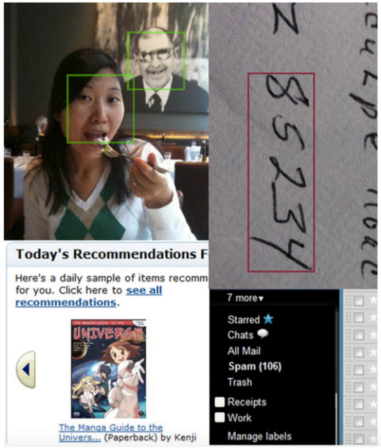
Introduction to Data Science

Hung-Hsuan Chen 陳弘軒
Computer Science and Information Engineering
National Central University
hhchen@ncu.edu.tw

Many slides are obtained from Prof. Shou-De Lin (NTU)

Examples of machine learning today



Today's Recommendations F

Here's a daily sample of items recomm
for you. Click here to see all recommendations.

The Manga Guide to the Univers... (Paperback) by Kenji

Traditional algorithm vs data driven algorithm

- How to detect a face?
  - Traditional algorithm
    - Round shape, with two black circles (eyes), …
    - Solve a problem based on your knowledge (prior information)
  - Data driven algorithm
    - Show many face/non-face photos to the machine, and let the machine identifies their differences
    - Solve the problem based on the data (and maybe some of the prior knowledge)

Types of Machine Learning

- Based on the input-output structure, ML can be categorized as:
  - Supervised Learning → most machure
  - Unsupervised Learning → have no poticular goal
  - Semi-supervised Learning
  - Reinforcement Learning
    ↳ like playing a game

We will mostly discuss the first type in this class

Supervised Learning

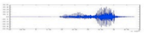
- Given: a set of <input, output> pairs
- Goal: given an unseen input, predict the corresponding output
- For example:
  1. Input: X-ray photo of chests, output: whether it is cancerous
  2. Input: a sentence, output: whether a sentence is grammatical
  3. Input: some indicators of a company, output: whether it will make profit next year
- Two typical types of outputs an ML system generates
  - Categorical: ***classification problem***
    - *Ordinal outputs: small, medium, large*
    - *Non-ordinal outputs: blue, green, orange*
  - Real values: ***regression problem***
- There are several other variations

ex: correct or incorrect

output a real number

Different types of outputs

- Speech Recognition

$f($  $)=$ "How are you"

- Image Recognition

$f($ $)=$ "Cat"

- Playing Go

$f($ $)=$ "5-5" (next move)

- Dialogue System

$f($ "Hi" (what the user said) $)=$ "Hello" (system response)

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

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)
  - We usually use *X* to represent features
  - Features are the "input" of a prediction task
- Target variable (a.k.a. outputs, dependent variables)
  - In classification, target variables are also called classes
  - We usually use *y* to represent target variables
  - Targets are the "output" of a prediction task

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

Unsupervised Learning

- Learning without teachers (presumably harder than supervised learning)
  - Learning "what normally happens"
  - Think of how babies learn their first language (unsupervised) comparing with how people learn their 2nd language (supervised).
- Given: a bunch of input *X* (there is no output *y*)
- Goal: depending on the tasks, for example
  - Estimate P(*X*) → then we can find augmax P(*X*)
  - Finding Sim(X_1,X_2) → then we can group similar *X*'s
  - Finding P(X_2|X_1)→ we can know whether some items can occur together.

A variety of ML Tasks

1. Classification
2. Regression number
3. Clustering grouping
4. Transfer learning
5. Multi-label learning
6. Multi-instance learning
7. Cost-sensitive leering
8. Active learning
9. Semi-supervised learning
10. Reinforcement learning

Classification (1/2)

- It is a supervised learning task that, given a feature vector x, predicts which class in C may be associated with x.
- |C|=2 → Binary Classification

  |C|>2 → Multi-class Classification
- Training and predicting of a binary classification problem:

Training set (Binary Classification)

	Feature Vector ($x_i \in R^d$)	Class
170.80, lw	x_1	+1
160, 50, 90	x_2	-1
	…	…
	x_{n-1}	-1
	x_n	+1

A new instance

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

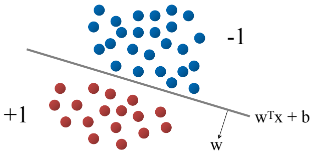
(1) Training

Classifier f(x)

(2) Predicting

Classification (2/2)

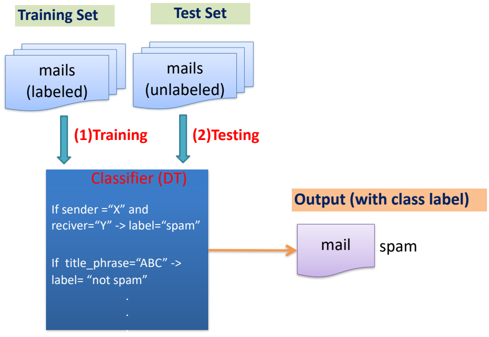
- A classifier can be either linear or non-linear
- The geometric view of a linear classifier



- Famous classification models:
  - k-nearest neighbor (kNN)
  - Decision Tree (DT)
  - Support Vector Machine (SVM)
  - …

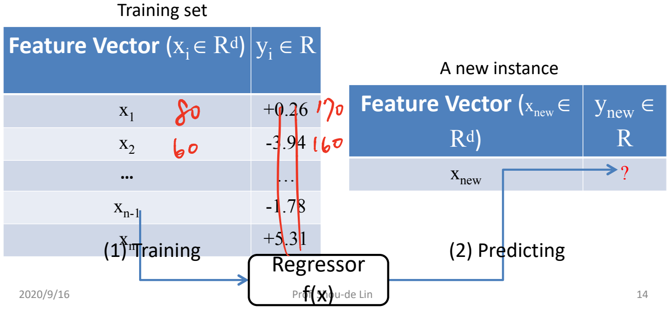
Real example: E-mail spam check

- Blocking the junk email and passing the normal email


Training Set
Test Set
mails (labeled)
mails (unlabeled)
(1)Training
(2)Testing
Classifier (DT)
If sender ="X" and reciver="Y" -> label="spam"
If title_phrase="ABC" -> label= "not spam"
.
.
.
Output (with class label)
mail
spam

Regression (1/2)

- A supervised learning task that, given a feature vector x, predicts the target value $y \in R$.
- Training and predicting of a regression problem:



Training set

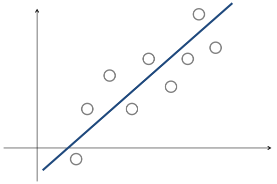
Feature Vector ($x_i \in R^d$)	$y_i \in R$
x_1 80	+0.26 170
x_2 60	-3.94 160
…	…
x_{n-1}	-1.78
x_n	+5.31

A new instance

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

Regression (2/2)

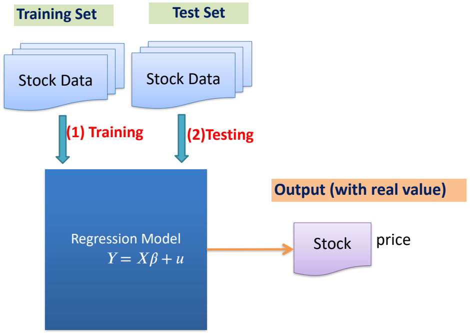
- The geometric view of a linear regression function



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

Real Example: Stock price prediction

- Predicting the price of stock



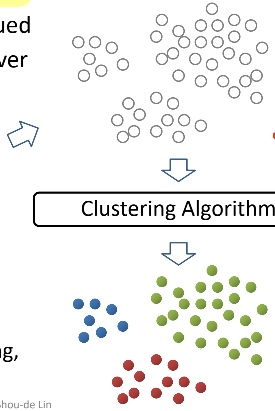
Clustering

✱ Cluster vs. Classification

- An unsupervised learning task
- Given a finite set of real-valued feature vector $S \subset R^d$, discover clusters in S

S

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



didn't tell the group or label

- K-Means,Hierarchical clustering, DBSCAN, etc

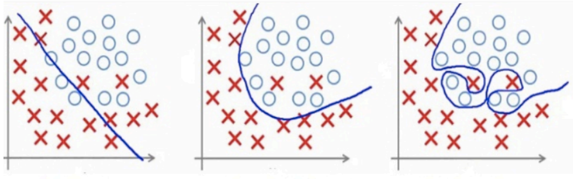
Problem Modeling is critical

- It is very important to know how to model the problem into a suitable ML task
  - Incorrect problem modeling leads you to nowhere

An example: click-through rate prediction

- Assuming you want to predict quality of an advertisement by estimating its click-ratio (i.e. how likely a person would buy the product after viewing this ad)
  - Since this click-ratio is a real number between 0 and 1, a natural way is to model it as a regression problem.
  - However, an experienced machine learning person would suggest decomposing this into a binary classification problem (i.e. 3/8 will be decomposed into 3 positive instances and 5 negative instances)

Overfitting vs underfitting



Under-fitting

(too simple to explain the variance)

didn't really idenerfy

Appropriate-fitting

better

Over-fitting

(forcefitting -- too good to be true)

Quiz

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