THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

DSAA2011 Machine Learning

L1 Logistics and Introduction

Dr. Zixin Zhong

Data Science and Analytics Thrust
Information Hub
Hong Kong University of Sceience and Technology (Guangzhou)

February 7, 2025

Outline

Outline

Course information

Weikai Yang
Assistant Professor

Visual analytics +
Machine learning

W1-316
weikaiyang@hkust-gz.edu.cn

Zixin Zhong
Assistant Professor

Reinforcement learning
Online learning

W1-308
zixinzhong@hkust-gz.edu.cn

- **Lecture time**: 12-3PM Fri
- **Lab time**: 8-9PM Fri
- **Venue**: Rm E1-134
- **Course material**:

  Slides, recommended materials

  Consultation: Office hour, email, appointment or Canvas, Teams

- **Teaching assistants (3/6)**:
  1. **Weiwen CHEN**
  2. **Guanghua LI**
  3. **Yang LUO**
  4. **Chunming MA**
  5. **Jingyi PAN**
  6. **Liangwei WANG**

Office hours (weekly, starting from 11 Feb)

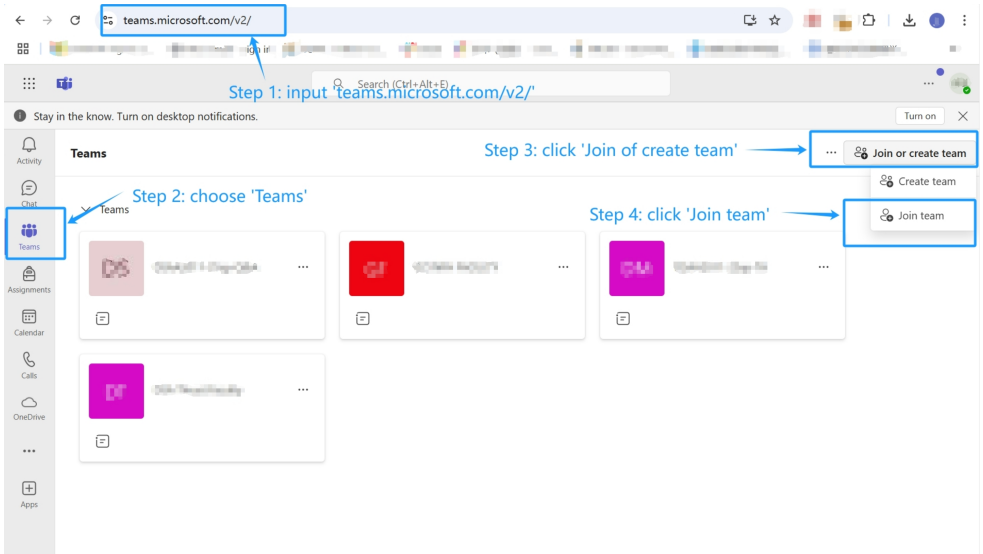
Time	Venue	Instructor/TA	Email
7-8PM Tue	Rm E2-301	Weiwen CHEN	wchen948@connect.hkust-gz.edu.cn
(7-8PM Wed)		Guanghua LI	gli945@connect.hkust-gz.edu.cn
		Yang LUO	yluo208@connect.hkust-gz.edu.cn
		Chunming MA	cma859@connect.hkust-gz.edu.cn
		Jingyi PAN	jpan305@connect.hkust-gz.edu.cn
		Liangwei WANG	lwang344@connect.hkust-gz.edu.cn
3-4PM Fri	Rm W1-316	Weikai Yang	weikaiyang@hkust-gz.edu.cn
	Rm W1-308	Zixin Zhong	zixinzhong@hkust-gz.edu.cn

♡ **Please show respect and appreciation to our TAs!**

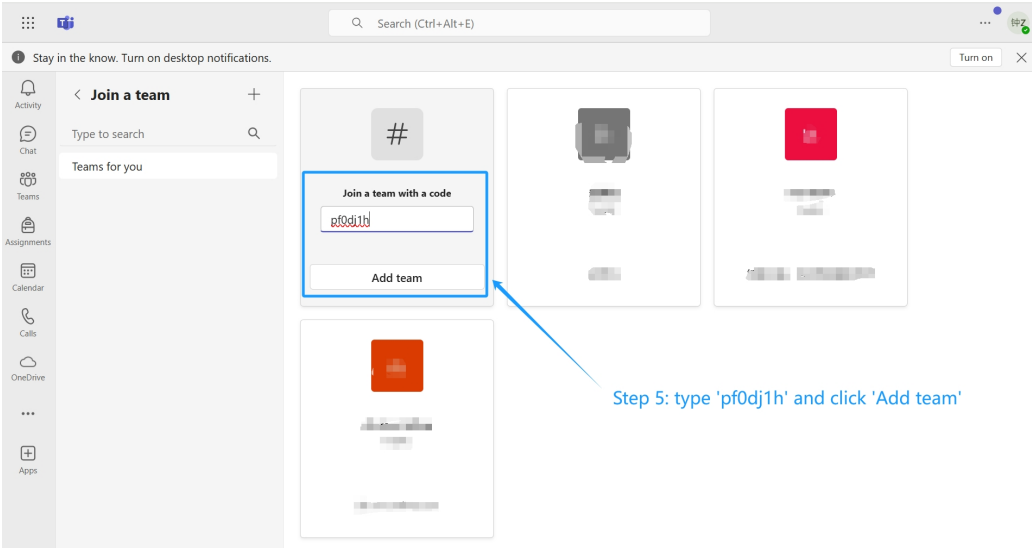
Q&A

- Office hour, **email**, appointment or Canvas, Microsoft Teams (code: pf0dj1h)

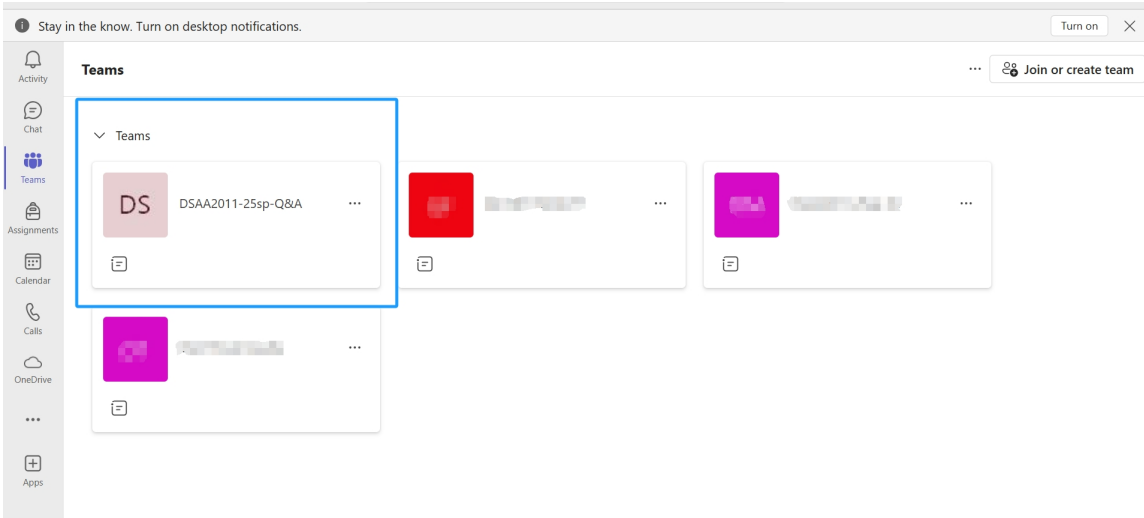
Q&A via Microsoft Teams: https://teams.microsoft.com/v2/



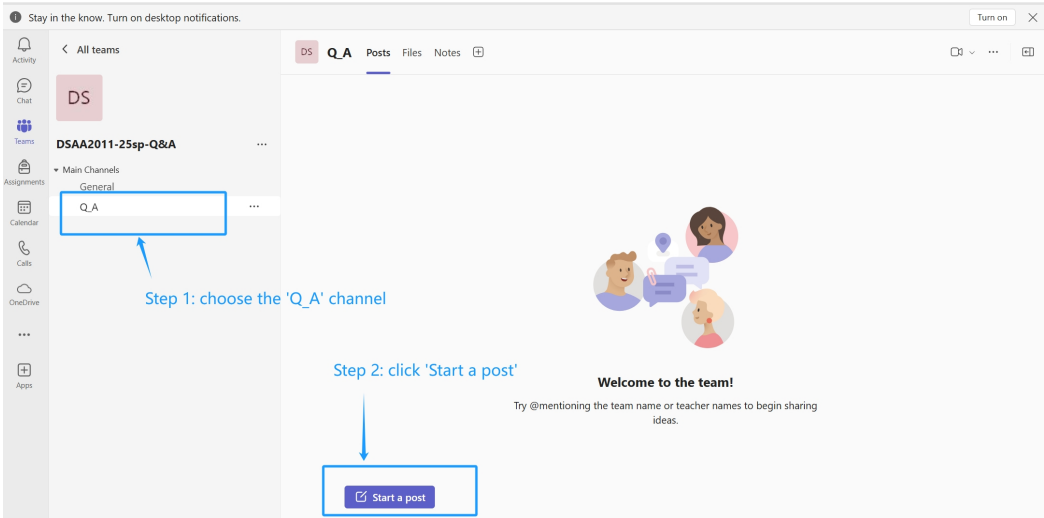
Q&A via Microsoft Teams: join with code pf0dj1h



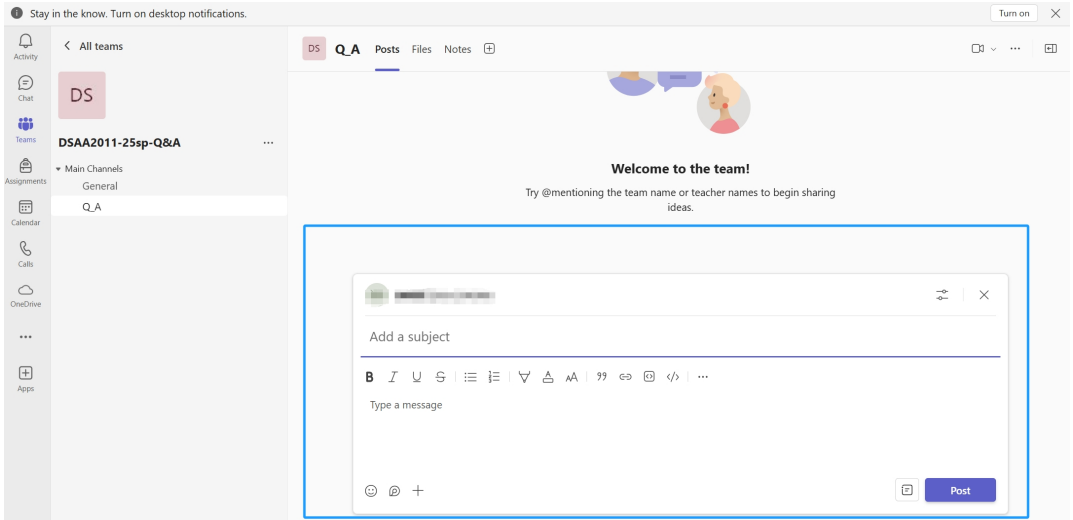
Q&A via Microsoft Teams: join with code pf0dj1h



Q&A via Microsoft Teams: ask questions in the 'Q_A' channel



Q&A via Microsoft Teams: ask questions in the 'Q_A' channel



Grading scheme

- **Midterm 20 Points**
- **Written assignments (Project) 30 Points**
- **Final exam 50 Points**

⊙ More details will be released later

Letter Grade	Raw Mark Range
A	$[85, 100]$
B	$[70, 85)$
C	$[55, 70)$
D	$[40, 55)$
E	< 40

Syllabus

Week #	Topic	Lecturer
1	Introduction + course Info	Zixin
2-3	Supervised learning: regression and classification	Zixin
4	Model evaluation and choice	Zixin
5	Feature selection	Zixin
6	Boosting methods	Weikai
	Midterm-29 March (Sat): save your day and mark it on calendar!	
7-8	Unsupervised learning: clustering	Weikai
9	Active learning	Weikai
10-11	Markov and graphical models	Weikai
12-13	Online learning	Zixin
14	Final exam	

Programming language policy

- We will do demos in **Python (Jupyter Lab)**.
- You're welcome to use any language you like (that your TAs can read) for homework or project.
- TAs will only support **Python**.

Collaboration policy

Project: yes of course! Clarify the contributions of each member.

- Give credit to the people who have helped you: write on your report the names of the people you worked with.
- Give credit to the other resources that have helped you: please write on your report the GenAI tools, textbooks, notes, or web pages you found useful.

Midterm/Final: Figure out your solutions by yourself.

- All of the text that you submit should be done by you.
- No collaboration/discussion here!

IP policy

Other course note websites:

- do not post any course materials in other websites. this makes the next rendition of the course worse for everyone.
- please report to us any course materials you find online (not on our websites).

How to do well in this course?

- Pay attention to lecture and lab
- Ask whenever
- Do exercise in lecture notes and lab notes
  - Treat each lab note as a quiz (good practise)
- Find a peer, describe a concept or teach each other
- Practice all you learn in the project

How to do well in this course?

- Pay attention to lecture and lab
- Ask whenever
- Do exercise in lecture notes and lab notes
  - Treat each lab note as a quiz (good practise)
- Find a peer, describe a concept or teach each other
- Practice all you learn in the project
- Be perserverant:

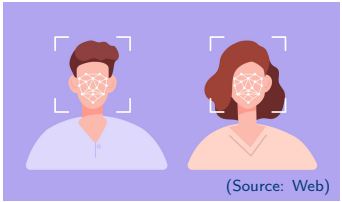
  "The brick walls are there for a reason. The brick walls are not there to keep us out. The brick walls are there to give us a chance to show how badly we want something. Because the brick walls are there to stop the people who don't want it badly enough. They're there to stop the other people."

  from *The Last Lecture* by Randy Pausch
- Brain power is like muscle power. It can only get better through heavy lifting.

Everyone can become smarter by training (practicing, analyzing, coding, ...)

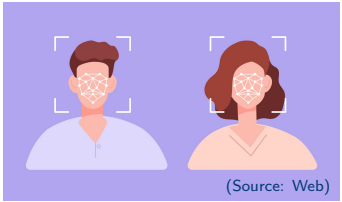
Outline

Machine Learning is everywhere



Identity verification

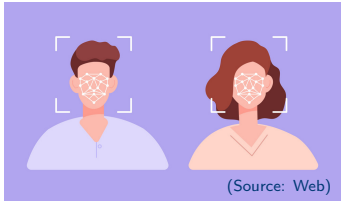
Machine Learning is everywhere



Identity verification

Large Language Model

Machine Learning is everywhere



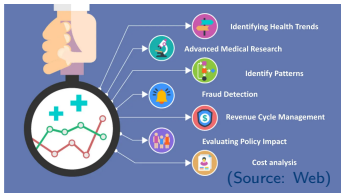
Identity verification

香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

HKUST(GZ) GPT Service System

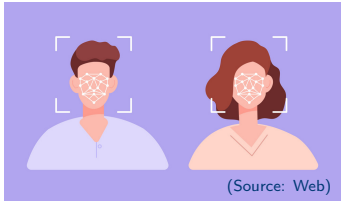
香港科技大学（广州）GPT服务系统

Large Language Model



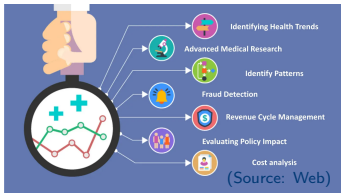
Healthcare

Machine Learning is everywhere

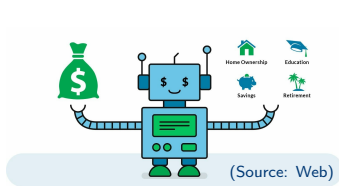


Identity verification

Large Language Model



Healthcare



Wealth management

What is Machine Learning (ML)?

Definition of Machine Learning

Prof. Tom M. Mitchell from CMU (Samuel, 1959):

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Definition of Machine Learning

Prof. Tom M. Mitchell from CMU (Samuel, 1959):

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

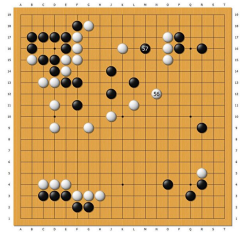
Definition of Machine Learning

Prof. Tom M. Mitchell from CMU (Samuel, 1959):

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.


(Source: Web)

- Experience E (data): games played by the program or human



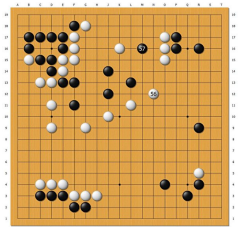
Definition of Machine Learning

Prof. Tom M. Mitchell from CMU (Samuel, 1959):

- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.


(Source: Web)

- Experience E (data): games played by the program or human
- Performance measure P: winning rate



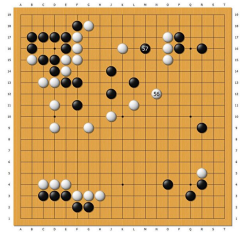
Definition of Machine Learning

Prof. Tom M. Mitchell from CMU (Samuel, 1959):

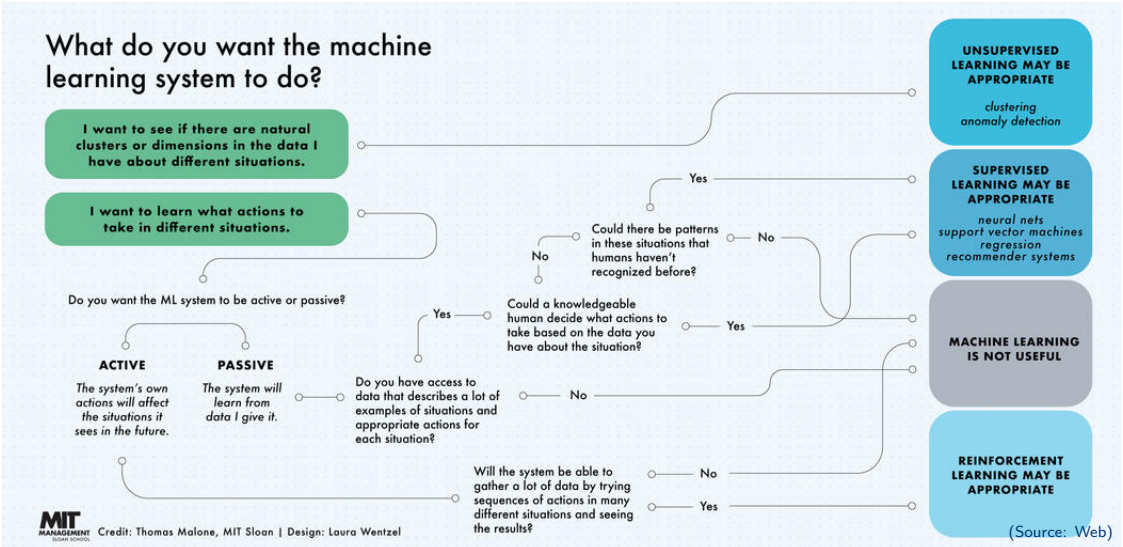
- A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.


(Source: Web)

- Experience E (data): games played by the program or human
- Performance measure P: winning rate
- Task T: to win



Taxonomy of Machine Learning



(Source: Web)

Taxonomy of Machine Learning

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Has outcome information (“labels”)	No outcome information available	Makes decisions based on trial and error
Finds patterns that relate to those outcomes	Analyzes or identifies groups without labels or human instruction	Decision-making algorithm is constantly refined based on “rewards”
Uses patterns to predict outcomes not yet known	Offers insights into characteristics that define groups	Excels in complex situations

Pecan

(Source: Web)

Taxonomy of Machine Learning (A Simplistic View)

- What type of data?
  - **Supervised learning** - labeled data: e.g., Prediction
  - Semi-supervised learning
  - **Unsupervised learning** - unlabeled data: e.g., clustering

Taxonomy of Machine Learning (A Simplistic View)

- What type of data?
  - **Supervised learning** - labeled data: e.g., Prediction
  - Semi-supervised learning
  - **Unsupervised learning** - unlabeled data: e.g., clustering
  - **Reinforcement learning** - environment feedback: e.g., multi-armed bandit

Taxonomy of Machine Learning (A Simplistic View)

f(x)≈y,
x is input data;
y is output data, denoted as 'label';

- What type of data?
  - **Supervised learning** - labeled data: e.g., Prediction -> all labelled
  - Semi-supervised learning -> partly labelled
  - **Unsupervised learning** - unlabeled data: e.g., clustering -> no labelled
  - **Reinforcement learning** - environment feedback: e.g., multi-armed bandit

classification

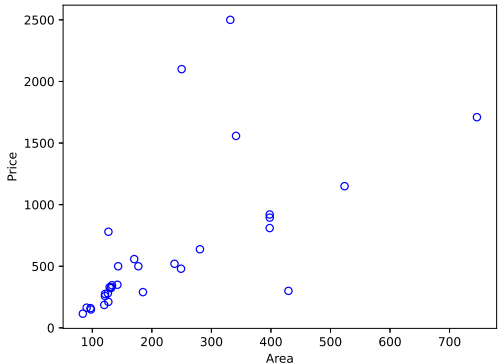
- When do we collect data?
  - **Offline learning**
  - **Online learning**

Supervised Learning

Task: housing price prediction

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

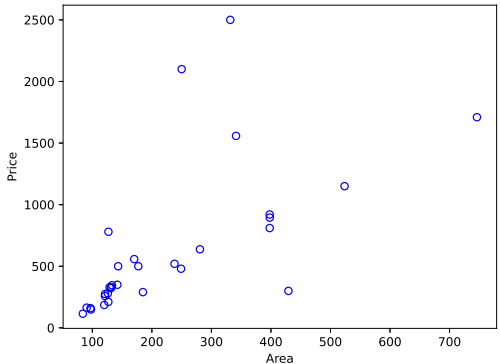


Task: housing price prediction

- Given a dataset containing n samples

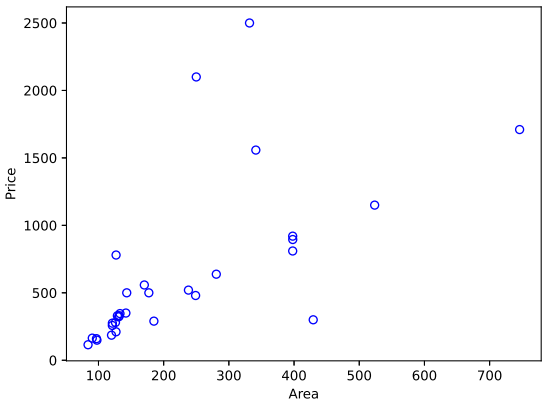
$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task: if a residence has $x = 320$ square kilometers, can we predict its price?



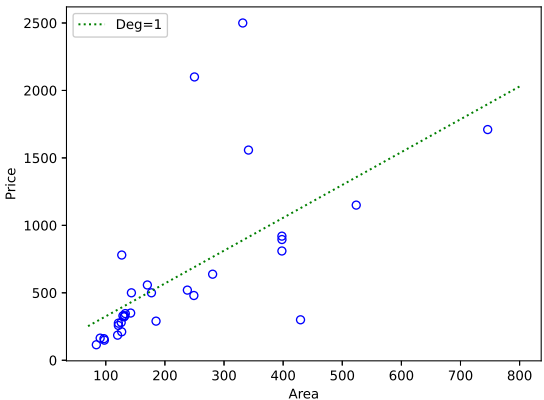
Task: housing price prediction

- Linear regression (polynomial regression with degree 1):
  Assume $y = p_1(x) = ax + b$ and guess the values of a and b.



Task: housing price prediction

- Linear regression (polynomial regression with degree 1):
  Assume $y = p_1(x) = ax + b$ and guess the values of a and b.

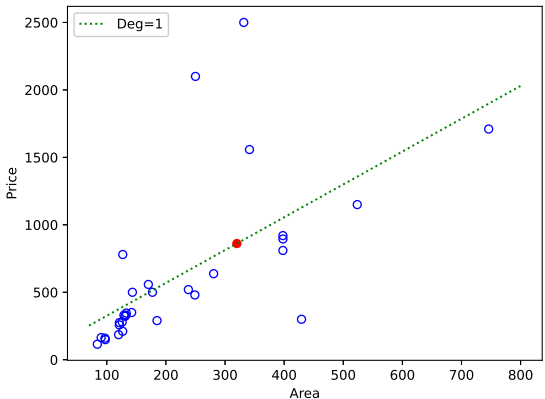


$a = 2.43, b = 82.36$

$p_1(x) = 2.43x + 82.36$

Task: housing price prediction

- Linear regression (polynomial regression with degree 1):
  Assume $y = p_1(x) = ax + b$ and guess the values of a and b.



$a = 2.43, b = 82.36$

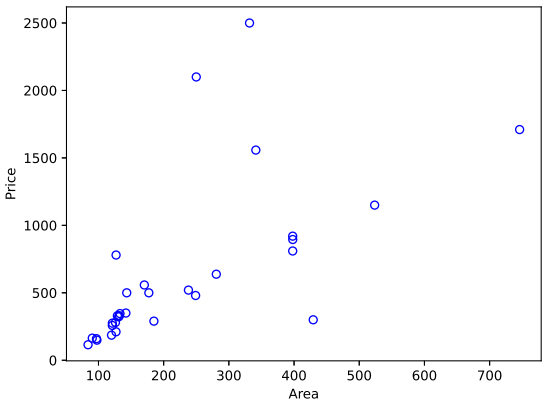
$p_1(x) = 2.43x + 82.36$

$p_1(320) = 861.29$

- If a residence has $x = 320$ square kilometers, we predict its price as $y = 861.29$ CNY.

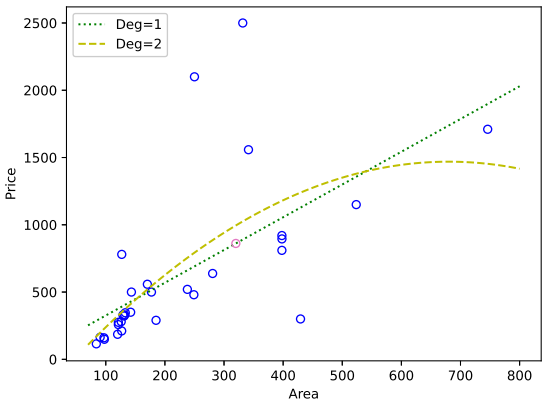
Task: housing price prediction

- Polynomial regression with degree 2:
  Assume $y = p_2(x) = ax^2 + bx + c$ and guess the values of a, b and c.



Task: housing price prediction

- Polynomial regression with degree 2:
  Assume $y = p_2(x) = ax^2 + bx + c$ and guess the values of a, b and c.

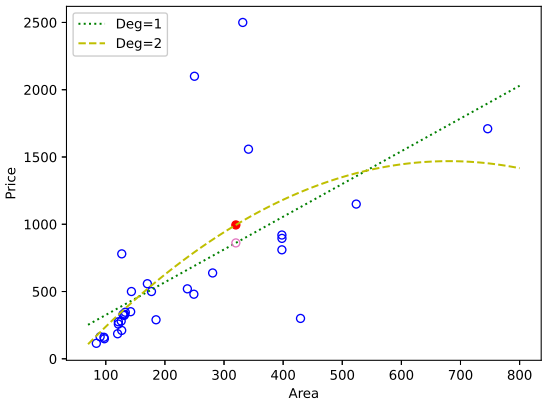


$a = -0.0037, b = 4.97, c = -221.8$

$p_2(x) = -0.0037x^2 + 4.97x - 221.8$

Task: housing price prediction

- Polynomial regression with degree 2:
  Assume $y = p_2(x) = ax^2 + bx + c$ and guess the values of a, b and c.



$a = -0.0037, b = 4.97, c = -221.8$

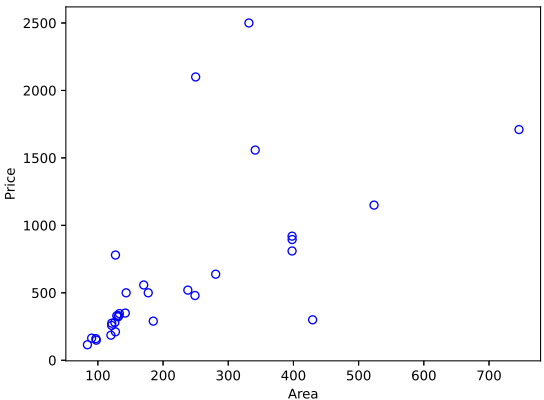
$p_2(x) = -0.0037x^2 + 4.97x - 221.8$

$p_2(320) = 994.94$

- If a residence has $x = 320$ square kilometers, we predict its price as $y = 994.94$ CNY.

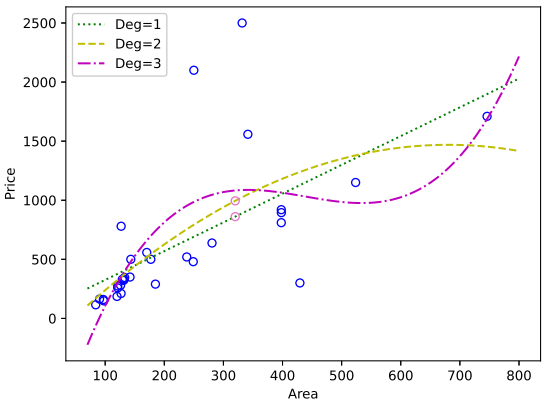
Task: housing price prediction

- Polynomial regression with degree 3:
  Assume $y = p_3(x) = ax^3 + bx^2 + cx + d$ and guess the values of a, b, c and d.



Task: housing price prediction

- Polynomial regression with degree 3:
  Assume $y = p_3(x) = ax^3 + bx^2 + cx + d$ and guess the values of a, b, c and d.



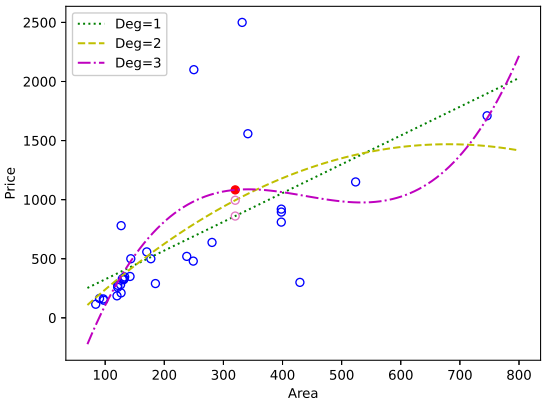
$a = 3.16 * 10^{-5}, b = -0.042,$

$c = 17.33, d = -1243$

$p_3(x) = 3.16 * 10^{-5}x - 0.042x + 17.33x - 1243$

Task: housing price prediction

- Polynomial regression with degree 3:
  Assume $y = p_3(x) = ax^3 + bx^2 + cx + d$ and guess the values of a, b, c and d.



$a = 3.16 * 10^{-5}, b = -0.042,$

$c = 17.33, d = -1243$

$p_3(x) = 3.16 * 10^{-5}x - 0.042x + 17.33x - 1243$

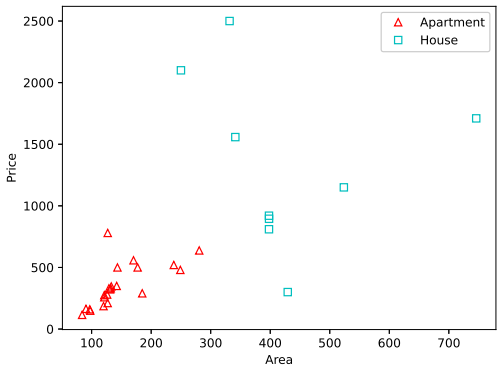
$p_3(320) = 1082.80$

- If a residence has $x = 320$ square kilometers, we predict its price as $y = 1082.80$ CNY.

Task: housing type prediction

- Given a dataset containing n samples ($z \in \{0, 1\}$)

$$\left(x^{(1)}, y^{(1)}, z^{(1)}\right), \left(x^{(2)}, y^{(2)}, z^{(2)}\right), \left(x^{(3)}, y^{(3)}, z^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}, z^{(n)}\right)$$

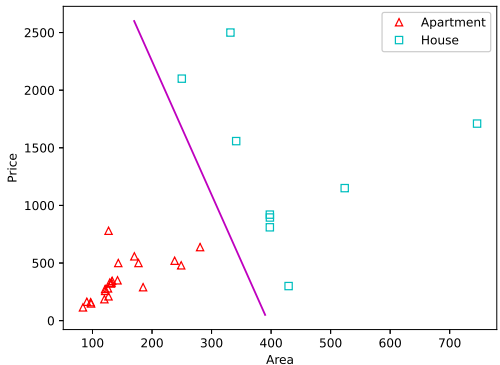


Task: housing type prediction

- Given a dataset containing n samples ($z \in \{0, 1\}$)

$$\left(x^{(1)}, y^{(1)}, z^{(1)}\right), \left(x^{(2)}, y^{(2)}, z^{(2)}\right), \left(x^{(3)}, y^{(3)}, z^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}, z^{(n)}\right)$$

- Task: if a residence has $x = 320$ square kilometers and worth $y = 1080.20$ CNY, can we predict its type z?

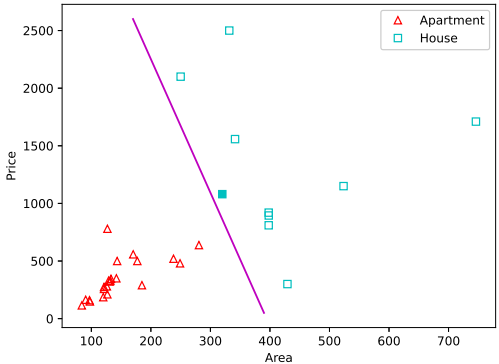


Task: housing type prediction

- Given a dataset containing n samples ($z \in \{0, 1\}$)

$$\left(x^{(1)}, y^{(1)}, z^{(1)}\right), \left(x^{(2)}, y^{(2)}, z^{(2)}\right), \left(x^{(3)}, y^{(3)}, z^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}, z^{(n)}\right)$$

- Task: if a residence has $x = 320$ square kilometers and worth $y = 1080.20$ CNY, can we predict its type z?



Task comparison: regression vs classification

- **Regression**: the prediction result is a continuous variable
  - e.g., price prediction
    $(x = \text{Area}) \rightarrow (y = \text{Price})$?
- **Classification**: the prediction result is a discrete variable
  - e.g., type prediction
    $(x = \text{Area}, y = \text{Price}) \rightarrow (z = \text{Type})$?

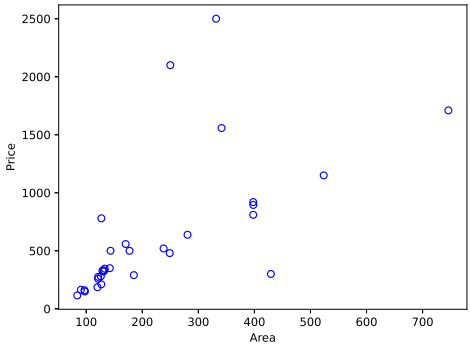
Unsupervised Learning

Task: housing type prediction

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task: can we learn the type of each residence?

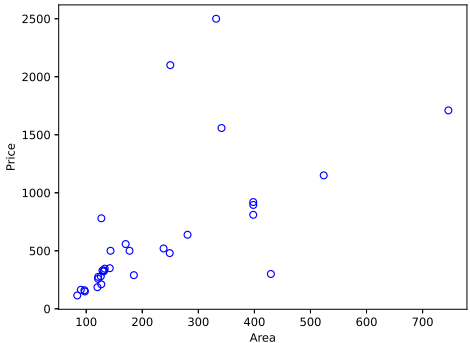


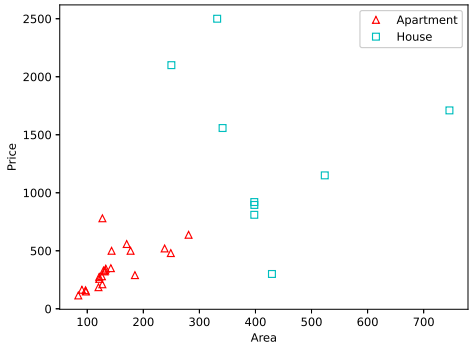
Task: housing type prediction

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task: can we learn the type of each residence?



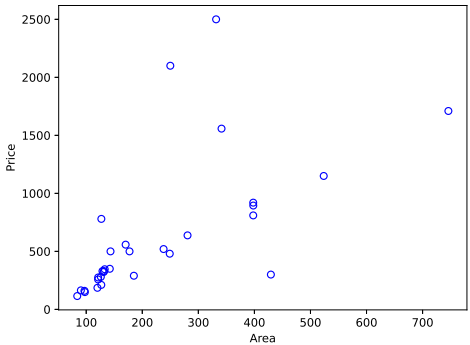


Clustering

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task (vague): find interesting structures in the data

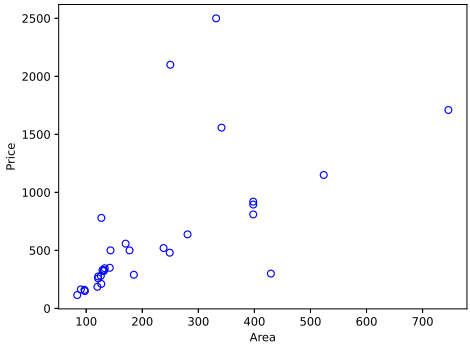


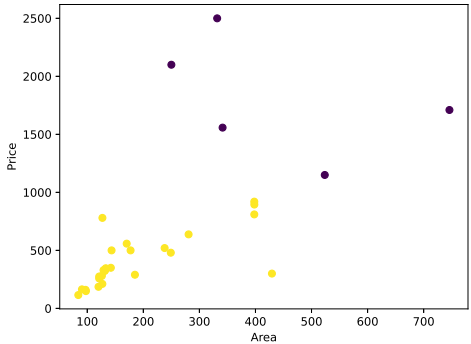
Clustering

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task (vague): find interesting structures in the data



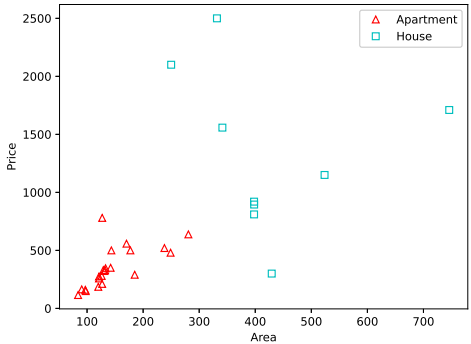


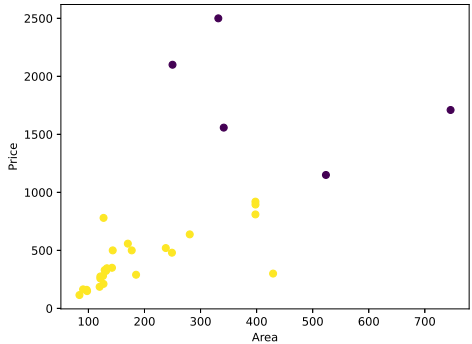
Clustering

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task (vague): find interesting structures in the data



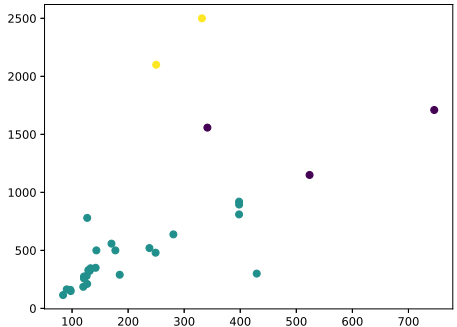


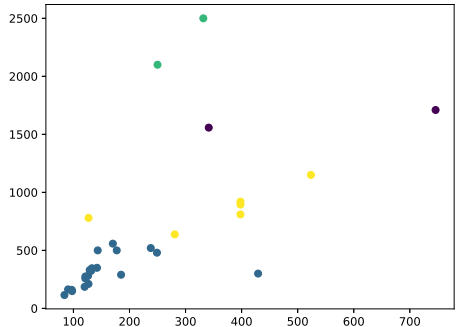
Clustering

- Given a dataset containing n samples

$$\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \left(x^{(3)}, y^{(3)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)$$

- Task (vague): find interesting structures in the data



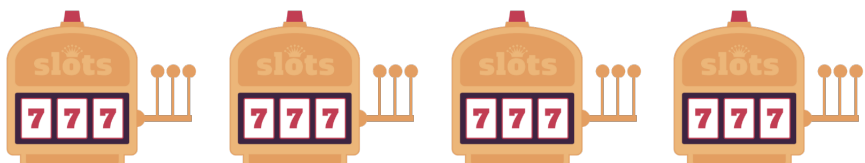


Reinforcement Learning

Task: win more

- We pull one of the 4 arms of a slot machine at each round t

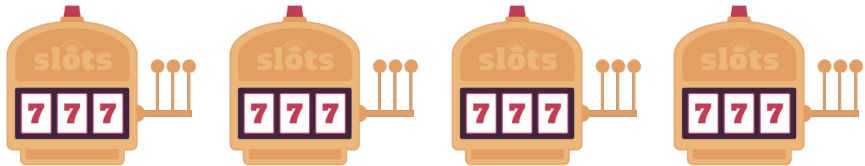
$L = 4$



Task: win more

- We pull one of the 4 arms of a slot machine at each round t
- Each arm i yields a random reward $W_t(i)$ at round t

$L = 4$

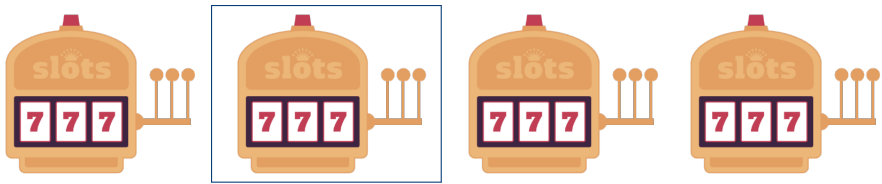


Reward $W_t(1)$ $W_t(2)$ $W_t(3)$ $W_t(4)$

Task: win more

- We pull one of the 4 arms of a slot machine at each round t
- Each arm i yields a random reward $W_t(i)$ at round t
- At time t, if we pull arm 2, we earn $W_t(2)$

$L = 4$

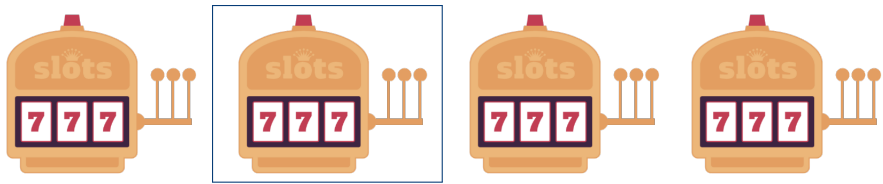


Reward	$W_t(1)$	$\boxed{W_t(2)}$	$W_t(3)$	$W_t(4)$

Task: win more

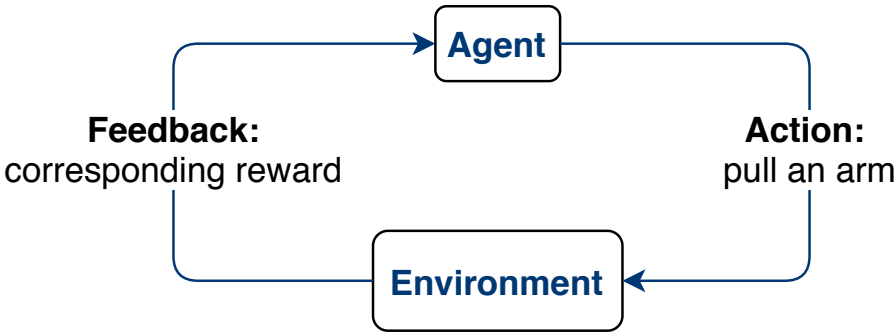
- We pull one of the 4 arms of a slot machine at each round t
- Each arm i yields a random reward $W_t(i)$ at round t
- At time t, if we pull arm 2, we earn $W_t(2)$
- Task: how can we earn as much as possible after 100 rounds?

$L = 4$

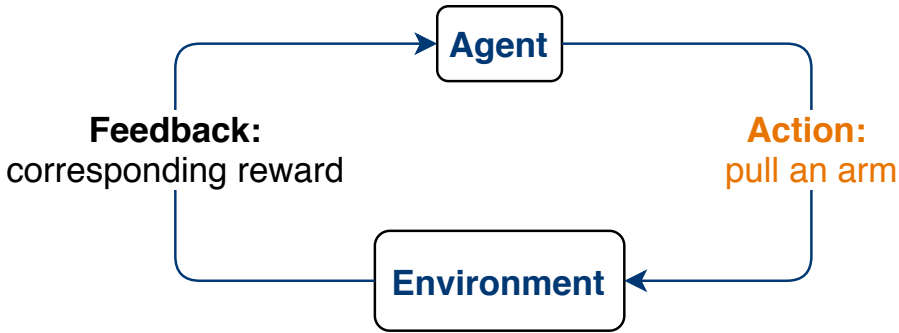


Reward	$W_t(1)$	$W_t(2)$	$W_t(3)$	$W_t(4)$

Multi-armed bandit problem (MAB)



Multi-armed bandit problem (MAB)



- Instead of labeled/unlabeled data, a reinforcement learning algorithm learns from environment feedback.

Taxonomy of Machine Learning (A Simplistic View)

- What type of data?
  - **Supervised learning** - labeled data: e.g., Prediction
  - Semi-supervised learning
  - **Unsupervised learning** - unlabeled data: e.g., clustering
  - **Reinforcement learning** - environment feedback: e.g., multi-armed bandit

- When do we collect data?
  - **Offline learning** provide all data at once
  - **Online learning** provide data step by step

both proper for supervised learning, it depends on the actual implementation

Outline

Sets

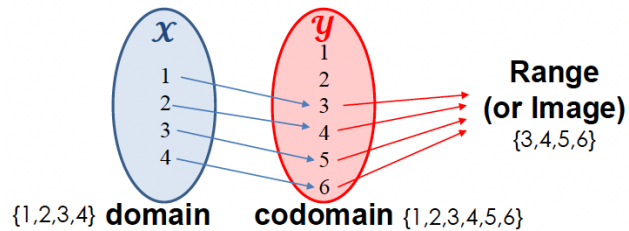
- A set S is an unordered collection of objects.
- $S = \{1, 2, 3, 4, 5, 6\}$ is the possible outcomes of the toss of a die.
- $S = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$ is the set of all numbers from a to b inclusive.
- $S = (a, b] = \{x \in \mathbb{R} : a < x \leq b\}$ is the set of all numbers from a to b, excluding a including b.
- $\mathbb{R}$ is the set of all real numbers.
- $\mathbb{R}^d$ is the set of all real vectors of length d.

Functions

- A function f is a map from a set X to another set Y. We write this as

$$f : X \to Y.$$

- For example, the function $f : \mathbb{R} \to [0, \infty)$ could be given by the recipe $f(x) = x^2$.
- The set of inputs is called the domain; the set of possible outputs is called the codomain; the set $\{f(x) : x \in X\}$ is called the range (or image).
- For example $f : \{1, 2, 3, 4\} \to \{1, 2, 3, 4, 5, 6\}$ given by the recipe $f(x) = x + 2$ has codomain $\{1, 2, 3, 4, 5, 6\}$ and range $\{3, 4, 5, 6\}$.



Functions

Definition 3.1

Let $f(x)$ be a continuous function defined on domain $\mathbb{R}$.

- $f(x)$ is **differentiable** if $f'(x) = \lim_{x \to c} \frac{f(x)-f(c)}{x-c}$ exists for all $x \in \mathbb{R}$.
- $f(x)$ is **convex** if $f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;
- $f(x)$ is **strictly convex** if $f(\lambda a + (1-\lambda)b) < \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;
- $f(x)$ is **concave** if $f(\lambda a + (1-\lambda)b) \geq \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;
- $f(x)$ is **strictly concave** if $f(\lambda a + (1-\lambda)b) > \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;

Functions

Definition 3.1

Let $f(x)$ be a continuous function defined on domain $\mathbb{R}$.

- $f(x)$ is **differentiable** if $f'(x) = \lim_{x \to c} \frac{f(x)-f(c)}{x-c}$ exists for all $x \in \mathbb{R}$.
- $f(x)$ is **convex** if $f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;
- $f(x)$ is **strictly convex** if $f(\lambda a + (1-\lambda)b) < \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;
- $f(x)$ is **concave** if $f(\lambda a + (1-\lambda)b) \geq \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;
- $f(x)$ is **strictly concave** if $f(\lambda a + (1-\lambda)b) > \lambda f(a) + (1-\lambda) f(b)$ for all $a, b \in \mathbb{R}$;

- $f(x)$ is differentiable and convex: x_0 such that $f'(x_0) = 0$ minimize function f.
- $f(x)$ is differentiable and concave: x_0 such that $f'(x_0) = 0$ maximize function f.

Linear functions

A function $f : \mathbb{R}^d \to \mathbb{R}$ is linear if it satisfies

- **(Homogeneity)** For any vector $\mathbf{x} \in \mathbb{R}^d$ and scalar $a \in \mathbb{R}$,

$$f(a\,\mathbf{x}) = a\,f(\mathbf{x})$$

- **(Additivity)** For any two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$$

Note that a linear function $f : \mathbb{R}^d \to \mathbb{R}$ must pass through the origin, i.e., $f(\mathbf{0}) = 0$ where $\mathbf{0} \in \mathbb{R}^d$ is the zero vector in d dimensions. Why?

Affine functions

- An affine function f is a linear function plus possibly a constant.
- More precisely, a function $f : \mathbb{R}^d \to \mathbb{R}$ is affine if it can be expressed as

$$f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$$

  for some vector $\mathbf{a} \in \mathbb{R}^d$ and some scalar $b \in \mathbb{R}$.
- The scalar b is called the bias or offset.

Example: The following function $f : \mathbb{R}^2 \to \mathbb{R}$ is affine. Why?

$$f(\mathbf{x}) = f(x_1, x_2) = -x_1 + 3x_2 + 7.$$

Exercise: Is a linear function affine? Is an affine function linear?

Linear and Affine Functions

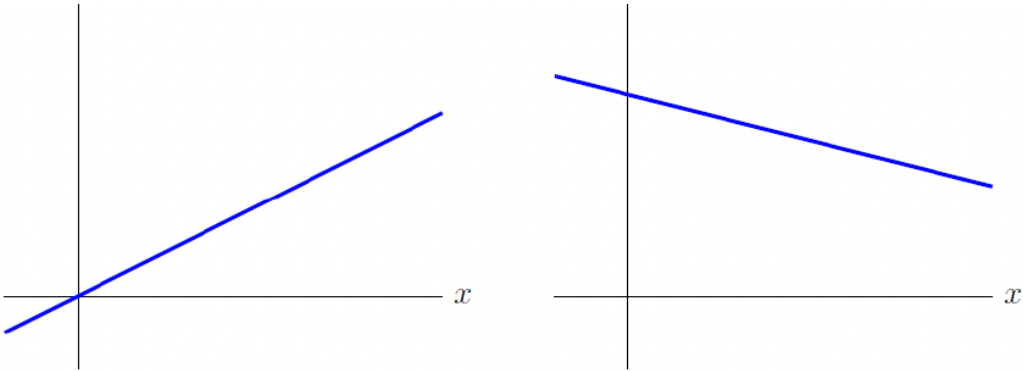


Figure: Left: Linear; Right: Affline but not linear.

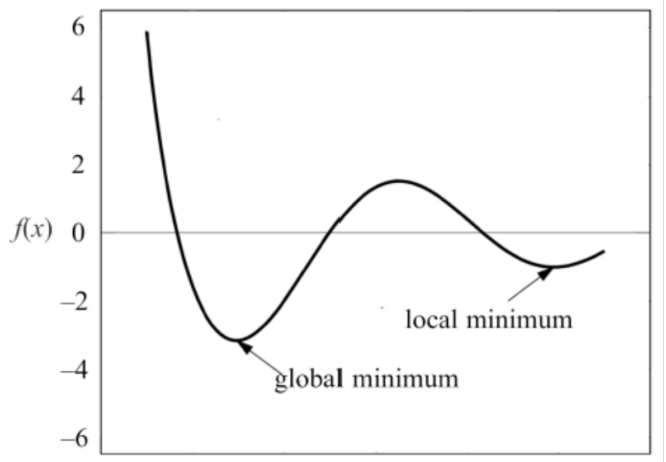
Local and global extrema

- Consider a function $f : [a, b] \to \mathbb{R}$.
- The function f has a local minimum at $c \in \mathbb{R}$ if $f(x) \geq f(c)$ for all x in an open neighborhood of c.
- The function f has a global minimum at $c \in \mathbb{R}$ if $f(x) \geq f(c)$ for all $x \in [a, b]$.

Exercise: If c is a local minimum of f, is it a global minimum? If c is a global minimum of f, is it a local minimum?

Exercise: How would you define local maximum and global maximum?

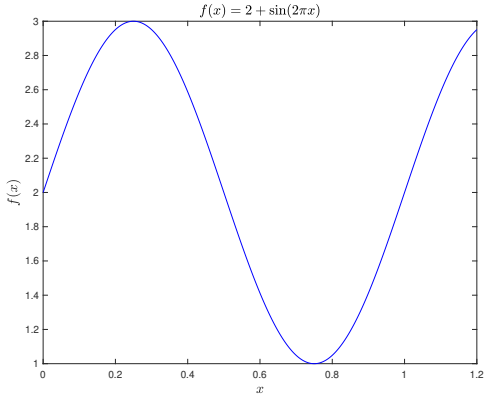
Local and global extrema



min **and** arg min

- For a function $f : X \to Y$, the minimum $\min_{x \in X} f(x)$ returns the smallest value among all elements in the set $\{f(x) : x \in X\}$.
- For a function $f : X \to Y$, the argmin $x^* = \arg\min_{x \in X} f(x)$ returns the value of $x \in X$ that minimizes $f(x)$, i.e.,
$$f(x^*) = \min_{x \in X} f(x)$$
- arg min returns a value from the domain of the function X and min returns from the range (codomain) Y of the function.
- Let $X = \{0, 1\}$ and $f(0) = \pi$ and $f(1) = e$. Then
$$\arg\min_{x \in X} f(x) = 1 \qquad \min_{x \in X} f(x) = e,$$
and
$$\arg\max_{x \in X} f(x) = 0 \qquad \max_{x \in X} f(x) = \pi.$$

min **and** arg min



- Let $f : X = [0, 1.2] \to \mathbb{R}$ be defined as $f(x) = 2 + \sin(2\pi x)$ (see plot above). Then

$$\arg\min_{x \in X} f(x) = 3/4 \qquad \min_{x \in X} f(x) = +1$$

Derivatives

- For a multivariable function $f : \mathbb{R}^d \to \mathbb{R}$, its gradient vector or derivative at $\mathbf{x} \in \mathbb{R}^d$ is the column vector

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{x}} = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} & \cdots & \dfrac{\partial f}{\partial x_d} \end{bmatrix}^\top$$

- Recall that $\frac{\partial f}{\partial x_i}$ is the partial derivative of f with respect to the scalar variable x_i.
- For example, if $f(x_1, x_2) = 2x_1^2 + 5x_1x_2 + 3x_2^3$, then

$$\frac{\partial f}{\partial x_1} = 4x_1 + 5x_2 \quad \text{and} \quad \frac{\partial f}{\partial x_2} = 5x_1 + 9x_2^2.$$

Important derivatives

- There are only two derivatives for functions $f : \mathbb{R}^d \to \mathbb{R}$ that take vectors to scalars you need to know for now.
- For a fixed vector $\mathbf{a} \in \mathbb{R}^d$, consider $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ (known as the inner or dot product). Then
$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{a}.$$
- For a fixed matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, consider $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ (known as the quadratic form). Then
$$\nabla_{\mathbf{x}} f(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}.$$
见微知著，先计算矩阵内每一项，找到规律之后再整合
定义入手证明
- In most applications, $\mathbf{A}$ is a symmetric matrix (i.e., $\mathbf{A} = \mathbf{A}^\top$) so
$$\nabla_{\mathbf{x}} f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}.$$
- This generalizes the basic fact that if $f(x) = ax^2$, then $\frac{\mathrm{d}f}{\mathrm{d}x} = 2ax$.

Important derivatives

Exercise: Show from the definition of $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$ that

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{a}.$$

Exercise: Show from the definition of $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ that

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = (\mathbf{A} + \mathbf{A}^\top)\mathbf{x}.$$

You'll be forgiven for having to exhibit a substantial amount of meticulous bookkeeping here.

Advice: It is difficult to remember a lot of derivative formulas of complicated multivariate functions. Usually, one consults the Matrix Cookbook
`https://www2.imm.dtu.dk/pubdb/edoc/imm3274.pdf`

Statistical machine learning deals with uncertainty

$\Downarrow$

Definitions of events, probabilities, joint probabilities and conditional probabilities

Notations

- Random variables (r.v.s): upper (e.g., X)
- Values r.v.s take on: lower case (e.g., x, x_i)
- Sets (in which the r.v.s assume their values): calligraphic font (e.g., $\mathcal{X}$)
- Probability, expectation and variance operators: $\Pr(\cdot)$, $\mathbb{E}[\cdot]$, $\mathrm{Var}(\cdot)$

Discrete random variables

Random variables X and Y:

- X can only take values in the finite set $\mathcal{X} = \{x_1, x_2 \ldots, x_M\}$.
- Y can only take values in the finite set $\mathcal{Y} = \{y_1, y_2 \ldots, y_L\}$.

A random variable is said to be **discrete** if the set of values it can take has a finite number of values.

$$\sum_{i=1}^{M} \Pr(X = x_i) = 1, \sum_{i=1}^{L} \Pr(Y = y_i) = 1.$$

Event

Consider m trials (of sampling X and Y) and let the number of trials for which $X = x_i$ and $Y = y_j$ be m_{ij}. Then, if m is large, we can assume that

$$\Pr(X = x_i,\ Y = y_j) = \frac{m_{ij}}{m}.$$

- The argument of $\Pr(\cdot)$ is known as an **event**.
- **Event**: A set in which a r.v. (or multiples r.v.s) assume some value(s) in some set. e.g., $\{X = x_i,\ Y = y_j\}$.

Sum rule and product rule

- **Sum rule**:

$$\Pr(X = x_i) = \sum_{j=1}^{L} \Pr(X = x_i, Y = y_j), \quad \Pr(Y = y_j) = \sum_{i=1}^{M} \Pr(X = x_i, Y = y_j).$$

$$\Pr(X = x_i, Y = y_j) = \Pr(X = x_i | Y = y_j) \cdot \Pr(Y = y_j).$$

- **Product rule**:

$$\Pr(X = x) = \sum_{y \in \mathcal{Y}} \Pr(X = x, Y = y).$$

$$\Pr(X = x, Y = y) = \Pr(X = x | Y = y) \cdot \Pr(Y = y).$$

Bayes' rule

- **Probability mass function (pmf)**:
  $p(x) := \Pr(X = x),\ p(x, y) := \Pr(X = x, Y = y),\ p(x|y) := \Pr(X = x|Y = y).$

- **Bayes' rule/Bayes' Theorem**:

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum_{y' \in \mathcal{Y}} p(x|y')p(y')}.$$

  - 'Invert the casucal relationship' between X and Y.
  - $p(x|y) \propto p(y|x)p(x)$.
  - $\text{posterior} = \dfrac{\text{likelihood} \times \text{prior}}{\text{model evidence}} \propto \text{likelihood} \times \text{prior}.$

Continuous random variables

- **Cumulative distribution function (cdf)**: $x \in \mathbb{R} \mapsto F_X(x) = \Pr(X \leq X)$ is differentiable on $\mathbb{R}$.
- **Probability density function (pdf)**:

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x), \quad x \in \mathbb{R}; \quad f_X(x) \geq 0 \quad \forall x \in \mathbb{R} \text{ and } \int_{\mathbb{R}} f_X(x)\mathrm{d}x = 1.$$

Continuous random variables

- **Cumulative distribution function (cdf)**: $x \in \mathbb{R} \mapsto F_X(x) = \Pr(X \leq X)$ is differentiable on $\mathbb{R}$.
- **Probability density function (pdf)**:

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x), \quad x \in \mathbb{R}; \quad f_X(x) \geq 0 \quad \forall x \in \mathbb{R} \text{ and } \int_{\mathbb{R}} f_X(x)\mathrm{d}x = 1.$$

- **Joint probability density function** of continuous r.v.s.:

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y), (x,y) \in \mathbb{R}^2.$$

- Conditional probability and independence can be defined analogously to the discrete case:

$$\Pr(a < X \leq b,\ c < Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y.$$

Independence

Definition 3.2 (Independence)

Two random variables X and Y are independent if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

- X and Y are independent if $p_{X|Y}(x|y) = p_X(x)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ such that $p_Y(y) > 0$ (what if $p_Y(y) = 0$?).

Expectation and variance

- **Expectation**:
  - Discrete r.v.: $\mathbb{E}X = \sum_{x \in \mathcal{X}} x p_X(x)$.
  - Continuous r.v.: $\mathbb{E}X = \int_{\mathbb{R}} x f_X(x) \mathrm{d}x$.

    If g is a function from the domain of X to $\mathbb{R}$, we can obtain the expectation of $Y = g(X)$ in the same way:

    $$\mathbb{E}Y = \mathbb{E}[g(X)] = \int_{\mathbb{R}} y f_Y(y) \mathrm{d}y = \int_{\mathbb{R}} g(x) f_X(x) \mathrm{d}x.$$

  - Linearity of expectation: if $g(X) = aX + b$ (for constants $a, b \in \mathbb{R}$), then $\mathbb{E}[g(X)] = a\mathbb{E}[X] + b = g(\mathbb{E}[X])$.

- **Variance**:
  - Discrete r.v.: $\mathrm{Var}(X) = \sum_{x \in \mathcal{X}} (x - \mathbb{E}X)^2 p_X(x)$.
  - Continuous r.v.: $\mathrm{Var}(X) = \int_{\mathbb{R}} (x - \mathbb{E}X)^2 f_X(x) \mathrm{d}x$.

Maximum likelihood estimation (MLE)

- In machine learning, we almost always do NOT have access to the underlying distributions that the data are generated from.
- Rather, we have access to sample data and we would like to use it to estimate some parameters.

Maximum likelihood estimation: Gaussian distribution

- If samples in $\mathcal{D} = \{X_1, \ldots, X_m\}$ are independently drawn from the univariate Gaussian distribution

$$f_X\left(x; \mu, \sigma^2\right) = \mathcal{N}\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad x \in \mathbb{R}, \tag{3.1}$$

  we can use the samples in $\mathcal{D}$ to estimate the parameter vector $\theta = \left(\mu, \sigma^2\right)$.

- Note: notation $f_X\left(x; \mu, \sigma^2\right)$ or $f_X(x; \theta)$ emphasize that the density is parametrized by the parameters $\left(\mu, \sigma^2\right)$ or θ.

Maximum likelihood estimation: Gaussian distribution

- If samples in $\mathcal{D} = \{X_1, \ldots, X_m\}$ are independently drawn from the univariate Gaussian distribution

$$f_X\left(x; \mu, \sigma^2\right) = \mathcal{N}\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad x \in \mathbb{R}, \tag{3.1}$$

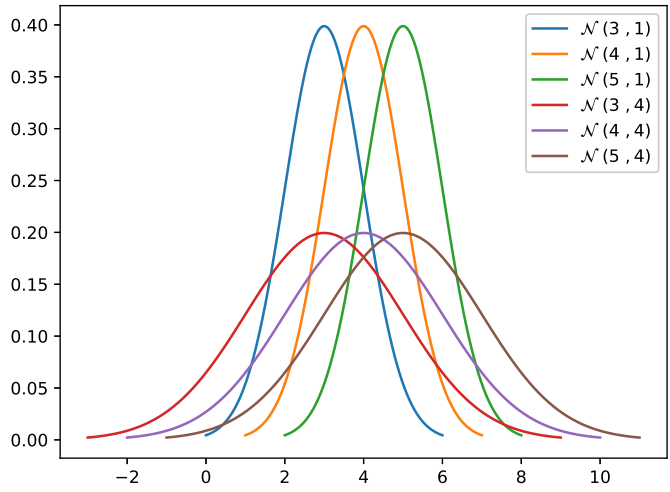
  we can use the samples in $\mathcal{D}$ to estimate the parameter vector $\theta = \left(\mu, \sigma^2\right)$.

- Note: notation $f_X\left(x; \mu, \sigma^2\right)$ or $f_X(x; \theta)$ emphasize that the density is parametrized by the parameters $\left(\mu, \sigma^2\right)$ or θ.

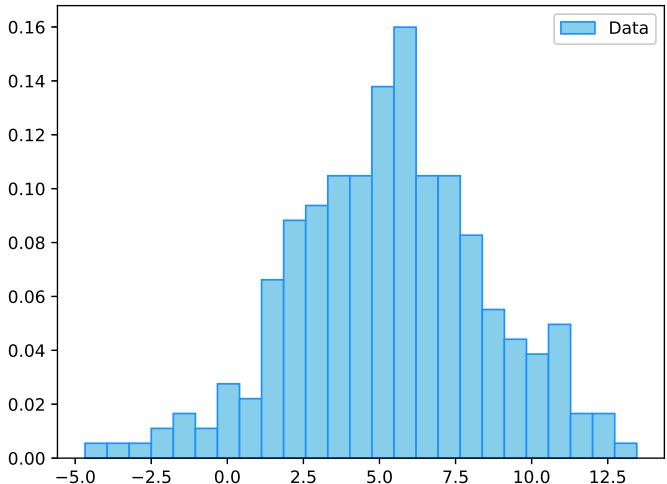
◎ Maximum likelihood estimates for the mean and variance:

$$\hat{\mu}_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m} X_i \quad \text{and} \quad \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{m}\sum_{i=1}^{m}\left(X_i - \hat{\mu}\right)^2. \tag{3.2}$$

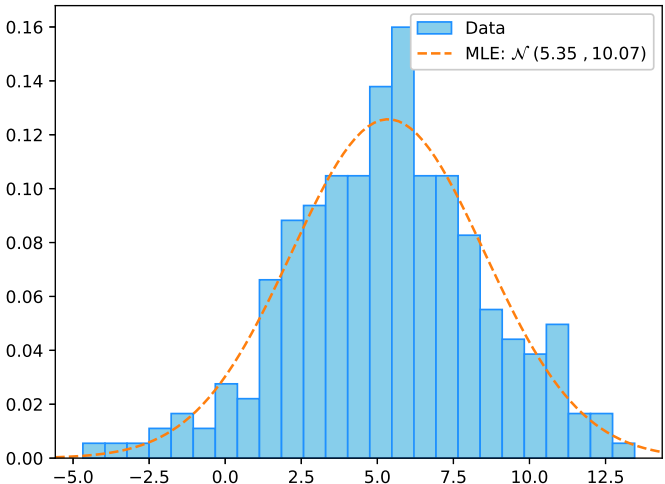
Maximum likelihood estimation: Gaussian distribution



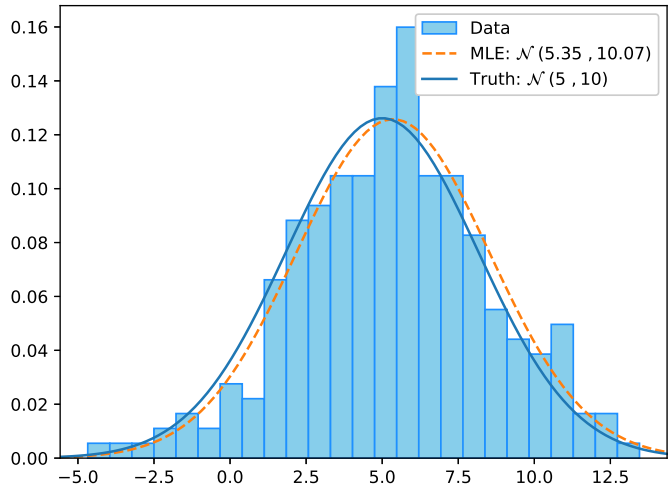
Maximum likelihood estimation: Gaussian distribution



Maximum likelihood estimation: Gaussian distribution



Maximum likelihood estimation: Gaussian distribution



Maximum likelihood estimation: Bernoulli distribution

- Samples in $\mathcal{D} = \{X_1, \ldots, X_m\}$ are generated independently from the Bernoulli (coin toss) distribution

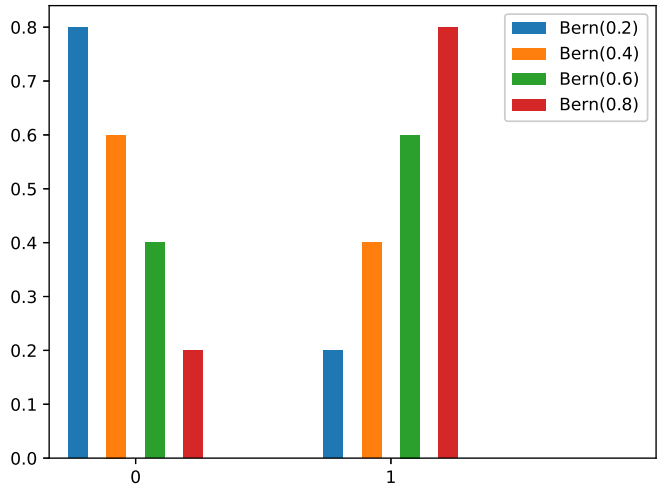
$$p_X(x;\theta) = \begin{cases} 1-\theta & x = 0 \\ \theta & x = 1 \end{cases} \tag{3.3}$$

or

$$p_X(x;\theta) = (1-\theta)^{1-x}\theta^x, \quad x \in \{0,1\}. \tag{3.4}$$

- How would we estimate θ from samples?

Maximum likelihood estimation: Bernoulli distribution



Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^{m} X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} \prod_{i=1}^{m} p_X(X_i; \theta) \tag{3.5}$$

Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^{m} X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in (0,1)} \prod_{i=1}^{m} p_X\left(X_i; \theta\right) \tag{3.5}$$

$$= \arg\max_{\theta \in (0,1)} \sum_{i=1}^{m} \log p_X\left(X_i; \theta\right) \tag{3.6}$$

Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^{m} X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta \in (0,1)} \prod_{i=1}^{m} p_X(X_i; \theta) \tag{3.5}$$

$$= \arg\max_{\theta \in (0,1)} \sum_{i=1}^{m} \log p_X(X_i; \theta) \tag{3.6}$$

$$= \arg\max_{\theta \in (0,1)} \sum_{i=1}^{m} [(1 - X_i) \log(1 - \theta) + X_i \log \theta] \tag{3.7}$$

$$= \arg\max_{\theta \in (0,1)} (m - M_1) \log(1 - \theta) + M_1 \log \theta. \tag{3.8}$$

Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^{m} X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} \underbrace{(m - M_1)\log(1-\theta) + M_1 \log\theta}_{\text{Objective function } g(\theta)}. \tag{3.9}$$

Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^{m} X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} \underbrace{(m - M_1)\log(1-\theta) + M_1 \log\theta}_{\text{Objective function } g(\theta)}. \tag{3.9}$$

- $g(\theta)$ is strictly concave $\Rightarrow$ differentiating and setting to zero yields the (unique) maximum:

Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^{m} X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} \underbrace{(m - M_1)\log(1-\theta) + M_1 \log\theta}_{\text{Objective function } g(\theta)}. \tag{3.9}$$

- $g(\theta)$ is strictly concave $\Rightarrow$ differentiating and setting to zero yields the (unique) maximum:

$$g'(\theta) = -\frac{m - M_1}{1-\theta} + \frac{M_1}{\theta}, \tag{3.10}$$

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} g(\theta) \implies g'(\hat{\theta}_{\mathrm{ML}}) = 0 \implies \hat{\theta}_{\mathrm{ML}} = \frac{M_1}{m}. \tag{3.11}$$

Maximum likelihood estimation: Bernoulli distribution

- $M_1 := \sum_{i=1}^m X_i$ denote the total number of ones in $\mathcal{D}$. Consider

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} \underbrace{(m - M_1)\log(1-\theta) + M_1 \log\theta}_{\text{Objective function } g(\theta)}. \tag{3.9}$$

- $g(\theta)$ is strictly concave $\Rightarrow$ differentiating and setting to zero yields the (unique) maximum:

$$g'(\theta) = -\frac{m - M_1}{1-\theta} + \frac{M_1}{\theta}, \tag{3.10}$$

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in (0,1)}{\arg\max} g(\theta) \implies g'(\hat{\theta}_{\mathrm{ML}}) = 0 \implies \hat{\theta}_{\mathrm{ML}} = \frac{M_1}{m}. \tag{3.11}$$

- Mean θ is estimated by the empirical mean M_1/m: agrees with common sense!

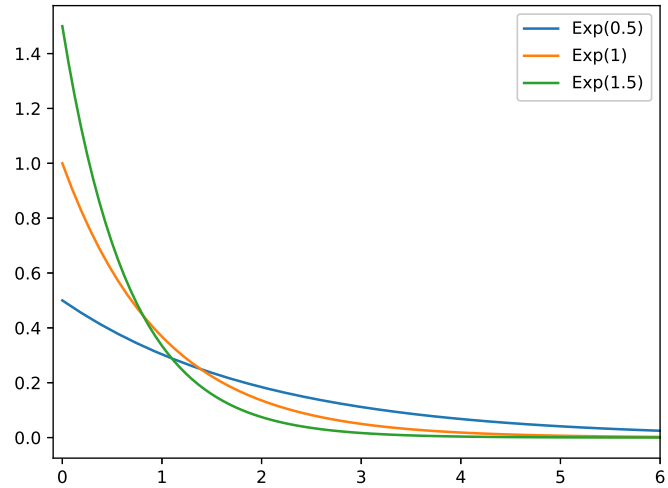
Maximum likelihood estimation: exponential distribution

- Samples in $\mathcal{D} = \{X_1, \ldots, X_m\}$ are generated independently from the exponential distribution (memory-less)

$$f_X(x; \theta) = \begin{cases} \theta \exp(-\theta x) & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

- $\theta > 0$ is the unknown (rate) parameter. In fact, $\mathbb{E}[X] = 1/\theta$.
- How would we estimate θ from samples?

Maximum likelihood estimation: exponential distribution



Maximum likelihood estimation: exponential distribution

- Consider

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta>0} \prod_{i=1}^{m} f_X\left(X_i;\theta\right) = \arg\max_{\theta>0} \sum_{i=1}^{m} \log f_X\left(X_i;\theta\right) \tag{3.12}$$

Maximum likelihood estimation: exponential distribution

- Consider

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta>0} \prod_{i=1}^{m} f_X(X_i;\theta) = \arg\max_{\theta>0} \sum_{i=1}^{m} \log f_X(X_i;\theta) \tag{3.12}$$

$$= \arg\max_{\theta>0} \sum_{i=1}^{m} (\log\theta - \theta X_i) = \arg\max_{\theta>0} m\log\theta - \sum_{i=1}^{m} \theta X_i. \tag{3.13}$$

Maximum likelihood estimation: exponential distribution

- Consider

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta>0} \prod_{i=1}^{m} f_X(X_i;\theta) = \arg\max_{\theta>0} \sum_{i=1}^{m} \log f_X(X_i;\theta) \tag{3.12}$$

$$= \arg\max_{\theta>0} \sum_{i=1}^{m} (\log\theta - \theta X_i) = \arg\max_{\theta>0} m\log\theta - \sum_{i=1}^{m} \theta X_i. \tag{3.13}$$

- $g(\theta) = m\log\theta - \sum_{i=1}^{m} \theta X_i$ is strictly concave $\Rightarrow$ differentiating and setting to zero yields the (unique) maximum:

Maximum likelihood estimation: exponential distribution

- Consider

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta>0} \prod_{i=1}^{m} f_X(X_i;\theta) = \arg\max_{\theta>0} \sum_{i=1}^{m} \log f_X(X_i;\theta) \tag{3.12}$$

$$= \arg\max_{\theta>0} \sum_{i=1}^{m} (\log\theta - \theta X_i) = \arg\max_{\theta>0} m\log\theta - \sum_{i=1}^{m} \theta X_i. \tag{3.13}$$

- $g(\theta) = m\log\theta - \sum_{i=1}^{m} \theta X_i$ is strictly concave $\Rightarrow$ differentiating and setting to zero yields the (unique) maximum:

$$g'(\theta) = \frac{m}{\hat{\theta}_{\mathrm{ML}}} - \sum_{i=1}^{m} X_i, \tag{3.14}$$

$$g'(\hat{\theta}_{\mathrm{ML}}) = 0 \implies \hat{\theta}_{\mathrm{ML}} = \hat{\theta}_{\mathrm{ML}} = \frac{m}{\sum_{i=1}^{n} X_i} = \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^{-1}. \tag{3.15}$$

Maximum likelihood estimation: exponential distribution

- Exponential distribution (this distribution models waiting times for buses)

$$f_X(x;\theta) = \begin{cases} \theta \exp(-\theta x) & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

- $\mathbb{E}[X] = 1/\theta$, $\hat{\theta}_{\mathrm{ML}} = \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^{-1}$.

Maximum likelihood estimation: exponential distribution

- Exponential distribution (this distribution models waiting times for buses)

$$f_X(x;\theta) = \begin{cases} \theta \exp(-\theta x) & x \geq 0 \\ 0 & x < 0 \end{cases}.$$

- $\mathbb{E}[X] = 1/\theta$, $\hat{\theta}_{\mathrm{ML}} = \left(\frac{1}{m}\sum_{i=1}^{m} X_i\right)^{-1}$.

◎ $\frac{1}{m}\sum_{i=1}^{m} X_i$ is the empirical mean:
it seems plausible that $m/\left(\sum_{i=1}^{m} X_i\right)$ is a "good" estimate of θ.

- Why would it not be "good"?

Systems of linear equations

Definition 3.3 (Rules)

A vector space over the reals consists of a set $\mathcal{V}$, a vector sum operation $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$ and a scalar multiplication operation $\cdot : \mathbb{R} \times \mathcal{V} \to \mathcal{V}$ satisfying the following properties.

- Commutativity: $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$;
- Associativity: $(\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z} = \boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z})$ for all $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{V}$;
- Identity element of addition: $\boldsymbol{x} + \boldsymbol{0} = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{V}$;
- Inverse element of addition: For every $\boldsymbol{x} \in \mathcal{V}$, there exists an element $-x \in \mathcal{V}$ such that $\boldsymbol{x} + (-\boldsymbol{x}) = \boldsymbol{0}$;
- Associativity of scalar multiplication: For all $a, b \in \mathbb{R}$ and $x \in \mathcal{V}$, $a(b\boldsymbol{x}) = (ab)\boldsymbol{x}$;
- Identity element of scalar multiplication: $1\boldsymbol{x} = \boldsymbol{x}$ for $\boldsymbol{x} \in \mathcal{V}$;
- Distributivity of scalar multiplication with respect to vector addition: $a(\boldsymbol{x} + \boldsymbol{y}) = a\boldsymbol{x} + a\boldsymbol{y}$ for all $a \in \mathbb{R}$ and $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$;
- Distributivity of scalar multiplication with respect to addition in $\mathbb{R}$: $(a + b)\boldsymbol{x} = a\boldsymbol{x} + b\boldsymbol{x}$ for all $a, b \in \mathbb{R}$ and $\boldsymbol{x} \in \mathcal{V}$.

Vector

Definition 3.4 (Linear independence)

A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ from a vector space $\mathcal{V}$ is linearly independent if

$$\beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \ldots + \beta_k \boldsymbol{x}_k = \mathbf{0} \tag{3.16}$$

implies that $\beta_1 = \beta_2 = \ldots = \beta_k = 0$.

- $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ are linearly independent $\Longleftrightarrow$ no vector $\boldsymbol{x}_i$ can be expressed as a linear combination of the other vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \ldots, \boldsymbol{x}_k\}$.

Matrix

Definition 3.5 (Basis of vector space)

A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ is a basis for a vector space $\mathcal{V}$ if

- $\mathcal{V} = \text{span}\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$;
- $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k\}$ is a **linearly independent set** of vectors.

Equivalently, every $\boldsymbol{x} \in \mathcal{V}$ can be uniquely written as $\sum_{i=1}^{k} \beta_i \boldsymbol{x}_i$ for some $\{\beta_i\}_{i=1}^{k} \subset \mathbb{R}$. The number of vectors in any basis of V is called the **dimension** of $\mathcal{V}$, written as $\dim(\mathcal{V})$.

Definition 3.6 (nullspace and range of matrix)

The **nullspace (also called kernel)** of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ is defined as

$$\mathcal{N}(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{x} = 0\}. \tag{3.17}$$

The range or column space of $\boldsymbol{A}$ is defined as

$$\mathcal{R}(\boldsymbol{A}) = \{\boldsymbol{A}\boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^f\} \subset \mathbb{R}^m. \tag{3.18}$$

Matrix

Definition 3.7 (Rank of matrix)

The **column rank or rank** of a matrix $\boldsymbol{A} \in \boldsymbol{R}^{m \times d}$ is defined as

$$\operatorname{rank}(\boldsymbol{A}) = \dim(\mathcal{R}(\boldsymbol{A})). \tag{3.19}$$

In other words, the column rank of $\boldsymbol{A}$ is the dimension of the column space of $\boldsymbol{A}$.

Matrix

Definition 3.7 (Rank of matrix)

The **column rank or rank** of a matrix $\boldsymbol{A} \in \boldsymbol{R}^{m \times d}$ is defined as

$$\operatorname{rank}(\boldsymbol{A}) = \dim(\mathcal{R}(\boldsymbol{A})). \tag{3.19}$$

In other words, the column rank of $\boldsymbol{A}$ is the dimension of the column space of $\boldsymbol{A}$.

- **Rank-nullity**: $\operatorname{rank}(\boldsymbol{A}) + \dim(\mathcal{N}(\boldsymbol{A})) = d$.
- It is always true that $\operatorname{rank}(\boldsymbol{A}) \leq \min\{m, d\}$.
- A matrix is full rank if $\operatorname{rank}(\boldsymbol{A}) = \min\{m, d\}$.
- A matrix is full column rank (resp. full row rank) if the set of columns (resp. rows) of the matrix is linearly independent.
- If the matrix A is square (i.e., $m = d$) and it is full rank, then the inverse $\boldsymbol{A}^{-1}$ exists.

Solutions to linear systems

Solve systems of equations of the form ($\underline{x}_i,\ 1 \le i \le d$: columns of $\boldsymbol{X}$)

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y} \quad \text{or} \quad [\underline{x}_1\ \underline{x}_2\ \dots\ \underline{x}_d] \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}, \text{ where } \underline{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{m,i} \end{bmatrix}. \tag{3.20}$$

Matrix $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and vector $\boldsymbol{y} \in \mathbb{R}^m$ are given and $\boldsymbol{w} \in \mathbb{R}^d$ is to be found.

Solutions to linear systems

Solve systems of equations of the form ($\underline{x}_i,\ 1 \le i \le d$: columns of $\boldsymbol{X}$)

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y} \quad \text{or} \quad [\underline{x}_1\ \underline{x}_2\ \dots\ \underline{x}_d] \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}, \text{ where } \underline{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{m,i} \end{bmatrix}. \tag{3.20}$$

Matrix $\boldsymbol{X} \in \mathbb{R}^{m \times d}$ and vector $\boldsymbol{y} \in \mathbb{R}^m$ are given and $\boldsymbol{w} \in \mathbb{R}^d$ is to be found.

- $m = d$: if matrix $\boldsymbol{X}$ is square and full rank, $\boldsymbol{X}^{-1}$ exists $\Longrightarrow$ solve for $\boldsymbol{w}$ by simple matrix inversion and multiplication $\boldsymbol{w} = \boldsymbol{X}^{-1}\boldsymbol{y}$.

Solutions to linear systems

Solve systems of equations of the form ($\underline{x}_i,\ 1 \le i \le d$: columns of $\boldsymbol{X}$)

$$\boldsymbol{X}\boldsymbol{w} = \boldsymbol{y} \quad \text{or} \quad [\underline{x}_1\ \underline{x}_2\ \dots\ \underline{x}_d] \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix}, \text{ where } \underline{x}_i = \begin{bmatrix} x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{m,i} \end{bmatrix}. \tag{3.20}$$

Matrix $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ and vector $\boldsymbol{y} \in \mathbb{R}^m$ are given and $\boldsymbol{w} \in \mathbb{R}^d$ is to be found.

- $m = d$: if matrix $\boldsymbol{X}$ is square and full rank, $\boldsymbol{X}^{-1}$ exists $\Longrightarrow$ solve for $\boldsymbol{w}$ by simple matrix inversion and multiplication $\boldsymbol{w} = \boldsymbol{X}^{-1}\boldsymbol{y}$.
- $m \neq d$: existence and uniqueness of solutions to the linear system?

Augmented matrix $\tilde{\boldsymbol{X}} = [\boldsymbol{X}\boldsymbol{y}] \in \mathbb{R}^{m\times(d+1)} : \operatorname{rank}(\boldsymbol{X}) \le \operatorname{rank}(\tilde{\boldsymbol{X}})$.

Solutions to linear systems

◎ **Method**: apply Guassian-elimination method to the augmented matrix $\tilde{\boldsymbol{X}}$.

Theorem 3.8 (Rouché-Capelli Theorem)

- *The system in* (3.20) *admits a unique solution if and only if* $\operatorname{rank}(\boldsymbol{X}) = \operatorname{rank}(\tilde{\boldsymbol{X}}) = d$;
- *The system in* (3.20) *has no solution if and only if* $\operatorname{rank}(\boldsymbol{X}) < \operatorname{rank}(\tilde{\boldsymbol{X}})$;
- *The system in* (3.20) *has infinitely many solutions if and only if* $\operatorname{rank}(\boldsymbol{X}) = \operatorname{rank}(\tilde{\boldsymbol{X}}) < d$.

Thanks for listening.

⊙ Slides credit: some slides are adapted from (alphabetical order)
Haiyun He (Cornell), Tengyu Ma and Chris Re (Stanford) and Vincent Y. F. Tan (NUS).

References

A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, 1959. doi: 10.1147/rd.33.0210.