機器學習基礎與演算法

陳弘軒 Hung-Hsuan Chen
https://www.ncu.edu.tw/~hhchen
中央大學資訊工程學系

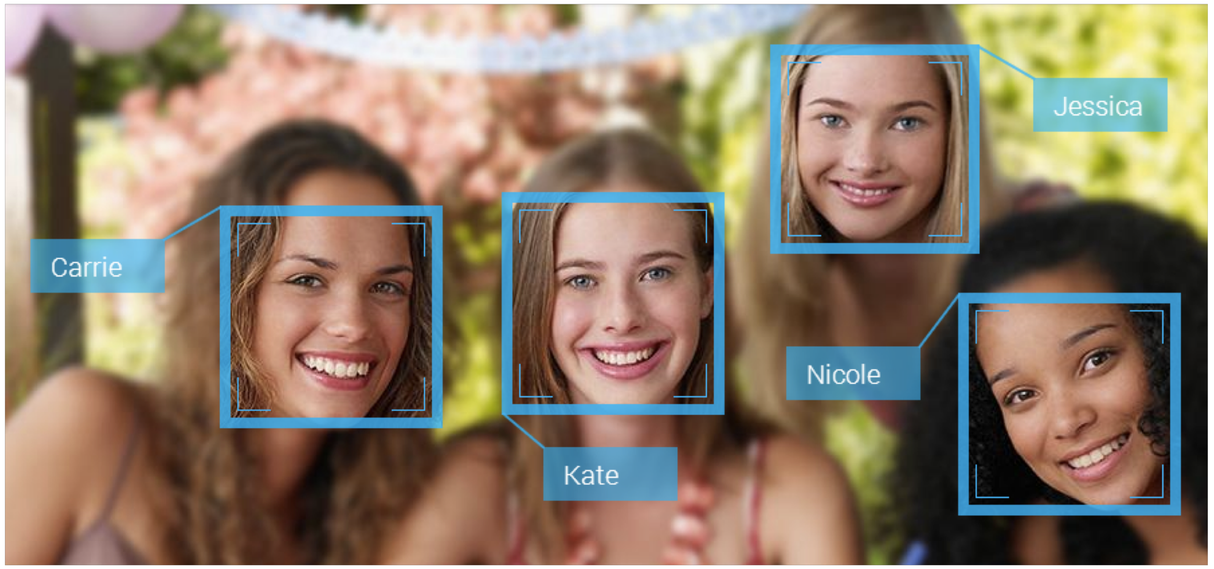
「版權聲明頁」

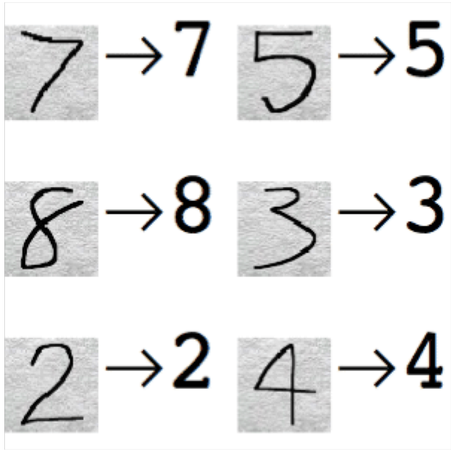
Outline

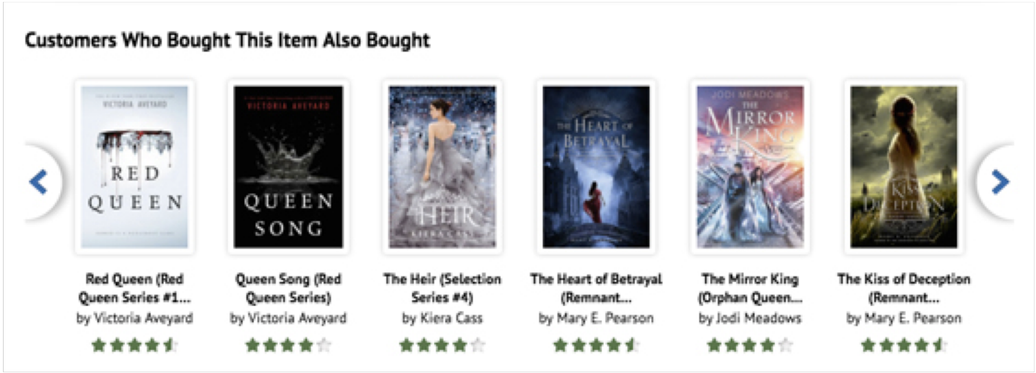
- Introduction to Machine Learning
- Regression
- Classification
- Support Vector Machine
- Decision trees and ensemble learning
- Dimension reduction

Introduction to Machine Learning

Examples of machine learning today





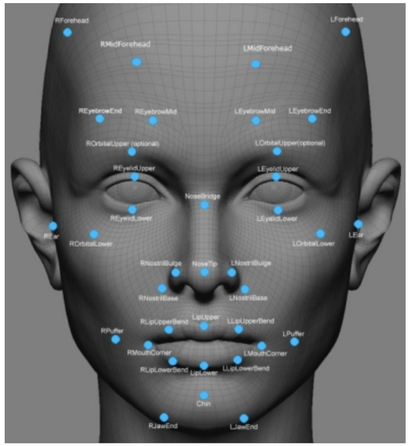


Traditional algorithm vs machine learning

- Traditional algorithm: requires “prior” knowledge of the target problem
  - E.g., sorting
- Some tasks are difficult to solve by prior knowledge
  - How to detect a spam mail? Answers may vary from person to person
- Machine learning: learning from examples
  - Data driven

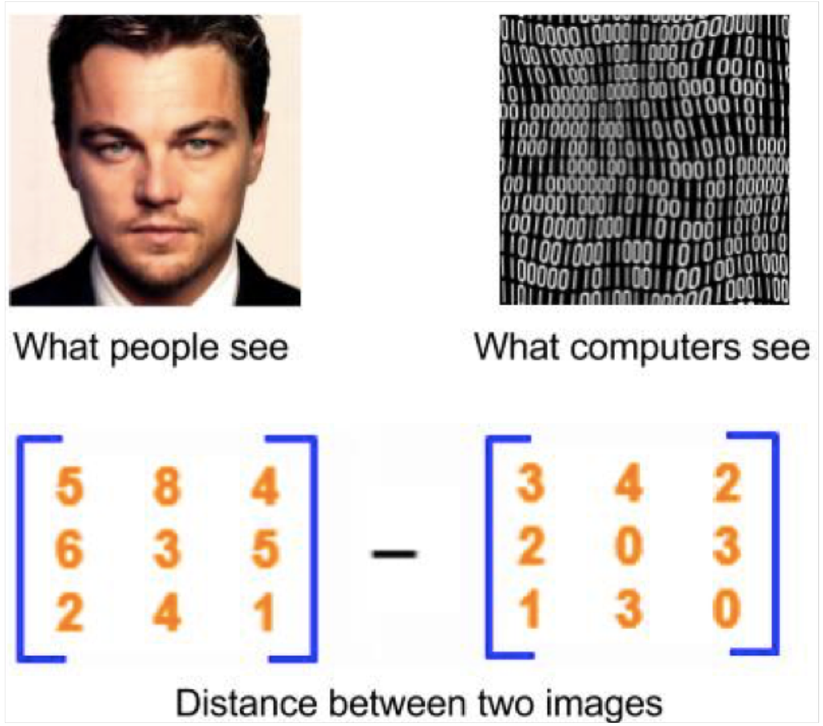
How to determine the photos with faces?

- Traditional algorithm
  - Round shape, with two black circles (eyes), …
  - Solve a problem based on your prior knowledge
- Data driven algorithm
  - Show many face/non-face photos to the machine, and let the machine identifies their differences
  - Let computer discover the patterns from the collected data



A naïve machine learning approach find faces

- From training data, find the neighboring photos of the new photo
- If most neighboring photos containing faces, the new photo probably contains faces
- If most neighboring photos have no faces, the new photo probably contains no faces



Types of Machine Learning

- Based on the input-output structure, ML can be categorized as (this is not a complete list):
  - Supervised Learning
  - Unsupervised Learning
  - Semi-supervised Learning
  - Reinforcement Learning
- We will mostly discuss supervised learning

Supervised Learning

- Given: a set of <input, output> pairs as the learning samples
- Goal: given an unseen input, predict the corresponding output
- Examples
  1. Input: X-ray photo of chests, output: whether it is cancerous
  2. Input: a sentence, output: whether a sentence is grammatical
  3. Input: some indicators of a company, output: whether it will make profit next year

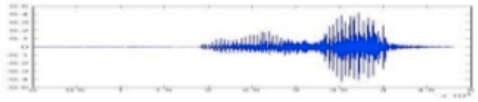
Output types

- Categorical: ***classification problem***
  - *Ordinal outputs: small, medium, large*
  - *Non-ordinal outputs: blue, green, orange*
- Real values: ***regression problem***
- Other types are possible (e.g., output a list), but the above two are most widely studied

Different types of outputs

Source: Hung-Yi Lee:
https://www.slideshare.net/tw_dsconf/ss-62245351

- Speech Recognition

$f($  $)=$ “How are you”

- Image Recognition

$f($  $)=$ “Cat”

- Playing Go

$f($  $)=$ “5-5” (next move)

- Dialogue System

$f($ “Hi” $)=$ “Hello”

(what the user said) (system response)

Terminology

- Training data: a set of data used to discover potentially predictive relationships
- Test data: the data that has been specifically identified for use in tests
- Features (a.k.a. attributes, independent variables)
  - “Input” of a prediction task, usually denoted by $\boldsymbol{x}$
- Target variable (a.k.a. outputs, dependent variables)
  - “Output” of a prediction task, usually denoted by y
  - In classification, target variables are also called classes

Example

Weight	Wingspan	Webbed feet?	Back color	Species
1000.1	125.0	No	Brown	Buteo jamaicensis
3000.7	200.0	No	Gray	Sagittarius serpentarius

Features

Target variable

Classification (1/3)

- It is a supervised learning task
- Goal: given a feature vector $\boldsymbol{x}$, predicts which class in C may be associated with $\boldsymbol{x}$.
- If $|C| = 2$ ➔ Binary Classification
  If $|C| > 2$ ➔ Multi-class Classification

Classification (2/3)

- Training and predicting of a binary classification problem:

Training set (binary classification)

Feature Vector ($x_i \in R^d$)	Class
x_1	$+1$
x_2	-1
…	…
x_{n-1}	-1
x_n	$+1$

A new instance

Feature Vector ($x_{\text{new}} \in R^d$)	Class
x_{new}	?

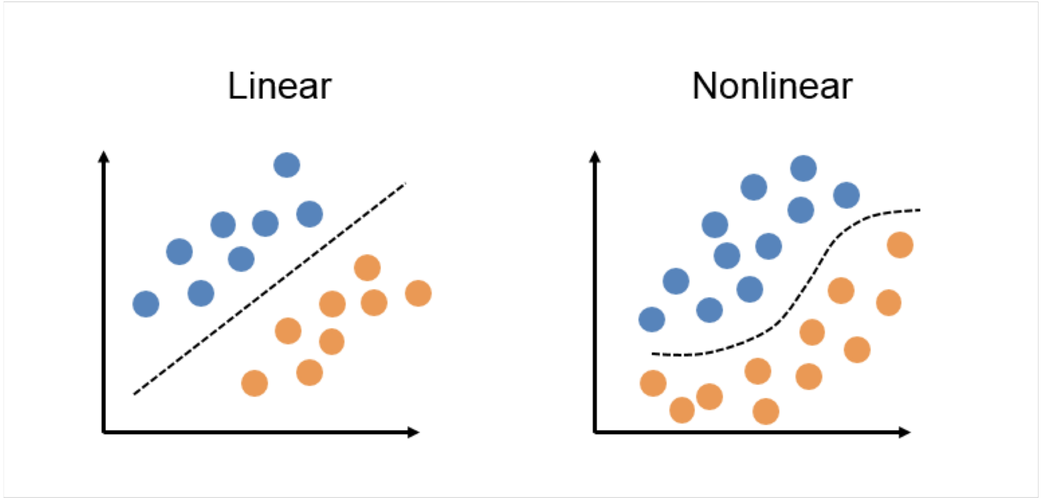
(1) Training

Classifier $f(x)$

(2) Predicting

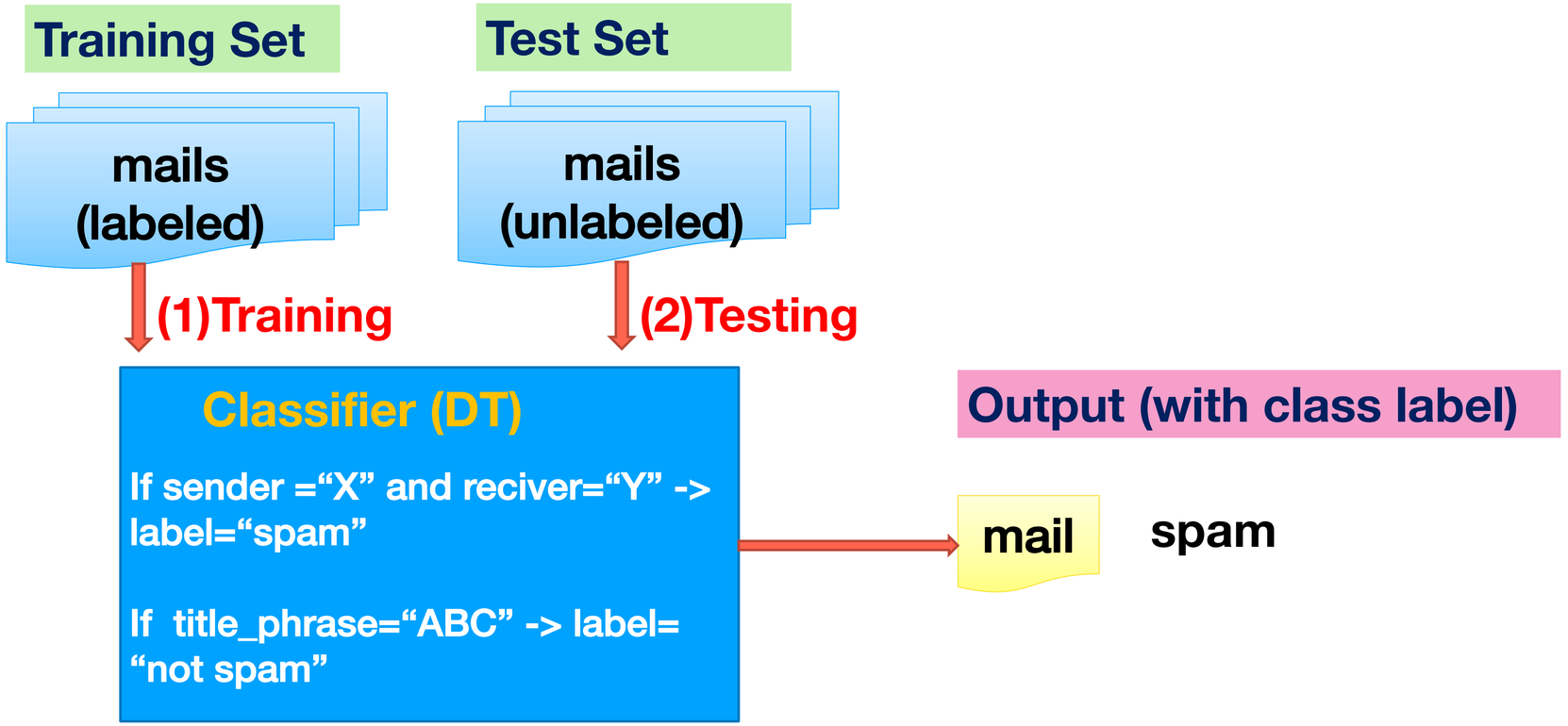
Classification (3/3)

- A classifier can be either linear or non-linear
- The geometric view of a linear and non-linear classifier
- Famous classification models:
  - k-nearest neighbor (kNN)
  - Decision Tree (DT)
  - Logistic regression
  - Support Vector Machine (SVM)



Real example: E-mail spam check

- Blocking the junk email and passing the normal email



Regression (1/2)

- A supervised learning task that, given a feature vector $\boldsymbol{x}$, predicts the target value $y \in \mathrm{R}$.

- Training and predicting of a regression problem:

Training set (regression)

Feature Vector ($\boldsymbol{x}_i \in R^d$)	y_i
$\boldsymbol{x}_1$	$+0.26$
$\boldsymbol{x}_2$	-3.94
…	…
$\boldsymbol{x}_{n-1}$	-1.78
$\boldsymbol{x}_n$	$+5.31$

A new instance

Feature Vector ($\boldsymbol{x}_{\text{new}} \in R^d$)	y_i
$\boldsymbol{x}_{\text{new}}$	?

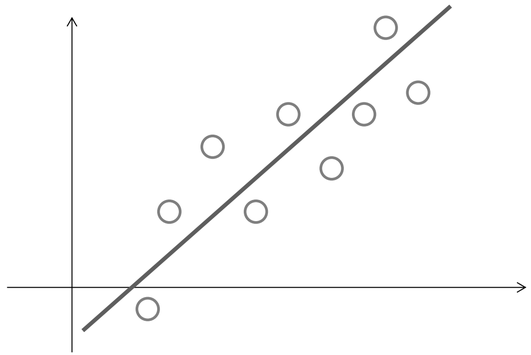
(1) Training

(2) Predicting

Regressor $f(\boldsymbol{x})$

Regression (2/2)

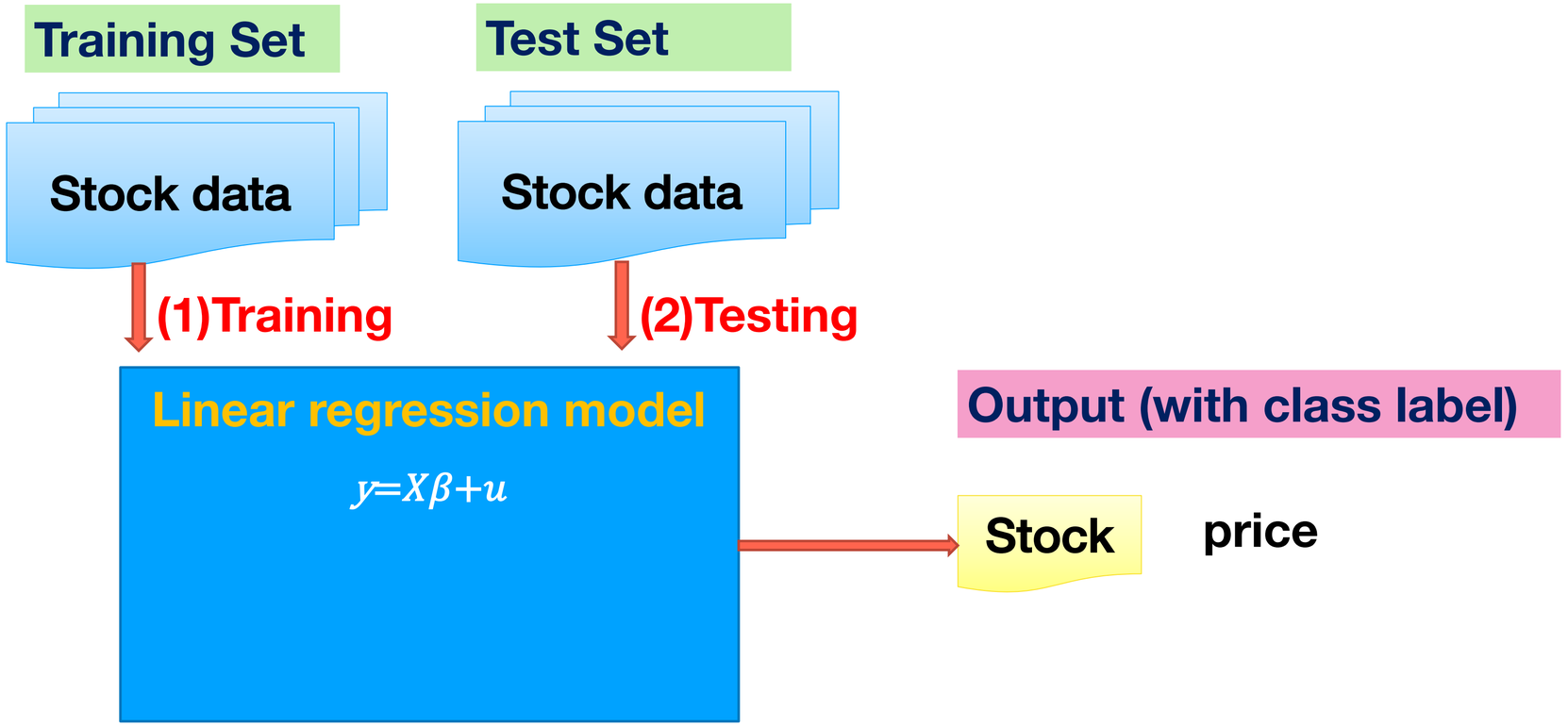
- The geometric view of a linear regression function



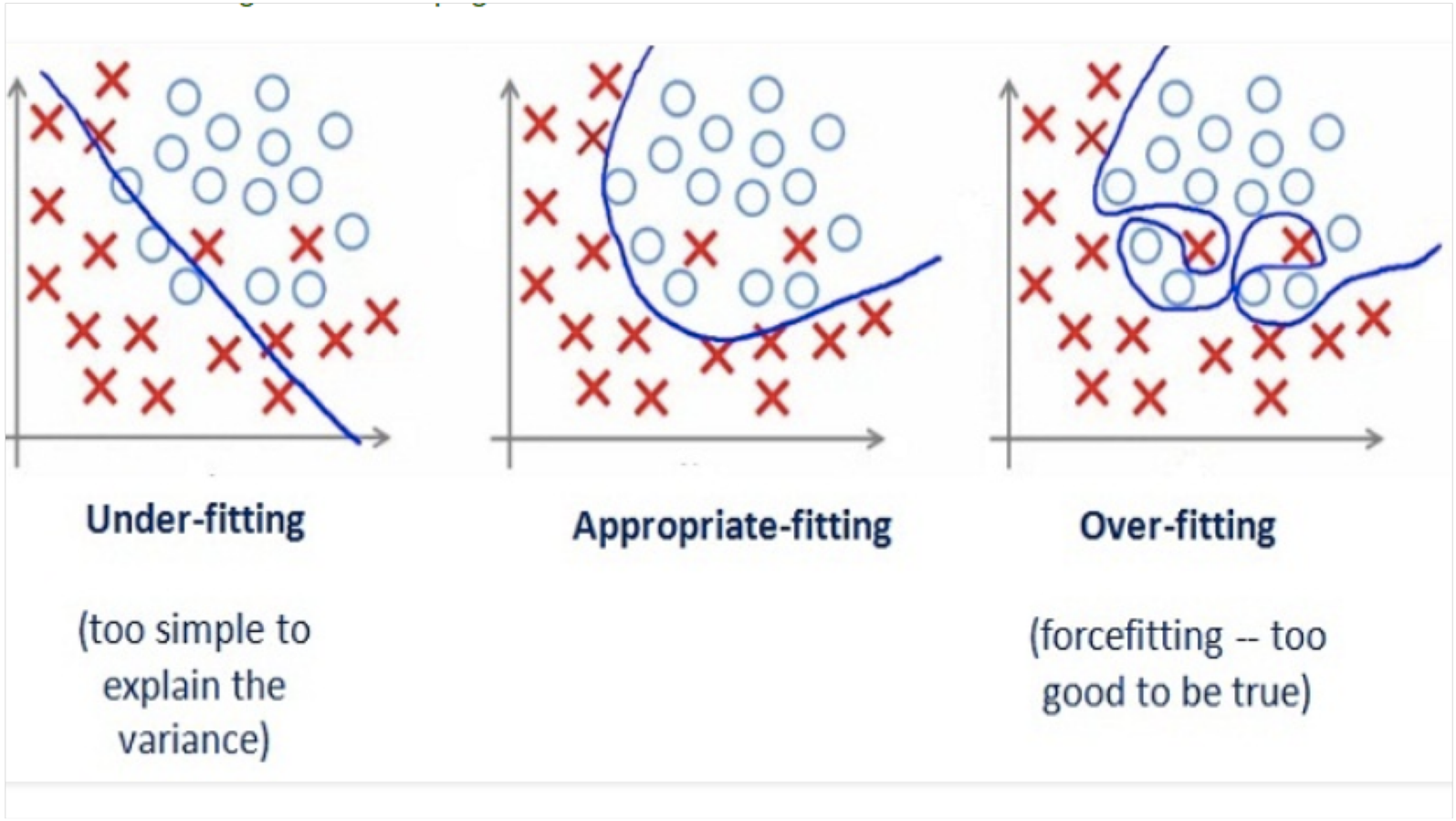
- Some types of regression: linear regression, support vector regression, …

Real Example: Stock price prediction

- Predicting the price of a stock



Overfitting vs underfitting



Quiz

- Explain the difference between a supervised and an unsupervised algorithm
- Explain the difference between classification and regression
- Classify the normal mails and the junk mail based on the labeled datasets
  - Supervised or unsupervised?
- Predicting the price of stock
  - Classification or regression?