CSED490Y: Optimization for Machine Learning

Week 02-2: Basics

Namhoon Lee

POSTECH

Spring 2022

Machine learning

Machine learning?

Machine learning

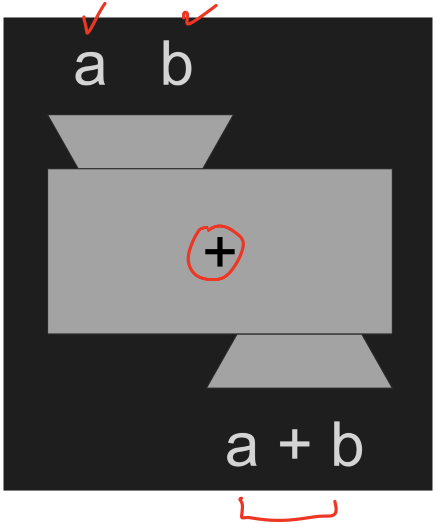
Machine learning?

- Machine learning gives computers the ability to learn without being explicitly programmed.

Learning to calculate

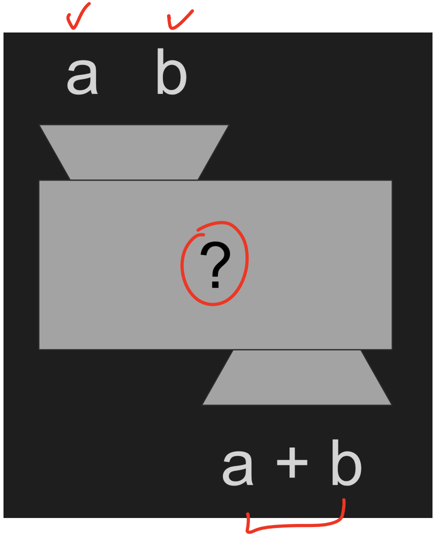
- How it works: An integrated circuit composed of transistors converts input numbers to binary strings, and performs a desired calculation by turning on and off transistors with electricity.
- Idea: Can we learn the ability to calculate? (by Hyung Jin Kim)

Learning to calculate



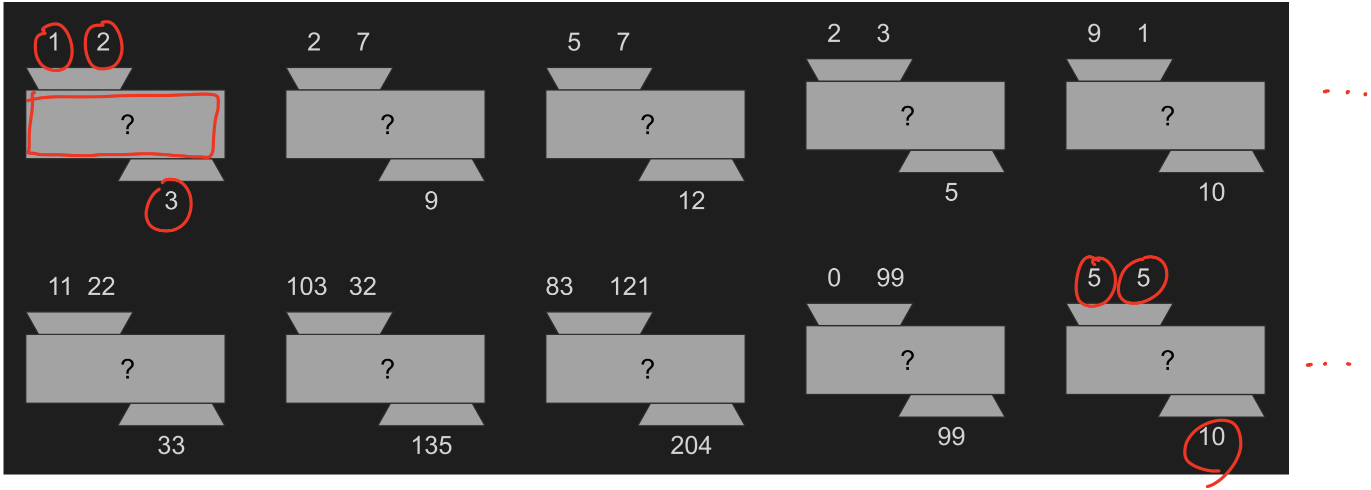
- For a "CS" approach, we assume that the logic of the operation to perform is known a priori, and therefore, it can be programmed.

Learning to calculate



- For an ML approach, we assume that we aren't given what is in the box, and therefore, it has to be figured out (from data).

Learning to calculate



Learning to calculate

hard-code

If you define "?"
→ CS

?

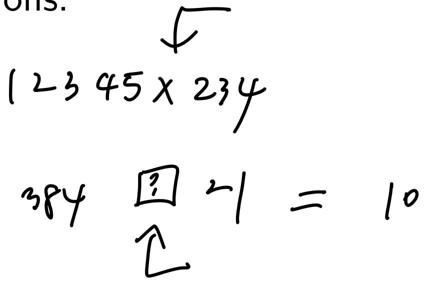
If a machine finds "?" from data
→ ML

Motivated by this idea ML is often considered as AI.

Machine learning

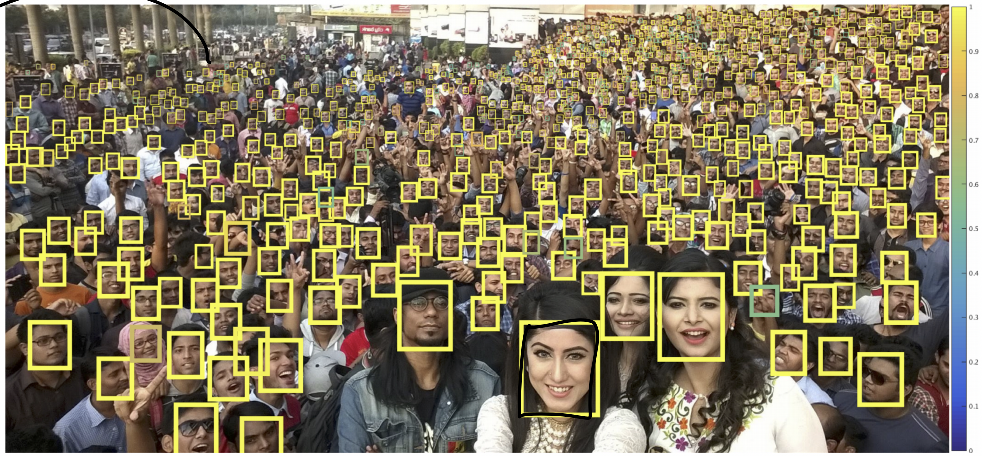
Machine learning?

- Machine learning gives computers the ability to learn without being explicitly programmed.
- Using computer to automatically detect patterns in data and use these to make predictions or decisions.



Face recognition

An automated face detection method developed at Carnegie Mellon University enables computers to recognize faces in images at a variety of scales, including tiny faces composed of just a handful of pixels (Byron Spice).



Finding tiny faces (Hu and Ramanan 2017)

Machine learning

Machine learning?

- Machine learning gives computers the ability to learn without being explicitly programmed.
- Using computer to automatically detect patterns in data and use these to make predictions or decisions.

(Tom Mitchell): "*A computer program is said to learn from experience $\mathcal{E}$ with respect to some class of tasks $\mathcal{T}$ and performance measure $\mathcal{P}$ if its performance at tasks in $\mathcal{T}$, as measured by $\mathcal{P}$, improves with experience $\mathcal{E}$.*"

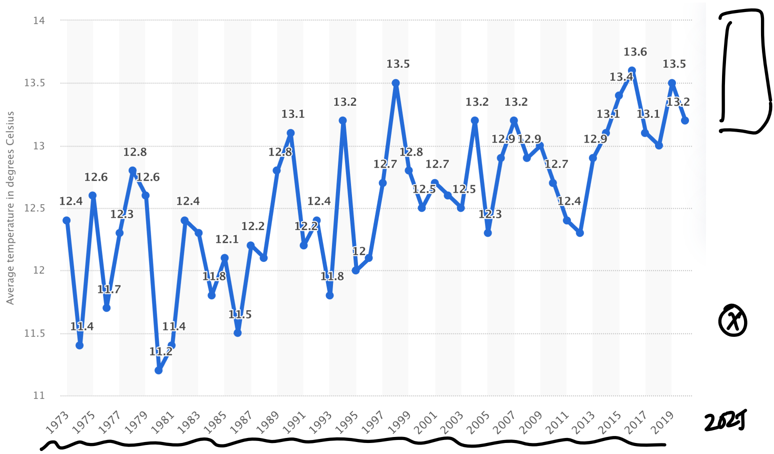
└ learning from data

Examples of machine learning tasks

- Spam filtering
- Weather forecasting
- Movie recommendation on Netflix
- Recognising faces from photos
- Translating English to Korean
- ✓ Discoverying new drugs
- Playing games

Examples of machine learning tasks

- Spam filtering
- Weather forecasting
- Movie recommendation on Netflix
- Recognising faces from photos
- Translating English to Korean
- Discoverying new drugs
- Playing games



Annual average temperature in South Korea from 1973 to 2020 (Statista 2021)

Relationship between numerical variables

We want to discover relationship between (numerical) variables.

- Does number of lung cancer deaths change with number of cigarettes?

Relationship between numerical variables

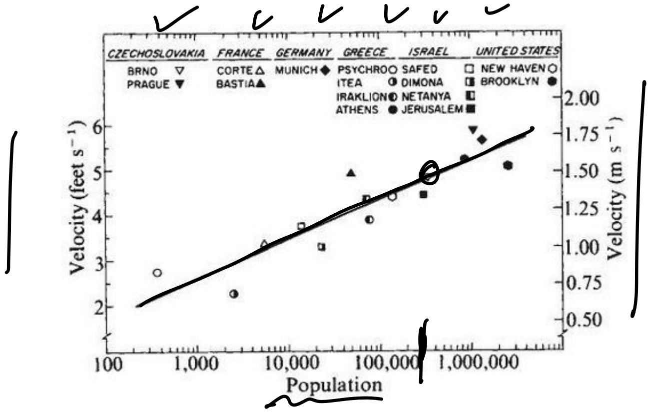
We want to discover relationship between numerical variables.

- Does number of lung cancer deaths change with number of cigarettes?
- Does number violent crimes change with violent video games?

Relationship between numerical variables

We want to discover relationship between numerical variables.

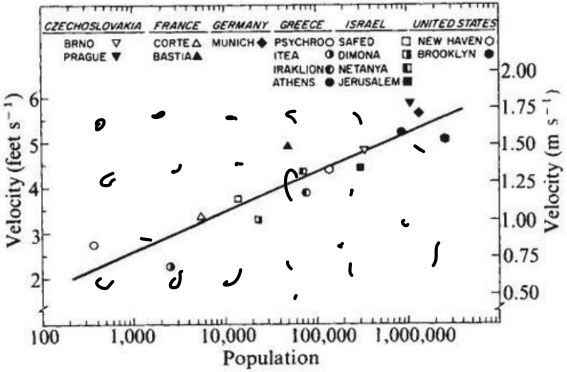
- Does number of lung cancer deaths change with number of cigarettes?
- Does number violent crimes change with violent video games?
- Do people in big cities walk faster?



Relationship between numerical variables

We want to discover relationship between numerical variables.

- Does number of lung cancer deaths change with number of cigarettes?
- Does number violent crimes change with violent video games?
- Do people in big cities walk faster?



- "Correlation does not imply causation".
  - (OK) "Higher velocity is correlated with higher population"
  - (BAD) "Higher population leads to higher velocity"

Scenario

Suppose a student is planning to take a machine learning course next semester and wondering how much time to study to receive good scores or grade. How do we address this problem?

Scenario

Suppose a student is planning to take a machine learning course next semester and wondering how much time to study to receive good scores or grade. How do we address this problem?

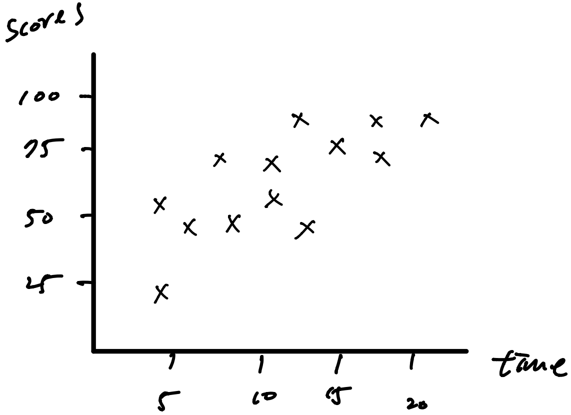
- An ML approach: collect data, train a prediction model, estimate scores.

Collect data

✓	✓	✓
Student ID	Time (hours)	Scores
1	5.3	70
2	2.3	62
3	11.8	88
4	4.9	67
5	15.1	93
..	..	..

Collecting students exam scores.

Collect data



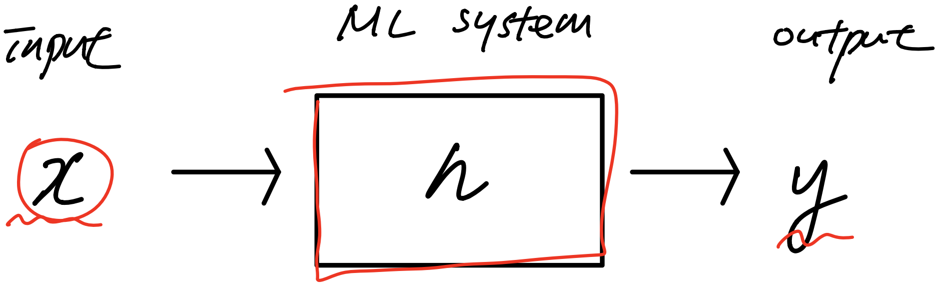
Plotting students exam scores.

Learning hypothesis function

How are we going to use such data?

Learning hypothesis function

How are we going to use such data?

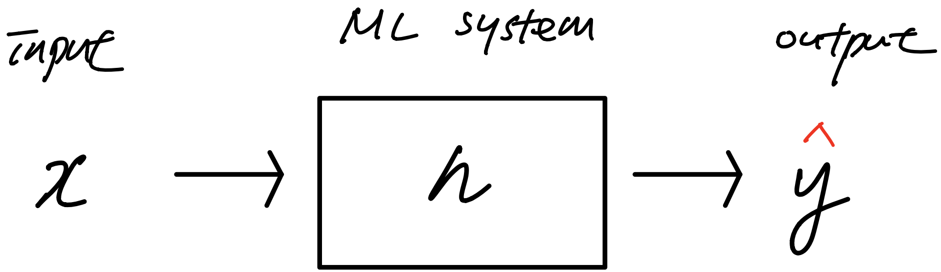


A typical ML system

- Learn a prediction function h which, given an input x, produces an output $h(x)$.

Learning hypothesis function

How are we going to use such data?

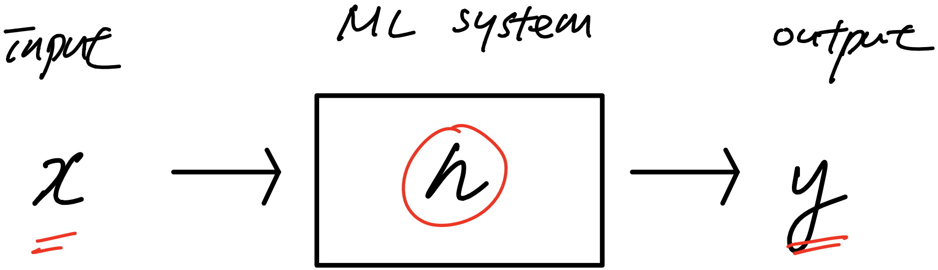


A typical ML system

- Learn a prediction function h which, given an input x, produces an output $h(x)$.
- For example, x is number of study hours, and $h(x)$ would be an estimate of scores.

Learning hypothesis function

How are we going to use such data?



A typical ML system

- Learn a prediction function h which, given an input x, produces an output $h(x)$.
- For example, x is number of study hours, and $h(x)$ would be an estimate of scores.
- But what hypothesis class h are we going to use?

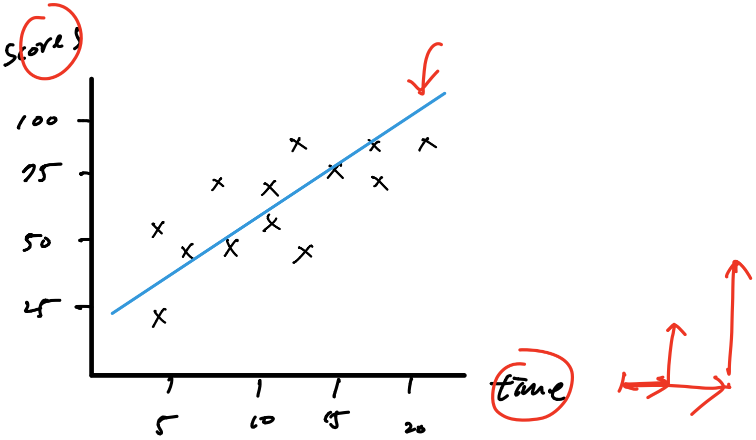
Linear hypothesis

Hypothesis of linear regression:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

- Linear regression is a machine learning model to solve a regression problem using a linear hypothesis.
- Here, $\theta = [\theta_0, \theta_1]^\top$ is a vector of parameters of the prediction function.
- Why do we use such model?

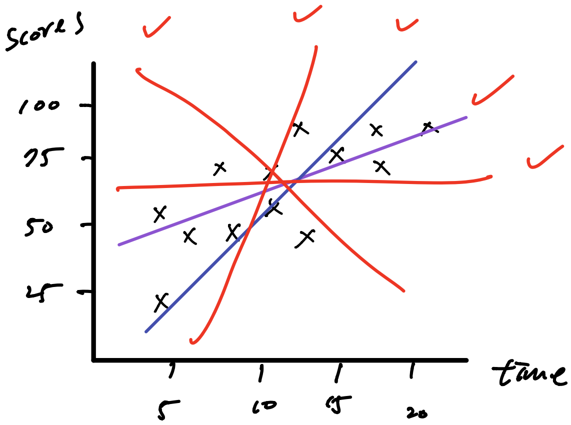
Fitting linear hypothesis



Fitting linear hypothesis

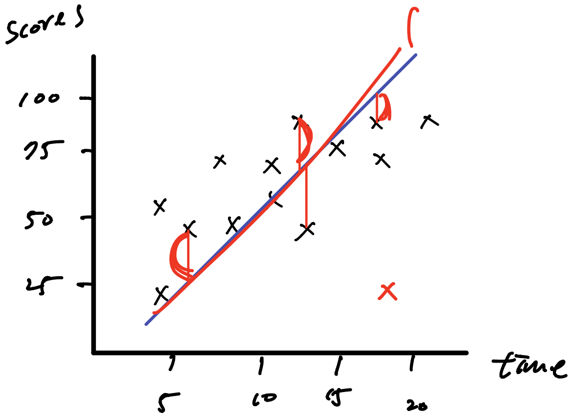
- What does this mean?

Many choices



Fitting many linear hypotheses

- Which one is better?



Measuring errors

- In order to evaluate how well a hypothesis fits data, we need to measure some erorrs as to how much it deviates from the true value.

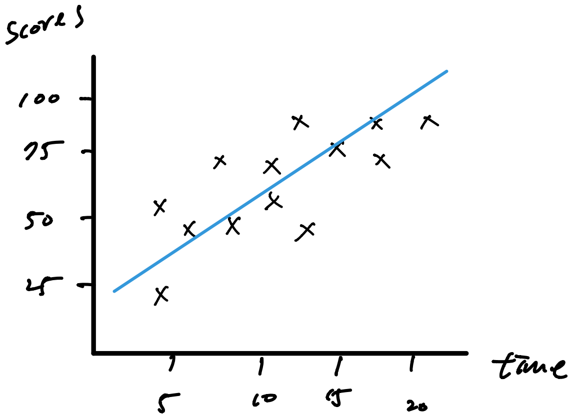
Least squares

One standard measure we can use is the squared error:

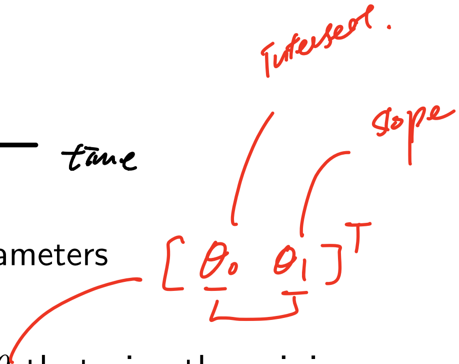
$$L_\theta = \sum_{i=1}^{n} (h_\theta(x_i) - y_i)^2$$

- We call it linear least squares.
- One can draw a probabilistic interpretation for this choice under some assumption.

Finding best hypothesis



Finding model parameters



- Find right values for the model parameters θ that give the minimum error.

Minimizing cost function

$\{x_i, y_i\}_{i=\{1,\dots,n\}} \Rightarrow$ data set

$h_\theta(x_i) = \hat{y}_i$

Find ones that minimize our cost function:

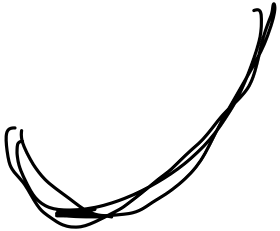
$(\hat{y}_i - y_i)^2$

$$\theta^\star = \arg\min_\theta \left[L_\theta \overset{def}{:=} \sum_i^n (h_\theta(x_i) - y_i)^2 \right]$$

$\theta^T x$

which is an optimization problem.

(min.)



$y = x^2$

$x = 0$

Thank you

Any questions?

Credits

A lot of material in this course is borrowed or derived from the following:

- Numerical Optimization, Jorge Nocedal and Stephen J. Wright.
- Convex Optimization, Stephen Boyd and Lieven Vandenberghe.
- Convex Optimization, Ryan Tibshirani.
- Optimization for Machine Learning, Martin Jaggi and Nicolas Flammarion.
- Optimization Algorithms, Constantine Caramanis.
- Advanced Machine Learning, Mark Schmidt.

References I

Hu, Peiyun and Deva Ramanan (2017). "Finding tiny faces". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.*