

香 港 大 學
THE UNIVERSITY OF HONG KONG

Optimization Problems in Data Science

Prof. Li Jiayang

DASE 7113 Optimization Methods for Intelligent Systems

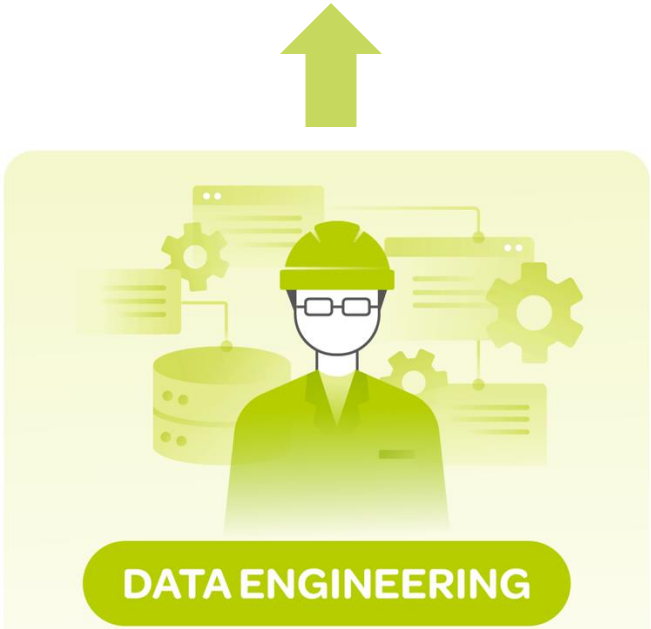
Data Science v.s. Data Engineering





Learning from data

The management of data



Data Compression

Definition

Data compression is the process of encoding data using fewer bits than the original representation.

Huffman coding → frequent characters → shorter codes

Char	Freq	Code
space	7	111
e	4	000
a	4	010
f	3	1101
t	2	0110
s	2	1011
n	2	0010
m	2	0111
i	2	1000

Run-length coding: AAAAABB ⇒ 5A2B

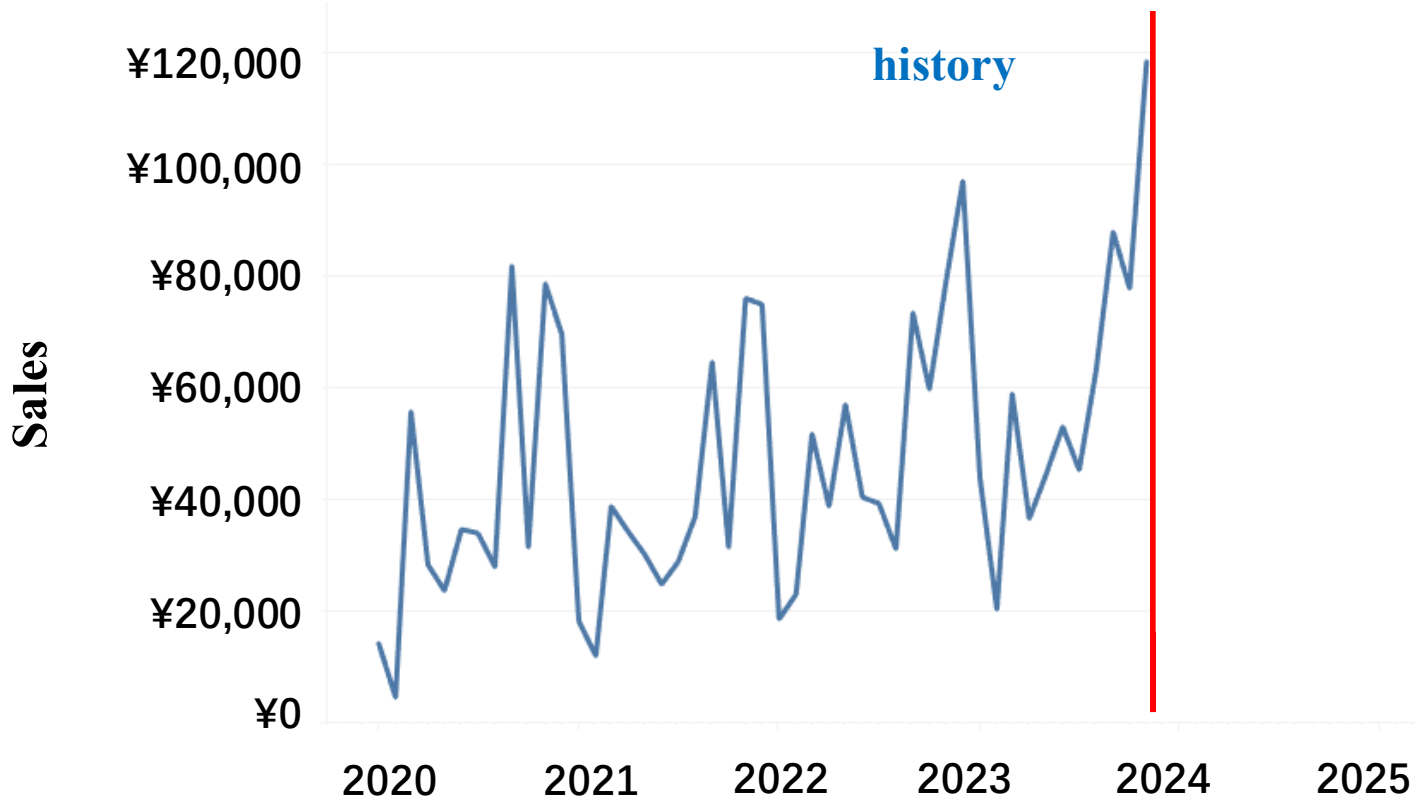


This is not data science, because by compressing the data, it does not extract any insights from the data

Time Series Forecasting

Definition

Time series forecasting is the use of a model to predict future values based on previously observed values.



Recommender System

Definition

A **recommender system** is a system that provides suggestions for items that are most pertinent to a particular user based on preference data.

					
Alice	**5**			**3**	
Bob		**4**			**5**
Chuck			**5**		

What games should be recommended to Alice, Bob and Chuck?

Clustering

Definition

Clustering is a machine learning technique designed to group unlabeled examples based on their similarity to each other.

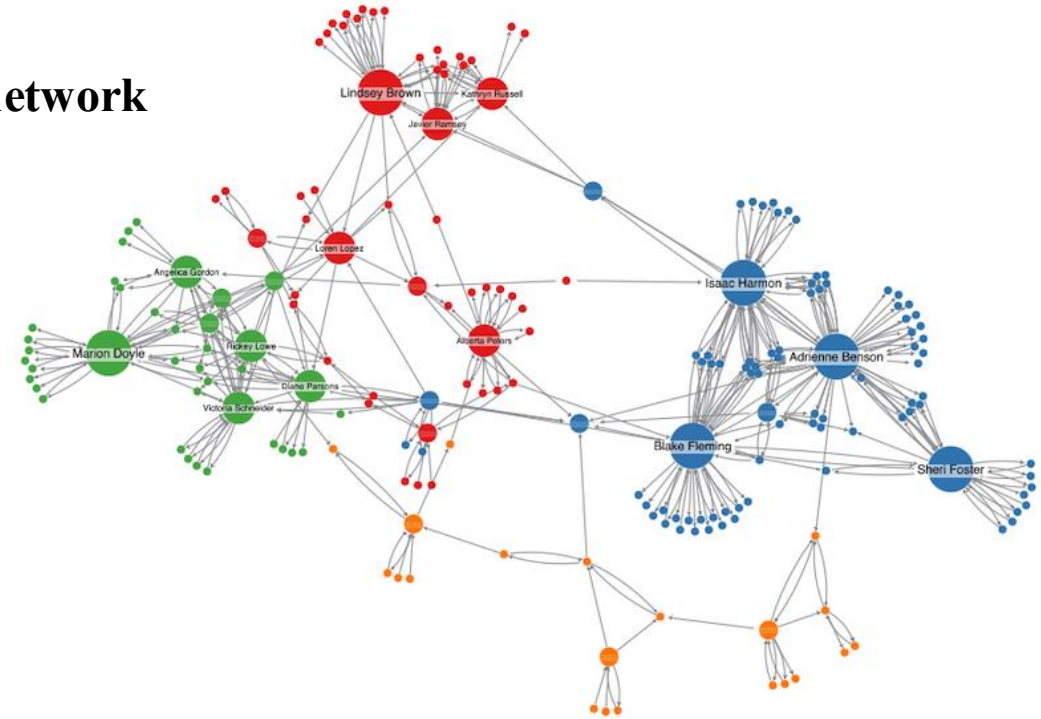


Clustering

Definition

Clustering is a machine learning technique designed to group unlabeled examples based on their similarity to each other.

Social network



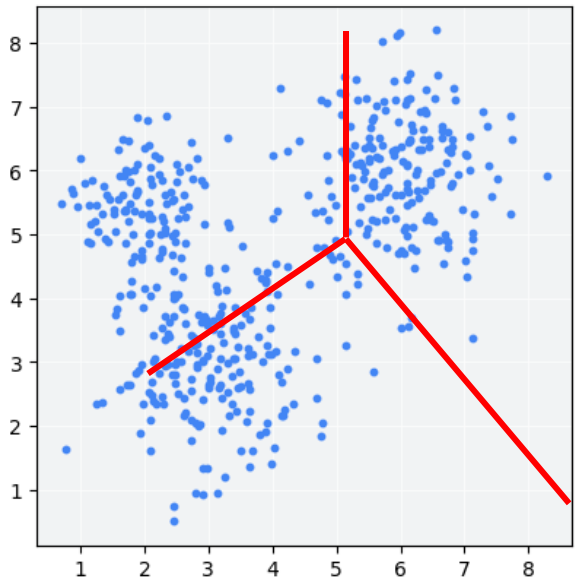
Individuals in the same cluster may have similar interests

Embedding

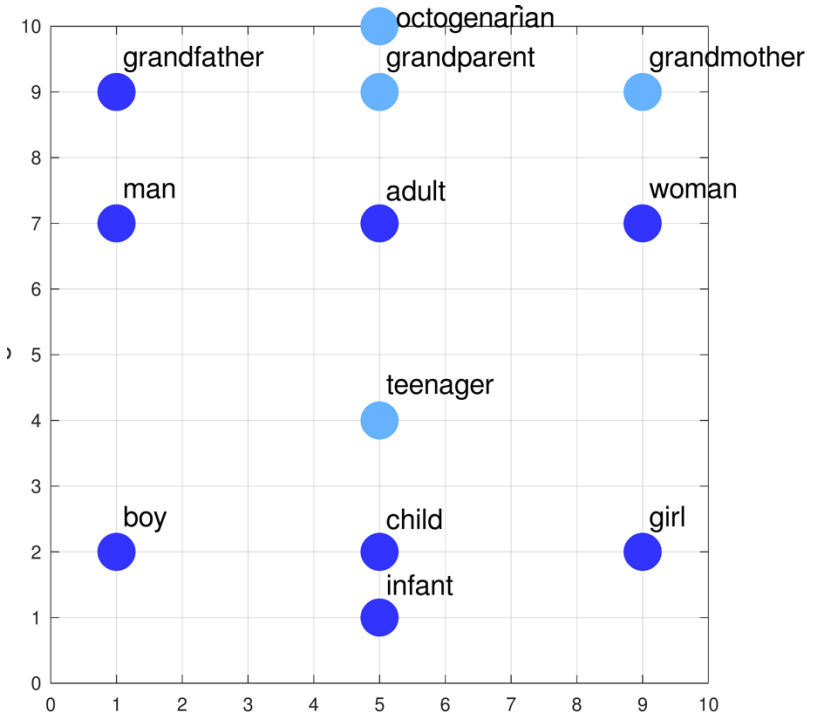
Definition

Embeddings are numerical representations of entities in high-dimension spaces (like words, items, or users) in a low-dimension vector space.

Low-dimension data points are easy to analyze…



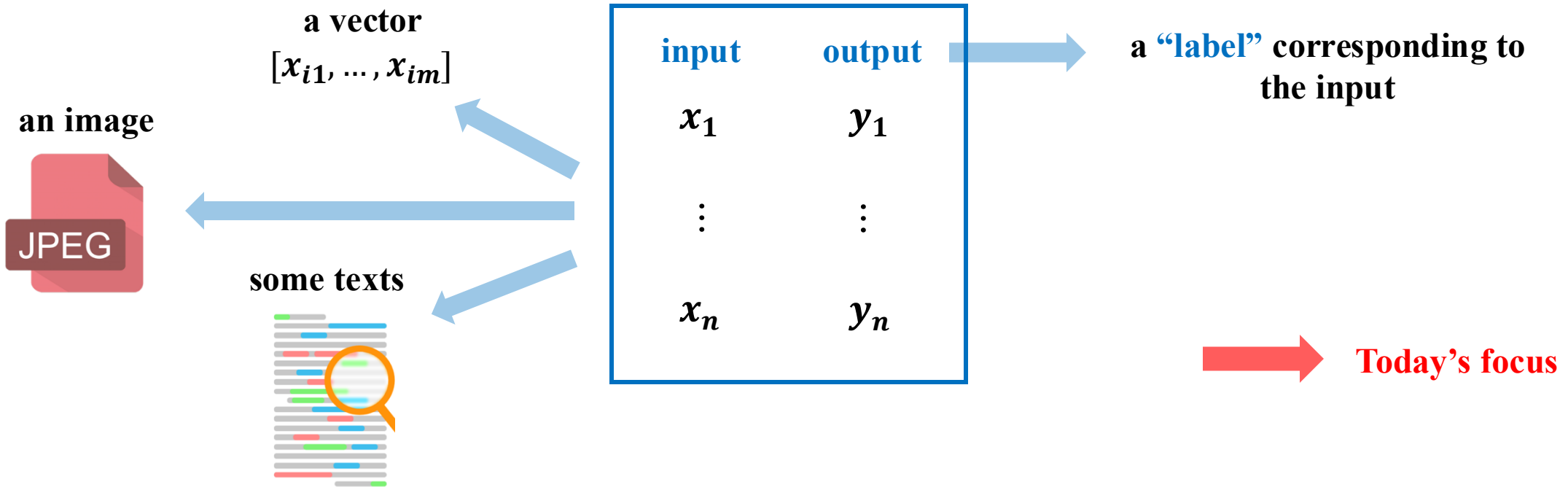
man
boy
adult
girl
infant
…



Supervised Learning

Definition

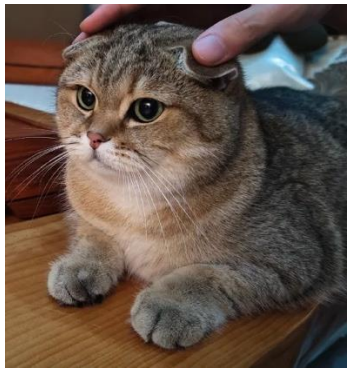
Supervised learning is a machine learning approach where models are trained for predicting outputs from inputs based on **example input-output pairs**.



Supervised Learning

A label can be a categorical variable that indicates which class the input belongs to.

input







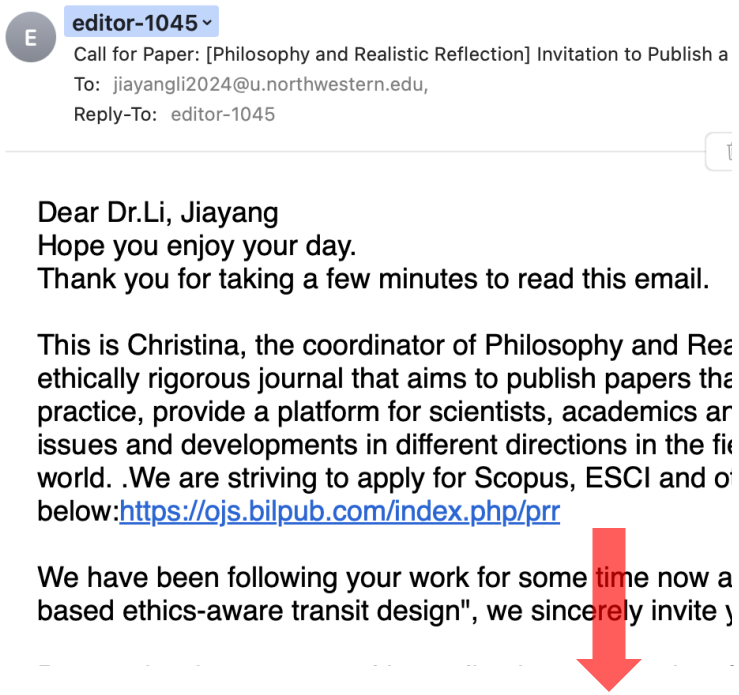
output

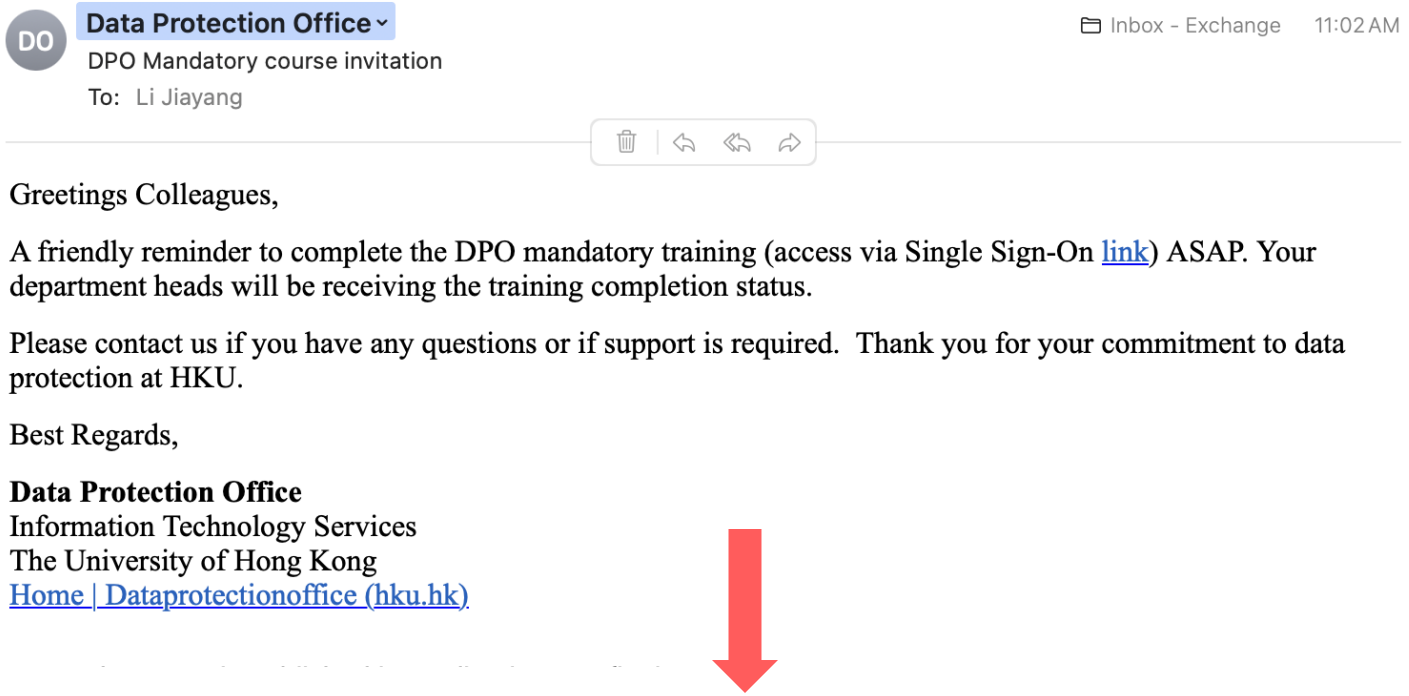
cat **dog** **cat**

Supervised Learning

A label can be a categorical variable that indicates which class the input belongs to.

input





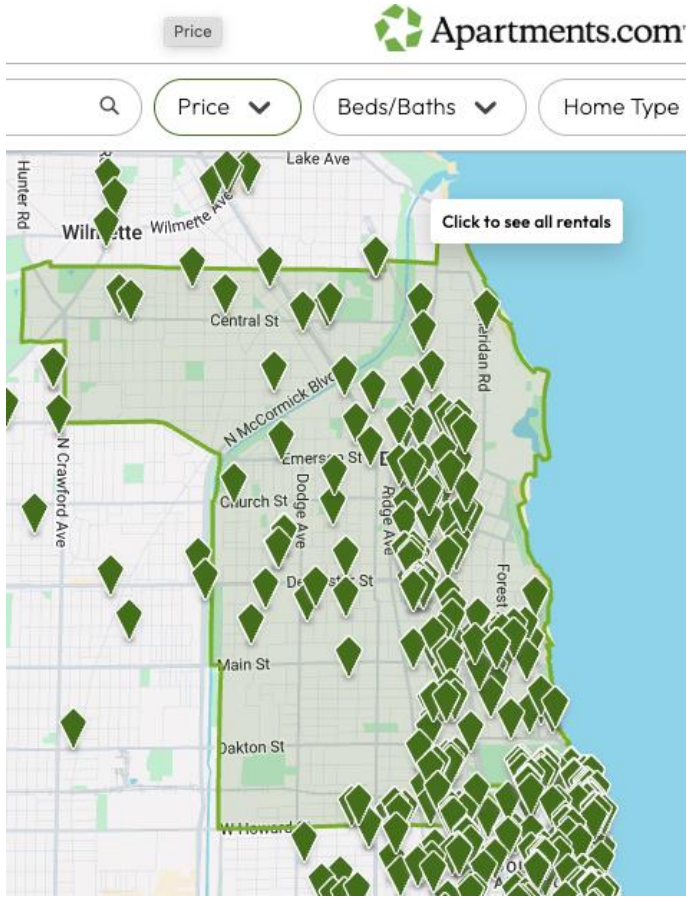
output

spam

not spam

Supervised Learning

A label can be a continuous variable that indicates a property of the input.



input — **each apartment's**

(1) size;
(2) age;
(3) number of bedrooms;
(4) number of bathrooms;
(5) whether with a washing machine

output — **their monthly rents**

Supervised Learning

Definition

Supervised learning is a machine learning approach where models are trained for predicting outputs from inputs based on **example input-output pairs**.

Input	Output
x_1	y_1
⋮	⋮
x_n	y_n

When y_i is a **categorical variable** → **Classification**

When y_i is a **continuous variable** → **Regression**

Supervised Learning

General Procedure

Step 1: Choose a model $f(x;\beta)$ that can represent the relationship between x and y,

where β represents parameters that can be adjusted.

Step 2: Adjust β such that the input-out pairs are fitted well, i.e.,

$$y_i \approx f(x_i;\beta), \qquad i = 1, \dots, n.$$

Input	Output
x_1	y_1
⋮	⋮
x_n	y_n

Given any new x, you can then predict the corresponding y as $f(x;\beta)$.

Linear Regression

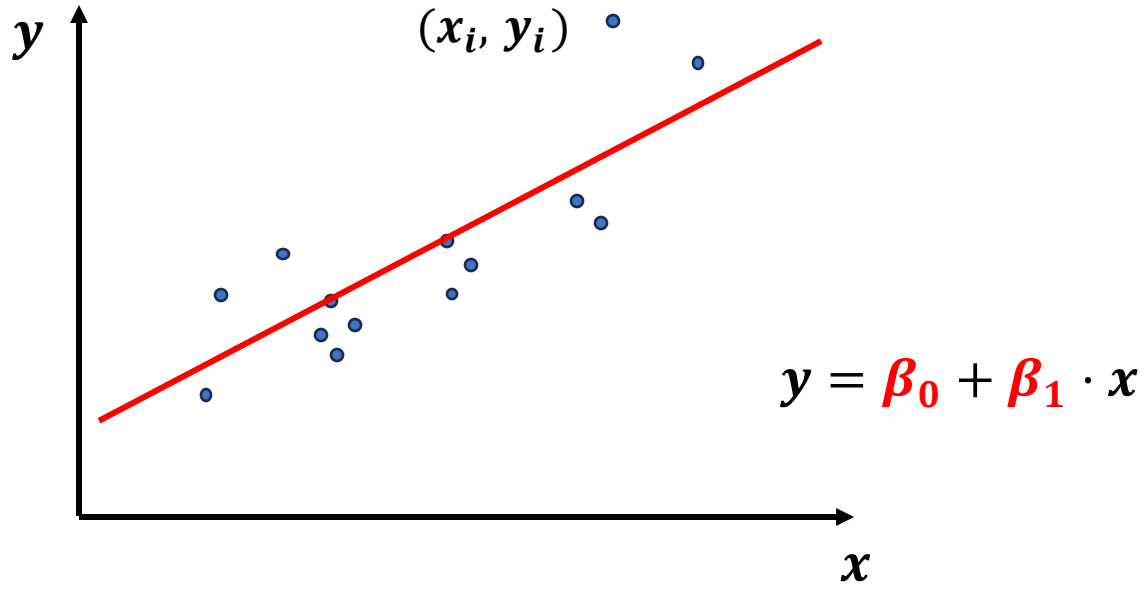
Definition

A model $f(x; \beta)$ for fitting input-output data pairs is called a linear regression model

if $f(x; \beta)$ is a linear function in β.

Question

Can the input-output pairs shown here be well-fitted through a linear regression model?



Linear Regression

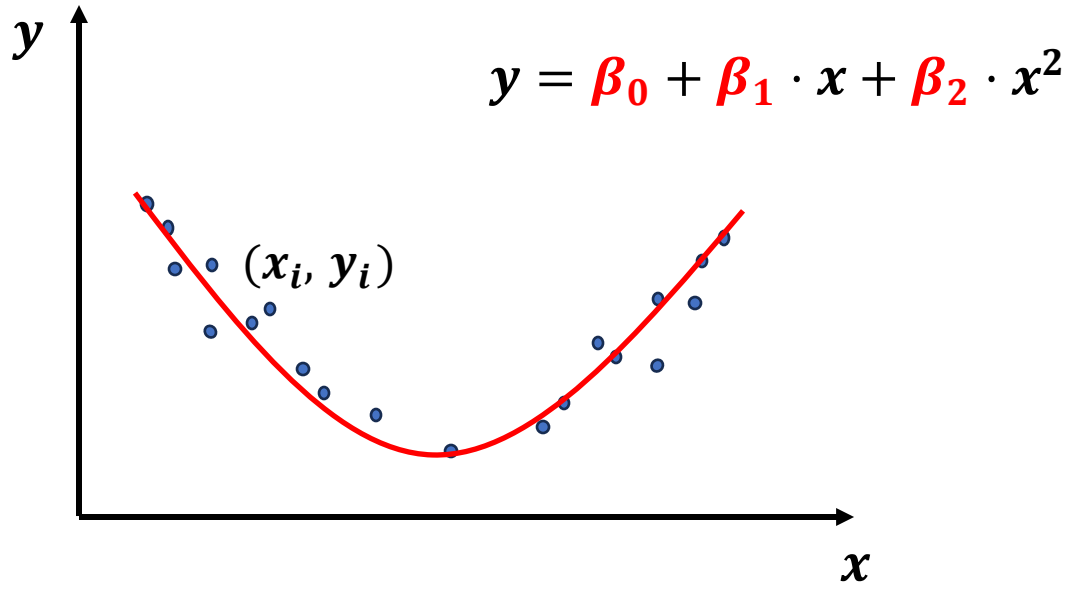
Definition

A model $f(x; \beta)$ for fitting input-output data pairs is called a linear regression model

if $f(x; \beta)$ is a linear function in β.

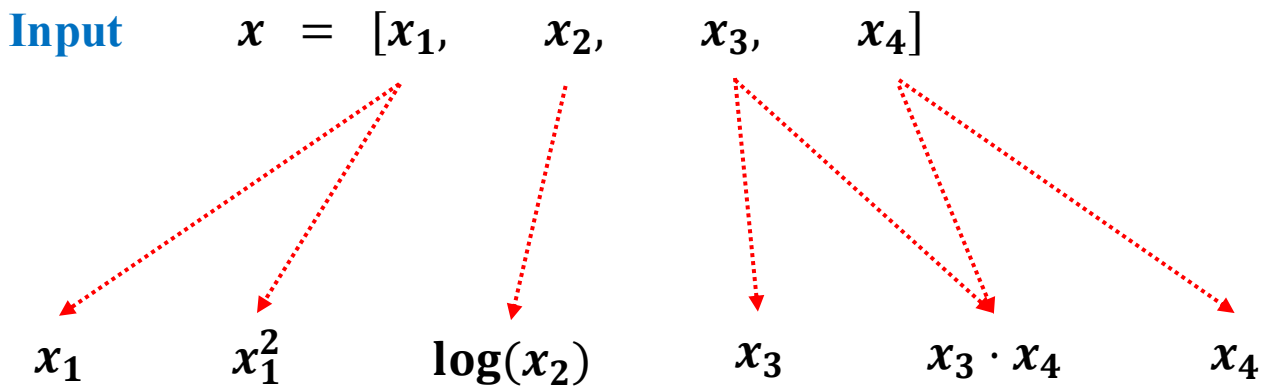
Question

Can the input-output pairs shown here be well-fitted through a linear regression model?



Linear Regression

A regression model is linear as long as the output is linear in the parameters;
there are no restrictions on how the output and the (original) input are related.



Output $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \beta_3 \cdot \log(x_2) + \beta_4 \cdot x_3 + \beta_5 \cdot x_3 \cdot x_4 + \beta_6 \cdot x_4$

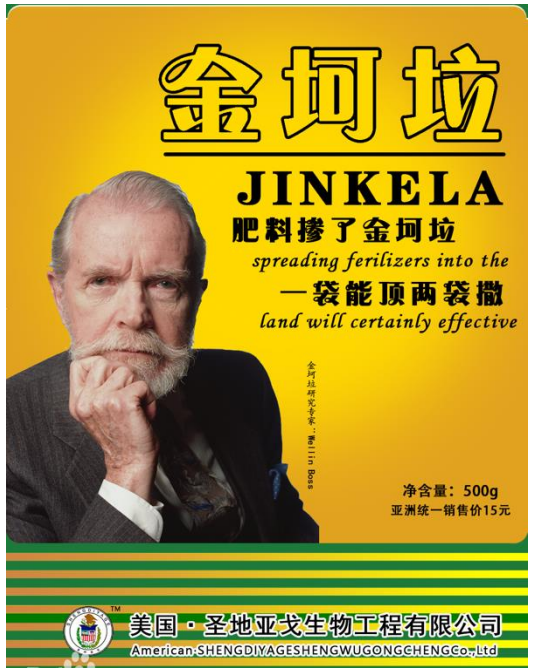
Linear Regression

Question

Can you identify a practical scenario where the inclusion of a quadratic term is fairly necessary?

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots$$

Example: when you are analyzing how varying amounts of fertilizer affect crop yield…





Linear Regression

Question

Suppose that you are analyzing how an individual's wage (y) is affected by the following factors:

(1) years of workforce experience (x_1);

(2) years with the current employer (x_2);

(3) years of formal education (x_3);

(4) parents' years of education (x_4).

How would you assume the relationship between y and x_1, x_2, x_3, x_4.

Baseline model: $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4$

Another model: $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_3 \cdot x_4$ **How this term can be motivated?**

Linear Regression

Question

Suppose that you are analyzing how an individual's **wage (y)** is affected by the following factors: (1) years of workforce experience (x_1); (2) years with the current employer (x_2); (3) **years of formal education (x_3)**; (4) **parent' years of education (x_4)**. How would you assume the relationship between y and x_1, x_2, x_3, x_4.

$$y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_3 \cdot x_4 \quad \Rightarrow \quad \frac{\partial y}{\partial x_3} = \beta_3 + \beta_5 \cdot x_4$$

How would you expect β_5?

$\beta_5 > 0$, if the social, economic, and cultural capital provided by educated parents amplifies the benefits of an individual's education

Linear Regression

Question

Suppose that you are analyzing how an individual's **wage (y)** is affected by the following factors: (1) years of workforce experience (x_1); (2) years with the current employer (x_2); (3) **years of formal education (x_3)**; (4) **parent' years of education (x_4)**. How would you assume the relationship between y and x_1, x_2, x_3, x_4.

$$y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_4 + \beta_5 \cdot x_3 \cdot x_4$$

$$\frac{\partial y}{\partial x_3} = \beta_3 + \beta_5 \cdot x_4$$

How would you expect β_5?

If you are not sure, then you should add $\beta_5 \cdot x_3 \cdot x_4$ and test whether $\beta_5 > 0$.

Linear Regression

Question

Suppose that you are analyzing how an individual's wage (y) is affected by the following factors:

(1) years of workforce experience (x_1);

(2) years of formal education (x_2);

(3) gender ($x_3 = 1$ for male and $x_3 = 0$ for female); (4) others.

How would you assume the relationship between y and x_1, x_2, x_3.

Baseline model: $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$

How would you interpret and expect β_3?

$$\Rightarrow \begin{cases} y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 \text{ for female} \\ y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \text{ for male} \end{cases}$$

Linear Regression

Question

Suppose that you are analyzing how an individual's wage (y) is affected by the following factors:

(1) years of workforce experience (x_1);

(2) years of formal education (x_2);

(3) gender ($x_3 = 1$ for male and $x_3 = 0$ for female).

How would you assume the relationship between y and x_1, x_2, x_3.

Another model: $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \beta_4 \cdot x_2 \cdot x_3$ How would you expect β_4?

$$\frac{\partial y}{\partial x_2} = \begin{cases} y = \beta_2 \text{ for female} \\ y = \beta_2 + \beta_4 \text{ for male} \end{cases}$$

Linear Regression

Question

When are we motivated to

(1) regress y on x;

(2) regress $\log(y)$ on x;

(3) regress y on $\log(x)$;

(4) regress $\log(y)$ on $\log(x)$?

$$\log(y) = \beta \cdot x + \cdots$$

$$y = \exp(\beta \cdot x + \cdots) \Rightarrow \frac{\partial y}{\partial x} = \beta \cdot \exp(\beta \cdot x + \cdots)$$

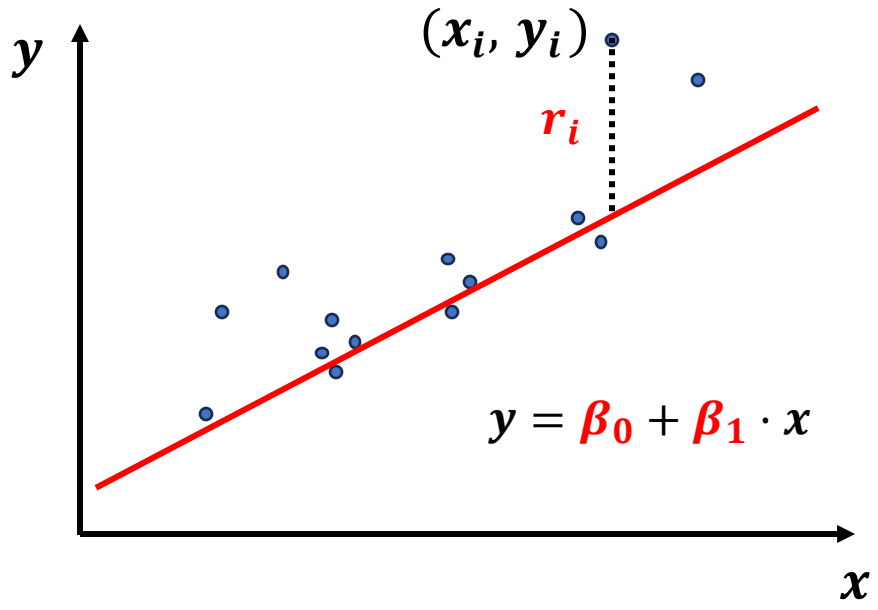
$$\frac{\partial y / \partial x}{y} = \beta \Rightarrow \frac{\Delta y}{y} = \beta \cdot \Delta x$$

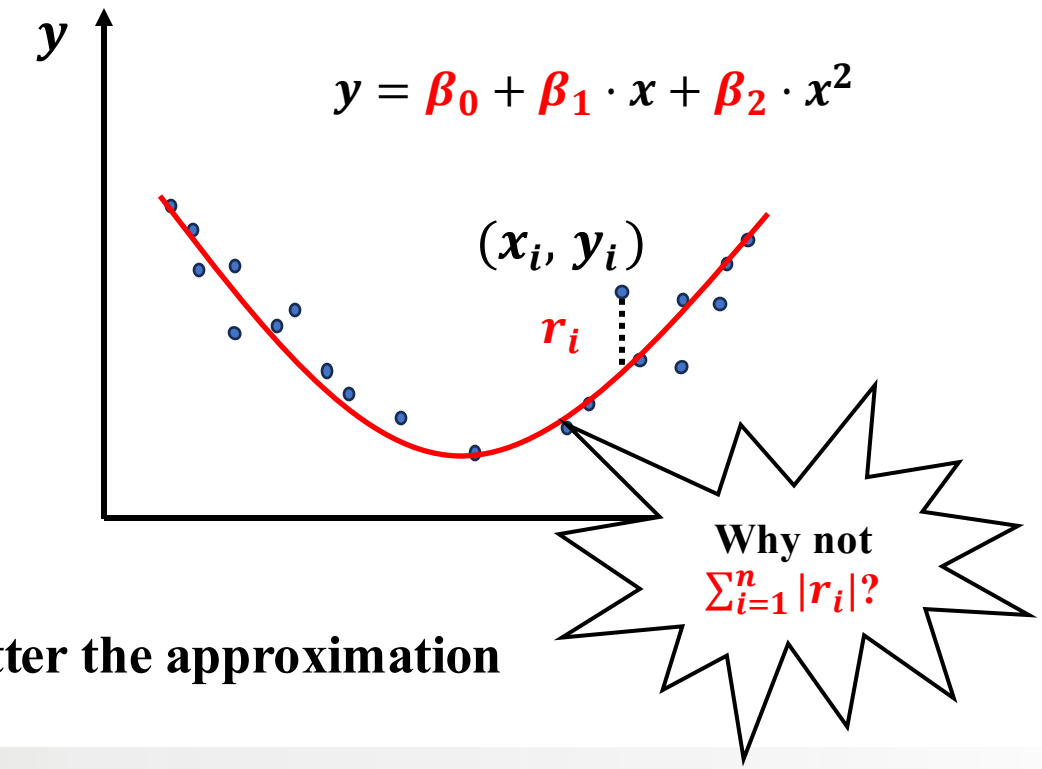
If x increases by Δx (Δx is small), y will increase by $(100 \cdot \beta \cdot \Delta x)\%$

Linear Regression

Question

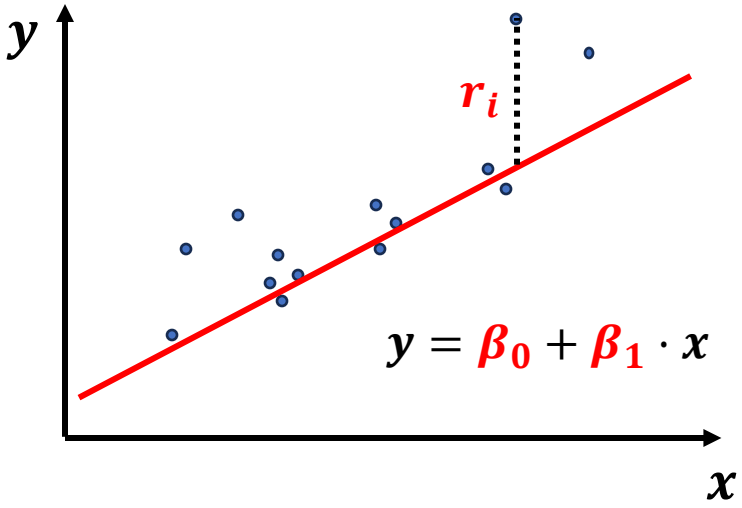
How to adjust $\beta = [\beta_0, \beta_1, \dots]$ so that the linear regression model fits the data well?



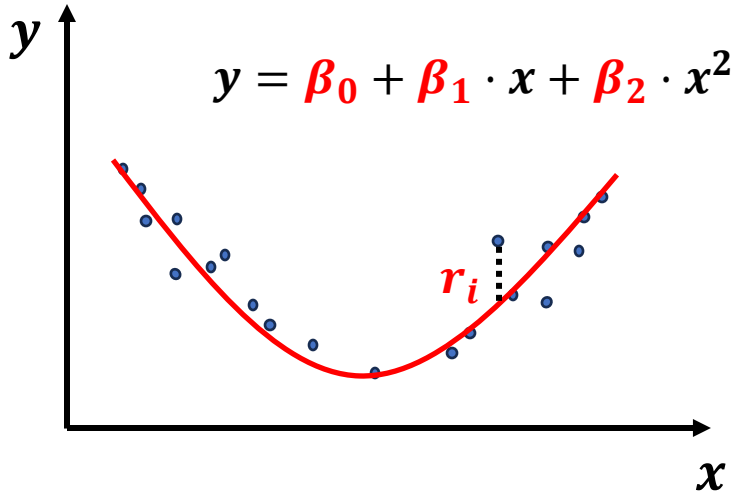


The smaller $\sum_{i=1}^{n} r_i^2$ is, the better the approximation

Linear Regression



$$\underset{\beta_0, \beta_1}{\text{minimize}} \quad \sum_{i=1}^{n} (\beta_0 + \beta_1 \cdot x_i - y_i)^2$$



$$\underset{\beta_0, \beta_1, \beta_2}{\text{minimize}} \quad \sum_{i=1}^{n} \left(\beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot x_i^2 - y_i\right)^2$$

Linear Regression

Optimization Formulation

A general linear regression model is trained by solving the following optimization problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} (f(\boldsymbol{x}_i; \boldsymbol{\beta}) - y_i)^2$$

where $f(\boldsymbol{x}_i; \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$



unconstrained convex quadratic program

- **The objective function is quadratic and convex**
- **There exist no constraints**

Linear Regression

Question

Consider the following linear regression model

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad l(\boldsymbol{\beta}) = \sum_{i=1}^{n} (\boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot x_{i1} + \cdots + \boldsymbol{\beta_m} \cdot x_{im} - y_i)^2.$$

Compute $\nabla l(\boldsymbol{\beta})$ and $\nabla^2 l(\boldsymbol{\beta})$.

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_0} \\ \vdots \\ \boldsymbol{\beta_m} \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$l(\boldsymbol{\beta}) = (\boldsymbol{X\beta} - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{X\beta} - \boldsymbol{y})$$

$$= \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta} - 2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X\beta} + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}$$

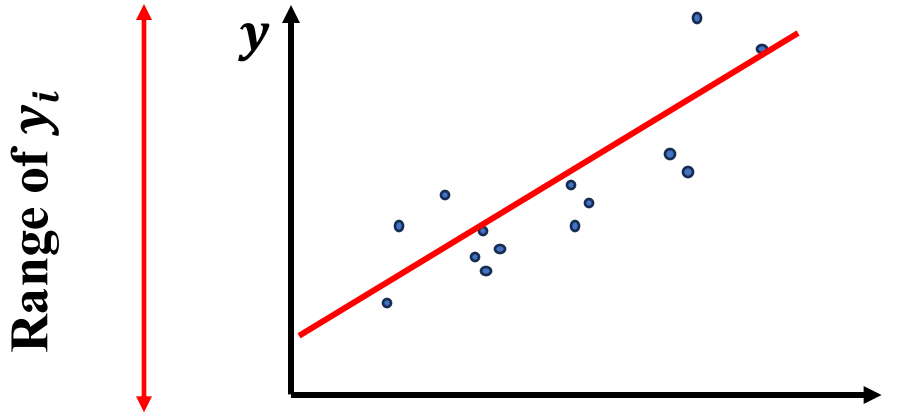
$$\Rightarrow \quad \nabla l(\boldsymbol{\beta}) = 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X\beta} - 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

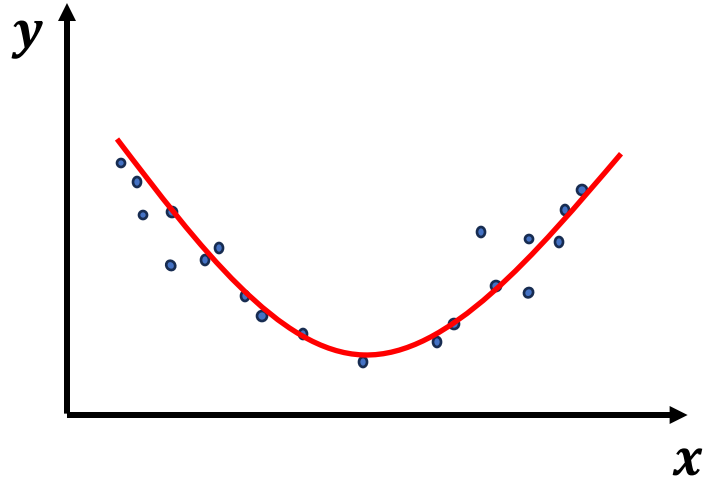
$$\Rightarrow \quad \nabla^2 l(\boldsymbol{\beta}) = \boxed{2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}}$$

A PSD matrix

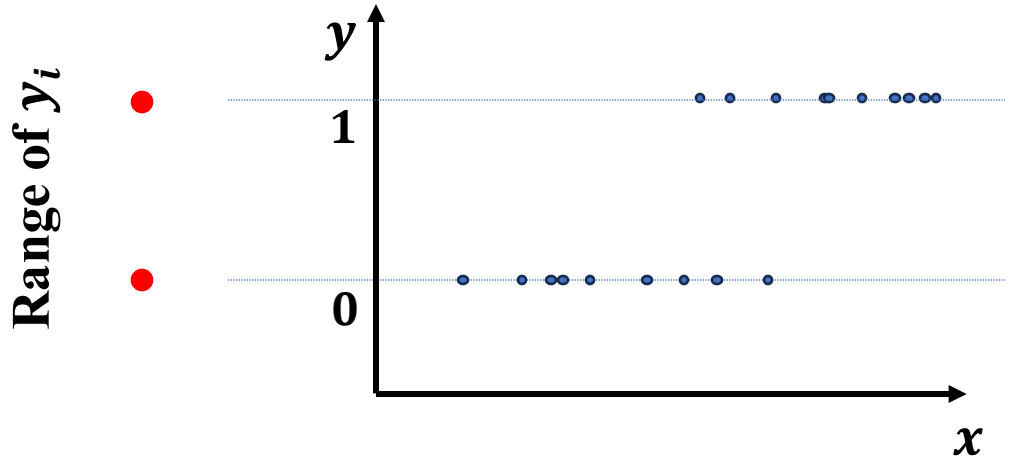
Logistic Regression

So far,





Next, we will move to the scenario when…



Logistic Regression

Question

How would you model the relationship between y and x_1, x_2, x_3 defined as follows?

y: whether or not a high school student is admitted by XX University

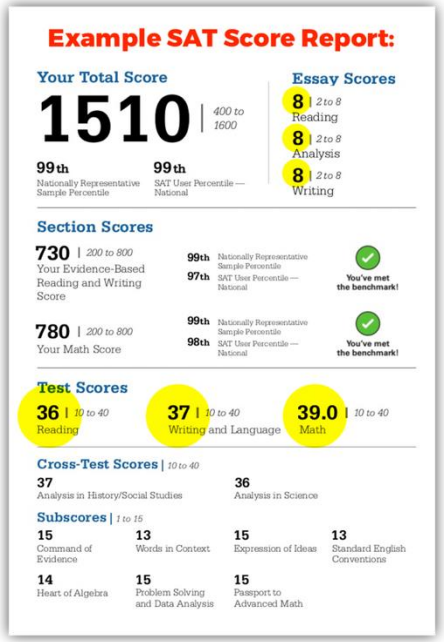
$y = 1$ if admitted

$y = 0$ if not admitted

x_1: the student's high school GPA

x_2: the student's SAT score

x_3: whether the student is a legacy





Logistic Regression

Question

How would you model the relationship between y and x_1, x_2, x_3 defined as follows?

y: whether or not a high school student is admitted by XX University

x_1: the student's high school GPA; x_2: the student's SAT score; x_3: whether a legacy

Is $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ is a reasonable model? **No, because $y \in \{0, 1\}$.**

Logistic Regression

Question

How would you model the relationship between y and x_1, x_2, x_3 defined as follows?

y: whether or not a high school student is admitted by XX University

x_1: the student's high school GPA; x_2: the student's SAT score; x_3: whether a legacy

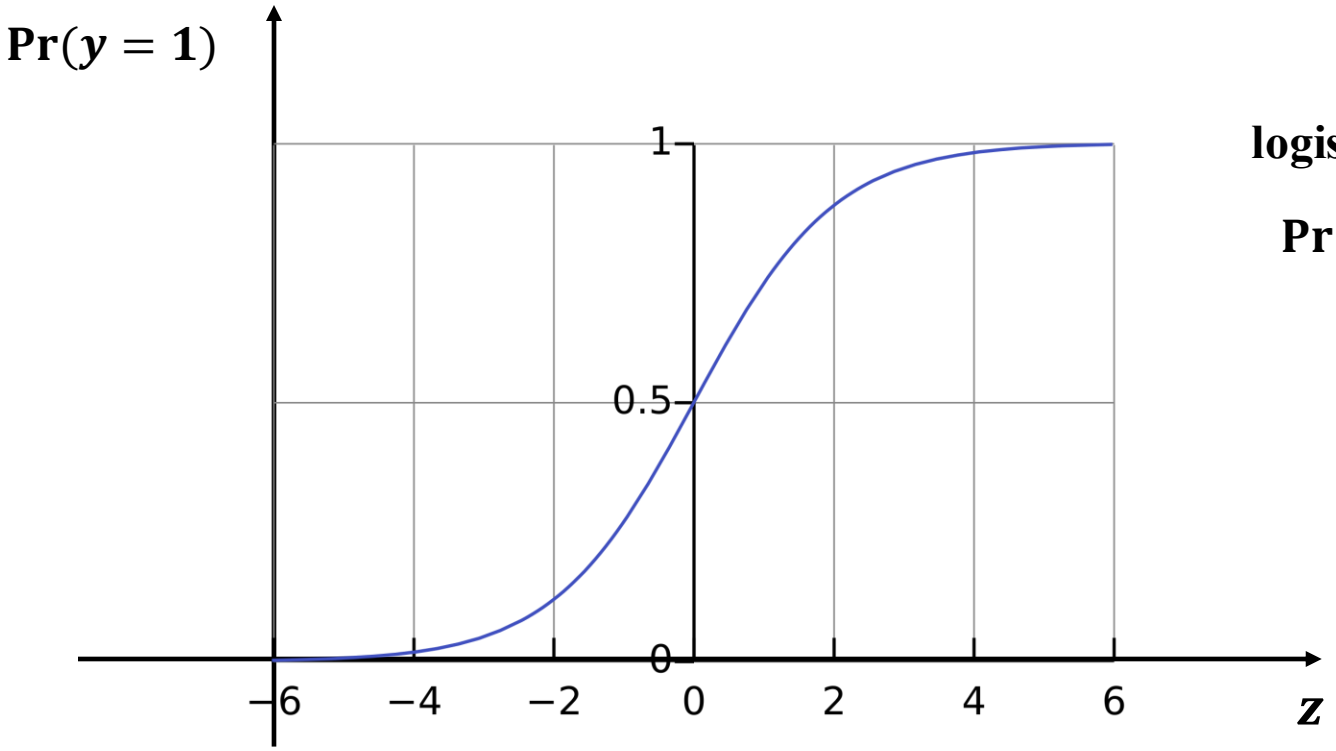
A continuous variable $z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ → The "potential" to be admitted

The higher z is, more likely is the student to be admitted

Logistic Regression

A continuous variable $z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ → The "potential" to be admitted

The higher z is, more likely the student to be admitted



logistic (sigmoid) function

$$\Pr(y = 1) = \frac{1}{1 + e^{-z}}$$

Logistic Regression

Question

By assuming $z = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$ and $p = \Pr(y = 1) = \frac{1}{1+e^{-z}}$,

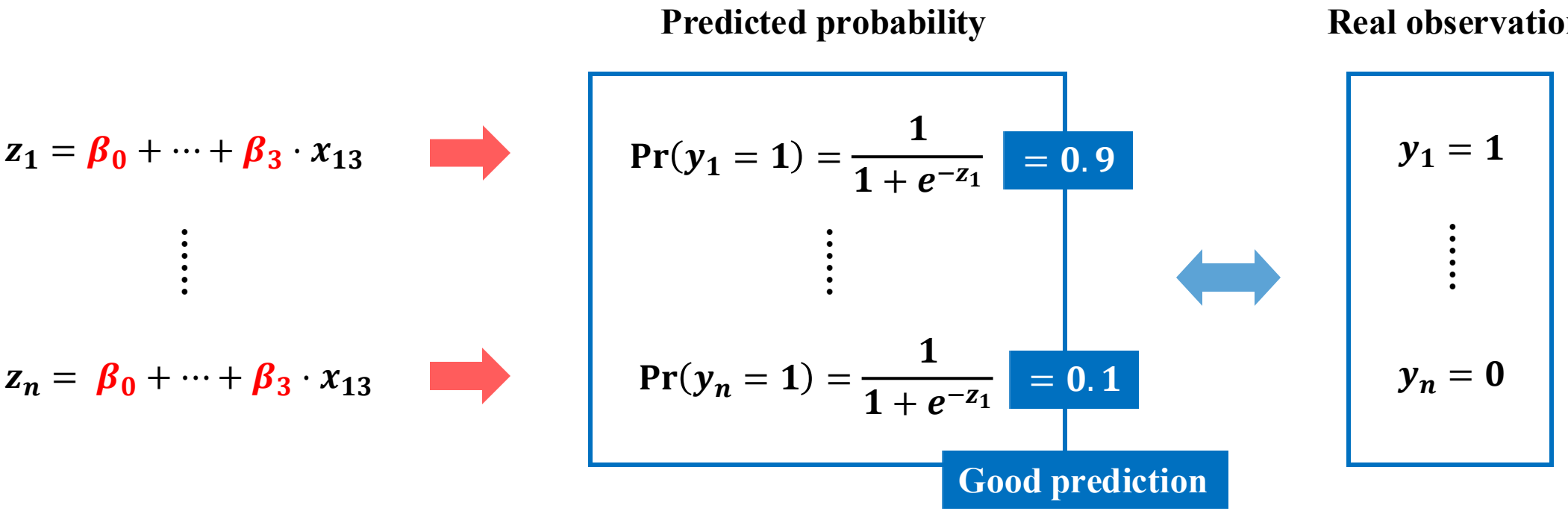
how can we explain β_1?

The "odd"

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{1}{e^{-z}}\right) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

If x_1 increases by Δx_1 (Δx_1 is small), the odd of $y = 1$ will increase by $(100 \cdot \beta_1 \cdot \Delta x_1)\%$

Logistic Regression



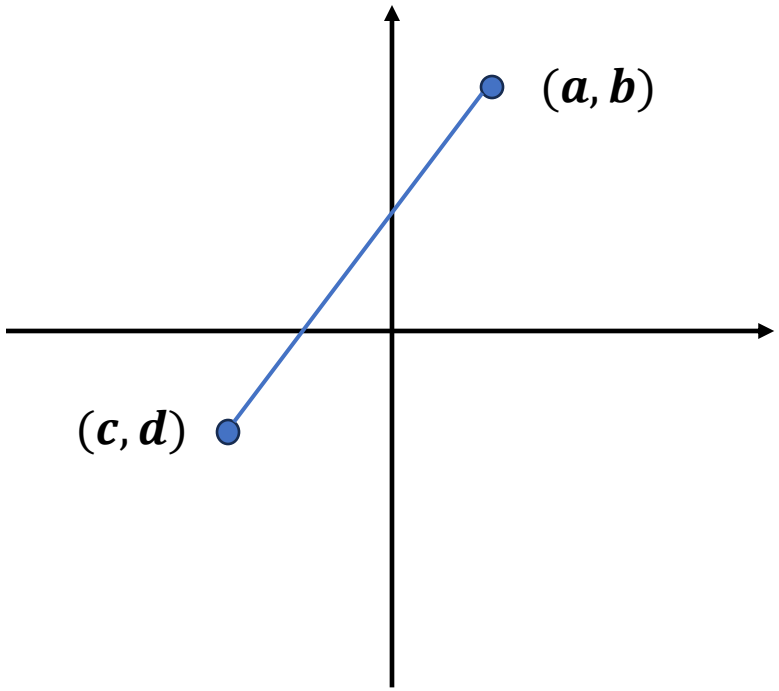
We need to adjust $\beta_1, \beta_2, \beta_3$ so that the predicted probability and the observation match

The smaller $\sum_{i=1}^{n} \left| \frac{1}{1+e^{-z_1}} - y_i \right|^2$, the better the approximation?

Logistic Regression

There exist different types of distance for points in different types of spaces

Euclidean distance



Kullback–Leibler (KL) divergence

$$\begin{matrix}(p, 1-p)\\(q, 1-q)\end{matrix} \in \text{Probability space}$$

$$d(p, q) = p \cdot \log\left(\frac{p}{q}\right) + (1-p) \cdot \log\left(\frac{1-p}{1-q}\right)$$

(1) $d(p, q) = 0$ if $p = q$

(2) $d(p, q) > 0$ if $p \neq q$

$$\frac{\partial d(p, q)}{\partial p} = \log\left(\frac{p}{q}\right) - \log\left(\frac{1-p}{1-q}\right)$$

Logistic Regression

$$z_1 = \beta_0 + \cdots + \beta_3 \cdot x_{13}$$

$$\vdots$$

$$z_n = \beta_0 + \cdots + \beta_3 \cdot x_{13}$$

Predicted probability

$$p_1 = \Pr(y_1 = 1) = \frac{1}{1 + e^{-z_1}}$$

$$\vdots$$

$$p_n = \Pr(y_n = 1) = \frac{1}{1 + e^{-z_1}}$$

Real observation

$$y_1$$

$$\vdots$$

$$y_n$$

$$\min_{\beta} \sum_{i=1}^{n} y_i \cdot \log\left(\frac{y_i}{p_i}\right) + (1 - y_i) \cdot \log\left(\frac{1-y_i}{1-p_i}\right).$$

$$\min_{\beta} \sum_{i=1}^{n} -y_i \cdot \log(p_i) - (1 - y_i) \cdot \log(1 - p_i).$$

Logistic Regression

Optimization Formulation

A general logistic regression model is trained by solving the following optimization problem

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \sum_{i=1}^{n} -y_i \cdot \log(p_i) - (1 - y_i) \cdot \log(1 - p_i)$$

where $p_i = \frac{1}{1+e^{-z_i}}$, $z_i = f(x; \boldsymbol{\beta})$, and $f(x; \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$

unconstrained convex nonlinear program

- The objective function is nonlinear but convex
- There exist no constraints

Logistic Regression

Question

Consider the following logistic regression model

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\ l(\boldsymbol{\beta}) = \sum_{i=1}^{n} -y_i \cdot \log(p_i) - (1-y_i) \cdot \log(1-p_i),$$

$$\text{where } p_i = \frac{1}{1+e^{-z_i}},\ z_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_m \cdot x_{im}.$$

Compute $\nabla l(\boldsymbol{\beta})$ and $\nabla^2 l(\boldsymbol{\beta})$.

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_m \end{bmatrix} \quad z = \begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix} \quad p = \begin{bmatrix} p_1 \\ \vdots \\ p_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$z = X\boldsymbol{\beta} \qquad p = \frac{1}{1+\exp(-z)}$$

$$l(\boldsymbol{\beta}) = -y^{\mathrm{T}}\log(p) - (1-y)^{\mathrm{T}}\log(1-p)$$

Logistic Regression

$$l = l(\beta) = -y^{\mathrm{T}}\log(p) - (1-y)^{\mathrm{T}}\log(1-p), \quad \text{where } p = \frac{1}{1+\exp(-z)} \text{ and } z = X\beta$$

$$\dot{p} = \left(\frac{\partial l}{\partial p}\right)^{\mathrm{T}} = -\frac{y}{p} + \frac{1-y}{1-p} \qquad \dot{z} = \left(\frac{\partial l}{\partial z}\right)^{\mathrm{T}} = \left(\frac{\partial p}{\partial z}\right)^{\mathrm{T}} \dot{p} = \mathrm{diag}(p \circ (1-p))\dot{p}$$

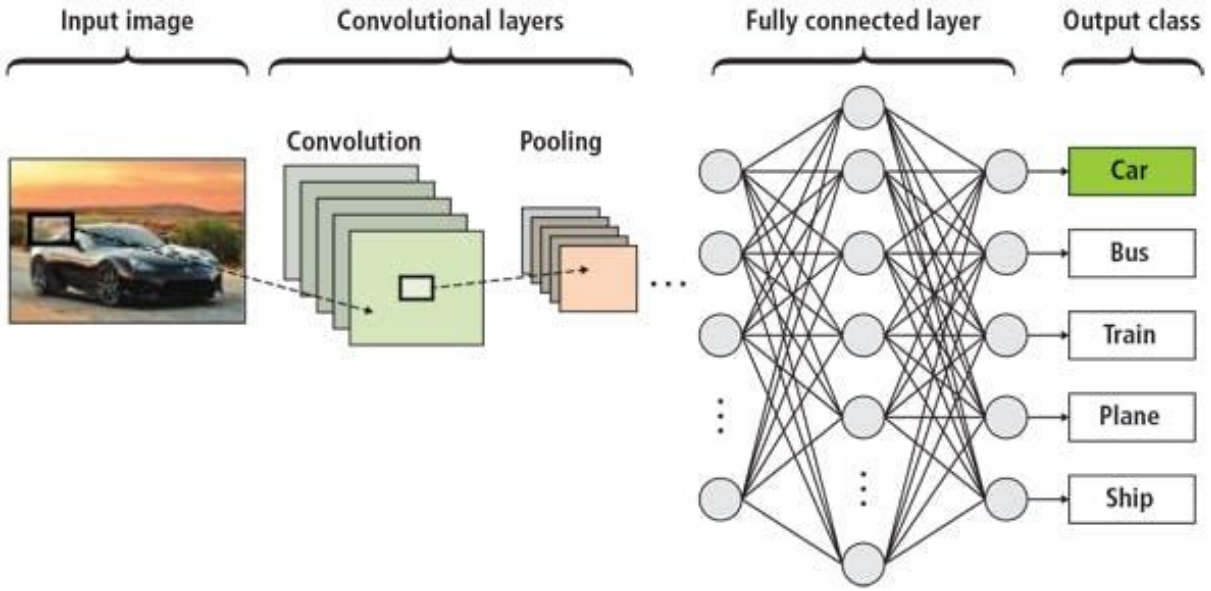
$$\nabla l(\beta) = \left(\frac{\partial l}{\partial z}\right)^{\mathrm{T}} = X^{\mathrm{T}}\dot{z} \qquad \nabla l(\beta) = \left(\frac{\partial l}{\partial z}\right)^{\mathrm{T}} = X^{\mathrm{T}}\mathrm{diag}\big(p \circ (1-p)\big)\left(-\frac{y}{p} + \frac{1-y}{1-p}\right)$$
$$= X^{\mathrm{T}}(p-y)$$

$$\nabla^2 l(\beta) = \frac{\partial \nabla l(\beta)}{\partial \beta} = \boxed{X^{\mathrm{T}}\mathrm{diag}\big(p \circ (1-p)\big)X}$$

A PSD matrix, noting that it equals $A^{\mathrm{T}}A$, where $A = \mathrm{diag}\big(p^{1/2} \circ (1-p)^{1/2}\big)X$

Deep Learning

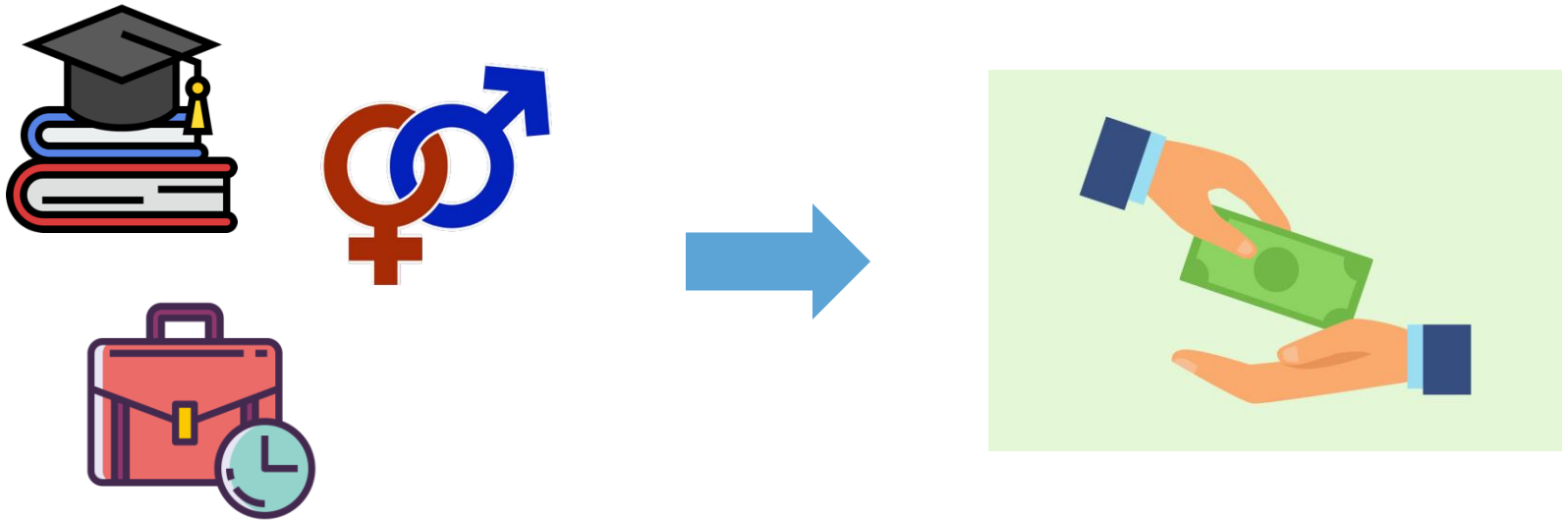
We are now in the era of deep learning…



While, linear and logistic regressions are still useful because they are explainable!

Labor Economics

When we regress wages on various factors…



- **It is usually not about building a good predictive model**
- **We are more curious about questions like whether female workers have been discriminated**