Entertainment: Recommender System (1/2)

data $\longrightarrow$ ML $\longrightarrow$ skill

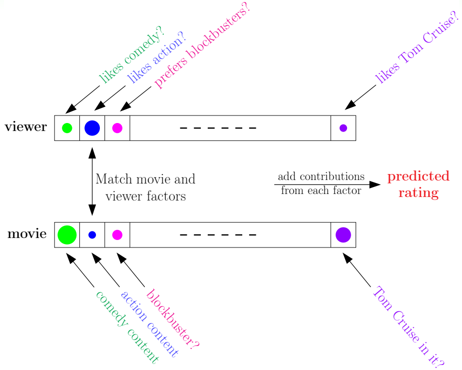
- data: how many users have rated some movies
- skill: predict how a user would rate an unrated movie

A Hot Problem

- competition held by Netflix in 2006
  - 100,480,507 ratings that 480,189 users gave to 17,770 movies
  - 10% improvement = **1 million dollar prize**
- similar competition (movies $\rightarrow$ songs) held by Yahoo! in KDDCup 2011
  - 252,800,275 ratings that 1,000,990 users gave to 624,961 songs

How can machines **learn our preferences**?

Entertainment: Recommender System (2/2)



A Possible ML Solution

- pattern:
  rating ← viewer/movie factors
- learning:
  known rating
  → learned factors
  → unknown rating prediction

key part of the **world-champion** (again!) system from National Taiwan Univ. in KDDCup 2011

Components of Machine Learning

Components of Learning: Metaphor Using Credit Approval

Applicant Information

age	23 years
gender	female
annual salary	NTD 1,000,000
year in residence	1 year
year in job	0.5 year
current debt	200,000

unknown pattern to be learned:

'approve credit card good for bank?'

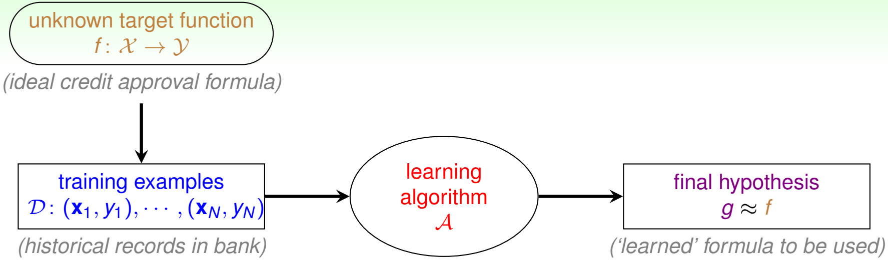
Formalize the Learning Problem

Basic Notations

- input: $\mathbf{x} \in \mathcal{X}$ (customer application)
- output: $y \in \mathcal{Y}$ (good/bad after approving credit card)
- unknown pattern to be learned $\Leftrightarrow$ target function: $f\colon \mathcal{X} \to \mathcal{Y}$ (ideal credit approval formula)
- data $\Leftrightarrow$ training examples: $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)\}$ (historical records in bank)
- hypothesis $\Leftrightarrow$ skill with hopefully good performance: $g\colon \mathcal{X} \to \mathcal{Y}$ ('learned' formula to be used)

$$\{(\mathbf{x}_n, y_n)\} \text{ from } f \longrightarrow \boxed{\text{ML}} \longrightarrow g$$

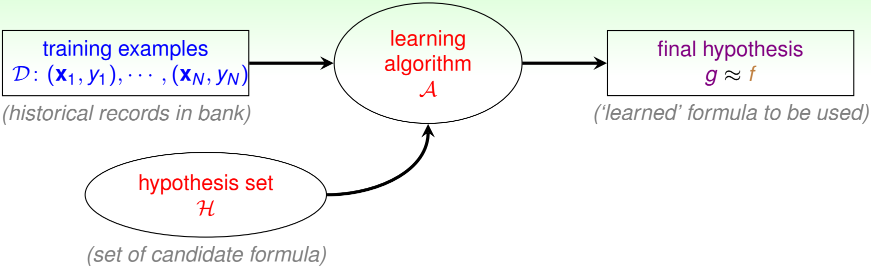
Learning Flow for Credit Approval



- target f **unknown**
  (i.e. no programmable definition)
- hypothesis g hopefully $\approx f$
  but possibly **different** from f
  (perfection 'impossible' when f unknown)

What does g look like?

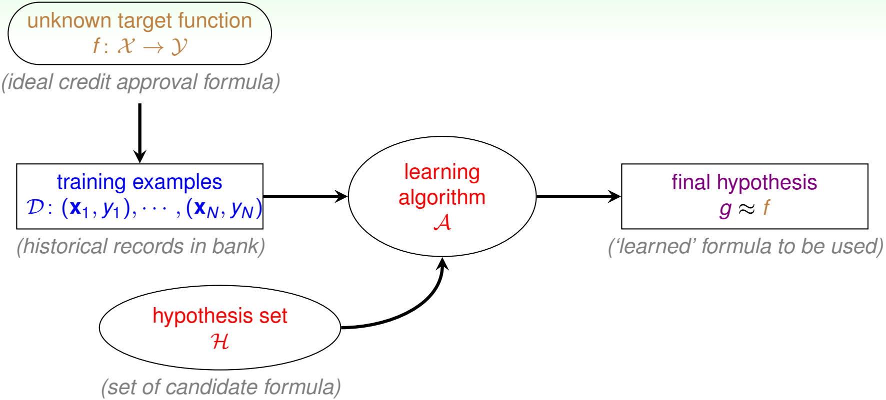
The Learning Model



- assume $g \in \mathcal{H} = \{h_k\}$, i.e. approving if
  - h_1: annual salary > NTD 800,000
  - h_2: debt > NTD 100,000 (really?)
  - h_3: year in job ≤ 2 (really?)
- hypothesis set $\mathcal{H}$:
  - can contain **good or bad hypotheses**
  - up to $\mathcal{A}$ to pick the 'best' one as g

learning model = $\mathcal{A}$ and $\mathcal{H}$

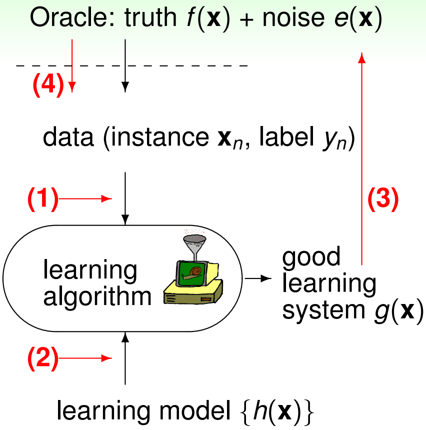
Practical Definition of Machine Learning



machine learning:
use data to compute hypothesis g
that approximates target f

Machine Learning Research in CLLab

Making Machine Learning **Realistic**: Now



CLLab Works: **Loosen the Limits of ML**

1. cost-sensitive classification: limited protocol (classification) + **auxiliary info. (cost)**
2. multi-label classification: limited protocol (classification) + **structure info. (label relation)**
3. active learning: limited protocol (unlabeled data) + **requested info. (query)**
4. online learning: limited protocol (streaming data) + **feedback info. (loss)**

next: (1) cost-sensitive classification

Which Digit Did You Write?

?

one (1)

two (2)

three (3)

a **classification** problem
—grouping "pictures" into different "categories"

Traditional Classification Problem

Oracle: truth $f(\mathbf{x})$ + noise $e(\mathbf{x})$

data (instance $\mathbf{x}_n$, label y_n)

learning algorithm  → good learning system $g(\mathbf{x})$

learning model $\{g_\alpha(\mathbf{x})\}$

1. input: a batch of examples (digit $\mathbf{x}_n$, intended label y_n)
2. desired output: some $g(\mathbf{x})$ such that $g(\mathbf{x}) \neq y$ **seldom** for future examples $(\mathbf{x}, y)$
3. evaluation for some digit

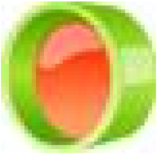
$(\mathbf{x} =$  $, y = 2)$

$$—g(\mathbf{x}) = \begin{cases} 1 : \textit{wrong}; \\ 2 : \textit{right}; \\ 3 : \textit{wrong} \end{cases}$$

Are all the **wrong**s equally bad?

What is the Status of the Patient?

?



H1N1-infected

cold-infected

healthy

another **classification** problem
—grouping “patients” into different “status”

Patient Status Prediction

error measure = society cost

actual \ predicted	H1N1	cold	healthy
H1N1	0	1000	**100000**
cold	100	0	3000
healthy	100	30	0

- H1N1 mis-predicted as healthy: **very high cost**
- cold mis-predicted as healthy: high cost
- cold correctly predicted as cold: no cost

human doctors consider costs of decision; **can computer-aided diagnosis do the same**?

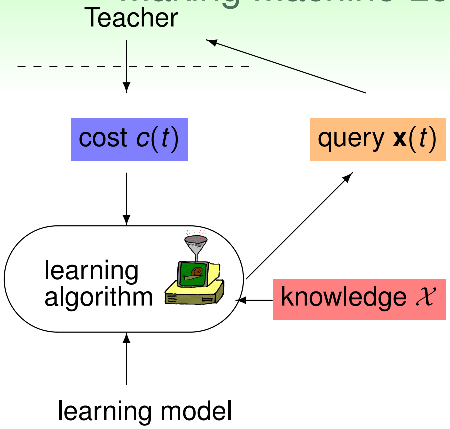
Our Contributions

	binary	multiclass
regular	well-studied	well-studied
cost-sensitive	known (Zadrozny, 2003)	ongoing (our works)

theoretic, algorithmic and empirical studies of cost-sensitive classification

- ICML 2010: a theoretically-supported algorithm with **superior experimental results**
- BIBM 2011: application to real-world **bacteria classification** with promising experimental results
- KDD 2012: a cost-sensitive **and error-sensitive** methodology (achieving both low cost and **few wrongs**)

Making Machine Learning Realistic: **Next**



Interactive Machine Learning

1. environment
2. exploration
3. dynamic
4. partial feedback

let us teach machines as "easily" as teaching students

Case: Interactive Learning for Online Advertisement

Traditional Machine Learning for Online Advertisement

- data gathering: system **randomly shows ads** to **some previous users**
- expert building: system **analyzes data gathered** to **determine best (fixed) strategy**

Interactive Machine Learning for Online Advertisement

- environment : system serves **online users with profile**
- exploration : system **decides to show an ad** to the user
- dynamic : system receives data from **real-time user click**
- partial feedback : system receives **reward only if clicking**

ICML 2012 Exploration & Exploitation Challenge

Interactive Machine Learning for Online Advertisement

- environment : system serves **online users with profile**
- exploration : system **decides to show an ad** to the user
- dynamic : system receives data from **real-time user click**
- partial feedback : system receives **reward only if clicking**

NTU beats two MIT teams to be the phase 1 winner!

NAME	AFFILIATION	LAST SCORE (CTR * 10 000)	BEST SCORE (CTR * 10 000)	RANK
Ku-Chun	NTU	882.9	**905.9**	1
tvirot	MIT	903.9	**903.9**	2
edjoesu	MIT	889.9	**903.4**	3

ongoing collaboration with **Appier** for online advertisement

More on KDDCup

What is KDDCup?

Background

- an annual competition on KDD (knowledge discovery and data mining)
- organized by ACM SIGKDD, starting from 1997, now **the most prestigious data mining competition**
- usually lasts 3-4 months
- participants include famous research labs (IBM, AT&T) and top universities (Stanford, Berkeley)

Aim of KDDCup

Aim

- bridge the gap between theory and **practice**, such as
  - scalability and efficiency
  - missing data and noise
  - heterogeneous data
  - unbalanced data
  - combination of different models
- define the **state-of-the-art**

KDDCups: 2008 to 2013 I

2008

- organizer: Siemens
- topic: breast cancer prediction (medical)
- data size: 0.2M
- teams: > 200
- NTU: **co-champion** with IBM (led by Prof. Shou-de Lin)

2009

- organizer: Orange
- topic: customer behavior prediction (business)
- data size: 0.1M
- teams: > 400
- NTU: **3rd place** of slow track

KDDCups: 2008 to 2013 II

2010

- organizer: PSLC Data Shop
- topic: student performance prediction (education)
- data size: 30M
- teams: > 100
- NTU: **champion** and **student-team champion**

2011

- organizer: Yahoo!
- topic: music preference prediction (recommendation)
- data size: 300M
- teams: > 1000
- NTU: **double champions**

KDDCups: 2008 to 2013 III

2012

- organizer: Tencent
- topic: webuser behavior prediction (Internet)
- data size: 150M
- teams: > 800
- NTU: **champion of track 2**

2013

- organizer: Microsoft Research
- topic: paper-author relationship prediction (academia)
- data size: 600M
- teams: > 500
- NTU: **double champions**

KDDCup 2011

from

YAHOO! LABS

Music Recommendation Systems

- host: Yahoo!
- **11 years** of Yahoo! music data
- **2 tracks** of competition
- official dates: **March 15 to June 30**
- 1878 teams submitted to track 1;
  1854 teams submitted to track 2

NTU Team for KDDCup 2011

- 3 faculties:
  Profs. Chih-Jen Lin, Hsuan-Tien Lin and Shou-De Lin
- 1 course (starting in 2010)
  Data Mining and Machine Learning: Theory and Practice
- 3 TAs and 19 students:
  most were **inexperienced in music recommendation in the beginning**
- official classes: April to June;
  actual classes: December to June

our motto: study state-of-the-art approaches
and then **creatively improve them**

Previously: How Much Did You Like These Movies?

http://www.netflix.com

(1M dollar competition between 2007-2009)

goal: use "movies you've rated" to automatically
predict your **preferences** on future movies

The Track 1 Problem (1/2)

Given Data

263M examples (user u, item i, rating r_{ui}, date t_{ui}, time τ_{ui})

user	**item**	**rating**	**date**	**time**
1	**21**	**10**	**102**	**23:52**
1	**213**	**90**	**1032**	**21:01**
4	**45**	**95**	**768**	**09:15**
...				

- u, i: abstract IDs
- r_{ui}: integer between 0 and 100, **mostly multiples of** 10

Additional Information: Item Hierarchy

- track (46.85%)
- album (19.01%)
- artist (28.84%)
- genre (5.30%)

The Track 1 Problem (2/2)

Data Partitioned by Organizers

- training: $253M$; validation: $4M$;
  test (w/o rating): $6M$
- per user, **training < validation < test in time**
  - $\geq$ 20 examples total
  - 4 examples in validation; 6 in test
- **fixed random half of test: leaderboard**;
  another half: award decision

Goal

predictions $\hat{r}_{ui} \approx r_{ui}$ on the test set, measured by

$$RMSE = \sqrt{\text{average}(\hat{r}_{ui} - r_{ui})^2}$$

— one submission allowed **every eight hours**

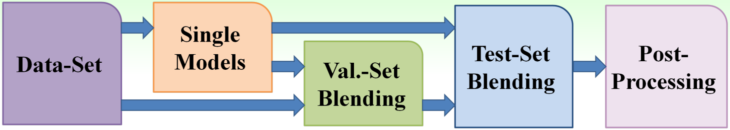
Three Properties of Track 1 Data

$$\mathbf{R} = \begin{array}{c|c|c|c|c|c|c} & \text{track}_1 & \text{track}_2 & \text{album}_3 & \text{author}_4 & \cdots & \text{genre}_I \\ \hline \text{user}_1 & 100 & 80 & 70 & ? & \cdots & - \\ \hline \text{user}_2 & - & 0 & ? & 80 & \cdots & - \\ \hline \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \hline \text{user}_U & ? & - & 20 & - & \cdots & 0 \end{array}$$

similar to Netflix data, but with the following differences......

- scale: larger data
  —study mature models that are **computationally feasible**
- taxonomy: relation graph of tracks, albums, authors and genres
  —**include as features** for combining models nonlinearly
- time: detailed; training earlier than test
  —**include as features** for combining models nonlinearly;
  respect time-closeness during training

Framework of Our Solution



System Architecture

- **improve standard models**: design **variants within 6 families of state-of-the-art models** (reaches RMSE 22.7915)
- **blend the models**: improve prediction power by **blending the variants carefully** (reaches RMSE 21.3598)
- **aggregate the blended predictors**: construct a linear ensemble with **test performance estimators** (reaches RMSE 21.0253)
- **post-process the ensemble**: add a final touch based on **observations from data analysis** (reaches RMSE 21.0147)

not only **hard work** (200+ models included), but also **key techniques**

That’s about all. Thank you!