Machine Learning Basics

葉建華

jhyeh@mail.au.edu.tw

http://jhyeh.csie.au.edu.tw/

What is Machine Learning?

T-800
www.toyhaven.blogspot.com
THE TERMINATOR
"Schwarzenegger creates an inimitable villain, an unstoppable killing machine, part metal, part man."
SCHWARZENEGGER
THE TERMINATOR
TERMINATOR

What is Machine Learning?

- A sub-domain of Artificial Intelligence (AI)
- Algorithms learn patterns by historical data
  - The learning result: model
- A simple example: spam mail filtering
  - Keyword: “online pharmacy”
  - Define a filter based on this keyword, then a simple model created
- Algorithms are mostly based on mathematics and statistics

Limitations of Machine Learning

- What about new patterns?
  - The historical data is not useful anymore
- Simplified filtering rule
  - Keywords without semantic understanding?
- Simple models may cause over-concluded results
  - A simple separator lead to imprecision

Real Life Examples

- Please find and understand:
  - Google (http://www.google.com/)
  - Amazon (http://www.amazon.com/)
  - Netflix (http://www.netflix.com/)
  - Pandora (http://www.pandora.com/), Last.fm (http://last.fm/)
  - Hollywood stock exchange (http://hsx.com)
  - eHarmony (http://www.eharmony.com/)

Other Uses for Learning Algorithms

- Biotechnology: sequencing and screening technology for DNA, protein, RNA, etc.
- Financial fraud detection: examining transactions
- Machine/Computer vision: intrusion detection, car/face detection
- Product marketing: consumer trend detection
- Supply chain optimization
- Stock market analysis
- National security

Again, what is machine learning?

- Gain insight from the dataset, ask the computer to make some sense from data
- Used in many more places than you’d expect
  - Search engine ranking
  - User click stream
  - Spam filtering
  - Related coupon or product recommendation when shopping
  - Handwriting recognition
  - …

Machine Learning

- Machine learning is helpful in
  - Improve business decisions
  - Increase productivity
  - Detect disease
  - Forecast weather
  - and more…

Not just looking at the raw data!

- Example: spam mail filtering
  - Suspicious word?
  - Single word for filtering?
  - Co-occurrence of words?
  - Length of mail?
  - Sender? Host?
- So?

Cross Discipline

- Machine learning contains intersection of many domains
  - Computer science, of course
  - Engineering
  - Statistics
  - Other specific application domain, from politics to geosciences

Statistics?

- Engineering problem are mostly deterministic
  - Design a flow for vending machine
  - Coin in, product out
  - Statistics?

Statistics!

- But even more problems are not deterministic
  - Ask which the best selling product is
  - Ask how much the customer is willing to pay
  - Ask when is the best selling period during a day, a week, a month, even a season, a year.
  - Ask the status of changes dispense
  - Too many statistical questions to ask!
- Maybe we don’t know enough about the problem
- Maybe we don’t have enough computing power to model the problem

Can we always properly model?

- Target is too complex to define
  - Happiness
- Assumption is not complete
- Data size is too big to model
- Data generation speed is too fast to model

An Example

In 1989, the Loma Prieta earthquake struck northern California, killing 63 people, injuring 3,757, and leaving thousands homeless. A similarly sized earthquake struck Haiti in 2010, killing more than 230,000 people. Shortly after the Loma Prieta earthquake, a study was published using low-frequency magnetic field measurements claiming to foretell the earthquake. A number of subsequent studies showed that the original study was flawed for various reasons. Suppose we want to redo this study and keep searching for ways to predict earthquakes so we can avoid the horrific consequences and have a better understanding of our planet.

- What would be the best way to go about this study?

An Example (cont.)

- Look at your smartphone
  - What sensors do it have?
  - Can you get readings of magnetometers?
  - The frequency? Hundreds of times a second!

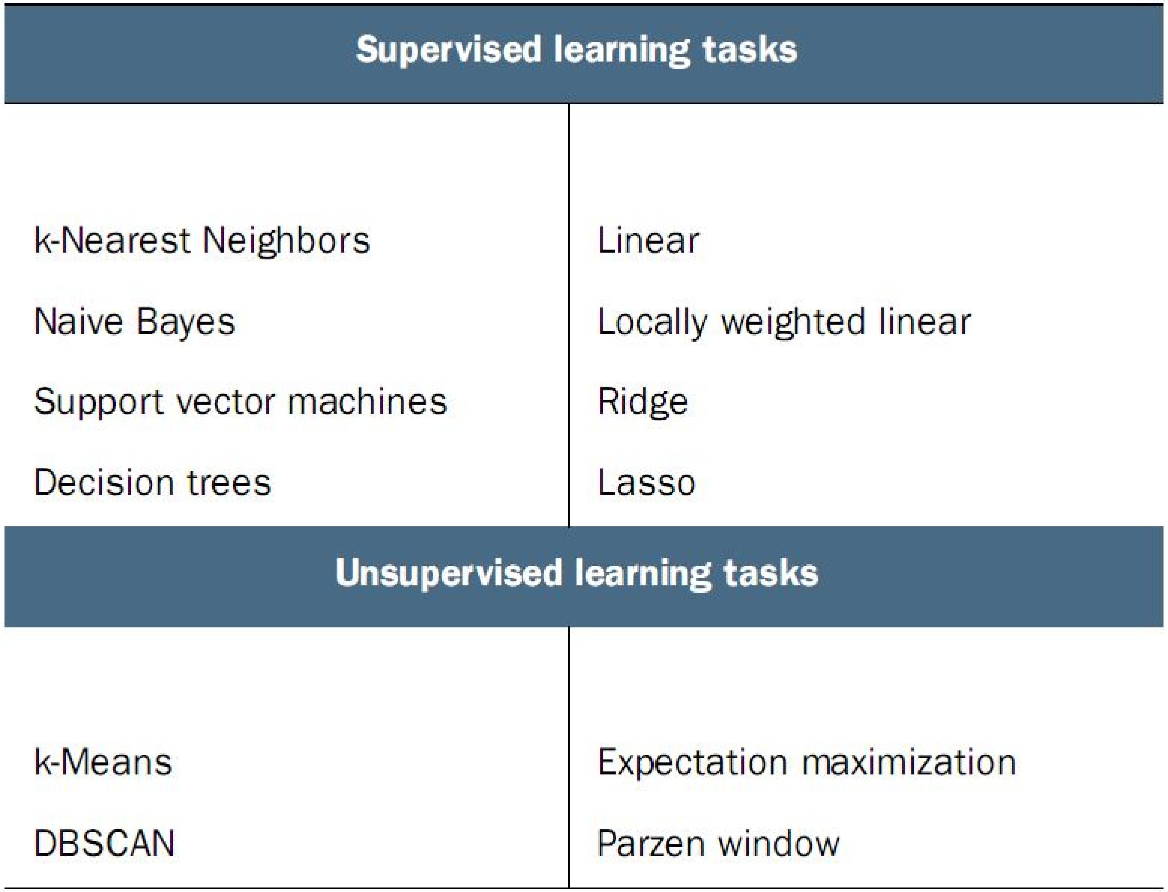
Machine Learning is getting Important

- *"I keep saying the sexy job in the next ten years will be statisticians."*
  - Hal Varian, chief economist of Google, 2009
- With so much of the economic activity dependent on information, you can't afford to be lost in the data
  - Machine learning will help

Take a Look!

http://www.youtube.com/watch?v=NLIGopyXT_g

Major Tasks

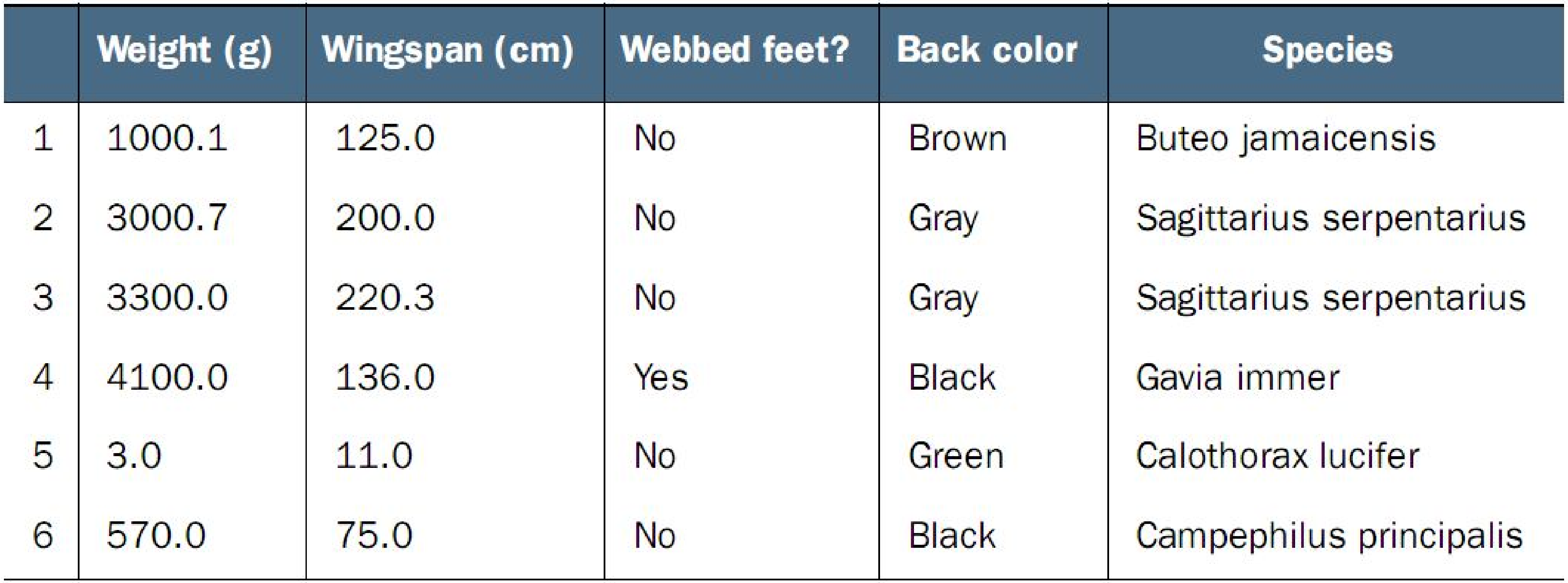


Supervised learning tasks	
k-Nearest Neighbors	Linear
Naive Bayes	Locally weighted linear
Support vector machines	Ridge
Decision trees	Lasso
Unsupervised learning tasks	
k-Means	Expectation maximization
DBSCAN	Parzen window

Table 1.2 Common algorithms used to perform classification, regression, clustering, and density estimation tasks

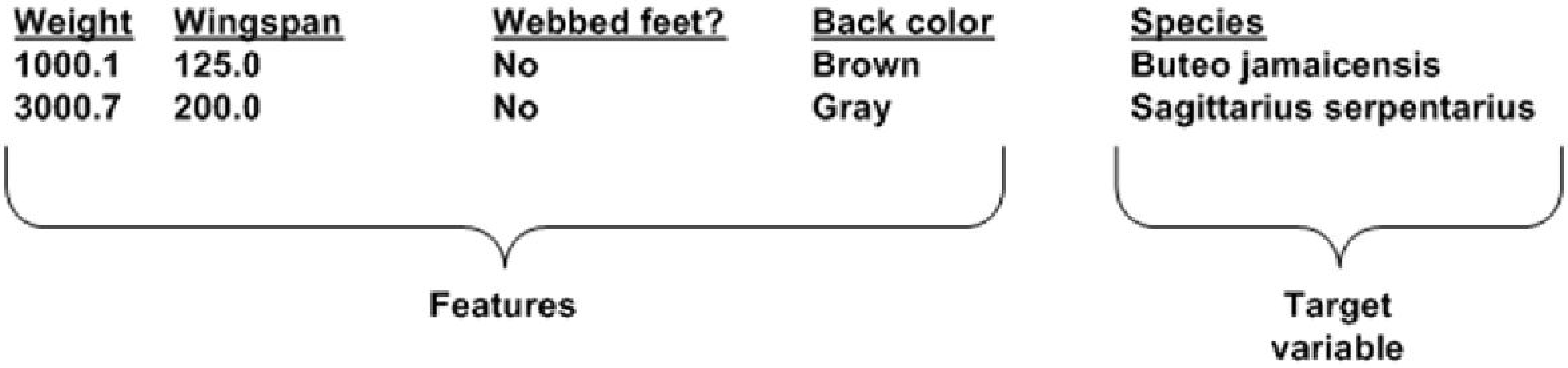
One More Example

Table 1.1 Bird species classification based on four features



	Weight (g)	Wingspan (cm)	Webbed feet?	Back color	Species
1	1000.1	125.0	No	Brown	Buteo jamaicensis
2	3000.7	200.0	No	Gray	Sagittarius serpentarius
3	3300.0	220.3	No	Gray	Sagittarius serpentarius
4	4100.0	136.0	Yes	Black	Gavia immer
5	3.0	11.0	No	Green	Calothorax lucifer
6	570.0	75.0	No	Black	Campephilus principalis

One More Example (cont.)



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

Figure 1.2 Features and target variable identified

Developing Machine Learning Applications

- Collect data
- Prepare the input data
- Analyze the input data
- Data screening, cleaning
- Train the algorithm
- Test the algorithm

Codes are Python-based

- Learn to read them!